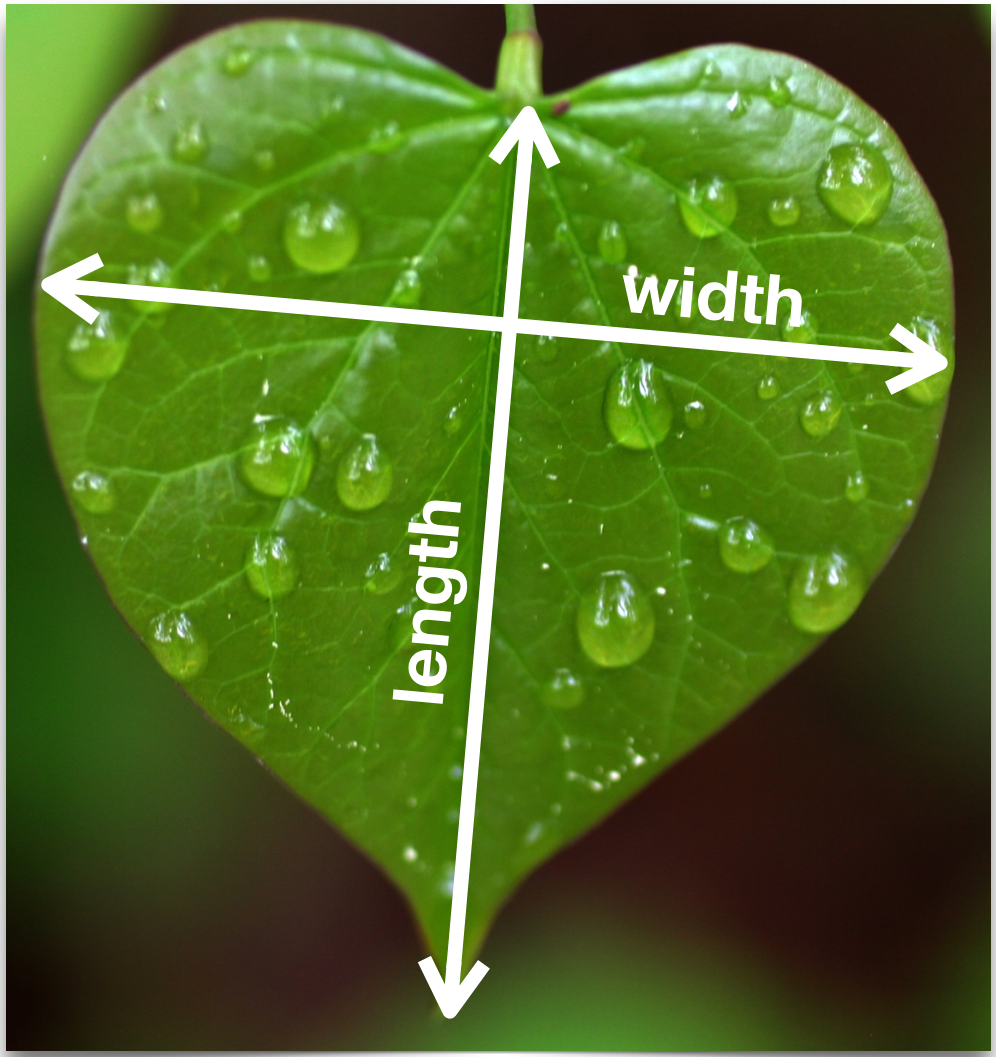
Phylogenetic Generalized Least Squares (PGLS)

