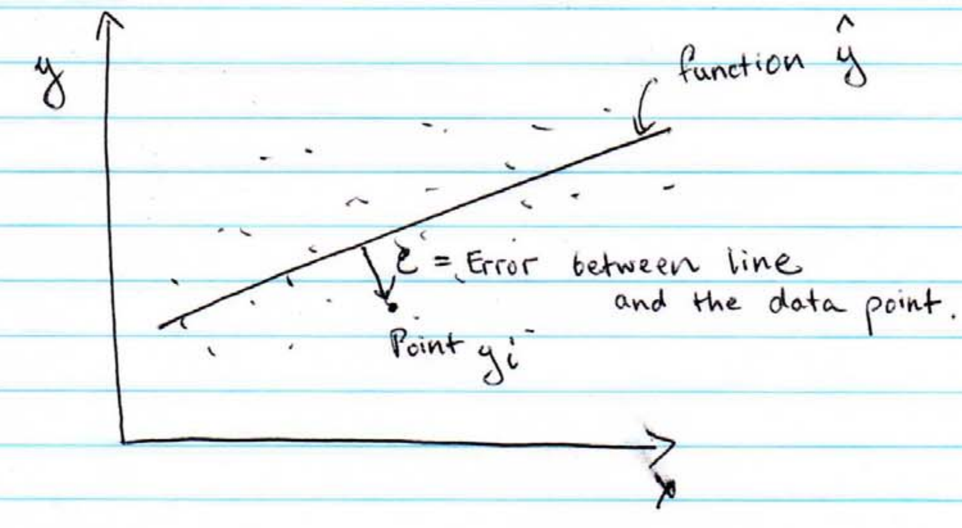
Regression and Correlation

Idea of regression: 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. → least squares linear regression.

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

$$Q = \sum_{i=1}^{N} (\hat{y} - y_i)^2 = \sum_{i=1}^{N} (a_1 x + a_0 - y_i)^2$$

$$\hat{y} = a_1 x + a_0$$

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.

$$\frac{dQ}{da_0} = a_1 \sum x_i + a_0 N - \sum y_i = 0$$

$$\frac{dQ}{da_1} = a_1 \sum x_i^2 + a_0 \sum x_i - \sum x_i y_i = 0$$

Suggest do the algebra to show this is true.

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

$$\frac{dQ}{da_0} = 0 = a_1 \bar{x} + a_0 - \bar{y} = 0$$

$$\frac{dQ}{da_1} = 0 = a_1 \overline{x^2} + a_0 \bar{x} - \overline{xy} = 0$$

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

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

(Bias)

→ Solutions to coefficients for linear regression.

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

$$\overline{xy} = \overline{(\bar{x}+x')(\bar{y}+y')}$$

$$= \bar{x}\bar{y} + \underbrace{\cancel{\overline{\bar{x}y'}}}_{\text{to } 0} + \underbrace{\cancel{\overline{x'\bar{y}}}}_{\text{to } 0} + \overline{x'y'}$$

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

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

So the solution to a_1 reduces to:

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

$\overline{x'y'}$ = Covariance of x and y

$\overline{x'^2}$ = Variance of x.

Total variance in y:

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

Explained variance

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

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

Unexplained variance

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

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...