Máster Interuniversitario en Matemáticas

Modelización. Procesos Estocásticos

Ejercicios Temas 3 y 4

Alejandro José Florido Tomé

Curso académico 2024/25

1 Ejercicio (1.5 puntos)

Para el modelo de regresión lineal simple,

$$Y = \beta_0 + \beta_1 X + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I_n) \tag{1}$$

a) Deduce la expresión de los estimadores de los parámetros por máxima verosimilitud.

Como los errores ϵ son normalmente distribuidos, la variable dependiente Y también seguirá una distribución normal. Para nuestro caso, la función de verosimilitud, dada la observación muestral $y = (y_1, y_1, .., y_n)^T$, es

$$L(\beta_0, \beta_1, \sigma/y) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n exp\left[-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right], \tag{2}$$

con $\beta_i \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$. Para deducir la forma de los estimadores de los parámetros por máxima verosimilitud, lo más conveniente es tomar el logaritmo neperiano de (2) y derivarlo. Sea el logaritmo

$$ln(L) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2, \tag{3}$$

sus respectivas derivadas serán las llamadas ecuaciones de verosimilitud. Derivando respecto a β_0

$$\frac{\partial}{\partial \beta_0} ln(L) = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \tag{4}$$

de donde se puede despejar β_0,

$$\sum_{i=1}^n \beta_0 = n\beta_0 = \sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i \rightarrow \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \tag{5}$$

con $\hat{\beta}_i$ los estimadores que se buscan, y $\overline{y} = \sum_{i=1}^n y_i/n$ la media de la observación muestral (análogo para $\overline{x}$). En segundo lugar, derivemos (3) respecto a β_1,

$$\frac{\partial}{\partial \beta_1} ln(L) = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \overline{y} + \beta_1 \overline{x} - \beta_1 x_i) x_i, \tag{6}$$

$$\rightarrow -\sum_{i=1}^n y_i x_i + \overline{y} \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n (-x_i)(\overline{x} - x_i) = 0 \rightarrow \beta_1 \sum_{i=1}^n (-x_i)(\overline{x} - x_i) = \sum_{i=1}^n y_i x_i - n\overline{xy} \tag{7}$$

$$\rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - n\overline{yx}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}. \tag{8}$$

Simplifiquemos el numerador y el denominador:

- Numerador: Reescribamos $n\overline{xy} = n\overline{x}\sum_{i=1}^n y_i/n = \sum_{i=1}^n \overline{x}y_i = \sum_{i=1}^n x_i\overline{y}$, siendo el numerador $\sum_{i=1}^n (y_i - \overline{y})x_i$. Si sumamos y restamos $\overline{x}\sum_{i=1}^n (y_i - \overline{y}) = \overline{x}(n\overline{y} - n\overline{y}) = 0$,

$$\sum_{i=1}^n y_i x_i - n\overline{yx} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) = nS_{xy}, \tag{9}$$

siendo $S_{xy} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})/n$.

- Denominador: Veamos cuál es el resultado de $\sum_{i=1}^n (x_i - \overline{x})^2$:

$$\sum_{i=1}^n (x_i - \overline{x})^2 = \sum_{i=1}^n (x_i^2 - 2x_i\overline{x} + \overline{x}^2) = \sum_{i=1}^n x_i^2 - 2\overline{x}(n\overline{x}) + n\overline{x}^2 = \sum_{i=1}^n x_i^2 - n\overline{x}^2 = nS_{xx}^2, \tag{10}$$

con $S_{xx}^2 = \sum_{i=1}^n (x_i - \overline{x})^2/n$. Así, nuestro denominador será claramente nS_{xx}^2.

Juntando todo,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}^2}, \qquad \hat{\beta}_0 = \overline{y} - \frac{S_{xy}}{S_{xx}^2}\overline{x}. \tag{11}$$

El último estimador es el asociado al parámetro σ, conque volvamos a derivar (3) respecto a σ:

$$\frac{\partial}{\partial\sigma} ln(L) = -\frac{n}{2}\frac{1}{2\pi\sigma^2}(4\pi\sigma) + \frac{1}{\sigma^3}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 - n\sigma^2}{\sigma^3}. \tag{12}$$

Despejando se obtiene el estimador resultante,

$$\hat{\sigma} = \frac{1}{n}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^n e_i^2, \tag{13}$$

habiéndose usado en la última igualdad la ecuación original (1).

b) Comprueba que la covarianza de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ es

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{x}\sigma^2}{nS_{xx}^2}. \tag{14}$$

Para ello, partamos de la definición de la covarianza en función de las esperanzas,

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = E[\hat{\beta}_0\hat{\beta}_1] - E[\hat{\beta}_0]E[\hat{\beta}_1] = E\left[\left(\overline{y} - \hat{\beta}_1\overline{x}\right)\frac{S_{xy}}{S_{xx}^2}\right] - \beta_0\beta_1. \tag{15}$$