Imaginary Problem

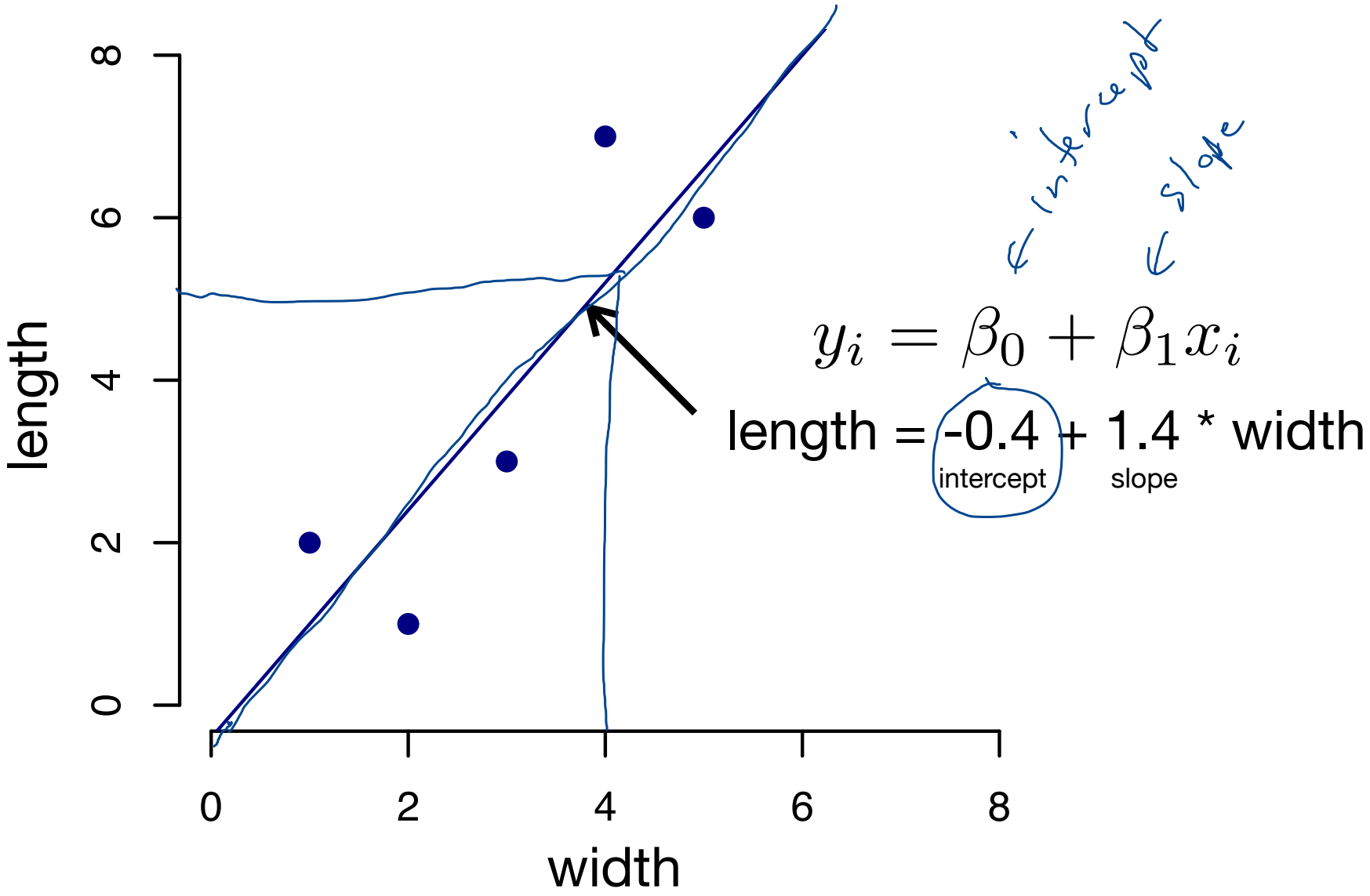


You have measured the average length and average width of a sample of leaves from 5 species of trees.

