Estadística II
Grado en Matemáticas, UAM, 2020-2021

Hoja 3 (Regresión lineal)

(Nota: salvo que se indique lo contrario, el modelo de regresión incluye la hipótesis de normalidad).

Regresión lineal simple

1. Dada una muestra (x_i, y_i), con $i = 1, \dots, n$, y con las notaciones habituales del modelo de regresión lineal, definimos el coeficiente de determinación

$$R^2 = \frac{\text{MSS}}{\text{TSS}}, \quad \text{donde MSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{y} \quad \text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

Comprueba que

$$|R| = \frac{|\text{cov}_{x,y}|}{\sqrt{V_x V_y}} \qquad \text{y que} \qquad \frac{|\hat{\beta}_1|}{s_R \sqrt{1/(nV_x)}} = \frac{\sqrt{n-2} \cdot |R|}{\sqrt{1-R^2}}.$$

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

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

Calcula un intervalo de confianza para la pendiente de la recta con un nivel de confianza 95 %. ¿Podemos rechazar, con un nivel de significación del 5 %, la hipótesis nula de que la variable X no influye linealmente en la variable Y?

3. Supongamos que la muestra $(x_1, Y_1), \dots, (x_n, Y_n)$ 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 $x_1 \neq \bar{x}$):

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

(a) ¿Es $\widehat{\beta}_1$ un estimador insesgado?

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

(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 $\widehat{\beta}_1$.

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

$$Y_i = \beta_1\, x_i + \varepsilon_i, \qquad \text{con } \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ independientes, para } i = 1, \dots, n.$$

(a) Calcula el estimador de mínimos cuadrados de β_1 y deduce su distribución.

(b) Sean $e_1, \dots, e_n$ los residuos del modelo. Comprueba si se cumplen o no las siguientes propiedades: $\sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} e_i\, x_i = 0$.

(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 $V(\varepsilon_i) = \sigma^2 x_i^2$; 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 $\widetilde{\beta}_1$ que se obtiene al minimizar la expresión $\sum_{i=1}^n \omega_i (y_i - \beta_1 x_i)^2$, donde $\omega_i = 1/x_i^2$. Calcula una fórmula explícita para $\widetilde{\beta}_1$ y, a partir de ella, deduce su esperanza y su varianza. Compara los estimadores $\widehat{\beta}_1$ y $\widetilde{\beta}_1$. ¿Cuál es mejor? (A $\widetilde{\beta}_1$ se le llama estimador de mínimos cuadrados ponderados).

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

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

c) 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?

7. 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, pues, 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 con la siguiente función de densidad:

$$f(x) = \begin{cases} x + 1 & \text{si } x \in [-1, 0], \\ 1 - x & \text{si } x \in (0, 1]. \end{cases}$$

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

Como estimador de β_0 elegimos el habitual de mínimos cuadrados:

$$\widehat{\beta}_0 = \overline{Y} - \frac{\overline{x}}{V_x} \operatorname{cov}_{x,Y}.$$

a) Comprueba si $\widehat{\beta}_0$ es un estimador insesgado de β_0 para muestras (x_i, Y_i) de tamaño n.

b) Calcula la varianza de $\widehat{\beta}_0$ para muestras (x_i, Y_i) de tamaño n.

Regresión lineal múltiple

8. En el modelo de regresión lineal múltiple, denotamos por MSS, RSS y TSS las sumas de cuadrados explicados, residuales y totales, respectivamente. Sea $R^2 = \text{MSS}/\text{TSS}$ el coeficiente de determinación. Comprueba que

$$\frac{\text{MSS}}{\text{RSS}} = \frac{R^2}{1 - R^2}.$$

9. Escribe la expresión explícita de la matriz $(X^\intercal X)$ para el caso $k = 2$ del modelo de regresión lineal. Escribe explícitamente qué supone, en términos de los estadísticos de las muestras de las dos variables regresoras, que la matriz $(X^\intercal X)$ sea definida positiva. Quizás te animes a escribir una expresión explícita para $(X^\intercal X)^{-1}$.