Centrémonos pues en el primer término, que se puede escribir como

$$E\left[\left(\overline{y} - \hat{\beta}_1\overline{x}\right)\hat{\beta}_1\right] = E[\overline{y}\hat{\beta}_1] - E[\hat{\beta}_1^2\overline{x}] = E[\overline{y}]E[\hat{\beta}_1] - E[\hat{\beta}_1^2]E[\overline{x}] = (\beta_0\beta_1 + \beta_1^2\overline{x}) - (Var(\hat{\beta}_1) + \beta_1^2)\overline{x}, \tag{16}$$

conque sustituyendo de nuevo en la expresión original da lugar a

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\overline{x}Var(\hat{\beta}_1). \tag{17}$$

Para obtener una expresión de $\hat{\beta}_1$, desarrollemos en primer lugar su expresión para escribirla de una manera más conveniente:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})(\beta_0 + \beta_1 x_i + \epsilon_i - (\beta_0 + \beta_1\overline{x}))}{\sum_{i=1}^n (x_i - \overline{x})^2} = \tag{18}$$

$$= \frac{\beta_1 \sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{i=1}^n (x_i - \overline{x})\epsilon_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \overline{x})\epsilon_i}{\sum_{i=1}^n (x_i - \overline{x})^2}. \tag{19}$$

Aplicándole la varianza, recordando que $Var(a) = 0$, $a \in \mathbb{R}$,

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum_{i=1}^n (x_i - \overline{x})\epsilon_i}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) = \frac{1}{(\sum_{i=1}^n (x_i - \overline{x})^2)^2} Var\left(\sum_{i=1}^n (x_i - \overline{x})\epsilon_i\right). \tag{20}$$

En la segunda igualdad se ha podido extraer el denominador por ser un término constante, tal que $Var(ax) = a^2 Var(x)$, con $a \in \mathbb{R}$. Para proceder, qué es la varianza que nos acaba de aparecer, teniendo en cuenta que la varianza de términos independientes es la suma de sus varianzas:

$$Var\left(\sum_{i=1}^n (x_i - \overline{x})\epsilon_i\right) = \sum_{i=1}^n Var\left((x_i - \overline{x})\epsilon_i\right) = \sum_{i=1}^n (x_i - \overline{x})^2 Var(\epsilon_i). \tag{21}$$

Por definición, $Var(\epsilon_i) = \sigma^2$, conque el resultado será

$$Var(\hat{\beta}_1) = \frac{1}{(\sum_{i=1}^n (x_i - \overline{x})^2)^2} \sigma^2 \sum_{i=1}^n (x_i - \overline{x})^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2 \cdot \frac{n}{n}} = \frac{\sigma^2}{n S_{xx}^2}. \tag{22}$$

Recopilando todo, la solución será:

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\overline{x} Var(\hat{\beta}_1) = -\overline{x} \frac{\sigma^2/n}{S_{xx}^2}. \tag{23}$$

Así, hemos demostrado este segundo apartado.

c) Verifica que los residuos satisfacen las dos restricciones siguientes:

$$\sum_{i=1}^n e_i = 0 \qquad y \qquad \sum_{i=1}^n e_i x_i = 0. \tag{24}$$

Los residuos, por definición, vienen dados por la diferencia entre el valor observado y el previsto,

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i. \tag{25}$$

Con (11) podemos simplificar la ecuación anterior de la siguiente manera:

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - (\overline{y} - \hat{\beta}_1 \overline{x}) - \hat{\beta}_1 x_i) = n\overline{y} - (n\overline{y} - n\hat{\beta}_1 \overline{x}) - \hat{\beta}_1 n \overline{x} = 0, \tag{26}$$

comprobándose así la primera restricción. Se ha usado que $\sum_{i=1}^n 1 = n$ y que $n\overline{y} = \sum_{i=1}^n y_i$.

Procediéndose de una manera análoga, se puede mostrar la segunda restricción:

$$\sum_{i=1}^n e_i x_i = \sum_{i=1}^n (y_i - \hat{y}_i) x_i = \sum_{i=1}^n y_i x_i - \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i = \sum_{i=1}^n y_i x_i - (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \tag{27}$$

2 Ejercicio (1.5 puntos)

Para el modelo de regresión cuadrática

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon, \tag{28}$$

determina los estimadores mínimo cuadráticos de los parámetros a partir de una muestra de tamaño n. Investiga y explica qué órdenes de R usarías para resolver un problema de regresión cuadrática, inventando una muestra.