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



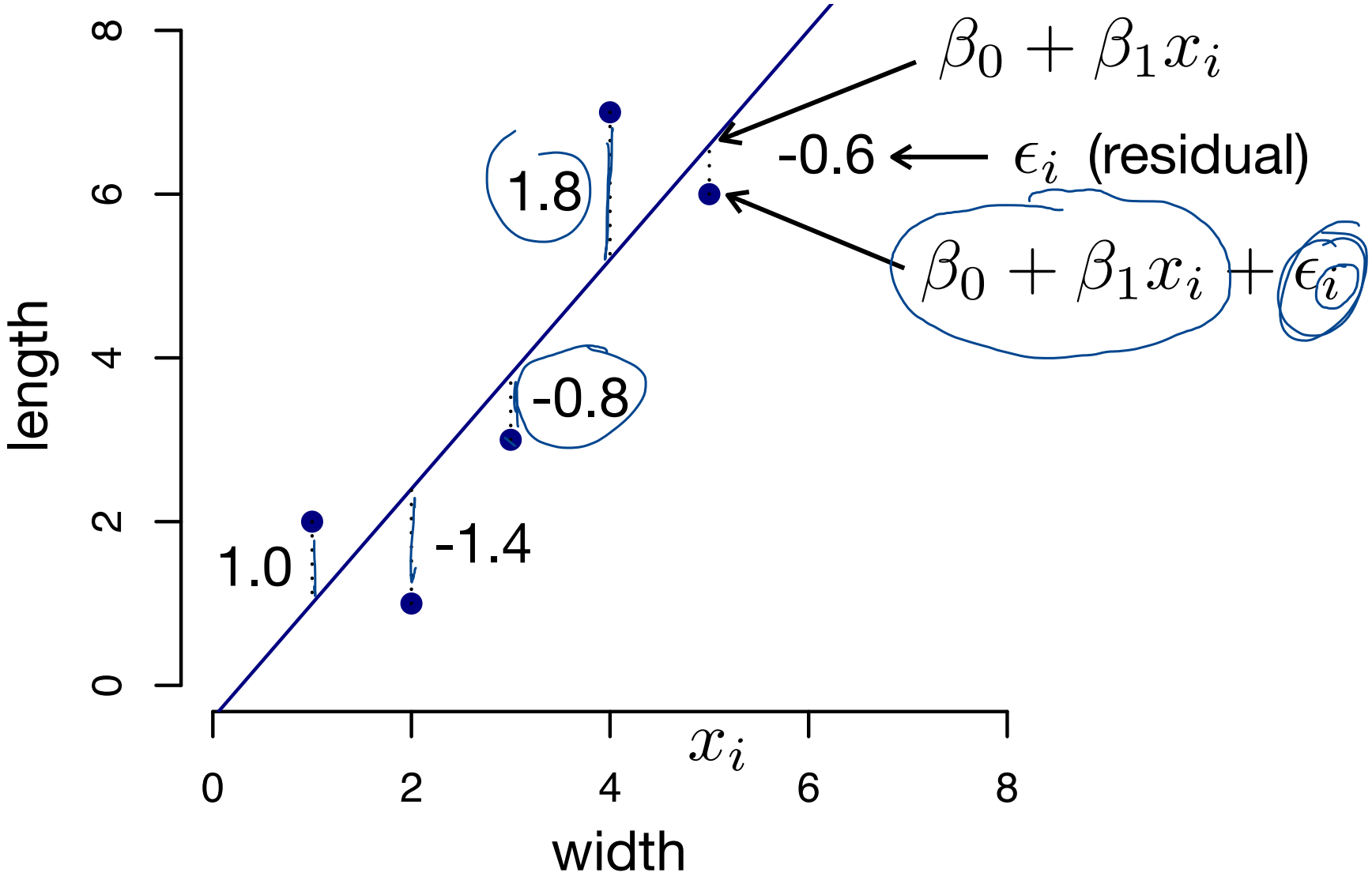
$$y_i = \beta_0 + \beta_1 x_i$$

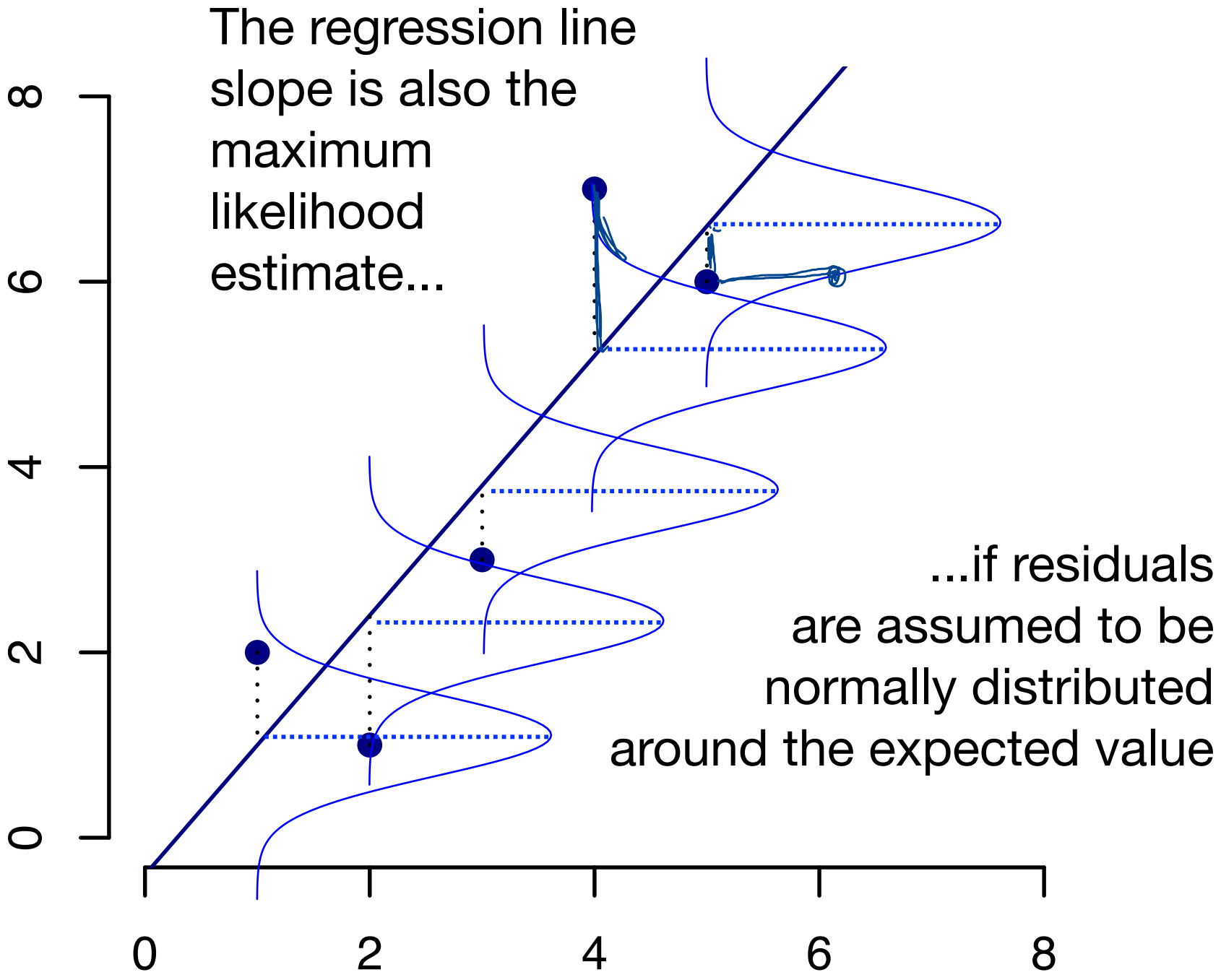
length = -0.4 + 1.4 * width

intercept

slope

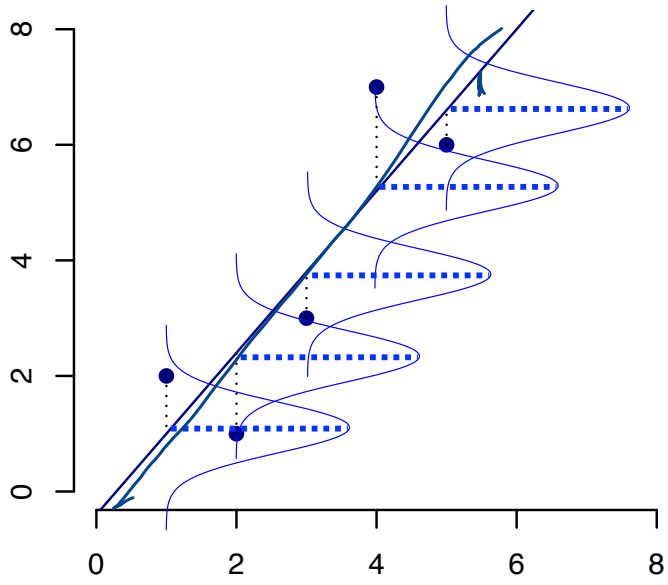
Regression line chosen to minimize the sum of squared residuals




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

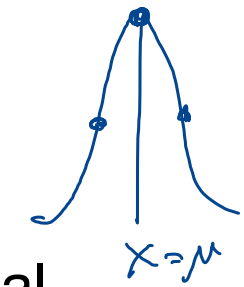
Assuming intercept = $\beta_0 = 0$ for a moment, let's find the maximum likelihood estimate of β_1