10. Supongamos que cierta variable respuesta Y depende linealmente de dos variables regresoras X_1 y X_2, de manera que se verifica el modelo:

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

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 $\widehat{\beta}_1$ de β_1 es en general sesgado, y determina bajo qué condiciones se anula el sesgo.

11. 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:

$$\begin{array}{rcccccccccc} \hat{y}_i & = & 5970 & + & 22.35\, x_{i1} & + & 2701.12\, x_{i2} & - & 67.67\, x_{i3} & \qquad & R^2 = 0.9843 \\ & & & & (2.55) & & (1820) & & (15.4) & & \end{array}$$

Entre paréntesis aparecen las desviaciones típicas (estimadas) de los estimadores.

(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 ($\alpha = 5\,\%$) la hipótesis nula: el número de habitaciones no influye en el precio.

(c) A nivel $\alpha = 5\,\%$, ¿puede afirmarse que la vivienda se encarece cuando disminuye la contaminación acústica?

(d) Contrasta con $\alpha = 5\,\%$ 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.

12. 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) ¿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) 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) 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?

13. Se plantea un modelo de regresión lineal múltiple con dos variables regresoras, X_1 y X_2, para muestras de tamaño n:

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

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

Se dispone de la siguiente muestra de tamaño $n = 4$:

X_1	X_2	Y
1	2	3
2	1	2
−1	1	0
1	3	2

Llamando X a la matriz de diseño, se han calculado los siguientes productos matriciales:

$$(X^\intercal \cdot X)^{-1} = \frac{1}{50}\begin{pmatrix} 69 & -3 & -31 \\ -3 & 11 & -3 \\ -31 & -3 & 19 \end{pmatrix}, \quad (X^\intercal \cdot X)^{-1} \cdot X^\intercal = \frac{1}{50}\begin{pmatrix} 4 & 32 & 41 & -27 \\ 2 & 16 & -17 & -1 \\ 4 & -18 & -9 & 23 \end{pmatrix}$$

Además, se ha obtenido que $R^2 = 0.730526$.

a) Halla un intervalo de confianza al 95 % para el parámetro β_1.

b) Se dispone de una nueva observación $(1, 1)$. Se pide calcular cuántas veces es mayor

- la longitud del intervalo de confianza, al 95 %, para predecir el *valor* de la variable respuesta que correspondería a esa observación,
- que la longitud del intervalo de confianza, al 95 %, para predecir el *valor medio* de la variable respuesta que correspondería a esa observación.

14. 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?

15. Considera la matriz H del modelo de regresión lineal múltiple, con n observaciones y k variables regresoras. Recuerda que H es la matriz ($n \times n$, simétrica, idempotente y con rango $k + 1$) dada por

$$H = X(X^\intercal X)^{-1} X^\intercal,$$

donde X es la matriz de diseño de la regresión (matriz $n \times (k + 1)$, de rango $k + 1$ y primera columna de unos).

Llamemos $h_{i,j}$ a las entradas de la matriz H. Comprueba que

(a) $\sum_{i=1}^{n} h_{i,i} = k+1$.

(b) $h_{i,i} > 0$.

(c) $h_{i,i} = h_{i,i}^2 + \sum_{j \neq i} h_{i,j}^2$.

(d) $h_{i,i} \leq 1$.

(e) $h_{i,j}^2 \leq 1/4$ si $i \neq j$.

16. Considera el modelo de regresión múltiple $\mathbb{Y} = X\beta + \boldsymbol{\epsilon}$, donde el vector de errores $\boldsymbol{\epsilon}$ verifica las hipótesis habituales. Hay n observaciones y k variables regresoras.

(a) Define el vector de valores ajustados $\hat{\mathbb{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\intercal$ 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 $\sum_{i=1}^{n} \mathbf{V}(\hat{Y}_i)$.

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

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

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

18. 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.

Preguntas:

(a) Expresa el problema de estimación como un modelo de regresión lineal múltiple, indicando claramente cuál es la matriz de diseño.

(b) Determina la distribución del estimador de mínimos cuadrados del vector de posiciones $(\beta_1, \beta_2, \beta_3)^\intercal$.

(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?