Para hallar los estimadores por mínimo cuadráticos de una muestra de tamaño n, se elegirán aquellos que minimicen la suma de los cuadrados de los residuos, escrita como

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2))^2. \tag{29}$$

Calculemos ahora las distintas derivadas:

$$\frac{\partial SS_E}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2)) = 0 \rightarrow \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} - \hat{\beta}_2 \overline{x}^2, \tag{30}$$

con $\overline{x}^2 = \sum_{i=1}^{n} x_i^2/n$.

$$\frac{\partial SS_E}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2))x_i = 0. \tag{31}$$

Siguiendo un proceso análogo al realizado en (7),

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}^2} - \hat{\beta}_2 \sum_{i=1}^{n}(x_i^3 - x_i \overline{x}^2) = \frac{S_{xy}}{S_{xx}^2} - \hat{\beta}_2 S_{xxx}. \tag{32}$$

$$\frac{\partial SS_E}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2))x_i^2 = 0, \tag{33}$$

se podría obtener una expresión para $\hat{\beta}_2$. Tras sustituir $\hat{\beta}_0$ y $\hat{\beta}_1$ en ella, se obtiene que

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(y_i x_i^2 - \overline{y} x_i^2 + S_{xy}(\overline{x} x_i^2 - x_i^3)/S_{xx}^2)}{\sum_{i=1}^{n}(S_{xxx}(\overline{x} x_i^3 - x_i^3) - x_i^2 \overline{x}^2 + x_i^4)}. \tag{34}$$

Expongamos a continuación un ejemplo sobre cómo resolver un problema de regresión cuadrática en R. Para ello, inventaremos una prueba en la que se relacionen dos variables independientes X y X^2, y una dependiente Y.

Implementando directamente el código en R:

```r
#Definimos en primer lugar la variable dependiente Y y la independiente X para representarlas y ver que se trata de una relacion cuadratica
datos <- data.frame(X = c(5, 10, 15, 20, 25, 30, 35, 45, 50, 55, 65),
          Y= c(10, 22, 35, 50, 65, 75, 80, 85, 70, 55, 40))

#Mostremos los datos en una tabla
datos

#Representamos ahora dichos datos en la grafica para apreciar la relacion cuadratica
plot(datos$X, datos$Y, pch = 16,
     xlab = "X", ylab = "Y",
     main = "Relacion entre X y Y")

#Agregamos a nuestros datos la otra variable independiente, X^2
datos$X2 <- datos$X^ 2

#Con un comando muy sencillo como es el siguiente ajustamos nuestra muestra con una regresion cuadratica
quadraticModel <- lm(Y ~ X + X2, data = datos)

# Mostramos el resumen de nuestro modelo
summary(quadraticModel)
```

Tras realizar esto, la función lm es capaz de decirnos si nuestros datos siguen una regresión cuadrática o no. Se ha obtenido un $R^2 = 95.67\,\%$, bastante cercano al $100\,\%$, asegurándonos que nuestro ajuste es bastante bueno, explicando un alto porcentaje de la variabilidad. Además, $\beta_0 = -19.07$ con un p-valor$= 0.0102 < 0.05$, indicando que dicho coeficiente es distinto de cero. Análogamente, $\beta_1 = 4.92$ con p-valor$= 1.45 \cdot 10^{-6} < 0.001$, habiendo una relación clara entre Y y X. De igual manera, $\beta_2 = -0.0623$ con p-valor$= 3.42 \cdot 10^{-6}$, siendo también un valor muy pequeño, asegurando que la relación presente entre Y y X^2 es cuadrática.

Recopilando todo lo anterior, el modelo de regresión cuadrática para el ejemplo presentado es de la forma:

