ESTADÍSTICA II
Grado en Matemáticas (2022/23)

Tema 3: REGRESIÓN LINEAL

3.1. La Comunidad de Madrid evalúa anualmente a los alumnos de sexto de primaria de todos los colegios sobre varias materias. Con las notas obtenidas por los colegios en los años 2009 y 2010 (fuente: diario *El País*) se ha ajustado el modelo de regresión simple:

$$\texttt{Nota2010} = \beta_0 + \beta_1 \texttt{Nota2009} + \epsilon,$$

en el que se supone que la variable de error ϵ satisface las hipótesis habituales. Los resultados obtenidos con `R` fueron los siguientes:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.40698   0.18832  7.471 1.51e-13
nota09      0.61060   0.02817 21.676 < 2e-16
---

Residual standard error: 1.016 on 1220 degrees of freedom
Multiple R-squared: 0.278,   Adjusted R-squared: 0.2774
F-statistic: 469.8 on 1 and 1220 DF, p-value: < 2.2e-16
```

También se sabe que en 2009 la nota media de todos los colegios fue 6.60 y la desviación típica fue 1.03 mientras que en 2010 la media y la desviación típica fueron 5.44 y 1.19, respectivamente.
Indicación: En este problema desviación típica se refiere a $s_{n-1} = (\sum_{i=1}^n (x_i - \bar{x})^2/(n-1))^{1/2}$.

a) ¿Se puede afirmar a nivel $\alpha = 0.05$ que existe relación lineal entre la nota de 2009 y la de 2010? Calcula el coeficiente de correlación lineal entre las notas de ambos años.
b) Calcula un intervalo de confianza de nivel 95 % para el parámetro β_1 del modelo.
c) Calcula, a partir de los datos anteriores, un intervalo de confianza de nivel 95 % para la nota media en 2010 de los colegios que obtuvieron un 7 en 2009.

3.2. Dada una muestra de 10 observaciones, se ha ajustado un modelo de regresión simple por mínimos cuadrados, resultando

$$\hat{y}_i = 1 + 3x_i, \quad R^2 = 0.9, \quad s_R^2 = 2.$$

Calcula un intervalo de confianza para la pendiente de la recta con un nivel de confianza 0.95. ¿Podemos rechazar, con un nivel de significación de 0.05, la hipótesis nula de que la variable x no influye linealmente en la variable Y?

3.3. Supongamos que la muestra $(x_1, Y_1), \ldots, (x_n, Y_n)$ procede de un modelo de regresión lineal simple en el que se verifican las hipótesis habituales. Consideramos el siguiente estimador de la pendiente del modelo (se supone $x_1 \neq \bar{x}$):

$$\tilde{\beta}_1 = \frac{Y_1 - \bar{Y}}{x_1 - \bar{x}}.$$

a) ¿Es $\tilde{\beta}_1$ un estimador insesgado?
b) Calcula la varianza de $\tilde{\beta}_1$.
c) Supongamos que la varianza de los errores del modelo, σ^2, es un parámetro conocido. Escribe la fórmula de un intervalo de confianza de nivel $1-\alpha$ para β_1 cuyo centro sea el estimador $\tilde{\beta}_1$.

3.4. Se considera el siguiente modelo de regresión lineal simple *a través del origen* (*simple linear regression through the origin*):

$$Y_i = \beta_1 x_i + \epsilon_i, \text{ donde } \epsilon_i \sim N(0, \sigma^2) \text{ independientes}, \quad i = 1, \ldots, n.$$

a) Calcula el estimador de mínimos cuadrados de β_1 y deduce su distribución.
b) Sean e_i, $i = 1, \ldots, n$ los residuos del modelo. Comprueba si se cumplen o no las siguientes propiedades: $\sum_{i=1}^n e_i = 0$ y $\sum_{i=1}^n e_i x_i = 0$.
c) Si la varianza de los errores σ^2 es conocida, deduce la fórmula de un intervalo de confianza de nivel $1-\alpha$ para el parámetro β_1.

3.5. En el modelo del problema anterior supongamos que $x_i > 0$ y que $\mathbb{V}(\epsilon_i) = \sigma^2 x_i^2$, es decir, no se cumple la hipótesis de homocedasticidad. Calcula en este caso la esperanza y la varianza del estimador de mínimos cuadrados $\hat{\beta}_1$. Consideremos ahora el estimador alternativo $\tilde{\beta}_1$ que se obtiene al minimizar la expresión $\sum_{i=1}^n w_i (y_i - \beta_1 x_i)^2$, donde $w_i = 1/x_i^2$. A $\tilde{\beta}_1$ se le llama *estimador de mínimos cuadrados ponderados* (*weighted least squares estimator*). Calcula una fórmula explícita para $\tilde{\beta}_1$ y, a partir de ella, deduce su esperanza y su varianza. Compara los estimadores $\widehat{\beta}_1$ y $\tilde{\beta}_1$. ¿Cuál es mejor?

3.6. Para estimar con R la regresión a través del origen utilizaremos los datos `cars`, ya cargados en R, y estudiaremos la distancia de frenado (`dist`) en función de la velocidad (`speed`).

