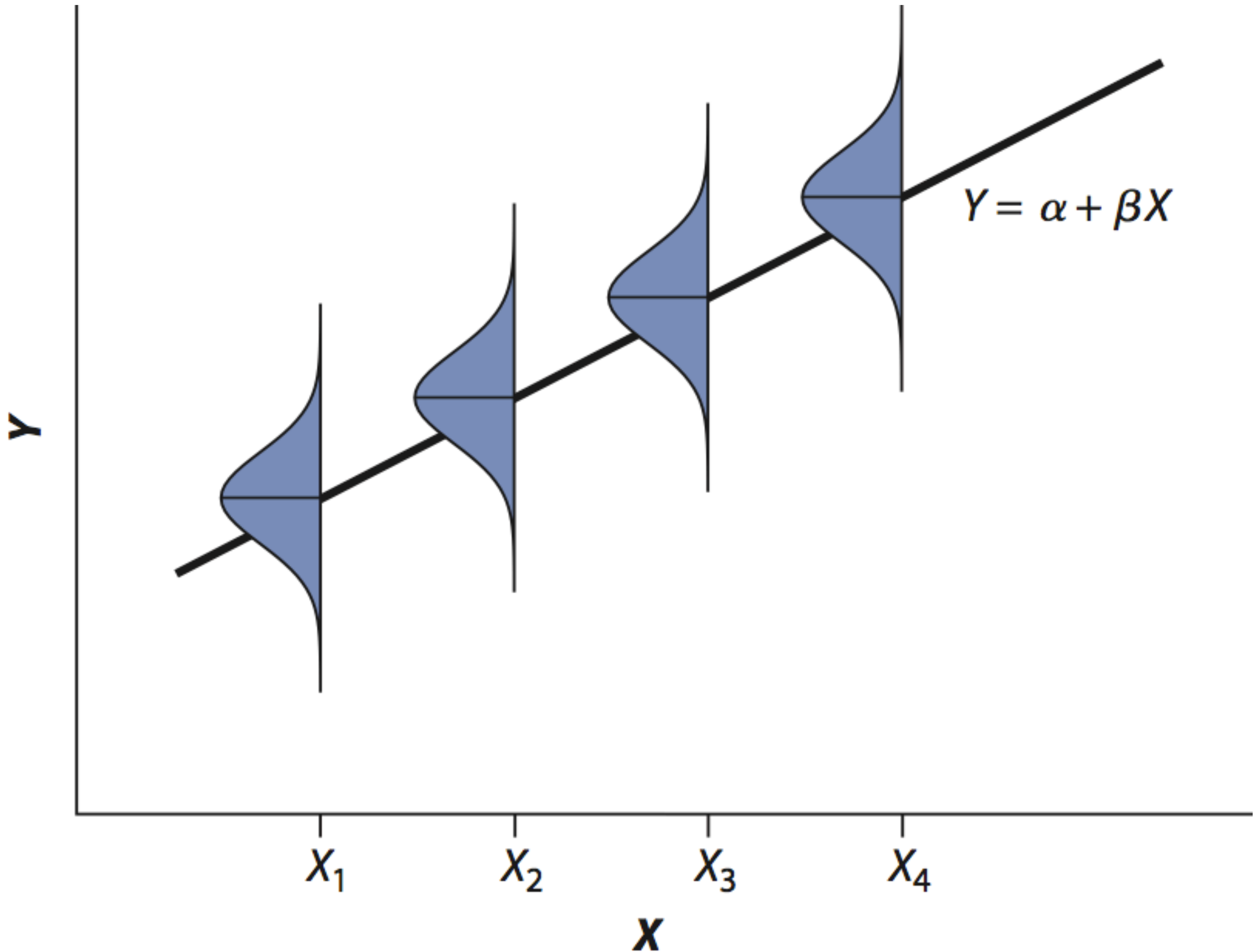
Violations in Assumptions of Regression Analysis

Assumptions of Regression Analysis:



Assumptions of Regression Analysis:

- For each X_i, there is a population of Y values whose mean lies on the ‘true’ regression line
  - For each X_i, the Y are a random sample
  - For each X_i, the Y are normally distributed