$$Y = -19.07 + 4.92 \cdot X - 0.0623X^2 + \epsilon. \tag{35}$$

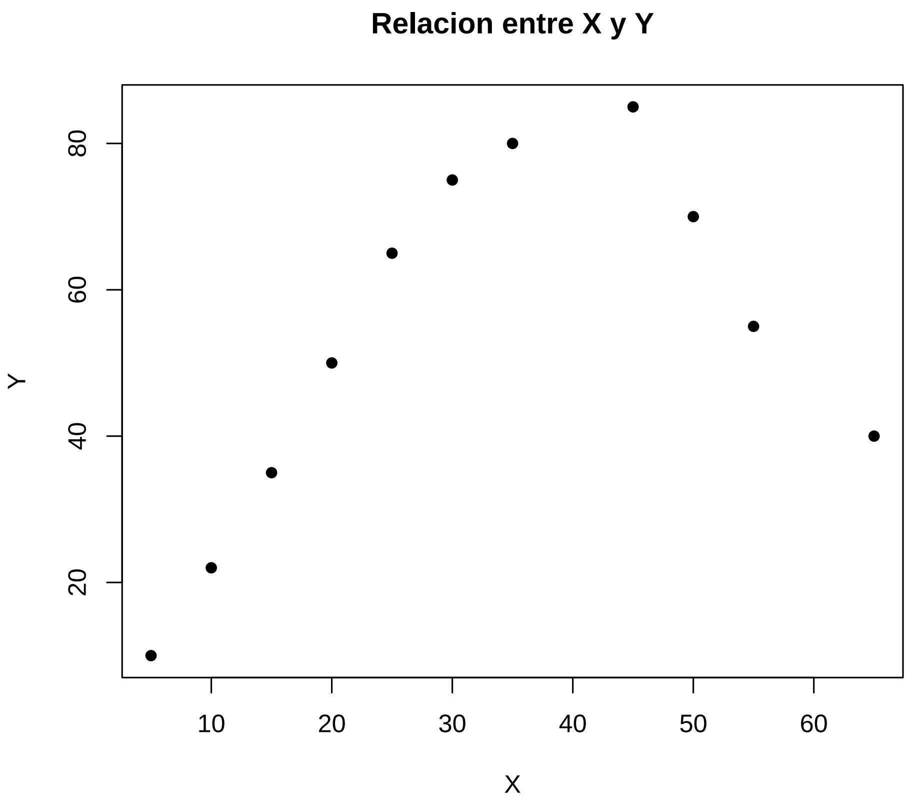


Figura 1: Gráfica obtenida en R al representar la relación entre X e Y.

3 Ejercicio (3 puntos)

Los datos recogidos en el fichero Sales.csv recogen información acerca de las ventas (en miles) a nivel nacional de una compañía de bolígrafos, y otras variables relacionadas, como número de anuncios en televisión, número de comerciales en el territorio y eficacia del distribuidor, según la escala 5= extraordinaria, 4=excelente, 3 = buena, 2 = media y 1 = pobre. Responder a las siguientes cuestiones:

a) Ajustar un modelo de regresión simple a los datos, para estimar las ventas en función del número de anuncios.

Para ello, primer se han tomado los datos del archivo Sales.csv y se han escrito en R. Una vez hecho esto, hacer la regresión lineal entre las ventas (variable dependiente) y el número de anuncios (variable independiente) se puede hacer directamente:

```
VENTAS <- c(260.3, 286.1, 279.4, 410.8, 438.2, 315.3, 565.1,
    570, 426.1, 315, 403.6, 220.5, 343.6, 644.6, 520.4, 329.5,
    426, 343.2, 450.4, 421.8, 245.6, 503.3, 375.7, 265.5, 620.6,
    450.5, 270.1, 368, 556.1, 570, 318.5, 260.2, 667, 618.3,
    525.3, 332.2, 393.2, 283.5, 376.2, 481.8)

NUMERO_ANUNCIOS <- c(5, 7, 6, 9, 12, 8, 11, 16, 13, 7, 8, 4, 9,
    17, 19, 9, 11, 8, 13, 14, 7, 16, 9, 5, 16, 18, 5, 7, 12, 13,
    8, 6, 16, 19, 17, 10, 12, 8, 10, 12)

NUMERO_COMERCIALES <- c(3, 5, 3, 4, 6, 3, 7, 8, 4, 3, 6, 4, 4,
    8, 7, 3, 6, 3, 5, 4, 4, 6, 3, 3, 6, 5, 3, 6, 6, 6, 4, 8, 8,
    7, 4, 4, 5, 3, 5, 5)

EFICACIA_DISTRIBUIDOR <- c(4, 2, 3, 4, 2, 4, 3, 2, 3, 5, 1, 1,
    3, 4, 2, 2, 4, 3, 4, 4, 4, 3, 3, 4, 3, 3, 2, 2, 1, 4, 3, 3,
    2, 2, 4, 3, 3, 4, 4, 2)

# Ajustamos el modelo de regresion lineal entre las ventas y el
    numero de anuncios
modelo <- lm(VENTAS ~ NUMERO_ANUNCIOS)

# Resumen del modelo
summary(modelo)
```

b) Calcular las pruebas t-Student sobre los coeficientes y analizar la bondad del modelo ajustado.

A partir de lo obteniedo con la última línea de código presentada, podemos obtener los valores t de $\beta_0 = 137.474$ y $\beta_1 = 25.353$ con $t = estimate/std.error$. Entonces, $t_{\beta_0} = 137.474/26.859 = 5.118$ y $t_{\beta_1} = 25.353/2.318 = 10.939$.

Como ambos $p-valor_{\beta_0} = 9.17\cdot 10^{-6}$ y $p-valor_{\beta_1} = 2.66\cdot 10^{-13}$ son menores que 0.05, podemos rechazar la hipótesis nula para ambos coeficientes. Es decir, ambos aportan al modelo, y deben ser no nulos para una descripción correcta del problema. Hay una relación importante entre el número de anuncios y las ventas.

Tanto $R^2 = 75.9\,\%$ como la desviación estándar de los residuos =61.6 indican que el problema no está descrito completamente, cosa que es clara porque todavía nos falta por añadir al problema otras variables independientes.

c) Predecir las ventas medias para los territorios con 18 anuncios y dar el intervalo de confianza al 95 %.