```
data("cars")
cars.lm <- lm(dist ~ speed, data = cars)
```

Podemos imponer que la ordenada β_0 sea igual a 0 de dos maneras: añadiendo un 0 como si fuera un regresor o poniendo -1 después del regresor:

```
cars.lm2 <- lm(dist ~ 0 + speed, data = cars)
summary(cars.lm2)
cars.lm3 <- lm(dist ~ speed -1, data = cars)
summary(cars.lm3)
```

Compara el modelo de regresión estimado con y sin β_0.

3.7. Supongamos que cierta variable respuesta Y depende linealmente de dos variables regresoras x_1 y x_2, de manera que se satisface el modelo:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \ldots, n,$$

donde los errores ϵ_i cumplen las hipótesis habituales. Se ajusta por mínimos cuadrados el modelo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$, sin tener en cuenta la segunda variable regresora. Demuestra que el estimador $\hat{\beta}_1$ es, en general, sesgado y determina bajo qué condiciones se anula el sesgo.

3.8. Se considera el siguiente modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n. \tag{1}$$

Se dispone de $n = 20$ observaciones con las que se ajustan todos los posibles submodelos del modelo (1), obteniéndose para cada uno de ellos las siguientes sumas de cuadrados de los errores (todos los submodelos incluyen un término independiente).

Variables incluidas en el modelo	RSS	Variables incluidas en el modelo	RSS
Sólo término independiente	42644.00	x_1 y x_2	7713.13
x_1	8352.28	x_1 y x_3	762.55
x_2	36253.69	x_2 **y** x_3	**32700.17**
x_3	36606.19	x_1,x_2 y x_3	761.41

(**Ejemplo en negrita:** Para el modelo ajustado $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$, la suma de cuadrados de los errores es 32700.17).

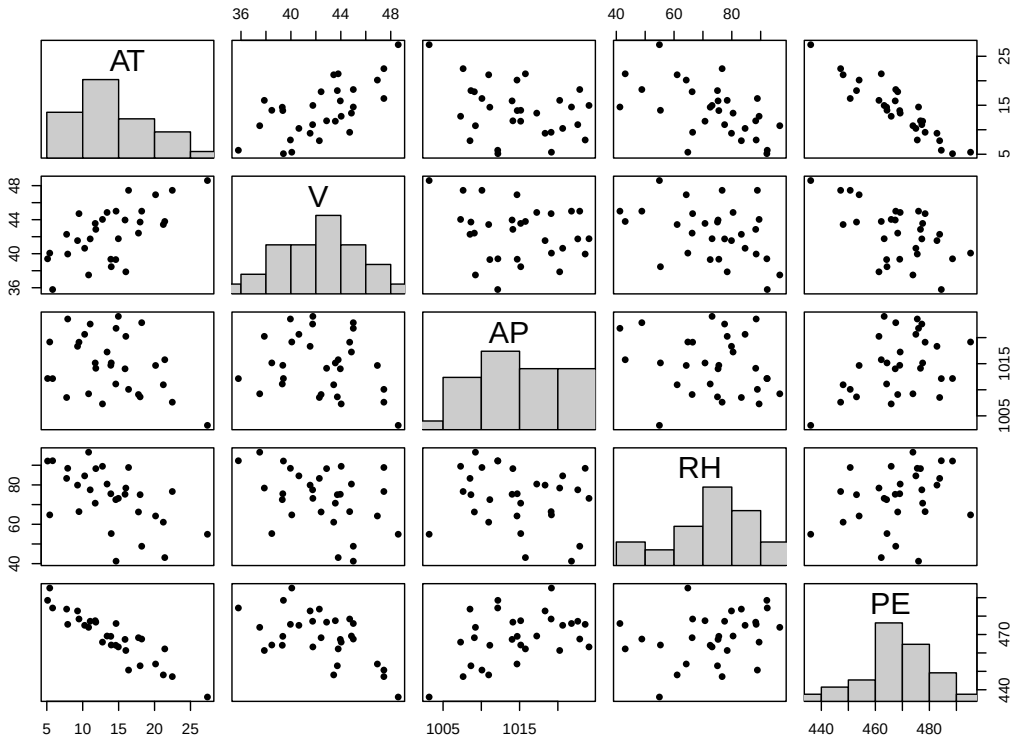
a) Calcula la tabla de análisis de la varianza para el modelo (1) y contrasta a nivel $\alpha = 0.05$ la hipótesis nula $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

b) En el modelo (1), contrasta a nivel $\alpha = 0.05$ las dos hipótesis nulas siguientes:

- $H_0 : \beta_2 = 0$
- $H_0 : \beta_1 = \beta_3 = 0$

c) Calcula el coeficiente de correlación entre la variable respuesta y la primera variable regresora sabiendo que es positivo.

3.9. Se han recogido datos de una central eléctrica de ciclo combinado. Las variables observadas son la temperatura ambiente (AT), la presión ambiente (AP), la humedad relativa (RH) y la presión del vapor V, con el objetivo de utilizarlas para predecir la producción de energía eléctrica (PE) de la planta. Al analizar los datos se han obtenido los siguientes resultados:



