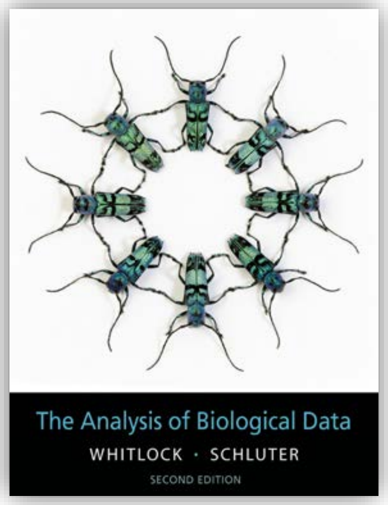
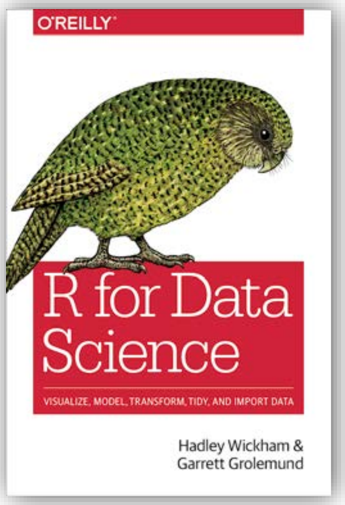
Data Science in Bioinformatics

Palle Villesen & Thomas Bataillon





Outline for week 11

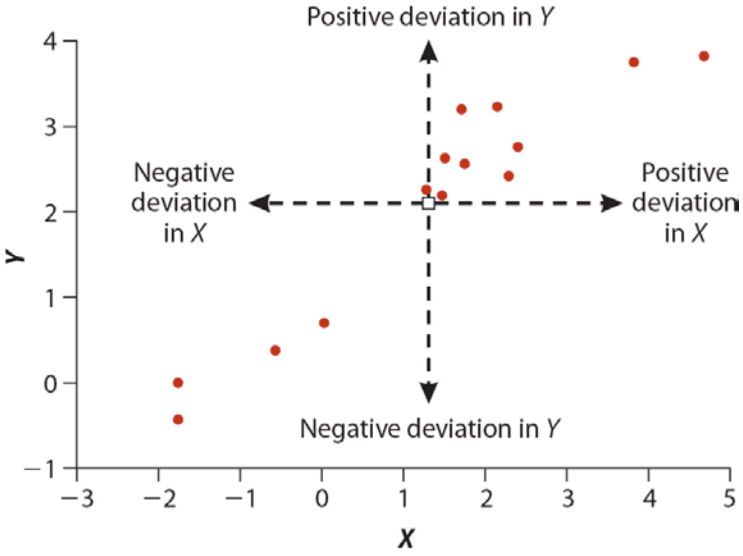
- Chapter 16+17
  - Open discussion & Exercises
- Thursday
  - Who will be responsible