Para ello, escribimos el siguiente código en R, teniéndose en cuenta que sólo hay un territorio con 18 anuncios:

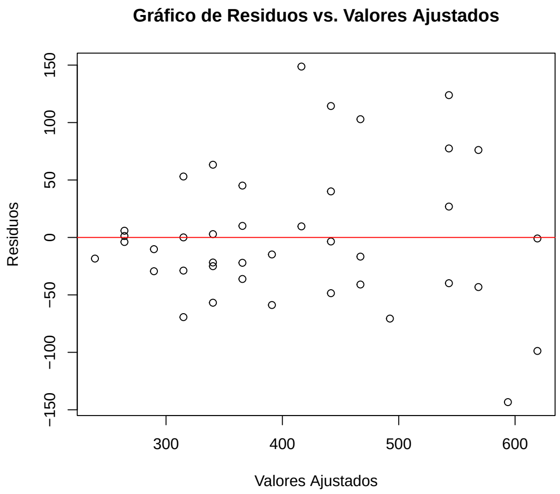
```
#Para el apartado c) definiremos un nuevo dato para el caso en
   el que el numero de anuncios sea 18
nuevo_dato <- data.frame(NUMERO_ANUNCIOS = 18)

#Predecimos las ventas y obtenemos el intervalo de confianza del
    95 por ciento
prediccion<-predict(modelo,nuevo_dato,interval = "confidence",
   level = 0.95)

# Mostrar la prediccion y el intervalo de confianza
prediccion
```

Las ventas medias da un valor de 593.8295, mientras que los límites inferiores y superiores del intervalo de confianza del 95 % son $[554.7143, 632.9447]$. Este intervalo es medianamente aceptable, sugiriendo, tal y como llevamos viendo del resto de apartados, que todavía necesitamos añadir las otras variables independientes al modelo para que sea más estrecho. A pesar de ello, no está nada mal el modelo así planteado, y tenemos un amplio margen pero tampoco tanto dentro de ciertos límites. A la hora de esperar unas ventas medias del 593.8295 y obtener valores menores es algo muy negativo que dicho rango sea tan amplio, y se desea que sea lo menor posible.

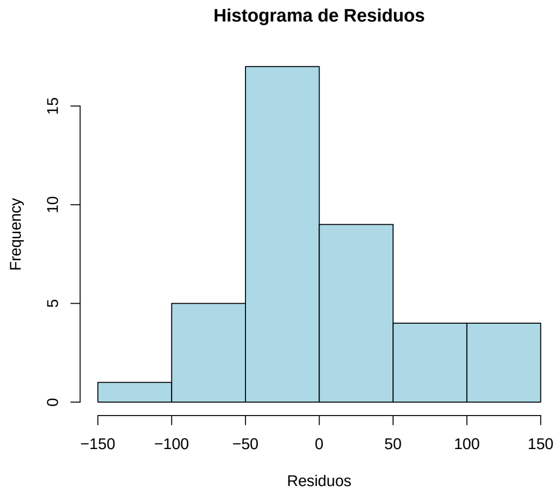
d) Observando los gráficos de los residuos, ¿qué puedes afirmar sobre la homocedasticidad, linealidad y normalidad de los datos?

```
# d) Calculamos los residuos
residuos <- residuals(modelo)

# Y tambien los valores ajustados. Primero veamos homocedastidad
    y linealidad
valores_ajustados <- fitted(modelo)

# Entonces, creamos un grafico de residuos vs. valores ajustados
plot(valores_ajustados, residuos,
     xlab = "Valores Ajustados",
     ylab = "Residuos",
     main = "Grafico de Residuos vs. Valores Ajustados")
abline(h = 0, col = "red") # Linea horizontal en 0
```

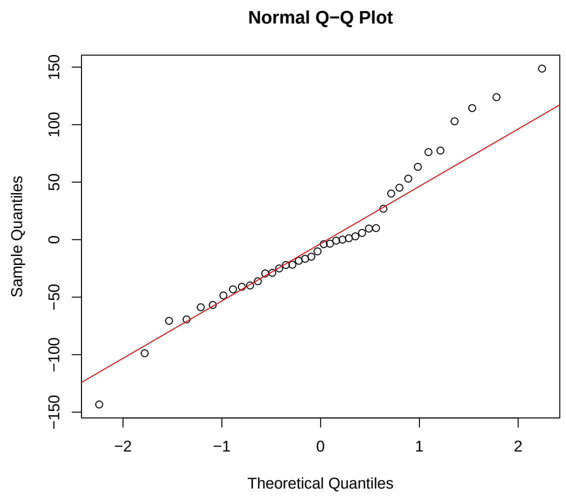
```r

#En segundo lugar veamos la normalidad creando un histrograma de los residuos
hist(residuos,
     breaks = 10,
     main = "Histograma de Residuos",
     xlab = "Residuos",
     col = "lightblue",
     border = "black")

#Y en ultimo lugar, veamos la normalidad de otra manera con un grafico Q-Q
qqnorm(residuos)
qqline(residuos, col = "red")  # Linea de referencia
```



(a)



(b)



(c)

Figura 2: En 2a se presentan los residuos frente a los valore ajustados. En 2b se presenta un histograma de los residuos. Y en 2c un gráfico Q-Q.