```
reg0 <- lm(PE~1,data=Datos3)
reg4 = lm(PE~AT+V+AP+RH,data=Datos3)
summary(reg4)
    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept) 542.06667 179.43412 3.021 0.005741
    AT          -2.74434   0.23845 -11.509 1.75e-11
    V            0.07701   0.32341  @@@@@ @@@@@@@@
    AP          -0.01714   0.17160 -0.100 0.921212
    RH          -0.28588   0.07394 -3.867 0.000697
    ---
    Residual standard error: 4.233 on 25 degrees of freedom
    Multiple R-squared: 0.9133,      Adjusted R-squared: @@@@
    F-statistic: @@@@ on @@ and @@ DF, p-value: @@@@

anova(reg4)
    Analysis of Variance Table
    Response: PE
              Df Sum Sq Mean Sq F value  Pr(>F)
    AT         1 4371.6 4371.6 244.0079 2.091e-14
    V          1    3.9    3.9   0.2155 0.6464988
    AP         1   75.6   75.6   4.2189 0.0505741
    RH         1  267.8  267.8  14.9502 0.0006974
    Residuals 25 447.9   17.9

reg2 = lm(PE~AT+RH,data=Datos3)
anova(reg2,reg4)
    Analysis of Variance Table
    Model 1: PE ~ AT + RH
    Model 2: PE ~ AT + V + AP + RH
      Res.Df   RSS Df Sum of Sq    F Pr(>F)
    1     27 449.13
    2     25 @@@@@@ 2    1.2311 @@@@@ 0.9663
```

a) Determinar qué modelo de regresión se ha ajustado en `reg4`. Plantear el contraste correspondiente al `F-statistic`, calculando el correspondiente estadístico del contraste y determinando la región de rechazo. ¿Es el p-valor menor que 0.01? ¿Es razonable rechazar la hipótesis nula?

b) ¿Cuál es la suma de cuadrados residual con `reg4`? ¿Y la suma de cuadrados total?

c) Calcular el coeficiente de determinación ajustado en `reg4`.

d) Para `reg4`, a nivel $\alpha = 0.05$, ¿puede afirmarse que la producción de energía eléctrica disminuye cuando `V` aumenta?

e) ¿Qué modelo de regresión se ajusta en `reg0`? ¿Cuál es la expresión de la respuesta prevista ($\hat{y}$) con este modelo? ¿Cuál es la suma de cuadrados residual con `reg0`? ¿Y la suma de cuadrados total?

f) ¿Qué contraste se resuelve con el comando `anova(reg2,reg4)`? Calcular los valores de `RSS` y `F` sustituidos por los símbolos `@`. ¿Qué nos permite concluir ese contraste?

g) Para el modelo ajustado a continuación, calcular un intervalo de confianza (al 95 %) para el valor esperado de PE cuando AT=15, conociendo el valor promedio $\overline{\text{AT}} = 13.9837$ y la varianza muestral (insesgada) $s^2_{\text{AT}} = 29.0013$.