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



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

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



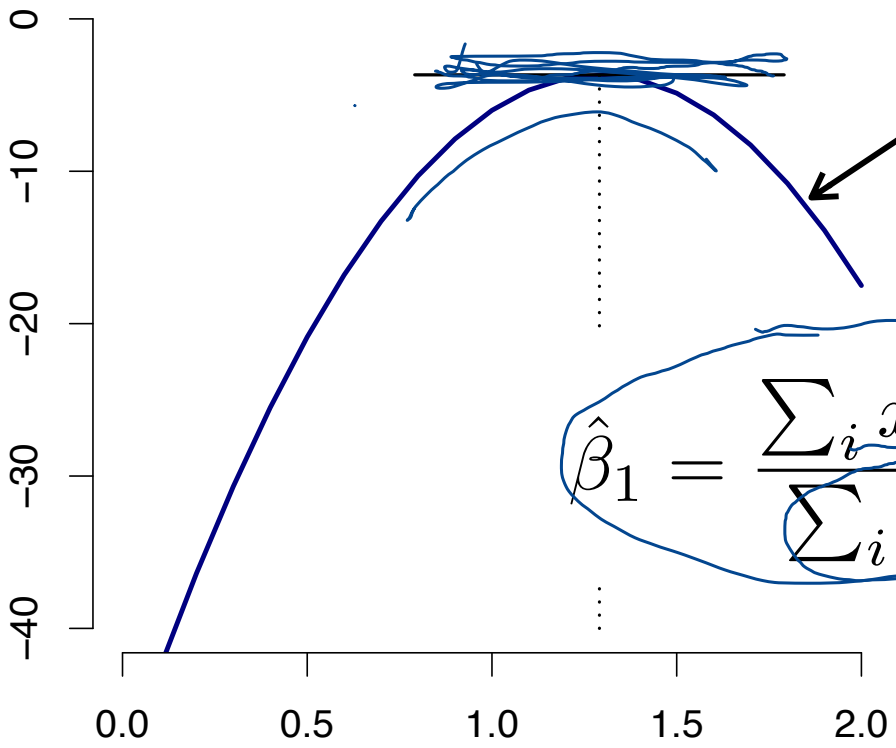
Log-likelihood function

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

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

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



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

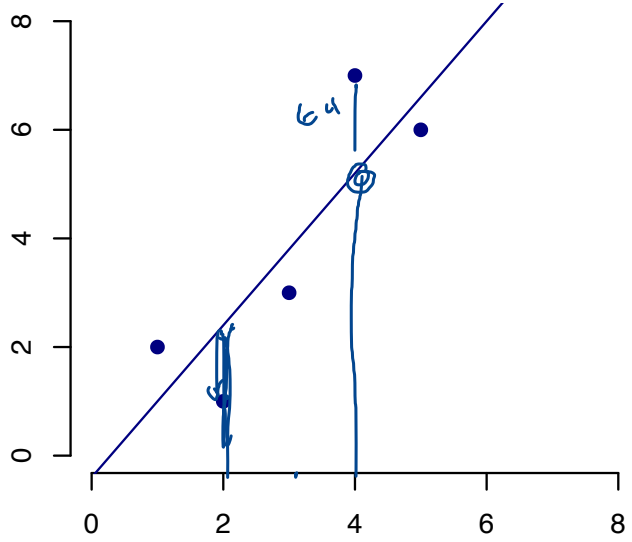
$$\hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

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



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

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

STOPPED HERE APR. 2, 2024

Residuals in matrix form

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$$

$$\begin{bmatrix} 1-2\beta_1 \\ 3-3\beta_1 \\ 2-1\beta_1 \\ 7-4\beta_1 \\ 6-5\beta_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 7 \\ 6 \end{bmatrix} - \begin{bmatrix} 2\beta_1 \\ 3\beta_1 \\ 1\beta_1 \\ 4\beta_1 \\ 5\beta_1 \end{bmatrix}$$

5 × 1 5 × 1 5 × 1

Sum of squared residuals

(Y - X β)' **(Y - X β)**

$$\begin{bmatrix} 1-2\beta_1 & 3-3\beta_1 & 2-1\beta_1 & 7-4\beta_1 & 6-5\beta_1 \end{bmatrix} \quad \begin{bmatrix} 1-2\beta_1 \\ 3-3\beta_1 \\ 2-1\beta_1 \\ 7-4\beta_1 \\ 6-5\beta_1 \end{bmatrix}$$

1 × 5 5 × 1

This matrix product yields a 1×1 matrix whose only element equals the sum of squared deviations. The prime symbol (which looks like an apostrophe) means that the matrix is transposed (i.e. the columns are changed to rows).

$$(\mathbf{Y} - \boldsymbol{\beta}\mathbf{X})' (\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}) = \sum_{i=1}^{5} (y_i - \beta_1 x_i)^2$$

matrix version non-matrix version

Log-likelihood

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

non-matrix

using matrices

$$\log L = -\frac{1}{2\sigma^2} (\mathbf{Y} - \boldsymbol{\beta}\mathbf{X})' (\mathbf{Y} - \boldsymbol{\beta}\mathbf{X})$$

Estimating the slope

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

non-matrix

using matrices

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\mathbf{Y})$$

Adding the intercept

The model just presented was this:

$$Y = \beta_1 X$$

Ordinarily, regressions also include an intercept term, β_0 (which is the predicted value of Y when X = 0):

$$Y = \beta_0 + \beta_1 X$$

Matrix representation with intercept

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{e}$$

$$\begin{bmatrix} 1 \\ 3 \\ 2 \\ 7 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix} = \begin{bmatrix} \beta_0 + 2\beta_1 + \epsilon_1 \\ \beta_0 + 3\beta_1 + \epsilon_2 \\ \beta_0 + 1\beta_1 + \epsilon_3 \\ \beta_0 + 4\beta_1 + \epsilon_4 \\ \beta_0 + 5\beta_1 + \epsilon_5 \end{bmatrix}$$