- Homoscedasticity
  - For every X_i, the variance of Y is equal

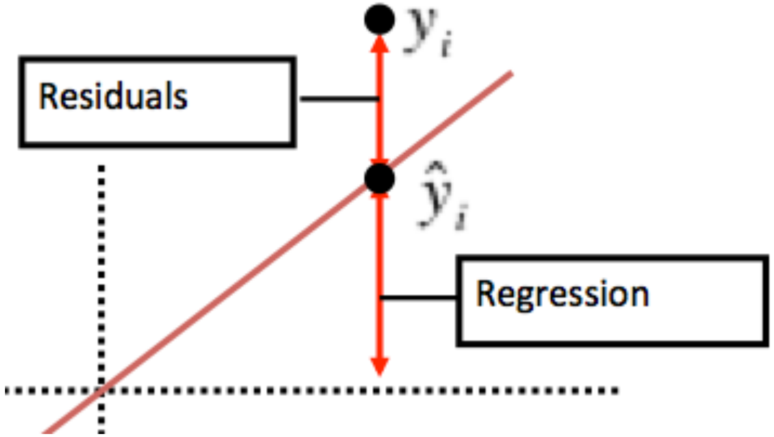
- Nothing is assumed about the distribution of X
  - It doesn’t need to be normally distributed or randomly sampled - they might be fixed by the experimenter

Major types of violation:

1. Outliers
   - Violates homoscedasticity
   - Violates normality of Y
   - May make regression inappropriate
     - Especially if they occur at the boundaries of X
   - Compare results of regression with and without outlier
   - Transformation of data ?

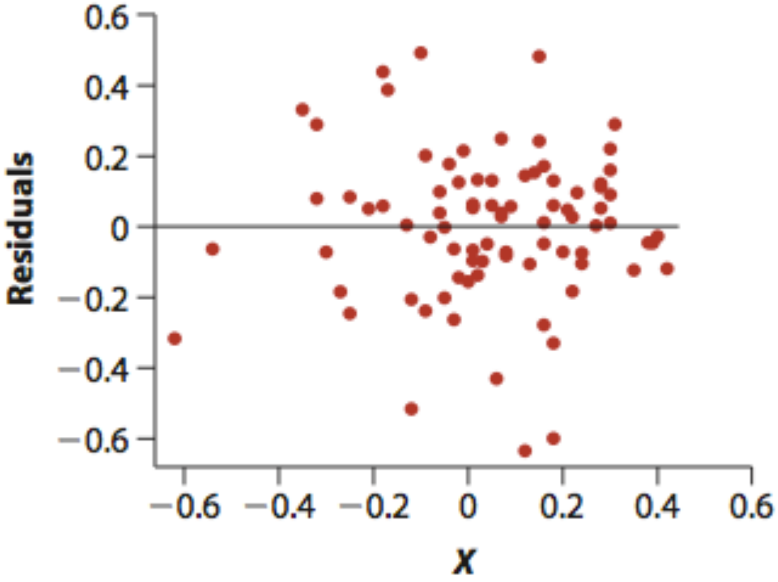
2. Non-linearity (we are dealing with linear regression)
   - Usually done by visual inspection of a scatterplot

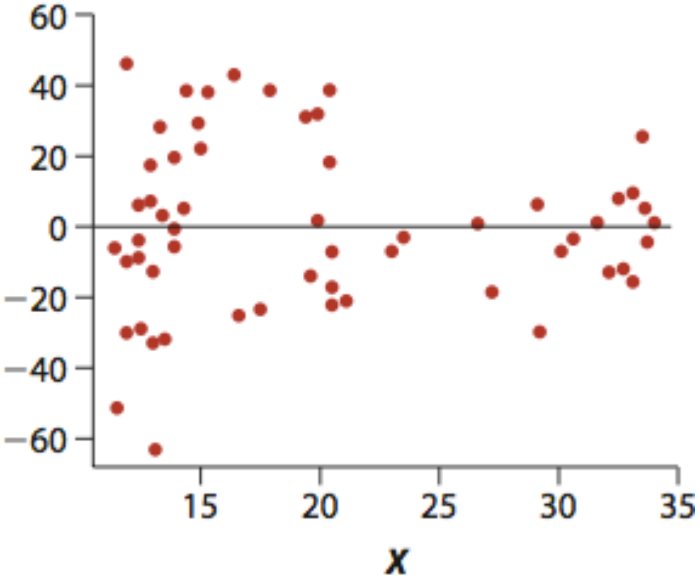
- Residual plot:
  - Help assess assumptions
  - Residual ($Y_i - \hat{Y}_i$) is plotted against X_i