Correlation and regression

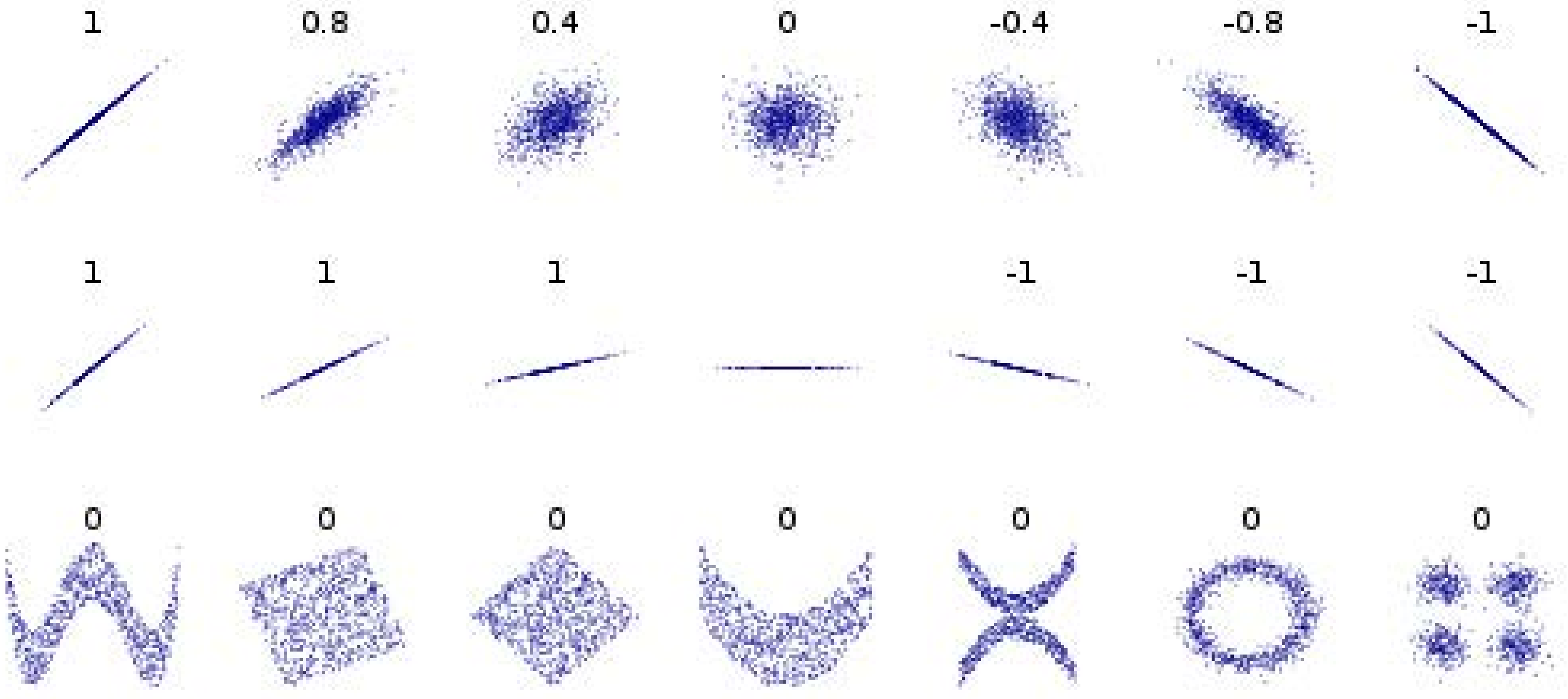
- Is two variables independent or not?
- How much of the variation in Y is explained by the variation in X?
- What is the standard error?
- How strong is the effect?
- How do we test it?

