Regresión

Tarea 1

Christian Badillo

Tabla de contenidos

1 Ejercicio 0 (clase).

Encuentra los Estimadores de Mínimos Cuadrados de los parámetros β_0 y β_1 minimizando la función:

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

1.1 Solución.

Derivamos la función $f(\beta_0, \beta_1)$ con respecto a β_0 y β_1 e igualamos a cero:

$$\frac{\partial f}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{1}$$

$$\frac{\partial f}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \tag{2}$$

Se divide entre -2 y se distribuyen las sumatorias en 1:

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i = 0$$

$$n\bar{y} - n\beta_0 - \beta_1 n\bar{x} = 0 \qquad \text{Utilizando } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ y } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0 \qquad \text{Factorizando } n$$

Realizamos el mismo procedimiento con 2:

$$\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \beta_0 x_i - \sum_{i=1}^{n} \beta_1 x_i^2 = 0 \qquad \text{Distribuyendo la sumatoria y } x_i$$

$$\sum_{i=1}^{n} y_i x_i - n\beta_0 \bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \qquad \text{Utilizando } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Obtenemos el sistema de ecuaciones:

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0 \tag{3}$$

$$\sum_{i=1}^{n} y_i x_i - n\beta_0 \bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{4}$$

Despejamos β_0 de 3:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Sustituimos β_0 en 4 y resolvemos para β_1:

$$\sum_{i=1}^{n} y_i x_i - n\left(\bar{y} - \beta_1 \bar{x}\right)\bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x} + n\beta_1 \bar{x}^2 - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \qquad \text{Distribuyendo } n$$

$$n\beta_1 \bar{x}^2 - \beta_1 \sum_{i=1}^{n} x_i^2 = -\sum_{i=1}^{n} y_i x_i + n\bar{y}\bar{x}$$

$$-\beta_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = -\sum_{i=1}^{n} y_i x_i + n\bar{y}\bar{x} \qquad \text{Factorizando } \beta_1$$

$$-\beta_1 s_X^2 = n\bar{x}\bar{y} - \sum_{i=1}^{n} y_i x_i \qquad \text{Se usa la relación } s_X^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$\beta_1 s_X^2 = s_{XY} \qquad \text{Donde } s_{XY} = n\bar{x}\bar{y} - \sum_{i=1}^{n} y_i x_i$$

$$\beta_1 = \frac{s_{XY}}{s_X^2}$$

Por último, sustituimos β_1 en 3 y despejamos β_0:

$$\beta_0 = \bar{y} - \frac{s_{XY}}{s_X^2}\bar{x}$$

Por lo tanto, los estimadores de mínimos cuadrados de los parámetros β_0 y β_1 son:

$$\hat{\beta}_0 = \bar{y} - \frac{s_{XY}}{s_X^2}\bar{x}$$

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

2 Ejercicio 2.

1. Menciona los supuestos sobre los errores ϵ_i para el modelo de regresiónn lineal simple.
    1. Su valor esperado es 0, $\mathbb{E}(\epsilon_i) = 0$.
    2. Homocedasticidad, es decir, la varianza de los errores es constante, $Var(\epsilon_i) = \sigma^2$.
    3. Los errores son no correlacionados entre sí, $Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$.

2. ¿Qué le da el carácter aleatorio al modelo de regresión lineal simple?

EL término ϵ del modelo dado que nos permite añadir un termino estocástico a un proceso determinístico como lo es la función del modelo (en este caso una función lineal).

3. ¿Cuáles de los siguientes pueden considerarse modelos de regresión lineal? ¿Por qué?

   - $f(\beta_0, \beta_1) = \beta_0 + e^{\beta_1 x_1}$.

   No es un modelo lineal dado que para tiene una no linealidad sobre uno de sus parámetros en este caso β_1 y no es posible eliminarla.

   - $f(\beta_0, \beta_1, \ldots, \beta_k) = \beta_0 + \sum_{i=1}^{k} \beta_i \cosh^i x_i$.

   Es un modelo lineal dado que es una combinación lineal de los parámetros.

   - $f(\beta_0, \beta_1) = \beta_0 + \beta_1 x_1 + \beta_0 \beta_1 x_2$.

   No es un modelo lineal dado que hay dos parámetros que interactuán de manera multiplicativa entre ellos.

3 Ejercicio 3.

Supongamos que la variable aleatoria ϵ sigue una distribución normal con media μ y varianza σ^2, y que, para $1 \leq i \leq n$, la variables aleatorias ϵ_i son independientes y cada una sigue una distribución normal con media μ_i y varianza σ_i.

- ¿Cómo podemos asociar ϵ a una variable aleatoria que sigue una distribución normal estándar?
- ¿Qué distribución siguen las siguientes variables aleatorias?

