Métodos avanzados en Estadística **Curso 2021/2022**
Máster en Ciencia de Datos

Ejercicios: regresión

1. Los datos del fichero Datos-geyser.txt corresponden al día de la observación (primera columna), el tiempo medido en minutos (segunda columna Y) y el tiempo hasta la siguiente erupción (tercera columna X) del geyser *Old Faithful* en el parque norteamericano de Yellowstone.

(a) Representa gráficamente los datos, junto con el estimador de Nadaraya-Watson de la función de regresión de Y sobre X.
(b) Representa gráficamente los datos, junto con el estimador localmente lineal de la función de regresión de Y sobre X.

2. 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?

3. 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) ¿Cuánto vale la suma de cuadrados 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.

4. 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 residuos (todos los submodelos incluyen un término independiente).

Variables incluidas en el modelo	SCR	Variables incluidas en el modelo	SCR
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 residuos 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:

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

5. 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 $(\beta_1, \beta_2, \beta_3)'$.

6. 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.

7. Demuestra que el estimador bootstrap ideal de la varianza de la pendiente de la recta de mínimos cuadrados en un modelo de regresión simple verifica:

$$\mathrm{Var}_{F_n}(\hat{\beta}_1^*) = \frac{n-2}{n} \frac{S_R^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

es decir, sólo difiere en un factor $(n-2)/n$ del estimador que se usa habitualmente.
INDICACIÓN: recuerda que la pendiente de la recta de mínimos cuadrados viene dada por:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i y_i,$$

donde $w_i = (x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2$.

8. El conjunto de datos star.txt corresponde a la temperatura y la intensidad de la luz en un conjunto de estrellas.

(a) Calcula la recta de mínimos cuadrados. Representa gráficamente la nube de puntos junto con la recta obtenida. Comenta el resultado.
(b) En lugar de mínimos cuadrados, otro criterio posible para ajustar una regresión es encontrar la recta que minimiza la **mediana** de los residuos al cuadrado. Esta recta se calcula con el comando `MASS::lmsreg` de `R`. Calcula la recta de mínima mediana de cuadrados para los datos de las estrellas. Representa gráficamente la nube de puntos junto con la recta obtenida.
(c) Usa el método bootstrap para calcular el error típico de la pendiente de la recta de mínima mediana de cuadrados para los datos de las estrellas. ¿Qué problema presenta el método bootstrap para este conjunto de datos concreto?

9. Genera aleatoriamente una variable regresora X y un vector aleatorio ϵ de longitud $n = 100$, con distribución normal estándar e independientes. Genera la variable respuesta de acuerdo con el modelo:

$$Y = X + X^2 + X^3 + \epsilon.$$

(a) Selecciona el modelo óptimo entre todos los submodelos que contienen como variables regresoras $X, X^2, X^3, \ldots, X^{10}$. ¿Cuál es el mejor modelo de acuerdo con los criterios C_p, BIC y R_a^2?

(b) Repite el apartado anterior usando el método iterativo hacia adelante.

(c) Aplica ahora lasso al modelo que incluye las variables regresoras $X, X^2, \ldots, X^{10}$. Selecciona el parámetro de regularización mediante validación cruzada y compara los resultados del ajuste con los de los apartados anteriores.

(d) Genera ahora las respuestas a partir del modelo

$$Y = X^7 + \epsilon$$

y aplica de nuevo el método lasso. Describe los resultados obtenidos.

10. Los datos fuel2001 del fichero `combustible.RData` (véase transparencias de clase) corresponden al consumo de combustible (y otras variables relacionadas) enlos estados de EE.UU. Se desea explicar la variable `FuelC` en función del resto de la información.

(a) Representa en un plano las dos primeras componentes principales de estos datos estandarizados (consulta la ayuda de `prcomp`). ¿Son suficientes estas dos componentes para explicar un alto porcentaje de la varianza?

(b) Ajusta el modelo completo con todas las variables. En este modelo completo, contrasta la hipótesis nula de que los coeficientes de las variables `Income`, `MPC` y `Tax` son simultáneamente iguales a cero.

(c) De acuerdo con el método iterativo hacia adelante y el criterio BIC, ¿cuál es el modelo óptimo?

(d) Ajusta el modelo usando lasso, con el parámetro de regularización seleccionado mediante validación cruzada.

(e) Ajusta el modelo usando ridge, con el parámetro de regularización seleccionado mediante validación cruzada.