En la Fig. 2 se presentan los 3 gráficos obtenidos con el código presentado más arriba. Veamos analizando uno a uno qué es lo que nos transmite:

- La Fig. 2a busca patrones en la dispersión de los residuos. En la figura se aprecia como los residuos están distribuidos de manera aleatoria en torno al cero (línea horizontal roja), sugiriendo que hay homocedasticidad y linealidad.
- La Fig. 2b se parece a una distribución normal con forma de campana, sugiriendo que los residuos son normales, aunque hay presente una asimetría entre la derecha y la izquierda, indicándonos y verificándonos que el modelo está incompleto y no es perfecto.
- En la Fig. 2c, los puntos se alinean cerca de la línea roja, sugiriendo que los residuos parecen normales. Al igual que en el histograma, hay diferencias que no se pueden despreciar, en este caso para altos valores de las cantidades teóricas donde los puntos se alejan significativamente de la diagonal.

e) Ajustar un modelo de regresión múltiple a los datos, para estimar las ventas en función del resto de las variables.

La regresión múltiple aplicada a nuestro caso para estimar las ventas en función del resto de variables será:

```
datos<-data.frame(
  VENTAS, NUMERO_ANUNCIOS, NUMERO_COMERCIALES, EFICACIA_
     DISTRIBUIDOR)

g = lm(VENTAS ~ ., data = datos)
summary(g)
```

Y el resultado se puede ver en la Fig. 3

```
> summary(g)

Call:
lm(formula = VENTAS ~ ., data = datos)

Residuals:
     Min       1Q   Median       3Q      Max 
-124.908  -27.904   -4.111   27.487  102.140 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            82.2332    46.0125   1.787  0.08233 .  
NUMERO_ANUNCIOS        20.1636     2.5814   7.811 2.92e-09 ***
NUMERO_COMERCIALES     22.9662     7.0912   3.239  0.00258 ** 
EFICACIA_DISTRIBUIDOR  -0.6121     9.4334  -0.065  0.94862    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.61 on 36 degrees of freedom
Multiple R-squared:  0.8206,	Adjusted R-squared:  0.8056 
F-statistic: 54.88 on 3 and 36 DF,  p-value: 1.655e-13
```

Figura 3: Datos obtenidos en R al realizar la regresión múltiple para los datos del Ejercicio 3.

f) ¿Eliminarías alguna de las variables del modelo? ¿Por qué? Realiza un modelo de selección de variables stepwise.

Tal y como se ha indicado en la Fig. 3, la eficacia del distribuidor tiene un p-valor $0.94862 > 0.05$. Es decir, se acepta que esta variable no aporta nada al modelo, y que las ventas serán independientes de la eficacia del distribuidor. Entonces, eliminamos este parámetro de nuestro modelo:

```r
#EFICACIA_DISTRIBUIDOR tiene el mayor p-valor con 0.94862>0.05,
    conque lo eliminamos del modelo
g = update(g, . ~ . - EFICACIA_DISTRIBUIDOR)
summary(g)
```

Obteniéndose lo que se observa en la Fig. 4.

```
> summary(g)

Call:
lm(formula = VENTAS ~ NUMERO_ANUNCIOS + NUMERO_COMERCIALES, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-125.600  -28.281   -3.546   26.771  101.776

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          79.910     28.508   2.803  0.00801 **
NUMERO_ANUNCIOS      20.132      2.501   8.050 1.19e-09 ***
NUMERO_COMERCIALES   23.137      6.494   3.563  0.00103 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.87 on 37 degrees of freedom
Multiple R-squared:  0.8205,	Adjusted R-squared:  0.8108
F-statistic: 84.59 on 2 and 37 DF,  p-value: 1.579e-14
```

Figura 4: Datos obtenidos en R al realizar la regresión múltiple al quitar la eficacia del distribuidor del modelo para los datos del Ejercicio 3.

En este caso, tal y como se ve en la Fig. 4, todos los p-valores son menores que 0.05, siendo este el modelo definitivo que describe nuestro problema. En forma de ecuación, sería

$$VENTAS = 79.910 + 20.132 \cdot NUMERO\_ANUNCIOS + 23.137 \cdot NUMERO\_COMERCIALES. \quad (36)$$

Queda una $R^2 = 0.8205$, que no está nada mal. Así que hemos acabado el ejercicio, demostrando como nuestro modelo queda al final dependiente del número de anuncios y de comerciales.