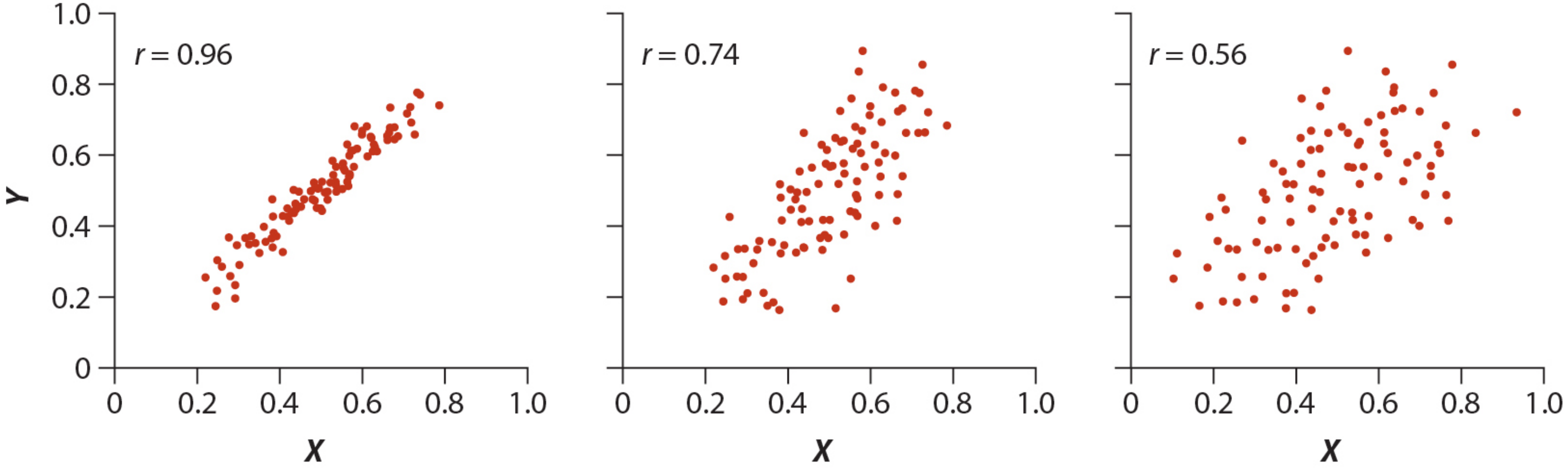
Pearson correlation

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$



Correlation





r = 0.96
r = 0.74
r = 0.56
1.0
0.8
0.6
0.4
0.2
0
Y
0
0.2
0.4
0.6
0.8
1.0
X
X
X

Sample size and correlation

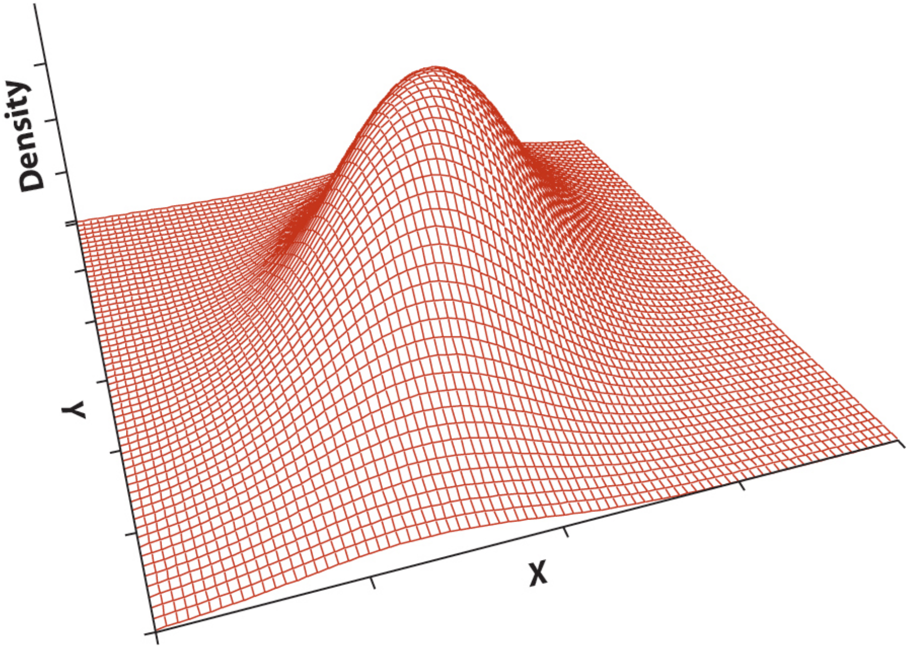
- $SE_r = \sqrt{\frac{(1-r^2)}{n-2}}$

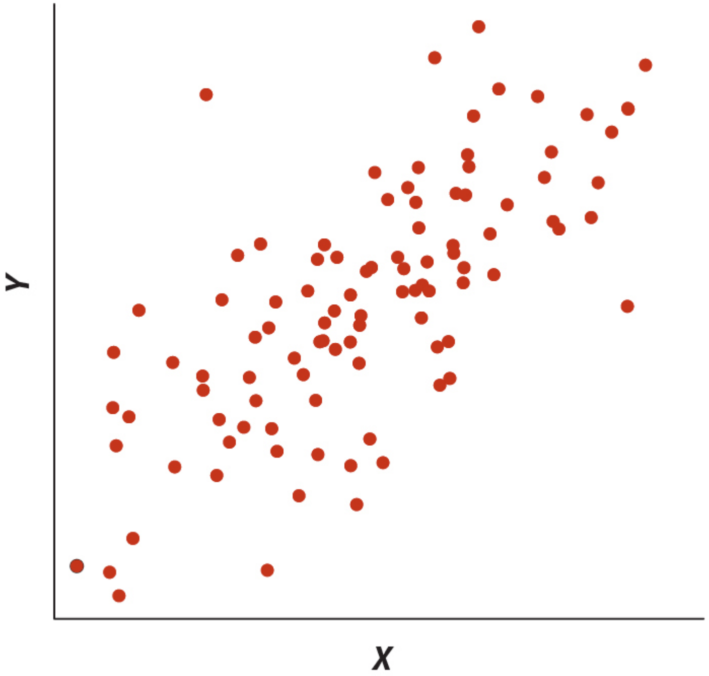
- **Confidence interval**
- Correlation coefficient is not normally distributed