```
reg1 <- lm(PE~AT,data=Datos3)
```

```
summary(reg1)

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept) 500.4293   2.7477  182.13 < 2e-16
    AT           -2.2799    0.1838 -12.41 6.76e-13

    Residual standard error: 5.329 on 28 degrees of freedom
    Multiple R-squared: 0.8461,      Adjusted R-squared: 0.8406
    F-statistic: 153.9 on 1 and 28 DF, p-value: 6.755e-13
```

3.10. Se desea estudiar la esperanza de vida `Y` en una serie de países como función de la tasa de natalidad `nat`, la tasa de mortalidad infantil `mortinf` y el logaritmo del producto nacional bruto `lpnb`. Para ajustar el modelo

$$Y_i = \beta_0 + \beta_1 \texttt{nat}_i + \beta_2 \texttt{mortinf}_i + \beta_3 \texttt{lpnb}_i + \epsilon_i,$$

donde los errores ϵ_i son v.a.i.i.d. $\mathrm{N}(0, \sigma^2)$, se ha utilizado el programa `R` con los resultados siguientes:

```
reg = lm(Y~nat+mortinf+lpnb)
summary(reg)
Call:
lm(formula = Y ~ nat + mortinf + lpnb)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 69.24045  2.90253 23.855 < 2e-16
nat         -0.17572   0.04244 -4.140   8e-05
mortinf     -0.14086   0.01370 -10.284 < 2e-16
lpnb         0.98901   0.29404  3.363  0.00115
---
Residual standard error: 2.788 on 87 degrees of freedom
Multiple R-Squared: 0.9303,  Adjusted R-squared: 0.9279
F-statistic: 386.9 on 3 and 87 DF, p-value: < 2.2e-16

anova(reg)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value Pr(>F)
nat        1 7602.7 7602.7 977.798 < 2.2e-16
mortinf    1 1334.2 1334.2 171.599 < 2.2e-16
lpnb       1   88.0   88.0  11.313 0.001146
Residuals 87 676.5    7.8
```

a) ¿De cuántos países consta la muestra utilizada?
b) ¿Cuál es la suma de cuadrados del modelo de regresión (MSS) que se utiliza para medir la variabilidad explicada por las tres variables regresoras?
c) ¿Cuánto vale la varianza muestral de la variable respuesta, $\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$?
d) Contrasta a nivel $\alpha = 0.05$ la hipótesis nula $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.
e) Determina cuál es la hipótesis nula y la alternativa correspondiente a cada uno de los tres estadísticos `F` que aparecen en la tabla de análisis de la varianza anterior.

3.11. Con el fin de evaluar el trabajo de los directores de los 30 departamentos de una gran empresa, se llevó a cabo una encuesta a los empleados a su cargo en la que se les pidió que valoraran varias afirmaciones con una nota de 1 (máximo acuerdo) a 5 (máximo desacuerdo). Algunas de las variables eran: Y, el trabajo del director es en general satisfactorio; x_1, el director gestiona correctamente las quejas de los empleados; x_2, el director trata equitativamente a los empleados; x_3, la asignación del trabajo es tal que los empleados pueden aprender cosas nuevas con frecuencia. El vector $(Y_i, x_{i1}, x_{i2}, x_{i3})$ contiene la suma de puntos de las respuestas en el departamento i, donde $i = 1, \ldots, 30$. Con estos datos se ajustó con `R` el modelo:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

donde las perturbaciones ϵ_i verifican las hipótesis habituales. Los resultados fueron los siguientes:

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.2583    7.3183   1.538   0.1360
x1           0.6824    0.1288   5.296 1.54e-05
x2          -0.1033    0.1293  -0.799   0.4318
x3           0.2380    0.1394   1.707   0.0997
---
Residual standard error: 6.863 on 26 degrees of freedom
Multiple R-squared: 0.715,   Adjusted R-squared: 0.6821
F-statistic: AAA on BBB and CCC DF, p-value: 2.936e-07
```

a) Calcula un intervalo de confianza de nivel 0.95 para el parámetro β_3. Contrasta la hipótesis $H_0 : \beta_3 \leq 0$.
b) Determina el valor de `AAA`, `BBB` y `CCC` en la última línea de la salida anterior. ¿A qué hipótesis nula corresponde el p-valor que aparece en esta última línea?