5 × 1 5 × 2 2 × 1 5 × 1 5 × 1

Variance-covariance matrix

$$\log L = -\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

variance replaced by variance matrix

$$\log L = -\frac{1}{2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

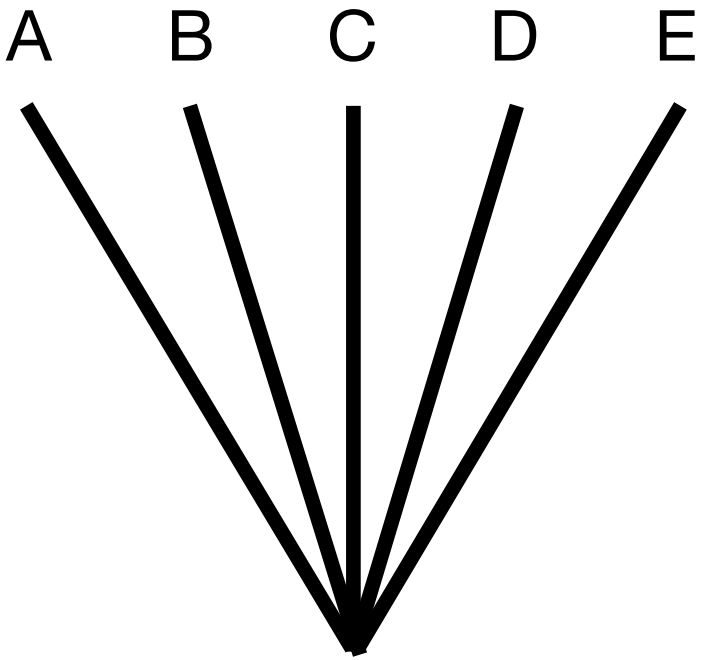
Variance matrix and its inverse

$$\mathbf{V} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

$$\mathbf{V}^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma^2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sigma^2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sigma^2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sigma^2} \end{bmatrix}$$

Variance-covariance matrix and phylogeny

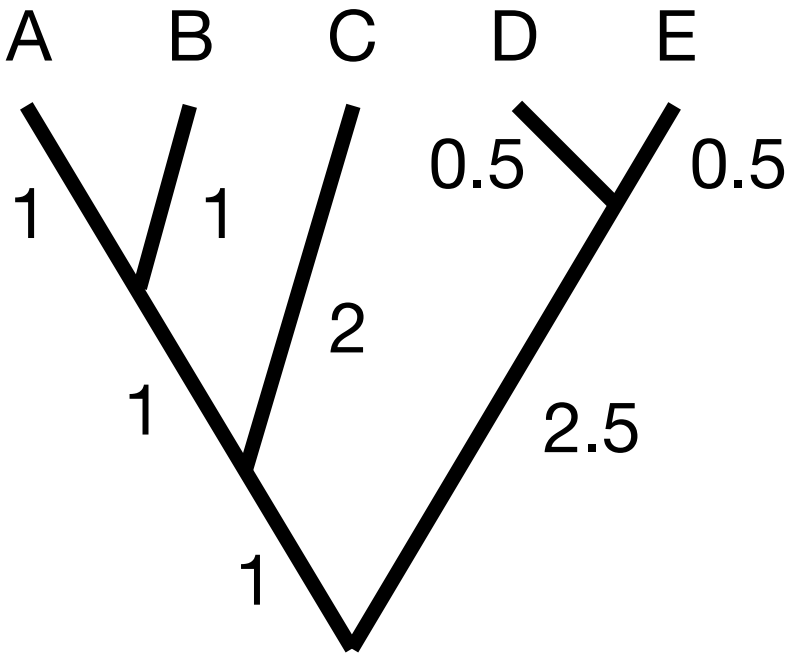
$$\mathbf{V} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$



The matrix V corresponds to a star phylogeny

Variance-covariance matrix and Brownian motion on a phylogeny

$$\mathbf{V} = \sigma^2 \begin{bmatrix} 3 & 2 & 1 & 0 & 0 \\ 2 & 3 & 1 & 0 & 0 \\ 1 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 2.5 \\ 0 & 0 & 0 & 2.5 & 3 \end{bmatrix}$$



