r_{xy} es un estimador (construido usando la muestra $(x_1,y_1),\ldots,(x_N,y_N)$) del coeficiente de correlación lineal de Pearson (Galton 1880)

$$R_{xy} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{VAR}(X)}\,\sqrt{\operatorname{VAR}(Y)}}$$

R_{xy} es una medida de la asociación lineal que pueda existir en las variables X y Y, un resultado que ilustra este hecho es

Proposición: $(r_{xy})^2 = 1$ si y sólo si existen dos constantes a y b tales que $y_i = a + bx_i$, $i=1,2,\ldots,N$. Además si $b>0$ entonces $r_{xy}=1$ y si $b<0$ $r_{xy}=-1$.

Dadas estas observaciones, se recomienda seguir los siguientes pasos para hacer un ajuste de un modelo lineal de regresión

1.- Producir un diagrama de dispersión de las parejas $(x_1,y_1),\ldots,(x_N,y_N)$, si no existe evidencia de una tendencia lineal se busca otro modelo ó se intenta alguna transformación de los variables[1]. En caso contrario, pasemos a 2 a continuación.

2.- Calcular el coeficiente de correlación lineal de Pearson. Si r_{XY}^2 no esta cercano (suficiente-mente cercano) a 1, el modelo lineal no es adecuado. En caso contrario se pasa a 3 a continuación

3.- Ajustar la linea recta de mínimos cuadrados e interpretar los valores de los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$.

Una pregunta inmediata con respecto a este algoritmo surge si al calcular r_{XY}^2 este valor es tal que $0 < r_{XY}^2 < 1$, ¿Cómo sabemos qué significa que r_{XY}^2 este cercano a 1 ó que r_{XY}^2 este cercano a 0?

(1) El capítulo 4 del libro "Applied Regression Analysis and Generalized Linear Models" de John Fox contiene una discusión (3ª edición, SAGE editores). También en la sección 6, capítulo 7 de "A second course in Statistics Regression Analysis" Mendenhall, Sincich (7ª edición, PEARSON editores).

En otras palabras ¿Cuándo está r_{XY}^2 cercano a 1 ó a 0?

Tenemos que $\Delta(\hat{\beta}_0,\hat{\beta}_1) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$ y que $\Delta(\hat{\beta}_0,\hat{\beta}_1) = \sum_{i=1}^{N} (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{N} (x_i - \bar{x})^2$. En consecuencia

$$\sum_{i=1}^{N} (y_i - \bar{y})^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \hat{\beta}_1^2 \sum_{i=1}^{N} (x_i - \bar{x})^2,$$

es decir

$$(SSQ) \ldots\ldots \frac{1}{N}\sum_{i=1}^{N} (y_i - \bar{y})^2 = \overbrace{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}^{A} + \overbrace{\frac{1}{N}\hat{\beta}_1^2 \sum_{i=1}^{N} (x_i - \bar{x})^2}^{B},$$

lo cual nos indica que la varianza de la variable de respuesta se descompone en dos partes:

A = la parte de la variabilidad de Y que la variable X <u>no</u> puede explicar a través del modelo $\boxed{\text{Lineal}}$

B = la parte de la variabilidad de Y que la variable X <u>sí</u> logra explicar a través del modelo Lineal

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

ó bien

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

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

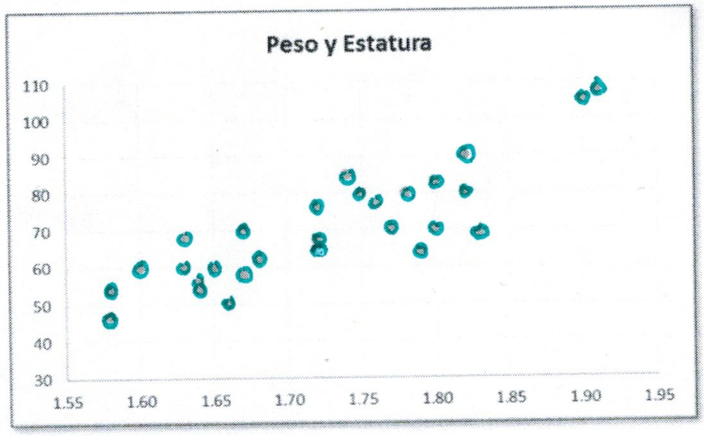
$$= r_{XY}^2 .$$

Entonces al dividir ambos lados de (SSQ) por $\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2$ tenemos

$$1 = \underbrace{\frac{\sum_{i=1}^{N}(y_i-\hat{y})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}}_{A'} + \underbrace{r_{XY}^2}_{B'}$$

