Estadística II
Tercero del grado en Matemáticas, UAM, 2019-2020

Examen parcial 2, 18-12-2019

Apellidos, nombre ....

Ejercicio 1. Para analizar la longevidad Y (en años) de una cierta especie de tortuga marina, se seleccionan las siguientes cuatro variables regresoras:

- X_1, el peso de cada individuo, en kilogramos,
- X_2, el sexo de cada individuo (macho$=1$, hembra $=0$),
- X_3, la concentración de calcio en la sangre del individuo (medida en mg/dl),
- X_4, la salinidad de las aguas en las que viven (niveles de salinidad: 1, 2, 3 y 4).

Se propone el habitual modelo de regresión lineal múltiple para muestras de tamaño n:

$$Y_i = \beta_0 + \beta_1\, x_{i,1} + \beta_2\, x_{i,2} + \beta_3\, x_{i,3} + \beta_4\, x_{i,4} + \varepsilon_i, \quad \text{para } i = 1, \ldots, n,$$

donde las ε_i son variables normales independientes, de media 0 y varianza σ^2.

Se ha analizado una población de 15 tortugas, y se han obtenido las siguientes estimaciones:

- $\widehat{\beta}_0 = 25$,
- $\widehat{\beta}_1 = 0.4$, con desviación típica estimada de 0.04,
- $\widehat{\beta}_2 = 10$, con desviación típica estimada de 3,
- $\widehat{\beta}_3 = -0.8$, con desviación típica estimada de 0.3,
- $\widehat{\beta}_4 = -3$, con desviación típica estimada de 1.

(a) (1 punto) ¿Hay evidencia estadística suficiente como para afirmar que la concentración de calcio en la sangre influye (linealmente) en la longevidad? Argumenta calculando el p-valor de la muestra para el contraste adecuado.

(b) (1 punto) Se ha contrastado la hipótesis nula de que las cuatro variables **no** influyen conjuntamente en la longevidad de la tortuga. Se ha obtenido un p-valor del 0.7 %. ¿Que valor de R^2 tiene el modelo de regresión?

(c) (0.5 puntos) De una tortuga que pesa 85 kg, con 9 mg/dl de calcio en sangre, y que vive en aguas de salinidad 3, se ha predicho una longevidad de 52.8 años. La tortuga, ¿era macho o hembra?

Ejercicio 2. 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 $\mu+2\lambda$ es la media de Y_3, donde $\lambda, \mu \in \mathbb{R}$.

(a) (0.5 puntos) Comprueba que el vector $\mathbb{Y} = (Y_1, Y_2, Y_3)^\intercal$ verifica un modelo de regresión múltiple $\mathbb{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Para ello, determina la matriz de diseño X, el vector de parámetros $\boldsymbol{\beta}$ y la distribución del vector de variables de error $\boldsymbol{\epsilon}$.

(b) (1 punto) Calcula los estimadores $\hat{\lambda}$ y $\hat{\mu}$ de máxima verosimilitud (equivalentemente, de mínimos cuadrados) de λ y μ.

(c) (1 punto) Determina la distribución conjunta del vector $\hat{\boldsymbol{\beta}} = (\hat{\lambda}, \hat{\mu})^\intercal$, formado por los estimadores calculados en el apartado anterior.

Ejercicio 3. (2.5 puntos) Considera las dos siguientes funciones de densidad, correspondientes a la distribución de un par de variables (X_1, X_2) en dos poblaciones π_0 y π_1:

- $f_0(x_1, x_2) = \frac{e}{\pi(e-1)}\, e^{-(x_1^2+x_2^2)}$, para $x_1^2 + x_2^2 \leq 1$.
- $f_1(x_1, x_2)$ es la función de densidad de un vector bidimensional que se distribuye uniformemente en el el disco $D(\mathbf{0}, 8/9)$ (centrado en el origen, y de radio $8/9$).

Ponemos $p_0 = p_1$. Identifica las regiones (óptimas) R_0 y R_1 de clasificación en π_0 y π_1, respectivamente.

Ejercicio 4. (2.5 puntos) Suponemos que una variable respuesta Y depende linealmente de una única variable regresora X. La muestra va a ser de tamaño n, del tipo $(x_1, y_1), \ldots, (x_n, y_n)$.

