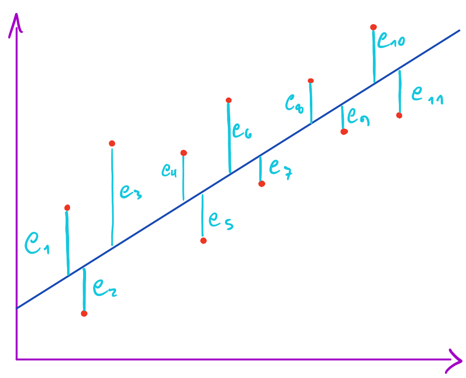
Regresión Lineal

Pensemos en la versión más sencilla del problema de Regresión Lineal:

Tenemos un conjunto de puntos en un plano: $[(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$ y queremos una recta que pase lo más cerca posible de todos los puntos



Así, nosotros queremos encontrar una recta

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

para poder predecir puntos que desconocemos

Así, en base a nuestros datos, queremos aprender los parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$

Sea $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ la predicción basada en el i-ésimo valor de X (x_i).

Llamaremos e_i al error $y_i - \hat{y}_i$, esto es, la diferencia entre la predicción y el valor real. Así, nosotros queremos minimizar la suma de los $|e_i|$ (recordemos que e_i puede ser negativo). Como no nos gusta el valor absoluto, minimizaremos:

$$\sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \dots + e_n^2$$

$$= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

¿Cómo encontramos los valores que minimizan la expresión?

$$\hookrightarrow \frac{\delta \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\delta \hat{\beta}_0} = \sum_{i=1}^{n} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Downarrow$$

$$0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} y_i - \underbrace{\sum_{i=1}^{n} \hat{\beta}_0}_{n\hat{\beta}_0} - \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

Así tenemos:

$$\boxed{\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

Con $\bar{Y}$ el promedio de los y_i

$\bar{X}$ el promedio de los x_i

$$\hookrightarrow \frac{\delta \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\delta \hat{\beta}_1} = \sum_{i=1}^{n} -2x_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$0 = \sum_{i=1}^{n} x_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} (x_i Y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2)$$

Pero recordemos que: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$$0 = \sum_{i=1}^{n} (x_i Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X}) x_i - \hat{\beta}_1 x_i^2)$$

$$0 = \sum_{i=1}^{n} (x_i Y_i - \bar{Y} x_i + \hat{\beta}_1 \bar{X} x_i - \hat{\beta}_1 x_i^2)$$

$$0 = \sum_{i=1}^{n} (x_i Y_i - \bar{Y} x_i) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i^2 - \bar{x} x_i)$$

Y de aquí obtenemos la famosa expresión:

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

Así, tenemos que la recta que buscamos es:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X \quad \text{donde}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

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

Ojo! También es frecuente ver $\hat{\beta}_1$ como:

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

Pero es fácil mostrar que $c\sum_{i=1}^{n}(z_i - \bar{z})$ es 0, así que llegaremos a que las identidades son iguales