- We convert it to “z” that is normal – calculates 95% CI and converts back from z to r
- Or we bootstrap the confidence interval

Testing correlation

- H0: the population correlation is 0
- HA: the population correlation is not 0
- t = r/SE, Use t distribution with n-2 df
- cor.test
- **Alternative way**
- Permutation test
  - Sample shuffle x and y, calculate r
  - How many are more extreme than observed r?

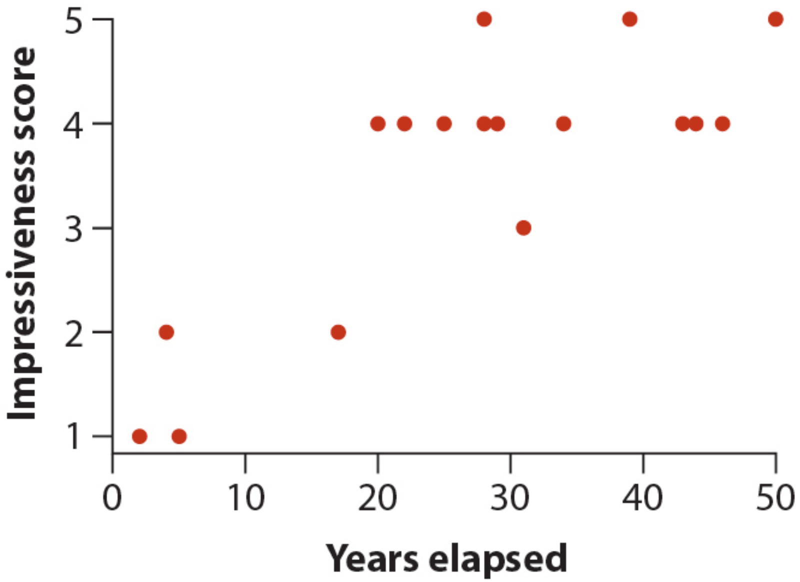
Assumptions
bivariate normal distribution