Martins and Hansen 1997

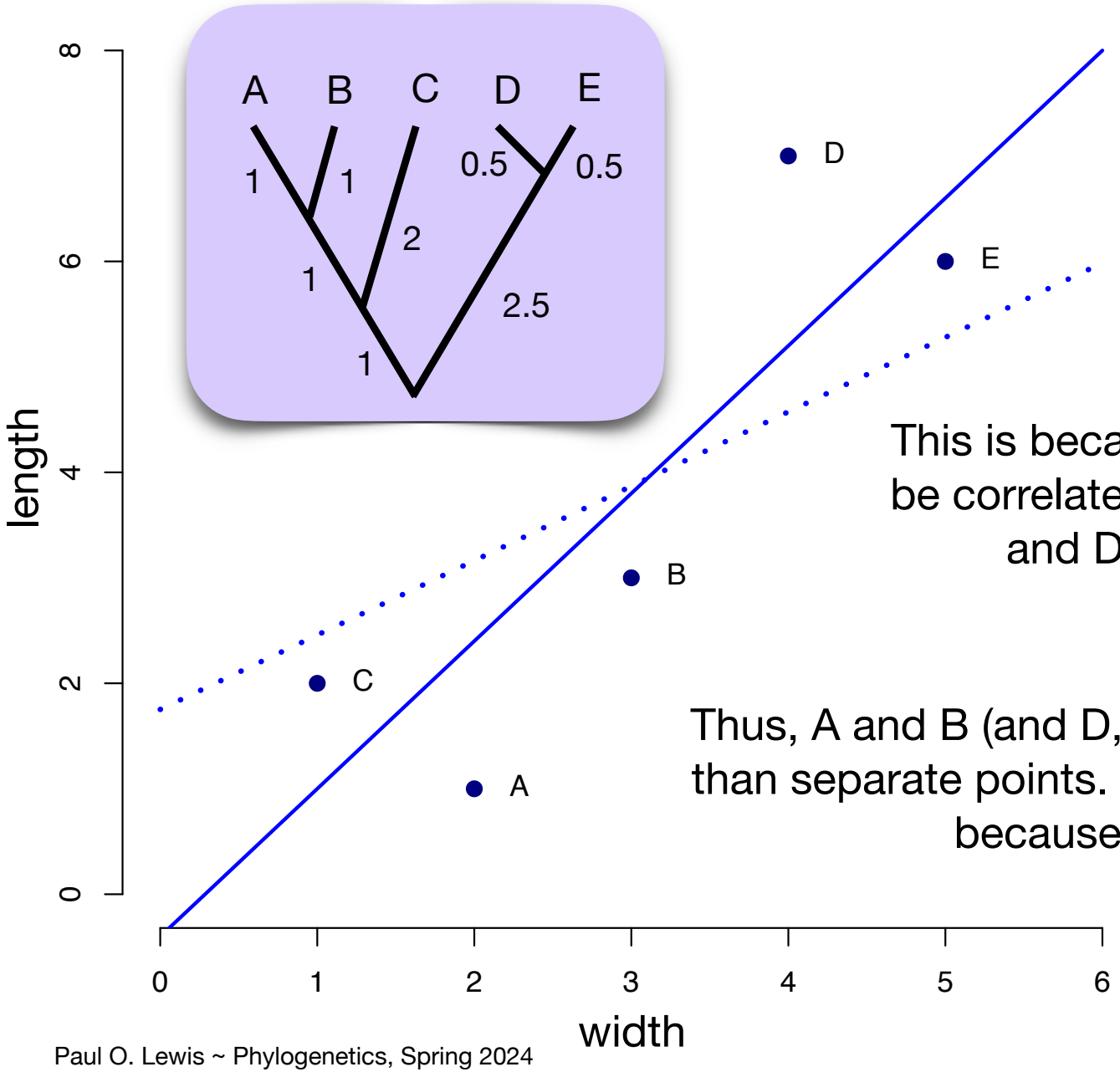
Phylogenetic Generalized Least Squares (PGLS) Regression

$$\underset{2\times 1}{\boldsymbol{\beta}} = \underbrace{\left(\underset{2\times 5}{\mathbf{X}'}\underset{5\times 5}{\mathbf{V}^{-1}}\underset{5\times 2}{\mathbf{X}}\right)^{-1}}_{2\times 2} \underbrace{\left(\underset{2\times 5}{\mathbf{X}'}\underset{5\times 5}{\mathbf{V}^{-1}}\underset{5\times 1}{\mathbf{Y}}\right)}_{2\times 1}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \begin{matrix} \text{intercept} \\ \text{slope} \end{matrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} 1.7521 \\ 0.7055 \end{bmatrix}$$

(for this example)

PGLS Regression



The PGLS regression (dotted) is less influenced by A and D than the non-phylogenetic regression (solid).

This is because A and B are expected to be correlated due to their shared history, and D and E share an even greater fraction of their history.