1. Para $a, b \in \mathbb{R}$, con $a \neq 0, X = a\epsilon + b$.

La variable aleatoria X sigue una distribución normal con media $a\mu_\epsilon + b$ y varianza $(a\sigma_\epsilon)^2$.

Demostración.

Teorema 2.1.5 en Casella y Berger (2002):

Si X es una variable aleatoria con pdf, $f_X(x)$, continua sobre su soporte $\mathcal{X}$ y $Y = g(X)$ con soporte $\mathcal{Y}$ y con g una función monótona. Y suponiendo que $g^{-1}(y)$ es diferenciable en $\mathcal{Y}$. Entonces:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right) \left| \frac{d}{dy} g^{-1}(y) \right|, & y \in \mathcal{Y} \\ 0, & \text{en otro caso.} \end{cases}$$

Usando los datos del problema:

$$f_X(x) = f_\epsilon(\frac{x-b}{a}) \left| \frac{d}{dx} \frac{x-b}{a} \right| = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} e^{-\frac{\left(\frac{x-b}{a} - \mu_\epsilon\right)^2}{2\sigma_\epsilon^2}} \frac{1}{|a|}$$

Con un poco de algebra:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_\epsilon^2 a^2}} e^{-\frac{(x-(a\mu_\epsilon+b))^2}{2\sigma_\epsilon^2 a^2}}$$

Con lo que llegamos a la conclusión de que la variable aleatoria X sigue una distribución normal con media $a\mu_\epsilon + b$ y varianza $(a\sigma_\epsilon)^2$.

2. $X = \epsilon_1 + \epsilon_2$.

La variable aleatoria X sigue una distribución normal con media $\mu_{\epsilon_1} + \mu_{\epsilon_2}$ y varianza $\sigma_{\epsilon_1}^2 + \sigma_{\epsilon_2}^2$.

Demostración.

Dada la función generadora de momentos de la distribución normal:

$$M_\epsilon(t) = e^{t\mu + \frac{t^2\sigma^2}{2}}.$$

Y que la función generadora de momentos de una combinación lineal de variables aleatorias independientes es el producto de sus funciones generadora de momentos individuales. Entonces:

$$\begin{aligned} M_{\epsilon_1+\epsilon_2}(t) &= \mathbb{E}\left(e^{(\epsilon_1+\epsilon_2)t}\right) \\ &= M_{\epsilon_1}(t) M_{\epsilon_2}(t) \\ &= e^{t\mu_{\epsilon_1} + \frac{t^2\sigma_{\epsilon_1}^2}{2}} e^{t\mu_{\epsilon_2} + \frac{t^2\sigma_{\epsilon_2}^2}{2}} \\ &= e^{t(\mu_{\epsilon_1}+\mu_{\epsilon_1}) + \frac{t^2(\sigma_{\epsilon_1}^2+\sigma_{\epsilon_2}^2)}{2}} \end{aligned}$$

Lo cual nos da la función generadora de momentos de la distribución ya mencionada.

3. $X = \epsilon_1 - \epsilon_2$.

La variable aleatoria X sigue una distribución normal con media $\mu_{\epsilon_1} - \mu_{\epsilon_2}$ y varianza $\sigma_{\epsilon_1}^2 + \sigma_{\epsilon_2}^2$.

Demostración.

Usando los hechos anteriormente demostrados podemos usar sus resultados para obtener que la distribución de $Y = -\epsilon_i$ es una distribución normal con media $-\mu_{\epsilon_i}$ (se utiliza $a = -1$ y $b = 0$) y varianza $(-1)^2\sigma_{\epsilon_i}^2 = \sigma_{\epsilon_i}^2$. Usando este resultado intermedio aplicado al resultado demostrado en el inciso anterior se puede concluir que la distribución de la variable aleatoria X es una distribución normal con media $\mu_{\epsilon_1} - \mu_{\epsilon_2}$ y varianza $\sigma_{\epsilon_1} + \sigma_{\epsilon_2}$.

4 Ejercicio 3.

La siguiente es una tabla que nos da las estaturas(en centímetros) y pesos(en kilogramos) de una muestra de 10 mujeres adolescentes de 18 años. Se busca intentar predecir el peso de acuerdo a la altura.

Tabla 1: Datos

Altura	Peso
169.6	71.2
166.8	58.2
157.1	56.0
181.1	64.5
158.4	53.0
165.6	52.4
166.7	56.8
156.5	49.2
168.1	55.6
165.3	77.8

Gráfica la altura contra el peso, ¿dirías que tiene sentido utilizar un modelo de regresión lineal para estos datos? En caso afirmativo, aproxima tal modelo con lo obtenido en el primer ejercicio, y luego dibuja tal aproximación sobre la gráfica previamente obtenida. ¿Consideras que el modelo es bueno? ¿Por qué?