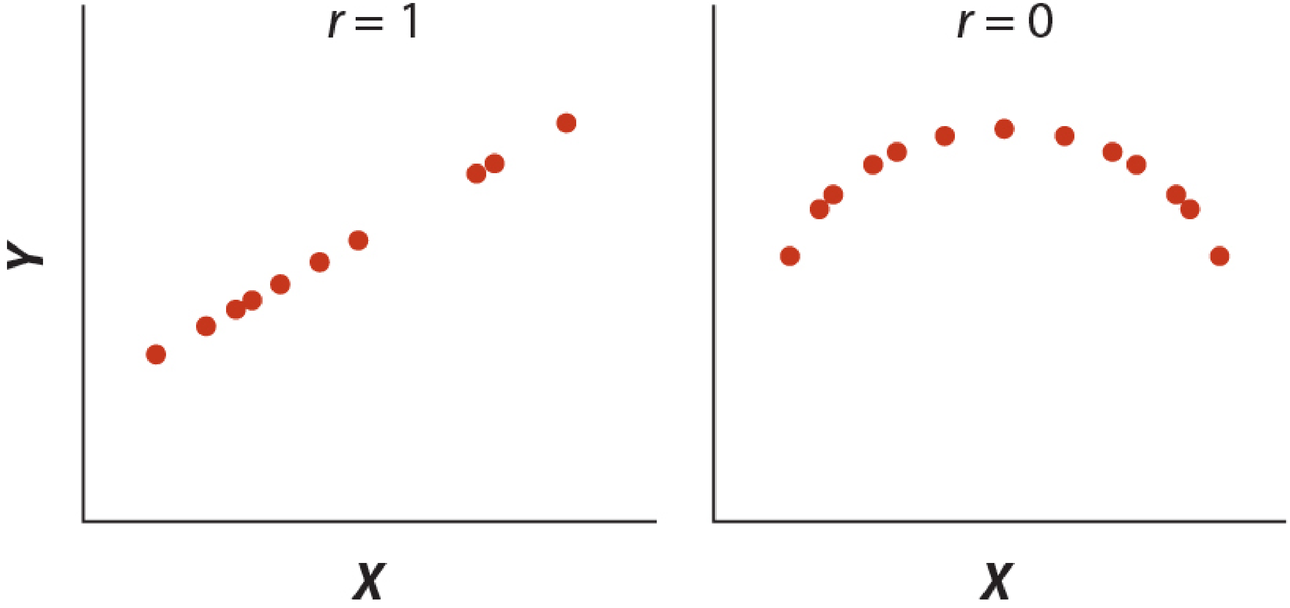


Other stuff

- Rank correlation
  - No assumption on distribution of x and y
  - It ranks all data and works on ranks



Low correlation?

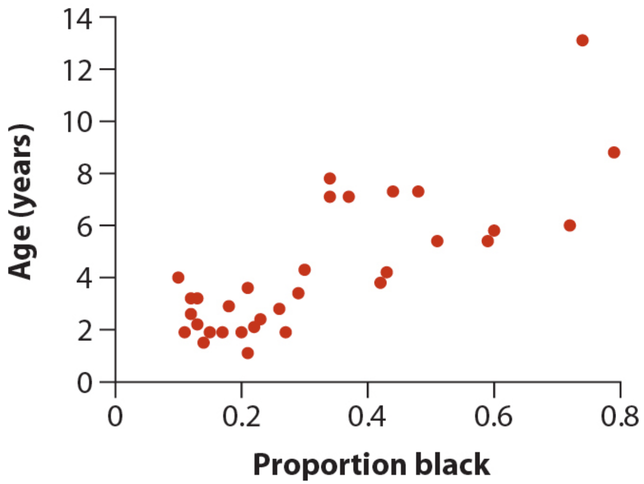


Linear regression

- $Y = \beta_o + \beta_1 X \ldots$
- $Y = \beta_o + \beta_1 X + \varepsilon$

- It is a **prediction** method
- Also called statistical learning (or machine learning)
  - Prediction and/or inference
- Supervised learning (we need known data to build the model)