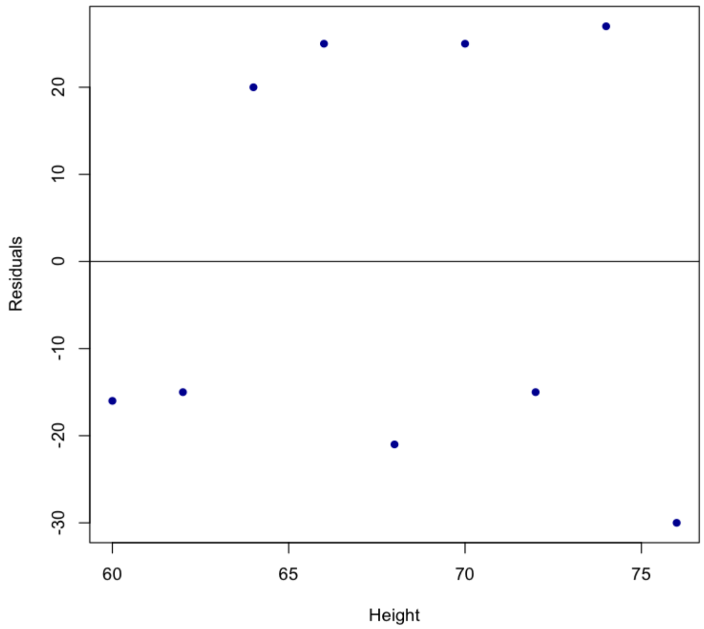


  - If assumptions about normality and homoscedasticity are correct:
    - Symmetric cloud of points above and below horizontal line
  - Use a computer