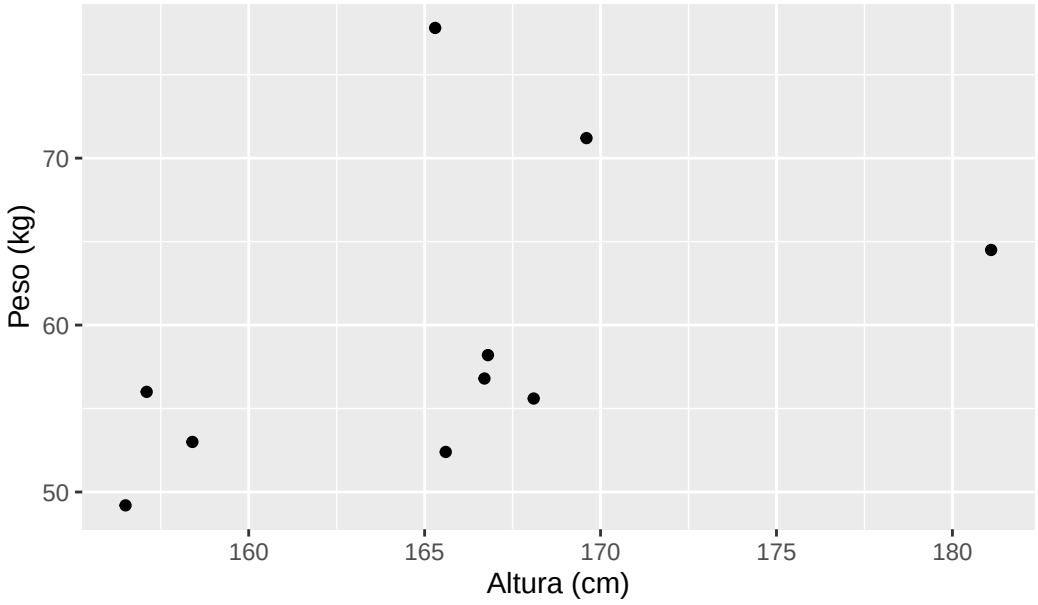

Figura 1: Gráfico de dispersión de los datos.

No parece que la relación entre el peso y la altura sea linealmente “fuerte”, dado que no se observa

una tendencia clara en los datos. De hecho parece que la relación es más bien cuadrática. Por lo tanto, un modelo de regresión lineal no parece adecuado para estos datos.

El modelo de regresión lineal ajustado es: Peso = -36.88 + 0.58 * Altura.

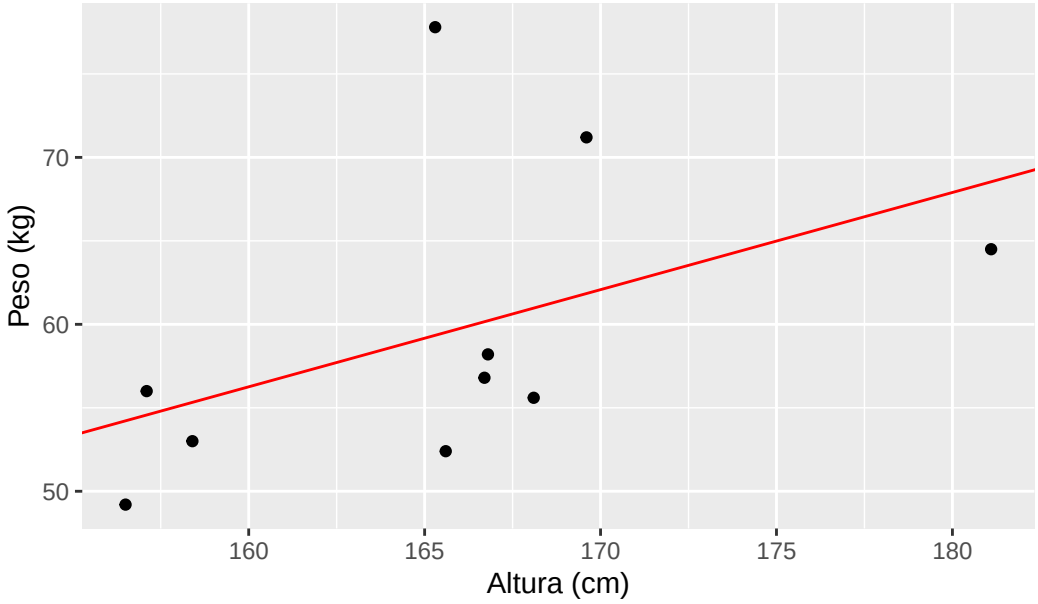

Figura 2: Ajuste del Modelo de Regresión Lineal Simple.

El ajuste del modelo es malo ya que se observan muchos puntos alejados de la recta de regresión, lo que se traduce que los errores de predicción son grandes para los datos observados y probablemente para datos no observados, por lo que se concluye que el modelo no es bueno para modelar la relación entre la altura y el peso.