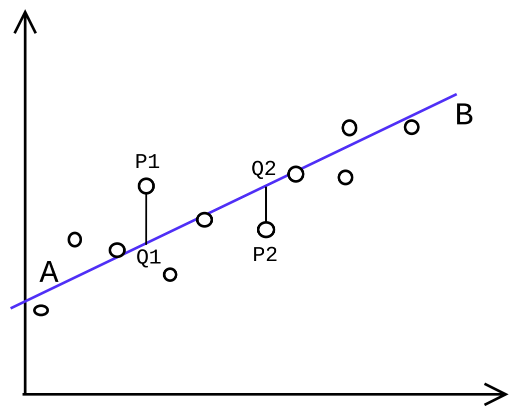
Cuadrados mínimos

Supongamos que tenemos un conjunto de N valores de una variable y, medidos como función de la variable x. Nos vamos a restringir al caso especial en el que toda la incertidumbre se limita a la dimensión y: esto es, los valores de x se conocen exactamente o, al menos, con una precisión tanto mayor que la de los valores de y, como para poder despreciar la incertidumbre en la dimensión x.

Queremos ajustar los datos por una recta. La pregunta por contestar es: cuál de todas las rectas en el plano x-y escogemos como la mejor y qué queremos decir con "la mejor"? El principio de mínimos cuadrados permite hacer esta elección con base en las desviaciones de los puntos en dirección vertical a partir de las rectas.

Sea AB una candidata a la categoría de "mejor recta". Consideremos todos los intervalos verticales entre los puntos y la línea, de los cuales P_2Q_2 es típico. Definimos como **mejor recta aquella que minimiza la suma de los cuadrados de las desviaciones** como P_2Q_2.



Se puede probar que el procedimiento de minimizar los cuadrados de las desviaciones da lugar a una menor varianza de los parámetros resultantes, como por ejemplo la pendiente, que al usar cualquier otro criterio.

⇒Definimos que la mejor recta es aquella que lleva a su valor mínimo a la suma:

$$\sum_i (P_i Q_i)^2$$

y queremos obtener los parámetros, pendiente m y ordenada al origen b, de esa mejor recta cuya ecuación es:

$$y = mx + b$$

La magnitud de la desviación P_iQ_i es el intervalo entre un cierto valor medido y_i y el valor de la función y en ese punto, para ese valor de x. Este valor y se puede calcular a partir del valor correspondiente de x como $mx_i + b$. Si llamamos δy_i a la diferencia tenemos:

$$\delta y_i = y_i - (mx_i + b)$$

⇒ Llamando

$$M = \sum_i \left[y_i - (mx_i + b)\right]^2$$

queremos que se cumplan las siguientes condiciones

$$\frac{\partial M}{\partial m} = 0, \qquad \frac{\partial M}{\partial b} = 0$$

$$\frac{\partial M}{\partial m} = 2\sum_i [y_i - (mx_i + b)](-x_i) = 0 \qquad \Rightarrow \quad b = \frac{\sum_i y_i x_i - \sum_i m x_i^2}{\sum_i x_i}$$

$$\frac{\partial M}{\partial b} = 2\sum_i [y_i - (mx_i + b)] = 0 \qquad \Rightarrow \sum_i y_i - \sum_i m x_i - Nb = 0$$

Reemplazando en la última ecuación lo que obtuvimos para b en función de m, podemos despejar m, que da:

$$m = \frac{N\sum_i (x_i y_i) - \sum_i x_i \sum_i y_i}{N\sum_i (x_i)^2 - (\sum_i x_i)^2} \tag{1}$$

Teniendo la expresión para la pendiente podemos insertarla en la ecuación que obtuvimos para b, despejando esta última:

$$b = \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i (x_i y_i)}{N\sum_i x_i^2 - (\sum_i x_i)^2} \tag{2}$$

La desviación estándar de la pendiente y de la ordenada al origen se calculan en términos de la desviación estándar de la distribución de valores de δy alrededor de la mejor recta, que llamaremos σ_y:

$$\sigma_y = \sqrt{\frac{\sum_i (\delta y_i)^2}{N-2}}$$

El $(N-2)$ es consecuencia de aplicar la definición de la desviación estándar a la posición de una recta en el plano.

