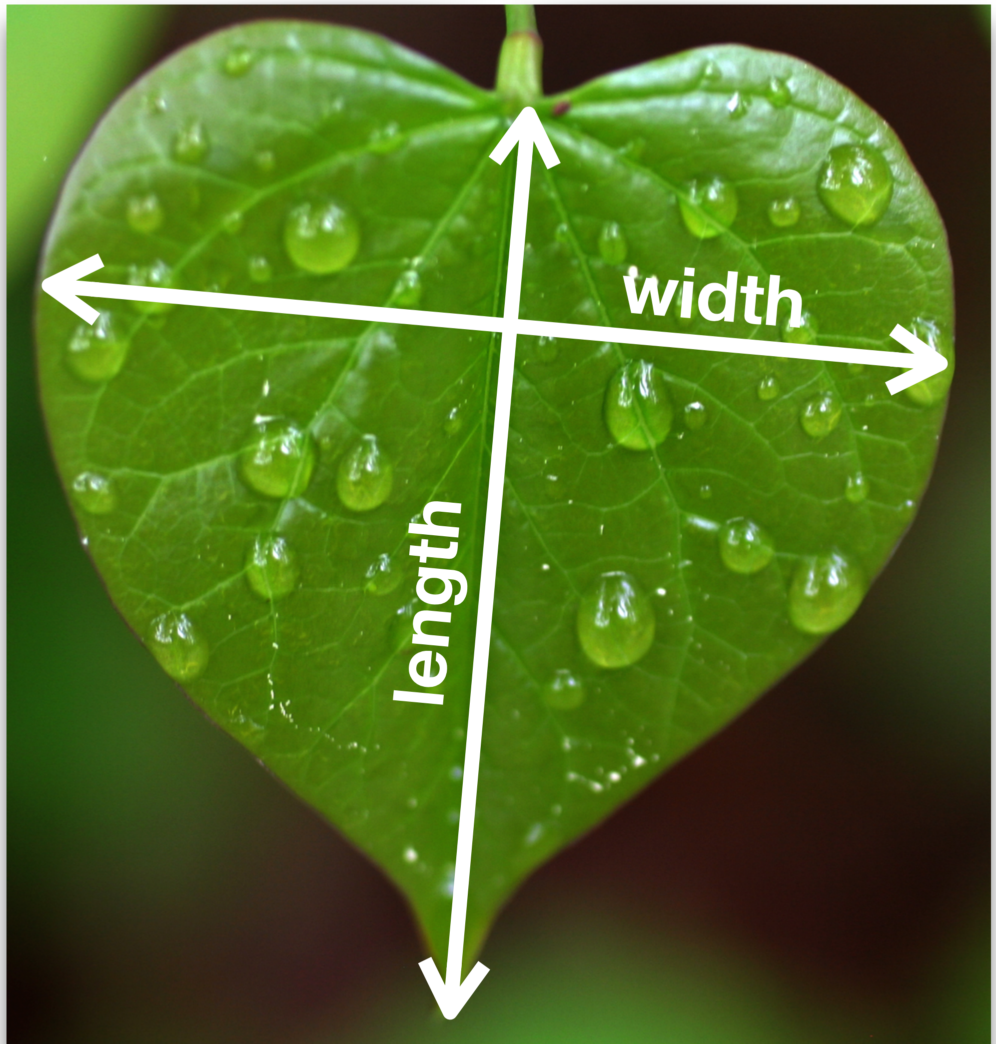
Phylogenetic Generalized Least Squares (PGLS)