4 Ejercicio (2 puntos)

a) Resolver el ejercicio anterior (apartado e) mediante una técnica de regularización Ridge. ¿Son los resultados muy diferentes?

Introducimos de nuevo todos los valores dados en el archivo .csv. Con esto, la intención será hallar el lamda óptimo que minimiza el error medio global en la validación cruzada. Para ello, se dividirán nuestras variables dependientes de las independientes, y se procederá con el ajuste del modelo Ridge. Con esto, procederemos con la búsqueda del lambda óptimo, estando asociado al valor mínimo de lambda en el modelo. Este se introducirá en el modelo, y se mostrarán sus coeficientes, parecidos a los ya obtenidos en la Fig. 3.

Se ha seguido un proceso análogo para el mayor valor de lambda, viéndose que se distancia mucho de lo esperado.

```r
datos<-data.frame(
  VENTAS, NUMERO_ANUNCIOS, NUMERO_COMERCIALES, EFICACIA_DISTRIBUIDOR)

#debemos descargar el siguiente paquete para usar tecnica de regularizacion Ridge
install.packages("glmnet")
library(glmnet)

# Separamos la variable dependiente y las independientes para que la tecnica de Ridge funcione correctamente
X <- as.matrix(datos[, -1])  # Variables independientes
Y <- datos$VENTAS              # Variable dependiente

# Ajustamos el modelo Ridge con alpha=0
modelo_ridge <- glmnet(X, Y, alpha = 0)
# Mostramos los coeficientes
print(coef(modelo_ridge))

#Esto nos da una gran cantidad de informacion. Vamos a buscar un valor optimo de lambda realizando una validacion cruzada
# Realizar validacion cruzada para seleccionar lambda
set.seed(123)  # Para reproducibilidad
cv_modelo_ridge <- cv.glmnet(X, Y, alpha = 0)

# Graficar el error de validacion cruzada
plot(cv_modelo_ridge)

# Obtener el valor optimo de lambda a partir de la lambda minima
lambda_optimo <- cv_modelo_ridge$lambda.min
print(lambda_optimo)

```

```r
# Ajustar el modelo final con el lambda optimo 1
modelo_final1 <- glmnet(X, Y, alpha = 0, lambda = lambda_optimo)

# Mostrar los coeficientes del modelo final
print(coef(modelo_final1))

#Se puede ver que si se toma el mayor valor de lambda, el modelo
    se distanciara mucho
lambda_maximo <- cv_modelo_ridge$lambda.1se
print(lambda_maximo)
modelo_final2 <- glmnet(X, Y, alpha = 0, lambda = lambda_maximo)
print(coef(modelo_final2))
```

```
> print(coef(modelo_final1))
4 x 1 sparse Matrix of class "dgCMatrix"
                              s0
(Intercept)             98.06115
NUMERO_ANUNCIOS         18.42528
NUMERO_COMERCIALES      23.55964
EFICACIA_DISTRIBUIDOR   -0.60433
```

(a)

```
> print(coef(modelo_final2))
4 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)             175.912994
NUMERO_ANUNCIOS          12.747725
NUMERO_COMERCIALES       21.039819
EFICACIA_DISTRIBUIDOR    -1.990597
```

(b)

Figura 5: Datos obtenidos en R al realizar la técnica de regularización Ridge para los datos del Ejercicio 3 con el valor óptimo de lambda (=10.65432) en 5a (se obtienen coeficientes análogos a los ya obtenidos en la Fig. 3) y con el valor máximo de lambda (=75.16411) en 5b.

Efectivamente, con el valor óptimo de lambda hemos obtenido unos coeficientes análogos tanto en la Fig. 3 como en la Fig. 5a.

b) Usando la técnica Lasso, ¿se seleccionan las mismas variables que en el apartado f)?

Se debe seguir un proceso análogo al anterior, pero esta vez tomando alpha=1 a la hora de usar la función glmnet. Con ello, todo se puede realizar prácticamente igual para obtener los coeficientes del modelo final obtenido con la técnica Lasso:

```r
#b) El modelo Lasso es analogo al anterior pero con alpha=1:
modelo_lasso <- glmnet(X, Y, alpha = 1)
print(coef(modelo_lasso))

# Realizamos de nuevo la validacion cruzada para seleccionar
    lambda optimo
set.seed(123)  # Para reproducibilidad
cv_modelo_lasso <- cv.glmnet(X, Y, alpha = 1)

# Graficamos el error de la validacion cruzada
plot(cv_modelo_lasso)

# Obtengamos ahora el valor optimo de lambda
lambda_optimo <- cv_modelo_lasso$lambda.min
print(lambda_optimo)

# Ajustamos el modelo final con el lambda optimo
```

```
modelo_final_lasso <- glmnet(X, Y, alpha = 1, lambda = lambda_
    optimo)

# Mostramos los coeficientes del modelo final
coeficientes_finales <- coef(modelo_final_lasso)
print(coeficientes_finales)
```

```
> print(coeficientes_finales)
4 x 1 sparse Matrix of class "dgCMatrix"
                            s0
(Intercept)           97.40958
NUMERO_ANUNCIOS       19.39101
NUMERO_COMERCIALES    21.20913
EFICACIA_DISTRIBUIDOR  .
```