Regression Assumption Violations

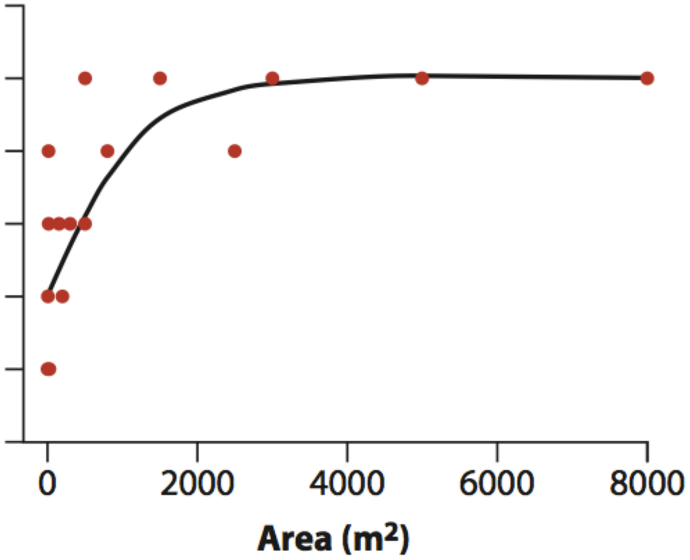


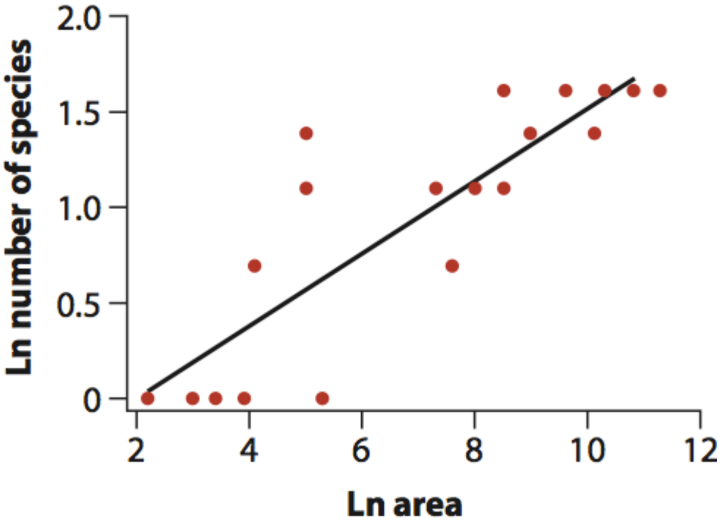




Regression Assumption Violations

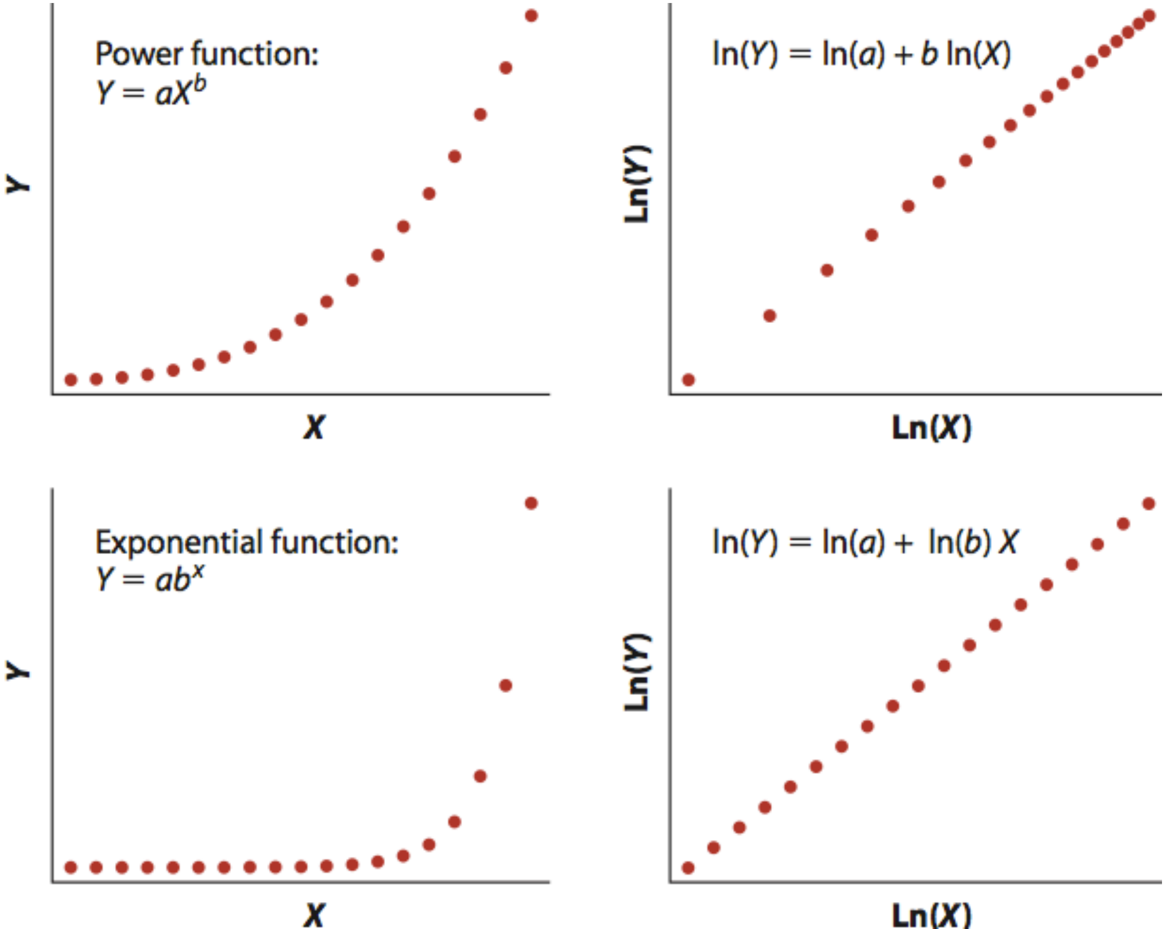
- **Transformations:**
  - Non-linear relationships can sometimes be forced into linearity