Proponemos el siguiente modelo:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \text{para } i = 1, \ldots, n,$$

donde las ε_i son variables aleatorias idénticas e independientes, cada una de las cuales se distribuye como una uniforme en el intervalo $[-\sigma, \sigma]$.

Los parámetros del modelo son $\beta_0, \beta_1 \in \mathbb{R}$ y $\sigma > 0$.

Como estimador de β_1 elegimos el habitual $\widehat{\beta}_1$ de mínimos cuadrados.

a) (1 punto) Comprueba si $\widehat{\beta}_1$ es un estimador insesgado de β_1, y en caso contrario, calcula su sesgo.

b) (1 punto) Calcula la varianza de $\widehat{\beta}_1$.

b) (0.5 puntos) Supongamos que hay solo dos observaciones, a saber, $x_1 = 1$ y $x_2 = 3$, y que $\sigma = 1$. ¿Cuál es la distribución de $\widehat{\beta}_1$ en este caso?

1. Percentiles de la t de Student con 10 grados de libertad

α	0.1 %	0.2 %	0.3 %	0.4 %	0.5 %	0.6 %	0.7 %	0.8 %	0.9 %	1.0 %
$t_{\{10;\alpha\}}$	4.144	3.716	3.472	3.301	3.169	3.062	2.972	2.894	2.825	2.764
α	1.1 %	1.2 %	1.3 %	1.4 %	1.5 %	1.6 %	1.7 %	1.8 %	1.9 %	2.0 %
$t_{\{10;\alpha\}}$	2.708	2.658	2.611	2.568	2.527	2.490	2.454	2.421	2.389	2.359
α	2.1 %	2.2 %	2.3 %	2.4 %	2.5 %	2.6 %	2.7 %	2.8 %	2.9 %	3.0 %
$t_{\{10;\alpha\}}$	2.331	2.303	2.277	2.252	2.228	2.205	2.183	2.161	2.140	2.120
α	3.1 %	3.2 %	3.3 %	3.4 %	3.5 %	3.6 %	3.7 %	3.8 %	3.9 %	4.0 %
$t_{\{10;\alpha\}}$	2.101	2.082	2.063	2.046	2.028	2.011	1.995	1.979	1.963	1.948
α	4.1 %	4.2 %	4.3 %	4.4 %	4.5 %	4.6 %	4.7 %	4.8 %	4.9 %	5. %
$t_{\{10;\alpha\}}$	1.933	1.919	1.904	1.890	1.877	1.863	1.850	1.837	1.825	1.812

2. Percentiles de la F de Fisher con 4 y 10 grados de libertad

α	0.1 %	0.2 %	0.3 %	0.4 %	0.5 %	0.6 %	0.7 %	0.8 %	0.9 %	1.0 %
$F_{\{4;10;\alpha\}}$	11.283	9.432	8.461	7.817	7.343	6.970	6.665	6.409	6.188	5.994
α	1.1 %	1.2 %	1.3 %	1.4 %	1.5 %	1.6 %	1.7 %	1.8 %	1.9 %	2.0 %
$F_{\{4;10;\alpha\}}$	5.823	5.669	5.530	5.402	5.286	5.178	5.078	4.985	4.898	4.816
α	2.1 %	2.2 %	2.3 %	2.4 %	2.5 %	2.6 %	2.7 %	2.8 %	2.9 %	3.0 %
$F_{\{4;10;\alpha\}}$	4.738	4.665	4.596	4.531	4.468	4.409	4.352	4.298	4.245	4.195
α	3.1 %	3.2 %	3.3 %	3.4 %	3.5 %	3.6 %	3.7 %	3.8 %	3.9 %	4.0 %
$F_{\{4;10;\alpha\}}$	4.147	4.101	4.056	4.013	3.972	3.931	3.893	3.855	3.818	3.783
α	4.1 %	4.2 %	4.3 %	4.4 %	4.5 %	4.6 %	4.7 %	4.8 %	4.9 %	5. %
$F_{\{4;10;\alpha\}}$	3.749	3.715	3.683	3.651	3.620	3.591	3.561	3.533	3.505	3.478