Figura 6: Datos obtenidos en R al realizar la técnica de regularización Lasso para los datos del Ejercicio 3 con el valor óptimo de lambda (=4.945297) (se obtienen coeficientes análogos a los ya obtenidos en la Fig. 4)

Con este método se ha vuelto a quitar del modelo final la dependencia de la variable dependiente con la eficacia del distribuidor, corroborando que lo que habíamos hecho en el apartado 3 f) es correcto. La diferencia entre lo realizado en dicho apartado y en este se aprecia considerablemente en β_0, en el valor de la intercepción, véanse las Fig. 4 y 6. Esto se debe a la diferencia entre los métodos utilizados.

5 Ejercicio (2 puntos)

Busca un ejemplo sencillo para resolver un problema de regresión logística explicando y comentando las órdenes de R que usarías. Usa este ejemplo al mismo tiempo para explicar en qué consiste la regresión logística, incluyendo las fórmulas más relevantes

La regresión logística es un método estadístico que modela la relación entre una variable dependiente que suele ser binaria (toma valores 0 o/y 1) y una o más variables independientes. Su finalidad es predecir la probabilidad de que un evento ocurra (el resultado será SÍ ocurre o NO ocurre, por eso el comportamiento binario).

La probabilidad de que la variable dependiente Y sea igual a 1 en función de $X_1, X_2, ..., X_n$ variables independientes es:

$$P(Y = 1|X) = \frac{1}{1 + exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)]}. \tag{37}$$

Al igual que en los otros ejercicios, β_0 es el intercepto del modelo, y $\beta_1, \beta_2, ..., \beta_n$ los coeficientes asociados a cada variable independiente.

Vamos a realizar una regresión logística para predecir si un cliente comprará un producto (=1) o no (=0) en función de la edad, los ingresos y el número de visitas al sitio web.

```r
#Creeemos primero un data frame con los datos. Haremos todos ellos aleatorios
set.seed(123)  # Para reproducibilidad
clientes <- data.frame(
  edad = sample(18:65, 100, replace = TRUE),  # Edad entre 18 y 65
  ingreso = sample(20000:100000, 100, replace = TRUE),  # Ingreso entre 20,000 y 100,000
  visitas = sample(1:20, 100, replace = TRUE),  # Numero de visitas al sitio web
  compra = sample(0:1, 100, replace = TRUE)  # 0 = no compra, 1 = compra
)
clientes

# Ajustamos el modelo de regresion logistica. Ponemos binamial para indicar que la compra se trata de una variable binomial
modelo_logistico <- glm(compra ~ edad + ingreso + visitas, family = binomial, data = clientes)

# Resumen del modelo
summary(modelo_logistico)

```

```
#Vamos a predecir las probabilidades para nuevos datos
nuevos_clientes <- data.frame(
  edad = c(25, 40, 55),
  ingreso = c(30000, 60000, 80000),
  visitas = c(5, 10, 15)
)

predicciones <- predict(modelo_logistico, nuevos_clientes, type
   = "response")
print(predicciones)

# Clasificamos segun 0.5 tal que mayor que 0.5 si comprara, y
   sino, no
clases <- ifelse(predicciones > 0.5, 1, 0)
print(clases)
```

Con este simple código, somos capaces de predecir si se comprará o no, un modelo que puede ser muy útil dentro del mundo del marketing.

```
> summary(modelo_logistico)

Call:
glm(formula = compra ~ edad + ingreso + visitas, family = binomial,
    data = clientes)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.905e-01  9.650e-01   0.716    0.474
edad        -5.319e-03  1.616e-02  -0.329    0.742
ingreso     -2.443e-06  9.177e-06  -0.266    0.790
visitas     -3.468e-02  3.572e-02  -0.971    0.332
```

Figura 7: Datos obtenidos en R al realizar la regresión logística para unos datos aleatorios creados dentro de un intervalo en el Ejercicio 5.

Los coeficientes asociados a cada variable independiente dentro de nuestro modelo logístico se pueden apreciar en la Fig. 7. Véase que todos los coeficientes son negativos, indicándonos que a medida que aumenta la edad, ingreso o visitar, la probabilidad de que el cliente compre el producto disminuye, debido a la relación presente en la ecuación (37). No hay que darle mucha mayor importancia a este modelo para un caso real para sacar conclusiones sobre las ventas de la página web.