Regression Assumption Violations

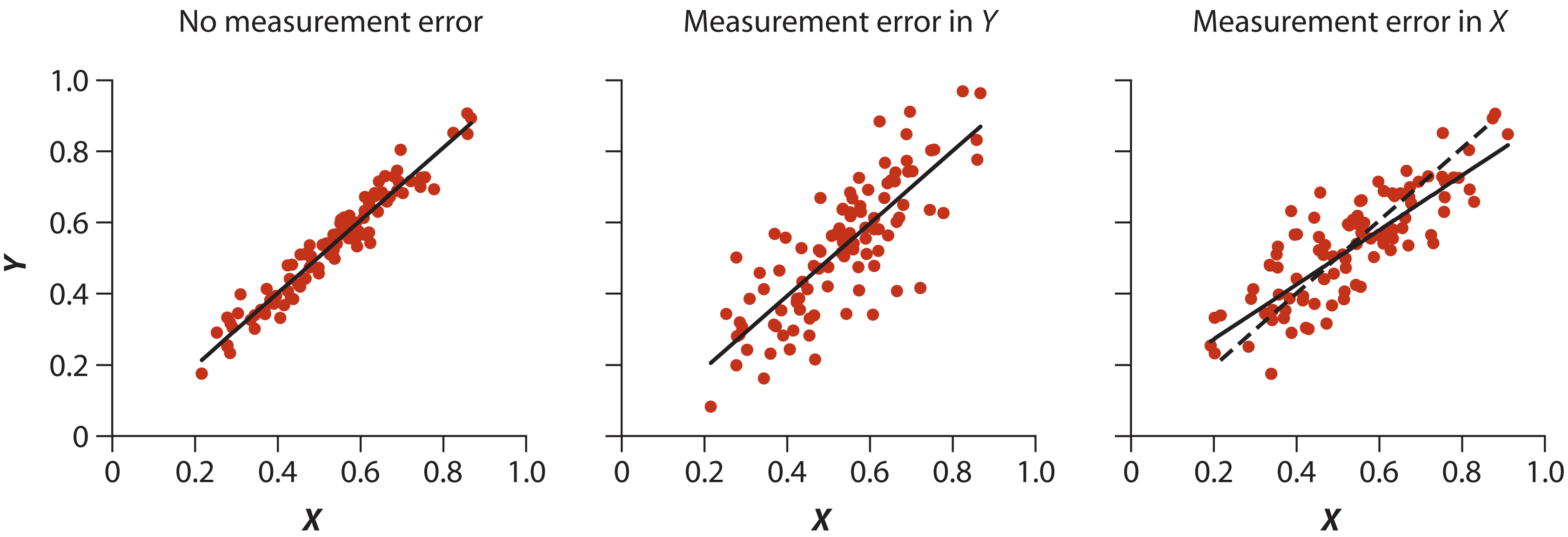
- **Transformations:**
  - The usual suspects:
    - log transformation for power and exponential relationships



- Measurement error:
  - Biological traits can be difficult to measure accurately
  - Effects of measurement error depends on the variable
    - **If measurement error occurs on Y**
      - Increase variance of residuals
      - Increases SE of slope
    - **If measurement error occurs on X**
      - Increases variance of residuals
      - **Causes bias in estimate of b (**underestimates slope)
        - » b will lie closer to 0 than β
        - » Remember: BIAS is really bad!

Regression Assumption Violations

- **Measurement error:**



Regression Assumption Violations

What happens if transformations don’t work?!

Or… linear regression is inappropriate?

Non-linear regression

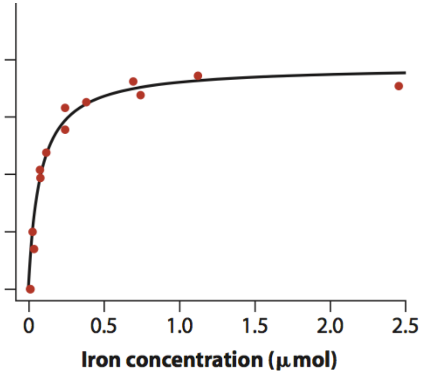
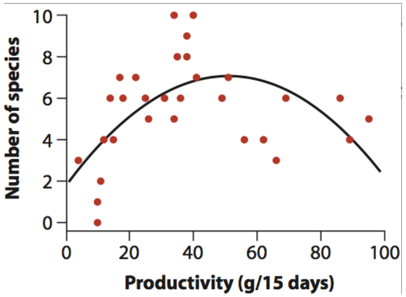
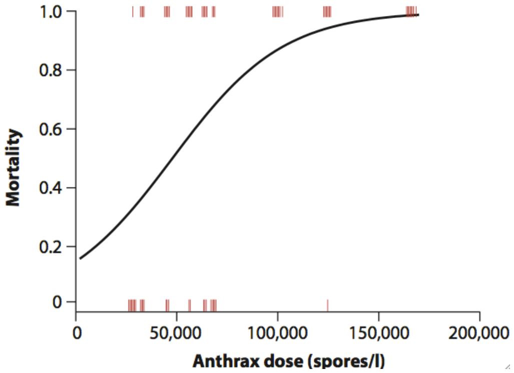
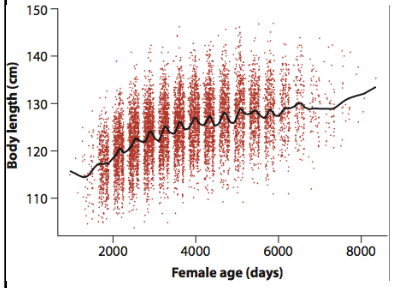
Non-linear Regression:

- Same assumptions are linear regression but, obviously, doesn’t assume a linear relationship

