ESTADÍSTICA II **Curso 2016/2017**
Grado en Matemáticas

Relación 3 de problemas

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 verifica 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 cuasidesviación típica fue 1,03 mientras que en 2010 la media y la cuasidesviación típica fueron 5,44 y 1,19, respectivamente.

(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.

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

$$\widehat{Y_i} = 1 + 3x_i, \;\; R^2 = 0{,}9, \;\; S_R^2 = 2.$$

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. 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 $\tilde{\beta}_1$.

4. Se considera el siguiente modelo de regresión simple *a través del origen*:

$$Y_i = \beta_1 x_i + \epsilon_i, \;\; \epsilon_i \equiv N(0,\sigma^2) \text{ independientes}, \;\; i = 1,\dots,n.$$

(a) Calcula el estimador de mínimos cuadrados de β_1 y deduce su distribución.
(b) Sean e_i, $i = 1,\dots,n$ 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.

5. En el modelo del problema anterior supongamos que $x_i > 0$ y que $V(\epsilon_i) = \sigma^2 x_i^2$, es decir, no se cumple la hipótesis de homocedasticidad. Calcula en este caso la esperanza y la varianza del estimador de mínimos cuadrados $\widehat{\beta_1}$. 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 $w_i = 1/x_i^2$. 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? (A $\tilde{\beta}_1$ se le llama *estimador de mínimos cuadrados ponderados*).

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

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

donde los errores ϵ_i verifican las hipótesis habituales. Se ajusta por mínimos cuadrados el modelo $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i1}$, sin tener en cuenta la segunda variable regresora. Demuestra que el estimador $\widehat{\beta_1}$ es, en general, sesgado y determina bajo qué condiciones se anula el sesgo.

7. En el Ayuntamiento de Madrid se estudió hace unos años la conveniencia de instalar mamparas de protección acústica en una zona de la M-30. Un técnico del Ayuntamiento piensa que si el ruido afecta mucho a los habitantes de la zona esto debe reflejarse en los precios de las viviendas. Su idea es que el precio de una casa en esa zona (y) depende del número de metros cuadrados (x_1), del número de habitaciones (x_2) y de la contaminación acústica, medida en decibelios, (x_3). Para una muestra de 20 casas vendidas en los últimos tres meses, se estima el siguiente modelo:

$$\begin{array}{rcccccc} \hat{y}_i & = & 5970 & + & \underset{(2,55)}{22,35\, x_{i1}} & + & \underset{(1820)}{2701,1\, x_{i2}} \;-\; \underset{(15,4)}{67,6730\, x_{i3}} \end{array}$$

$$R^2 = 0,9843,$$

donde las desviaciones típicas (estimadas) de los estimadores de los coeficientes aparecen entre paréntesis.
(a) Calcula el efecto que tendría sobre el precio un descenso de 10 decibelios, si el resto de variables en el modelo permanecieran constantes.
(b) Contrasta con $\alpha = 0,05$ la hipótesis nula de que el número de habitaciones no influye en el precio.
(c) A nivel $\alpha = 0,05$, ¿puede afirmarse que la vivienda se encarece cuando disminuye la contaminación acústica?

(d) Contrasta con $\alpha = 0{,}05$ la hipótesis nula de que las tres variables no influyen conjuntamente en el precio.
(e) Estima el precio medio de las casas (no incluidas en la muestra) que tienen 100 metros cuadrados, dos habitaciones y una contaminación acústica de 40 decibelios.

8. Se ajusta el modelo de regresión $Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i$, a los datos $(x_{1,1}, x_{1,2}, Y_1) = (1, 2, 19)$, $(x_{2,1}, x_{2,2}, Y_2) = (2, 1, 13)$ y $(x_{3,1}, x_{3,2}, Y_3) = (0, 0, 16)$.

(a) Escribe la matriz de diseño X. Determina el subespacio vectorial $V \subset \mathbb{R}^3$ al que, de acuerdo con el modelo, pertenece el vector de medias de las respuestas (Y_1, Y_2, Y_3).
(b) Calcula el vector de valores ajustados $(\hat{Y}_1, \hat{Y}_2, \hat{Y}_3)$ y el vector de residuos (e_1, e_2, e_3).
(c) En este ejemplo se observa que $e_1 + e_2 + e_3 \neq 0$. ¿Cómo habría que modificar el modelo para que la suma de residuos se anule?

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

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

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	SCE	Variables incluidas en el modelo	SCE
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 $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3}$, 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.

10. A partir de una muestra de $n = 20$ observaciones se ha ajustado el modelo de regresión lineal simple $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ con los siguientes resultados:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.29016    1.66161   0.175    AAA
x            1.01450    0.03246  31.252   <2e-16

Residual standard error: 0.1717 on 18 degrees of freedom
Multiple R-squared: 0.9819,     Adjusted R-squared: 0.9809
F-statistic: BBB on 1 and 18 DF,  p-value: < 2.2e-16

> vcov(reg)
            (Intercept)     x
(Intercept)  2.761      -0.054
x           -0.054       0.001
```

(a) Determina si el p-valor AAA es mayor o menor que 0.1. Escribe la hipótesis nula a la que corresponde este p-valor y determina si esta hipótesis se rechaza o no a nivel $\alpha = 0{,}1$.
(b) Contrasta la hipótesis nula $H_0 : \beta_0 + \beta_1 = 2$ a nivel $\alpha = 0{,}05$.
(c) Calcula el valor BBB que se ha omitido en los resultados anteriores.

11. 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 de la regresión (SCR) que se utiliza para medir la variabilidad

explicada por las tres variables regresoras?
(c) ¿Cuánto vale la cuasivarianza 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.

12. Considera el modelo de regresión múltiple $Y = X\beta + \epsilon$, donde el vector de errores ϵ verifica las hipótesis habituales.

(a) Define el vector de valores ajustados $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)$ y calcula su distribución.
(b) En general, ¿son las variables $\hat{Y}_1, \ldots, \hat{Y}_n$ independientes? ¿Son idénticamente distribuidas?
(c) Calcula el valor de $\sum_{i=1}^{n} \mathrm{Var}(\hat{Y}_i)$ si el modelo incluye un término independiente y 3 variables regresoras.

13. 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 los errores aleatorios 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?

14. 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 dando 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?

15. 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 $Y = (Y_1, Y_2, Y_3)'$ verifica el modelo de regresión múltiple $Y = X\beta + \epsilon$. Para ello, determina la matriz de diseño X, el vector de parámetros β y la distribución de las variables de error ϵ.

(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.

16. La siguiente tabla contiene información sobre los resultados de un examen en cuatro grupos de una misma asignatura:

	Alumnos	Media	Cuasi-varianza
Grupo 1	104	4.99	4.19
Grupo 2	102	4.63	5.75
Grupo 3	69	4.53	5.15
Grupo 4	80	4.79	5.35

Se supone que se satisfacen las hipótesis del modelo unifactorial. Escribe la tabla de análisis de la varianza y contrasta la hipótesis de que las notas medias son iguales en los cuatro grupos, con un nivel de significación $\alpha = 0{,}05$.