Thus, A and B (and D,E) act more like single points than separate points. C now wields more influence because of its relative independence.

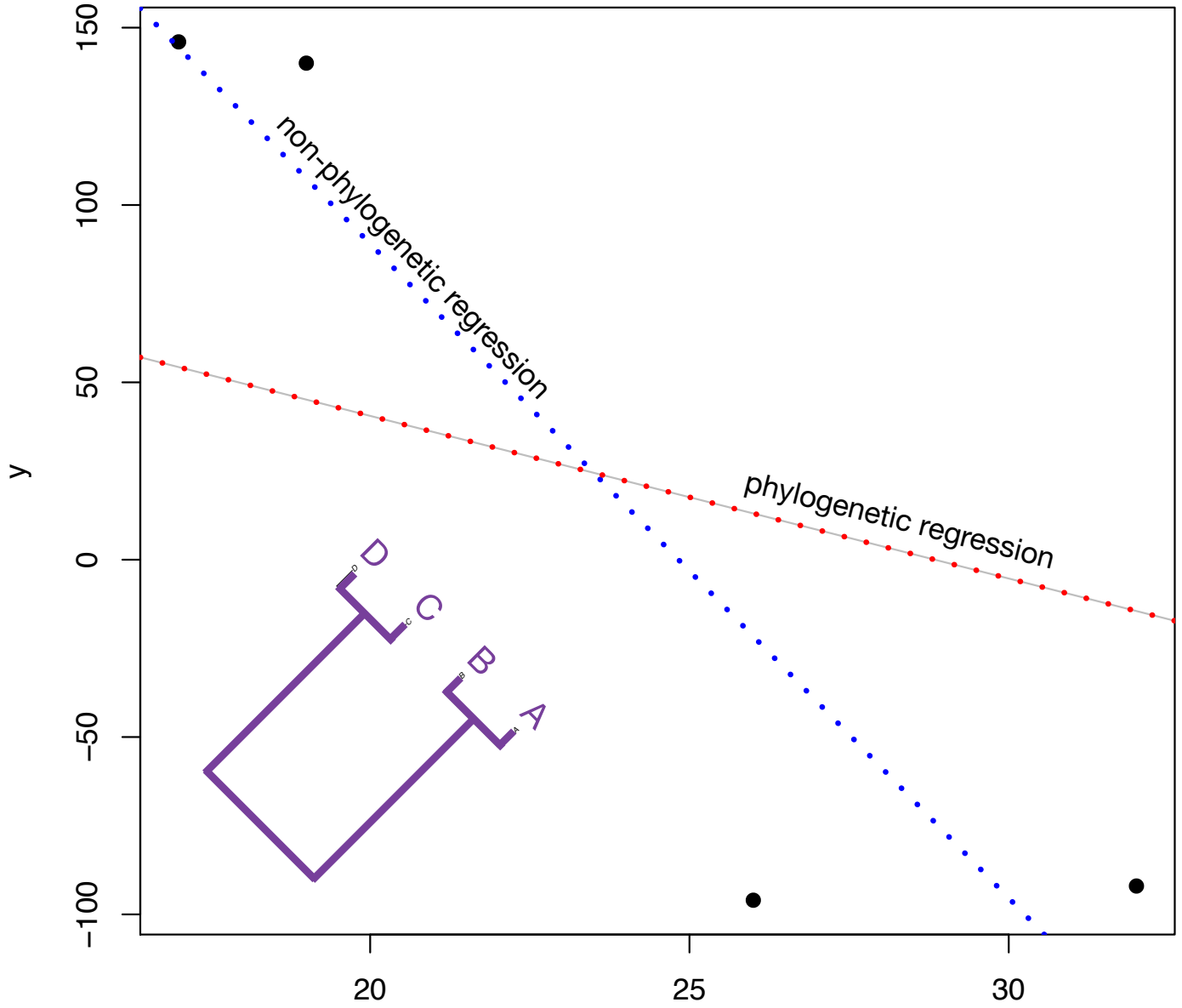
PGLS vs. PIC

- Under some conditions, Phylogenetic Independent Contrasts (PIC) and Phylogenetic Generalized Least Squares (PGLS) regression are identical
- PGLS:
  - Obtain intercept and slope assuming a Brownian motion model
- PIC:
  - Obtain slope for PIC using a linear regression of contrasts with intercept equal to 0
  - Obtain intercept for PIC using the formula:

    intercept = mean(y) - mean(x)*slope

Blomberg et al. 2012

PIC = PGLS


non-phylogenetic regression
phylogenetic regression
D
C
B
A
150
100
50
0
−50
−100
20
25
30
y
x