3.12. Tres vehículos se encuentran situados en los puntos $0 < \beta_1 < \beta_2 < \beta_3$ de una carretera recta. Para estimar la posición de los vehículos se toman las siguientes medidas (todas ellas sujetas a errores aleatorios de medición independientes con distribución normal de media 0 y varianza σ^2):

- Desde el punto 0 medimos las distancias a los tres vehículos, denotándolas Y_1, Y_2 e Y_3.
- Nos trasladamos al primer vehículo y medimos las distancias a los otros dos, dando dos nuevas medidas Y_4 e Y_5.
- Nos trasladamos al segundo vehículo y medimos la distancia al tercero, dando una medida adicional, Y_6.

a) Expresa el problema de estimación como un modelo de regresión múltiple indicando claramente cuál es la matriz de diseño.
b) Calcula la distribución del estimador de mínimos cuadrados del vector de posiciones $(\beta_1, \beta_2, \beta_3)'$.
c) Se desea calcular un intervalo de confianza de nivel 95 % para la posición del primer vehículo β_1 a partir de 6 medidas (obtenidas de acuerdo con el método descrito anteriormente) para las que la varianza residual resultó ser $s_R^2 = 2$. ¿Cuál es el margen de error del intervalo?

3.13. Sean Y_1, Y_2 e Y_3 tres variables aleatorias independientes con distribución normal y varianza σ^2. Supongamos que μ es la media de Y_1, λ es la media de Y_2 y $\lambda + \mu$ es la media de Y_3, donde $\lambda, \mu \in \mathbb{R}$.

a) Demuestra que el vector $\mathbf{Y} = (Y_1, Y_2, Y_3)'$ satisface el modelo de regresión múltiple $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Para ello, determina la matriz de diseño $\mathbb{X}$, el vector de parámetros $\boldsymbol{\beta}$ y la distribución de las variables de error $\boldsymbol{\epsilon}$.
b) Calcula los estimadores de máxima verosimilitud (equivalentemente, de mínimos cuadrados) de λ y μ.
c) Calcula la distribución del vector $(\hat{\lambda}, \hat{\mu})'$, formado por los estimadores calculados en el apartado anterior.

3.14. Durante la producción de celulosa de madera en la industria papelera la lignina de la madera se separa de la celulosa, y ésta conforma las fibras del papel. El licor negro es la combinación del residuo de la lignina con agua y los químicos usados para extracción de la lignina. El vertido de licor negro tiene un pH elevado que contamina los acuíferos. En la tabla siguiente vemos el pH de medidas efectuadas en tres tubos piezométricos, cada uno de los cuales está conectado a un acuífero.

Tubo (i)	pH						$\sum_j y_{ij}$	$\bar{y}_{i\cdot}$	s_i^2
1	7.0	7.2	7.5	7.7	8.7	7.8	45.9	7.650	0.355
2	6.3	6.9	7.0	6.4	6.8	6.7	40.1	6.683	0.078
3	8.4	7.6	7.5	7.4	9.3	9.0	49.2	8.200	0.676

a) Con el objetivo de contrastar si el nivel medio de pH es igual en los tres tubos piezométricos, se determina con R la tabla ANOVA para los datos del enunciado. Se reproducen a continuación las correspondientes salidas del programa, de las que se han borrado seis resultados (sustituidos por `!!`).

```
Datos = read.table("pHtubo.txt", header=TRUE)
pH <- Datos$pH
Tubo = factor(Datos$Tubo)
resultado = aov(pH ~ Tubo)
summary(resultado)
            Df Sum Sq Mean Sq F value Pr(>F)
Tubo        !!     !!      !!   9.572 0.00209 **
Residuals   !!     !!      !!
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Determinar los valores borrados de la tabla. Contrastar si el nivel medio de pH depende del tubo piezométrico, indicando claramente el contraste de hipótesis realizado.

b) Escribe la expresión del modelo unifactorial y estima los parámetros del modelo.