De forma que r_{XY}^2 representa la proporción de la variabilidad de Y, que logra explicar la variable X a través del modelo. La cantidad r_{XY}^2 recibe el nombre de "Coeficiente de Determinación" y representa el porcentaje de la varianza de la variable de respuesta que explica el modelo

Para el caso de los datos de Peso y estatura tenemos



Los datos presentan una tendencia monótona creciente y una posible relación lineal entre X y Y no parece inapropiada. Para calcular el coeficiente de correlación lineal

$$\bar{X} = 1.72 \quad \bar{Y} = 70.3$$

$$r_{XY} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}} = \frac{31.309}{\sqrt{0.225}\sqrt{6010.37}}$$

$$= 0.8521$$

cuya magnitud no resulta lejana de 1, de hecho $r_{XY}^2 = 0.726$ de manera que una regresión lineal de Y como función de X explica un 72% de la variabilidad de la variable Y = Peso

Como ya se mencionó, existen parejas (x_i, y_i) tales que este par ordenado no yace sobre la línea recta $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$. Lo anterior nos sugiere considerar el modelo

$$Y = \overbrace{\beta_0 + \beta_1 X}^{(i)} + \overbrace{\varepsilon}^{(ii)},$$

donde ε es un término que da cuenta del efecto en Y de todos los factores distintos a X, los cuales <u>no</u> se controlan durante el estudio. Este efecto aparece en forma aditiva en el modelo, es decir se asume que todos los factores que pueden influir[(1)] en Y producen un efecto que queda incorporado por el término ε en el modelo, en forma aditiva.

(i) = componente determinista
(ii) = componente aleatoria

(1) Que no sean X!

RESULTADO

Sea $f: \mathbb{R}^2 \to \mathbb{R}$ una función continua y con derivadas

$$\frac{\partial^2 f(x,y)}{\partial x^2}, \quad \frac{\partial^2 f(x,y)}{\partial y^2}, \quad \frac{\partial^2 f(x,y)}{\partial y \partial x}, \quad \frac{\partial^2 f(x,y)}{\partial x \partial y}$$

que son continuas. La función $f(x,y)$ tiene un mínimo relativo en el punto (a,b) si

1.- $\dfrac{\partial f}{\partial x}(a,b) = 0$ y $\dfrac{\partial f}{\partial y}(a,b) = 0$

2.- Para $D(a,b) = \dfrac{\partial^2 f}{\partial x^2}(a,b) \cdot \dfrac{\partial^2 f}{\partial y^2}(a,b) - \left(\dfrac{\partial^2 f}{\partial x \partial y}(a,b)\right)^2$

se tiene $D(a,b) > 0$ y además $\dfrac{\partial^2 f}{\partial x^2}(a,b) > 0$.

D es el determinante de la Matriz Hessiano H

$$H = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x \partial y} \\[2ex] \dfrac{\partial^2 f}{\partial y \partial x} & \dfrac{\partial^2 f}{\partial y^2} \end{pmatrix}$$

★ Si $\Lambda(\beta_0, \beta_1) \equiv \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i))^2$, entonces

$$\left.\begin{array}{ll} \dfrac{\partial^2 \Lambda(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0^2} = 2N > 0 & \dfrac{\partial^2 \Lambda(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_1^2} = 2\sum_{i=1}^{N} x_i^2 > 0 \\[3ex] \dfrac{\partial^2 \Lambda(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_1 \partial \beta_0} = 2\sum_{i=1}^{N} x_i & \dfrac{\partial^2 \Lambda(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0 \partial \beta_1} = 2\sum_{i=1}^{N} x_i \end{array}\right\}$$

DONDE:

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Calculando $D(\hat{\beta}_0, \hat{\beta}_1)$

$$D(\hat{\beta}_0, \hat{\beta}_1) = 4N \sum_{i=1}^{N} x_i^2 - 4\left(\sum_{i=1}^{N} x_i\right)^2$$

$$= 4N \sum_{i=1}^{N} x_i^2 - 4N^2 \bar{x}^2 = 4N\left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right)$$

$$= 4N \sum_{i=1}^{N} (x_i - \bar{x})^2 \geq 0$$

$D(\hat{\beta}_0, \hat{\beta}_1)$ es 0, sólo cuando $\sum_{i=1}^{N} (x_i - \bar{x})^2 = 0$, que como ya hemos discutido no es un caso de interés práctico. $\therefore D(\hat{\beta}_0, \hat{\beta}_1) > 0$

y usando el resultado de cálculo diferencial multivariado que se enunció tenemos que el valor $\Delta(\hat{\beta}_0, \hat{\beta}_1)$ es un mínimo.