Imaginary Problem

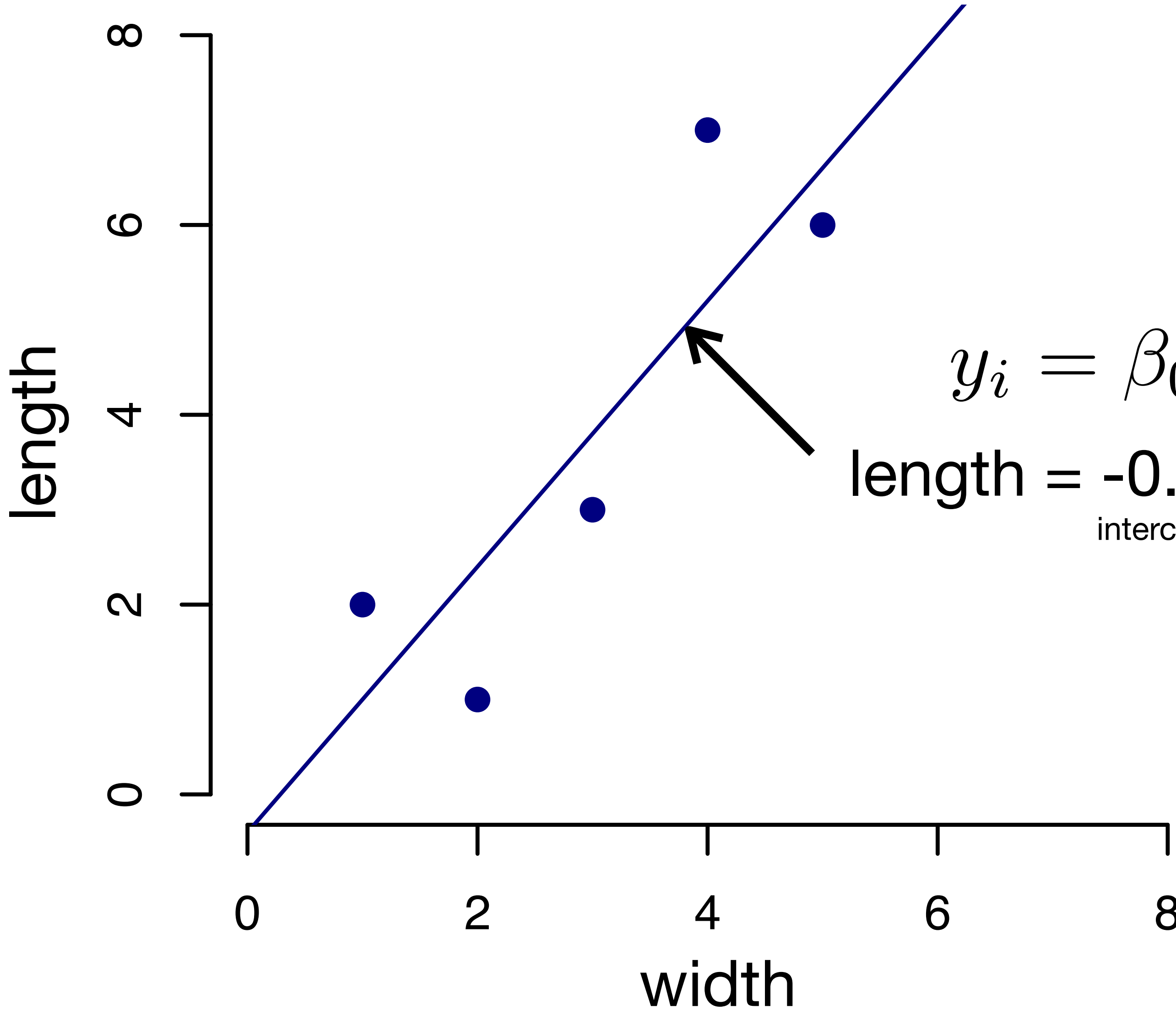


You have measured the average length and average width of a sample of leaves from 5 species of trees.

Question: Is leaf length correlated with leaf width?

If there is a correlation across species, is the correlation due only to the phylogeny?