Los valores de σ_m y σ_b están dados por:

$$\sigma_m = \sigma_y \sqrt{\frac{N}{N\sum_i x_i^2 - (\sum_i x_i)^2}} \tag{3}$$

$$\sigma_b = \sigma_y \sqrt{\frac{\sum_i x_i^2}{N\sum_i x_i^2 - (\sum_i x_i)^2}} \tag{4}$$

Confiabilidad

Los valores de x_i e y_i guardan efectivamente una relación lineal? Para poder responder ésto se define el coeficiente de correlación o de regresión lineal

$$R = \frac{N\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sqrt{N[\sum_i x_i^2 - (\sum_i x_i)^2}\sqrt{N[\sum_i y_i^2 - (\sum_i y_i)^2}} \tag{5}$$

El coeficiente R varía entre 0 y 1. En una correlación perfecta (recta perfecta) $R = 1$ y en ausencia de correlación $R = 0$. En la práctica, una correlación “buena” está, en general, dada por $R > 0.9$. Si $R < 0.3$ se considera que las dos magnitudes x e y no están correlacionadas.

Ponderación en los cálculos estadísticos

Las ecuaciones que obtuvimos para ajustar la recta por el principio de mínimos cuadrados son válidas sólo cuando todas las medidas son igualmente precisas. Si estas medidas son de precisión desigual, en una falacia considerar que éstas contribuyen de igual modo a la respuesta final. Las medias más precisas deben desempeñar un papel más importante que las menos precisas, en los cálculos. Para lograr ésto, asignamos a las medidas "**ponderaciones**" que son i**nversamente proporcionales a los cuadrados de las desviaciones estándar de las observaciones correspondientes**.

Media ponderada de un conjunto de medidas

Si tenemos un conjunto de cantidades x_i medidas independientes y conocemos las desviaciones estándar, σ_i, de cada una de las $x_i \Rightarrow$la media ponderada del conjunto de valores x y su desviación estándar estarán dadas por

$$\bar{x} = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}, \qquad \sigma^2 = \frac{\sum_i \frac{(x_i - \bar{x})^2}{\sigma_i^2}}{(N-1)\sum_i \frac{1}{\sigma_i^2}}$$

Ajuste de línea recta por mínimos cuadrados ponderados

Si tenemos un conjunto de valores de una variable y medidos como función de x y suponemos que los valores de x son precisos y que toda la incertidumbre se limita a los valores de $y \Rightarrow$ las ecuaciones por las cuales podemos calcular la pendiente m y la ordenada al origen b de la mejor recta puede expresarse como:

$$m = \frac{\sum_i w_i y_i \sum_i w_i x_i^2 - \sum_i w_i x_i \sum_i w_i x_i y_i}{\sum_i w_i \sum_i w_i x_i^2 - (\sum_i w_i x_i)^2}$$

$$b = \frac{\sum_i w_i y_i \sum_i w_i x_i y_i - \sum_i w_i x_i \sum_i w_i y_i}{\sum_i w_i \sum_i w_i x_i^2 - (\sum_i w_i x_i)^2}$$

Utilizamos el término w_i para la ponderación de cada par de valores medidos (x_i, y_i):

$$w_i = \frac{1}{(\sigma_{y_i})^2}$$

Las mejores estimaciones de las desviaciones estándar de m y b se pueden expresar desde el punto de vista de la desviación de los puntos medidos respecto de la mejor recta. Para un ajuste de mínimos cuadrados ponderados, estas desviaciones deberán ponderarse a su vez, y el valor ponderado de σ_y estará dado por

$$\sigma_y = \sqrt{\frac{\sum_i w_i \delta y_i^2}{N-2}}$$

$$\Rightarrow \sigma_m^2 = \frac{\sigma_y^2}{W}, \qquad \sigma_b^2 = \sigma_y^2 \left(\frac{1}{\sum_i w_i} + \frac{x^2}{W} \right)$$

donde

$$W = \sum_i w_i (x_i - \bar{x})^2, \qquad \bar{x} = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

$\bar{x}$ es el promedio ponderado de los valores de las x.