- Keep it simple

- Don’t **over fit**
  - It is possible to get a curve that fits each and every point ($MS_{residual}$ = 0) but it will not predict future points since the curve ***doesn’t describe a general trend***

Regression Assumption Violations

Curve with Asymptote	**Quadratic curve**	**Binary response Variable**	**Smoothing**
$Y = \dfrac{aX}{b+X}$	$Y=a+bX+cX^2$	Log-odds(Y)=a+bX	• depends on data
Michaelis-menten eq$\underline{^n}$	Parabolic relationships	**Dose response curve**	Diagnosis of exclusion
			

Interleaf 11: Using species as data points

- Species are not ***independent*** because they share a common evolutionary history

- Phylogenies illustrate (hypothesized) ancestor-descendant relationship
    - Phylogenetically independent contrasts (Felsenstein, 1985) – a little like a paired t-test for each node:
    - http://ib.berkeley.edu/courses/ib200b/lect/ib200b_lect08_Ginger_Jui_PICs.pdf
    - https://slideplayer.com/slide/7838671/
    - https://biology.ucr.edu/people/faculty/Garland/Garland_JoeFest_1_Upload_Post.pdf

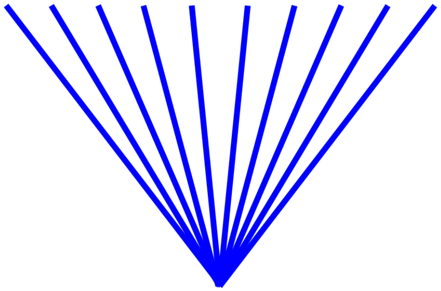
    - Many computer programs to deal with this issue but they all have their own baked-in assumptions that you should understand
    - Most commonly based on “random walk”/”Brownian motion”

Interleaf 11: Using species as data points

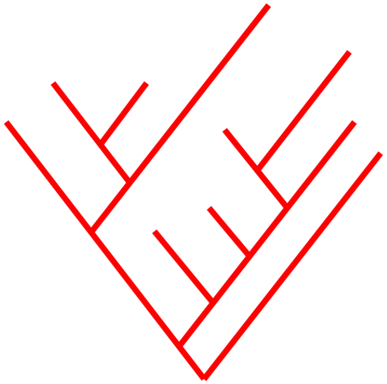
– Species are not ***independent*** because they share a common evolutionary history

https://biology.ucr.edu/people/faculty/Garland/Garland_JoeFest_1_Upload_Post.pdf

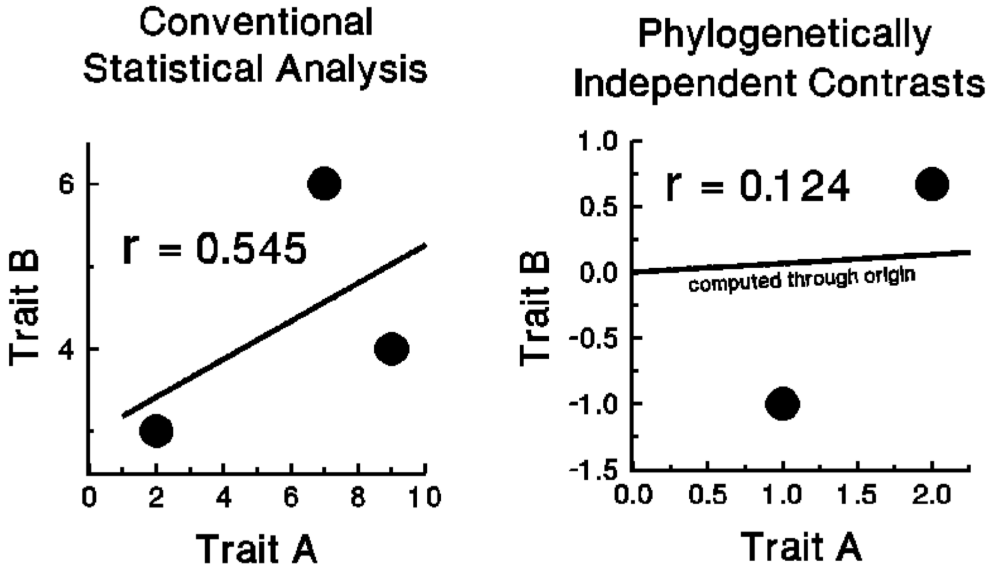
What Conventional Statistical Methods Assume



What Evolution Provides



It can make a big difference!



d.f. = 1

d.f. = 1