Linear Regression

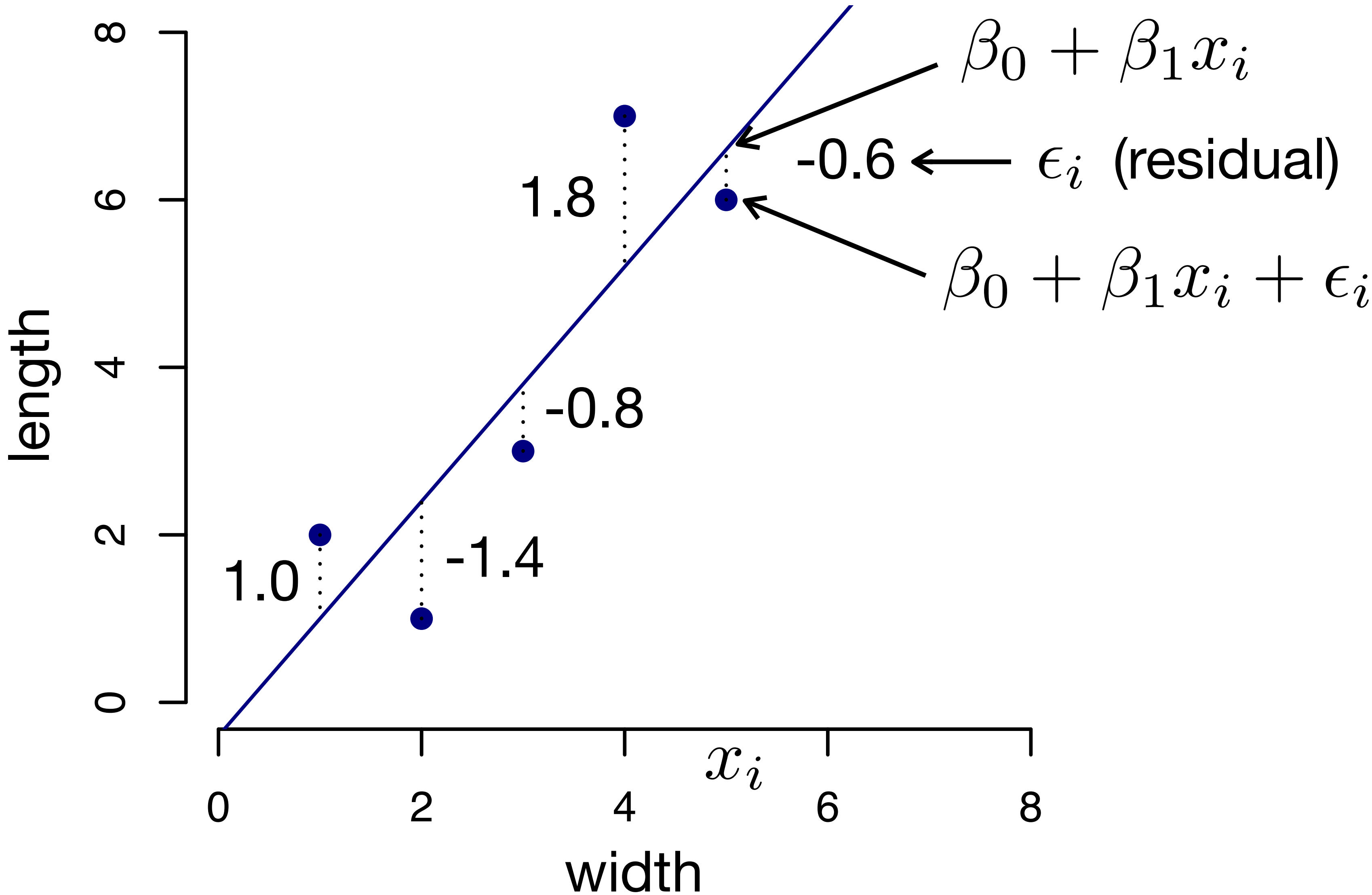


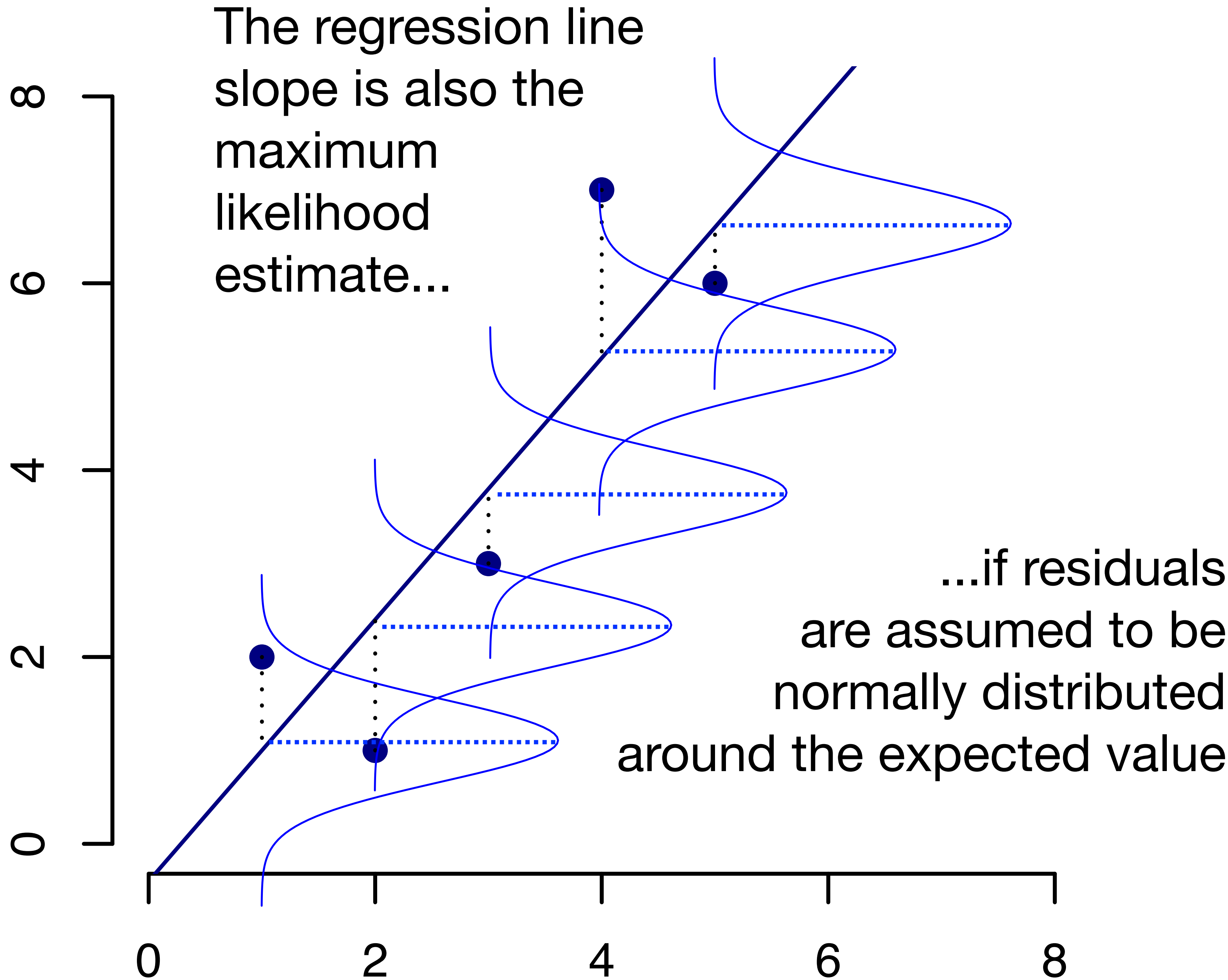
$$y_i = \beta_0 + \beta_1 x_i$$

length = -0.4 + 1.4 * width

intercept slope

Regression line chosen to minimize the sum of squared residuals

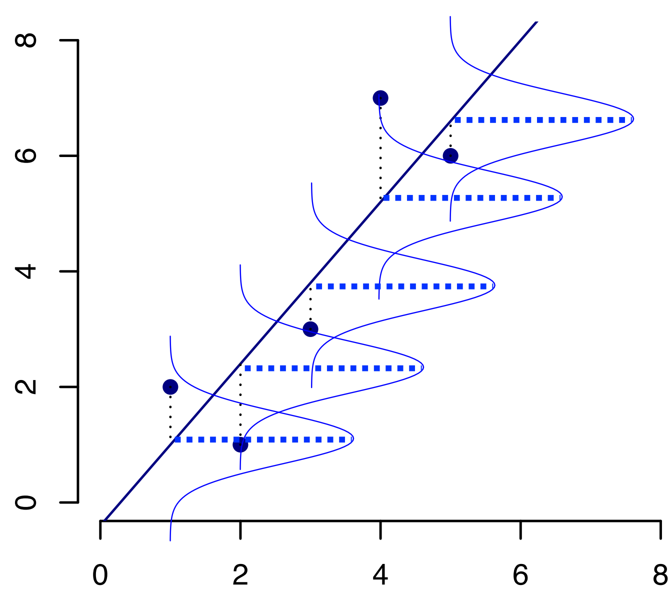




The regression line slope is also the maximum likelihood estimate...
...if residuals are assumed to be normally distributed around the expected value
0
2
4
6
8
0
2
4
6
8

Assuming intercept = β_0 = 0 for a moment, let's find the maximum likelihood estimate of β_1

$$p(\mathbf{y}|\mathbf{x},\beta_1) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_1-\beta_1 x_1)^2}{2\sigma^2}}\right] \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_2-\beta_1 x_2)^2}{2\sigma^2}}\right]$$
$$\cdot \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_3-\beta_1 x_3)^2}{2\sigma^2}}\right] \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_4-\beta_1 x_4)^2}{2\sigma^2}}\right]$$
$$\cdot \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_5-\beta_1 x_5)^2}{2\sigma^2}}\right]$$



The likelihood is a product of 5 normal densities, each corresponding to a different residual

