Linear Regression

Given n points $(x_1, y_1) \ldots (x_n, y_n)$ and an assumed relation $y = f(x) + \epsilon, \epsilon \sim N(\mu, \sigma)$ we want to find a model $\tilde{y_i} = ax_i + b$ such that the residual squared error

$$\mathrm{RSS}(a, b) = \sum_{i=1}^{n} (\tilde{y}_i - y_i)^2$$

is minimized.

RSS is a function of the line parameters a and b. To minimize it we set both partial derivatives to zero. (This could technically find a maximum – but it's reasonably clear this function has no maximum value because the error can always be increased.)

Take partial derivatives

$$\begin{aligned}
\frac{\partial RSS}{\partial a} &= 2 \sum (\tilde{y}_c - y_i) \frac{\partial}{\partial a} (\tilde{y}_i - y_i) \\
&= 2 \sum (\tilde{y}_i - y_i) (x_i) \\
\frac{\partial RSS}{\partial b} &= 2 \sum (\tilde{y}_i - y_i) \frac{\partial}{\partial b} (\tilde{y}_i - y_i) \\
&= 2 \sum (\tilde{y}_i - y_i) (1)
\end{aligned}$$

Since

$$\frac{\partial}{\partial a} \tilde{y}_i = \frac{\partial}{\partial a} (ax_i + b) = x_i$$

$$\frac{\partial}{\partial b} \tilde{y}_i = \frac{\partial}{\partial b} (ax_i + b) = 1$$

And solve

$$\begin{cases} \dfrac{\partial RSS}{\partial a} = 0 \\ \\ \dfrac{\partial RSS}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \sum(\tilde{y}_i - y_i)\, x_i = 0 \\ \\ \sum\left(\tilde{y}_i - y_i\right) = 0 \end{cases}$$

Since $\tilde{y}_i = ax_i + b$

$$\sum\left(ax_i + b - y_i\right) x_i = 0 \Rightarrow a \sum x_i^2 + b \sum x_i = \sum x_i y_i$$

and

$$\sum\left(ax_i + b - y_i\right) = 0 \Rightarrow a \sum x_i + b \sum 1 = \sum y_i$$

by Cramer's rule

$$\begin{aligned} a &= \begin{vmatrix} \sum x_i y_i & \sum x_i \\ \sum y_i & n \end{vmatrix} / \begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix} \\ b &= \begin{vmatrix} \sum x_i^2 & \sum x_i y_i \\ \sum x_i & \sum y_i \end{vmatrix} / \begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix} \end{aligned}$$

since $\sum_{i=1}^{n} 1 = n$

Taking determinants,

$$a = \frac{n \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

$$b = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

Interpretation as a ratio of variances

Students of statistics may appreciate the following manipulations

Definition of covariance

$$E(xy) - E(x)E(y) = \mathrm{Cov}(x, y)$$

Definition of variance

$$\mathrm{Var}(x) = E\left[(x - \mu)^2\right]$$

Lemma

$$
\begin{aligned}
\operatorname{Var}(x) &= E\left[(x-\mu)^2\right] \\
&= E\left(x^2\right) - 2\mu E[x] + E[\mu]^2 \\
&= E\left[x^2\right] - 2E[x]^2 + \mu^2 \\
&= E\left[x^2\right] - E[x]^2
\end{aligned}
$$

Manipulating the denominator of the equation for a on the previous page,

$$
\begin{aligned}
n \sum x_i^2 - \left(\sum x_i\right)^2 &= n^2 \left(\frac{1}{n} \sum x_i^2 - \left(\frac{\sum x_i}{n} \right)^2 \right) \\
&= n^2 \left(E\left[x^2\right] - E[x]^2 \right) \\
&= n^2 \operatorname{Var}(x)
\end{aligned}
$$

And the numerator

$$
\begin{aligned}
n \sum x_i y_i - \sum x_i \sum y_i & \\
&= n^2 \left(\frac{1}{n} \sum x_i y_i - \frac{1}{n} \sum x_i \cdot \frac{1}{n} \sum y_i \right) \\
&= n^2 \left(E\left[xy\right] - E[x]E[y] \right) \\
&= n^2 \left(E[xy] - \mu_y \mu_y \right) \\
&= n^2 \operatorname{Cov}(x, y)
\end{aligned}
$$

so

$$
a = \frac{E[xy] - \mu_x \mu_y}{E[x] - \mu_x^2} = \frac{\operatorname{Cov}(x, y)}{\operatorname{Vav}(x)}
$$