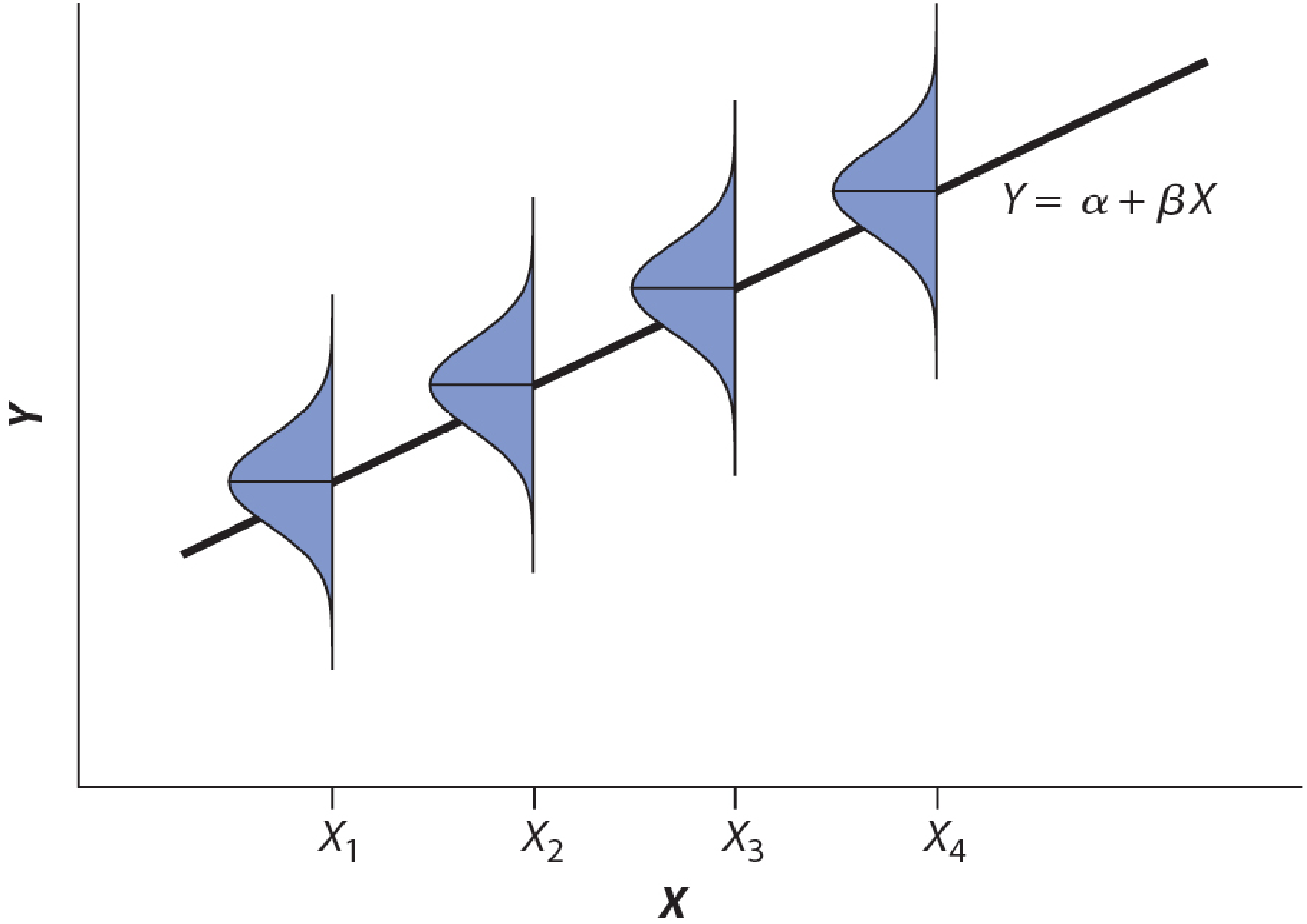
Predict lion age from nose



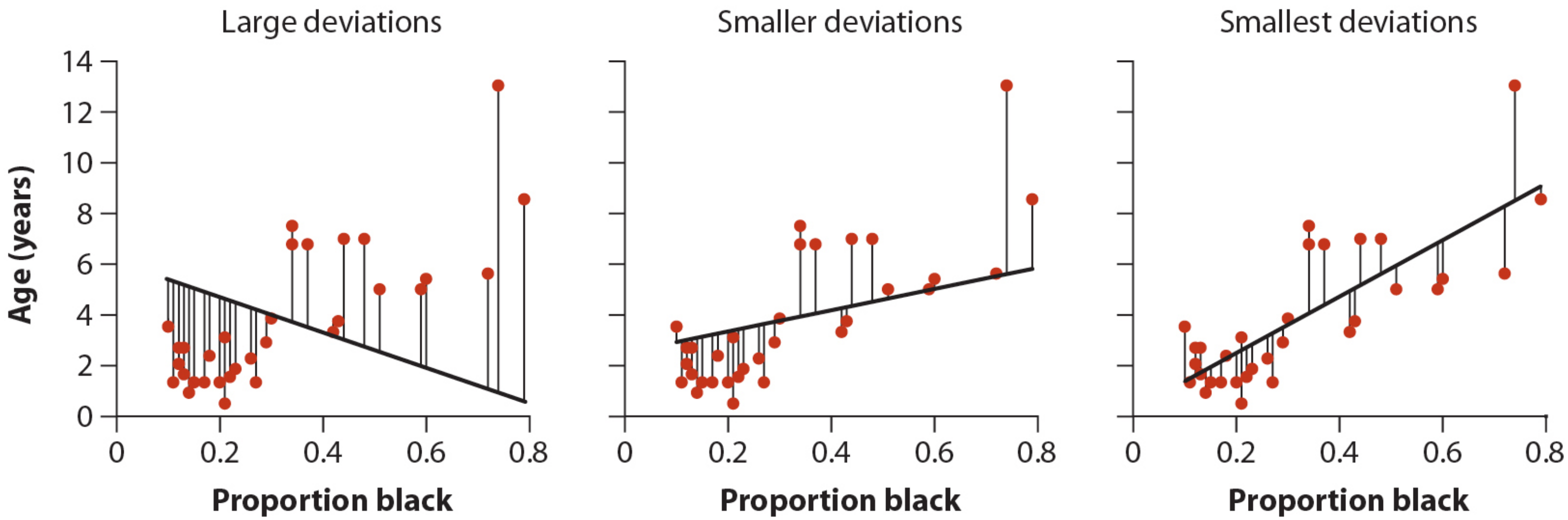
Essentials of regression

- Assumptions
- Tests & Measure of Fit
  - Parametric
  - Resampling methods
- Validation / Inspection
  - Visual
  - Test on residuals
  - Influence of outliers, colinearity

Assumptions…



A visual on least squares…



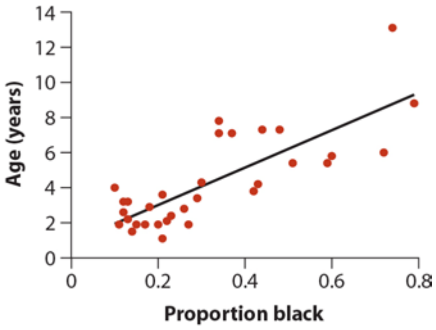
$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x},$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

$$= r_{xy}\frac{s_y}{s_x}.$$

minimum prediction error



The model for lions

- Age = 0.88 + 10.65 * proportion black

- Standard error of slope
- Confidence interval for the slope

- Also possible by bootstrapping
- CI for slope: 7.56 < slope < 13.73

What can we do

- We have shown that nose patterns change with age (inference)
- We can estimate how much it changes
- We can calculate how much of the variation is explained by this linear relation ship (R^2)
- We can predict the age of lions

R^2 = SS(regression) / SS(total)