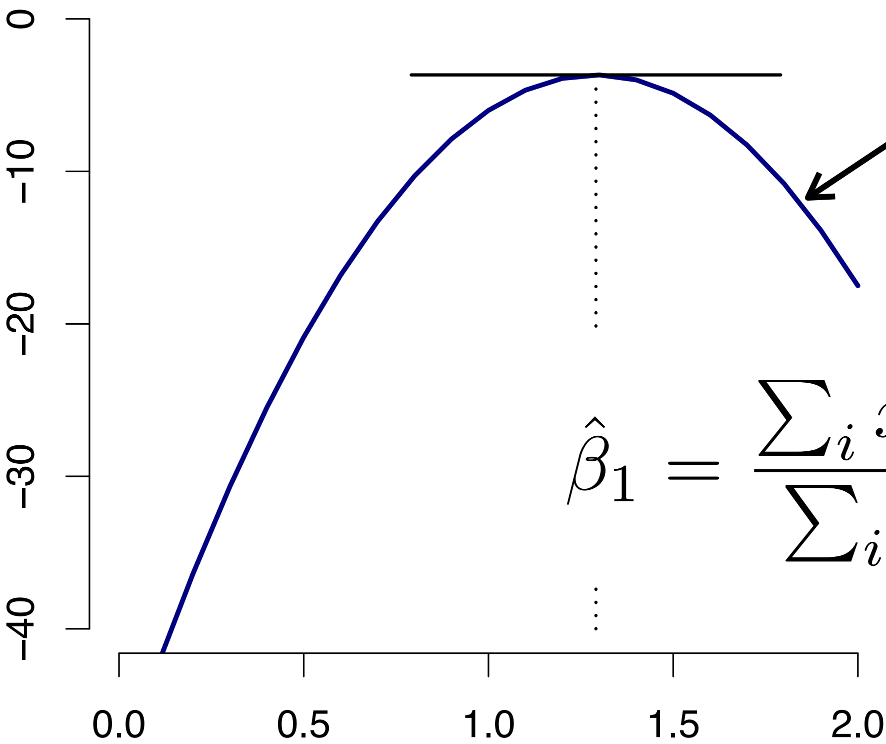
Log-likelihood function

$$p(\mathbf{y}|\mathbf{x}, \beta_1) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right]^5 e^{-\frac{\sum_{i=1}^{5}(y_i - \beta_1 x_i)^2}{2\sigma^2}}$$

$$\log p(\mathbf{y}|\mathbf{x}, \beta_1) = 5 \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{\sum_{i=1}^{5}(y_i - \beta_1 x_i)^2}{2\sigma^2}$$

Note that this term is a constant for our purposes because it does not contain β_1

We only need to find where this term is maximum to find the MLE of the slope



$$\log L = -\frac{\sum_{i=1}^{5}(y_i - \beta_1 x_i)^2}{2\sigma^2}$$

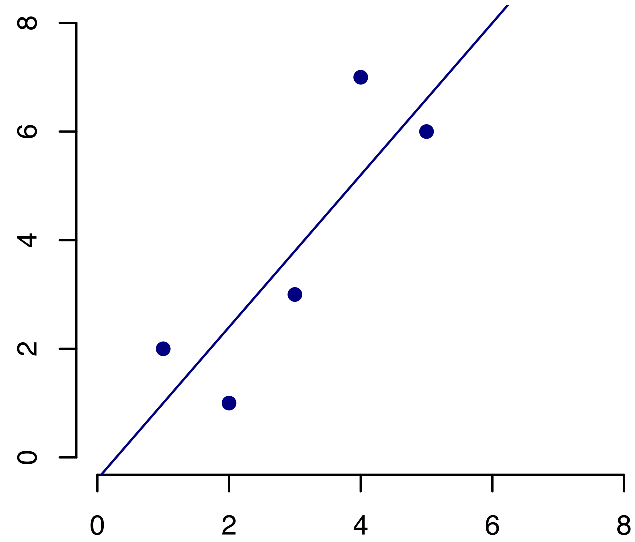
$$\hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

Maximum occurs where derivative of log-likelihood equals zero

$$\frac{d \log L}{d\beta_1} = -\frac{1}{2\sigma^2}\sum_{i=1}^{5} 2(y_i - \beta_1 x_i)(-x_i)$$

$$= \frac{1}{\sigma^2}\left\{\left(\sum_i x_i y_i\right) - \beta_1 \sum_i x_i^2\right\} = 0$$

Matrix representation



$$\mathbf{Y} = \mathbf{X} \quad \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} 1 \\ 3 \\ 2 \\ 7 \\ 6 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 1 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix} = \begin{bmatrix} 2\beta_1 + \epsilon_1 \\ 3\beta_1 + \epsilon_2 \\ 1\beta_1 + \epsilon_3 \\ 4\beta_1 + \epsilon_4 \\ 5\beta_1 + \epsilon_5 \end{bmatrix}$$

$5 \times 1 \qquad 5 \times 1 \qquad 1 \times 1 \qquad 5 \times 1 \qquad 5 \times 1$