

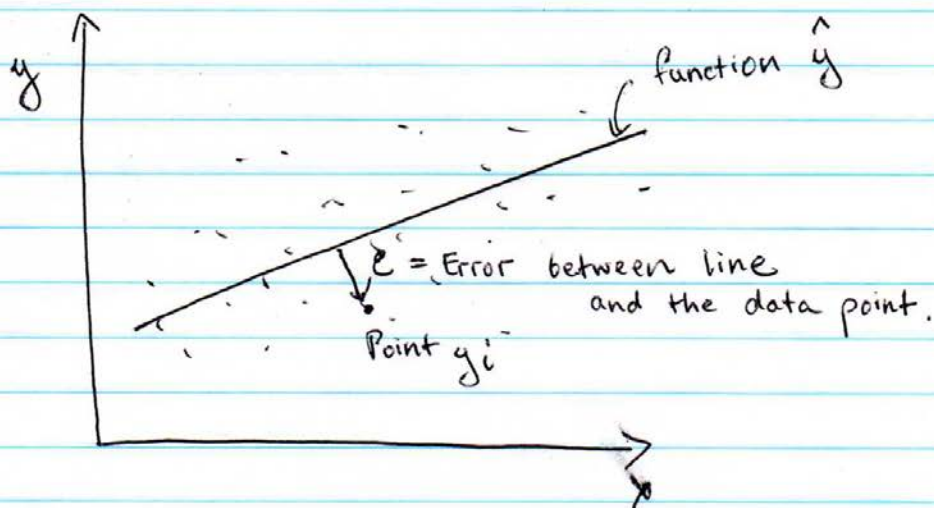
Regression and Correlation

Idea: Fit some function to a collection of data points (x_i, y_i) to approximate the relationship between them.

x_i = Predictor

y_i = Predictand.

Simplest function is a straight line.



The line that best fits the data is that which minimizes the sum of the squares of the error. $\rightarrow$ least squares linear regression.

Simple idea, but serves as the basis for the more complicated matrix methods later (EOF, PCA, SVD)

$$Q = \sum_{i=1}^N (\hat{y}_i - y_i)^2 = \sum_{i=1}^N (a_1 x_{i0} - y_i)^2$$

$$\hat{y}_i = a_1 x_{i0}$$

To minimize the sum of squares of the errors, must take the derivatives of Q with respect to a_1 and a_0 and set them equal to 0.

$$\frac{dQ}{da_0} = a_1 \sum x_i - \sum y_i = 0$$

$$\frac{dQ}{da_1} = \sum x_i^2 - \sum x_i y_i = 0$$

Suggest do the algebra to show this is true.

We can divide by N to get the above expressions in terms of means

$$\frac{dQ}{da_0} = 0 = a_1 \bar{x} - \bar{y} = 0$$

$$\frac{dQ}{da_1} = 0 = \sum x_i^2 - \sum x_i y_i = 0$$

$$\begin{aligned} \bar{y} &= a_1 \bar{x} \\ \bar{xy} &= a_1 \overline{x^2} \end{aligned} \quad \left. \vphantom{\begin{aligned} \bar{y} &= a_1 \bar{x} \\ \bar{xy} &= a_1 \overline{x^2} \end{aligned}} \right\} \text{Solve for } a_1 \text{ and } a_0.$$

Rewrite in matrix form:

$$\begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix} = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$$

→ Get used to this way

$$a_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \quad a_0 = \bar{y} - a_1\bar{x}$$

(Bias)

→ Solutions to coefficients for linear regression.

$$x = \bar{x} + x' \quad y = \bar{y} + y'$$

$$\begin{aligned} \overline{xy} &= \overline{(\bar{x} + x')(\bar{y} + y')} \\ &= \bar{x}\bar{y} + \underbrace{\overline{\bar{x}y'}}_{\rightarrow 0} + \underbrace{\overline{\bar{x}'y}}_{\rightarrow 0} + \overline{x'y'} \end{aligned}$$

(Middle terms to zero by Reynold's avg.)

$$\overline{x^2} = \overline{(\bar{x} + x')^2} = \bar{x}^2 + \underbrace{\overline{2\bar{x}x'}}_{\rightarrow 0} + \overline{x'^2}$$

So the solution to a_1 reduces to:

$$a_1 = \frac{\overline{x'y'}}{\overline{x'^2}}$$

$$\begin{aligned} \overline{x'y'} &= \text{Covariance of } x \text{ and } y \\ \overline{x'^2} &= \text{Variance of } x. \end{aligned}$$

Total variance in y :

$$\frac{1}{N} \sum (y - \bar{y})^2$$

Explained variance

$$\frac{1}{N} \sum (\hat{y} - \bar{y})^2$$

Where $\hat{y}$ is the regression line.

Unexplained variance

$$\frac{1}{N} \sum (y - \hat{y})^2$$

Logic holds no matter what the form of $\hat{y}$ is (e.g. line or something more complicated).

$$\frac{\text{Explained variance}}{\text{Total variance}} = \% \text{ variance explained.}$$

From this, we're almost to the definition of the correlation coefficient...