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

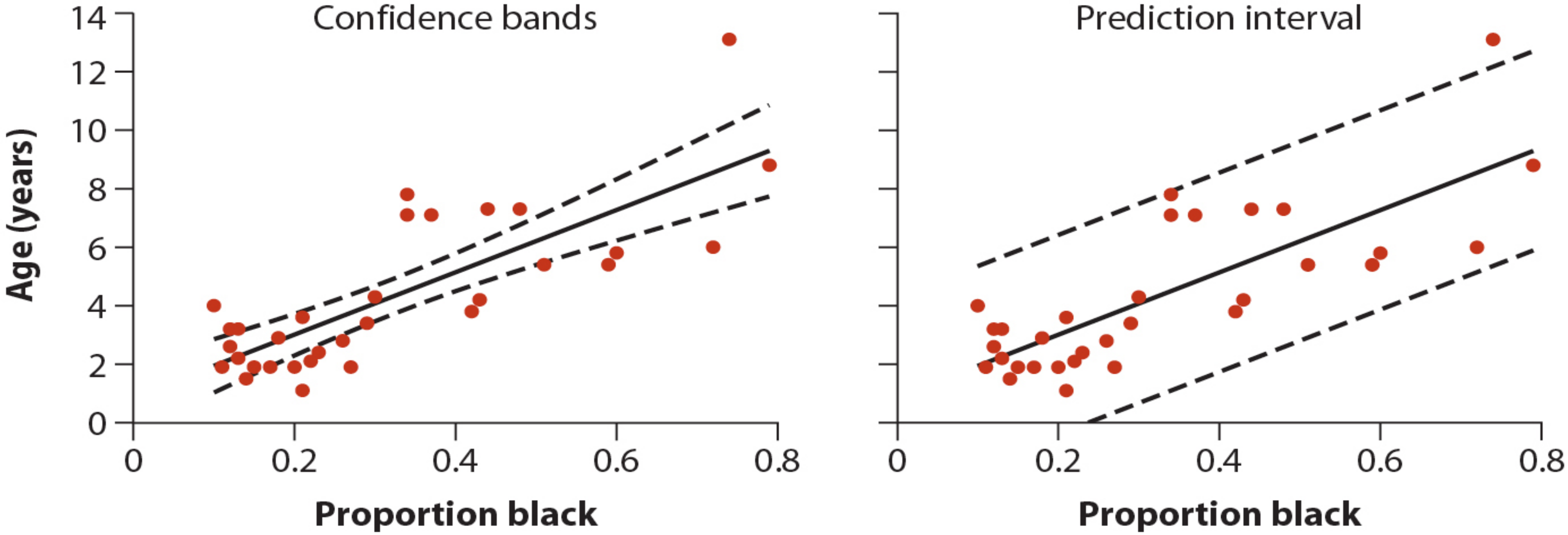
$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

$$RSS = \sum_{i=1}^{n} (\varepsilon_i)^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Testing the slope

- H0: slope = 0
- HA: slope != 0
- T.test – but for any practical purposes R will do an ANOVA test
- Permutation test? How?

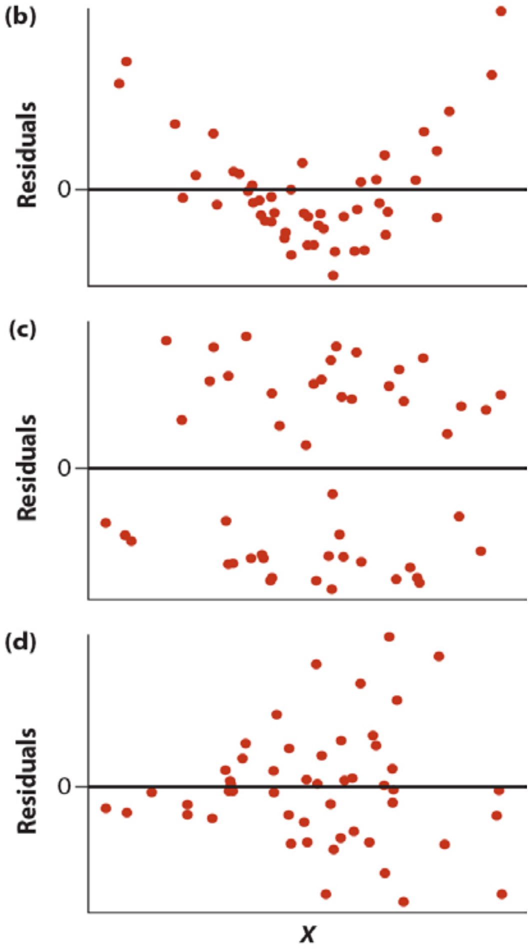
Predictions



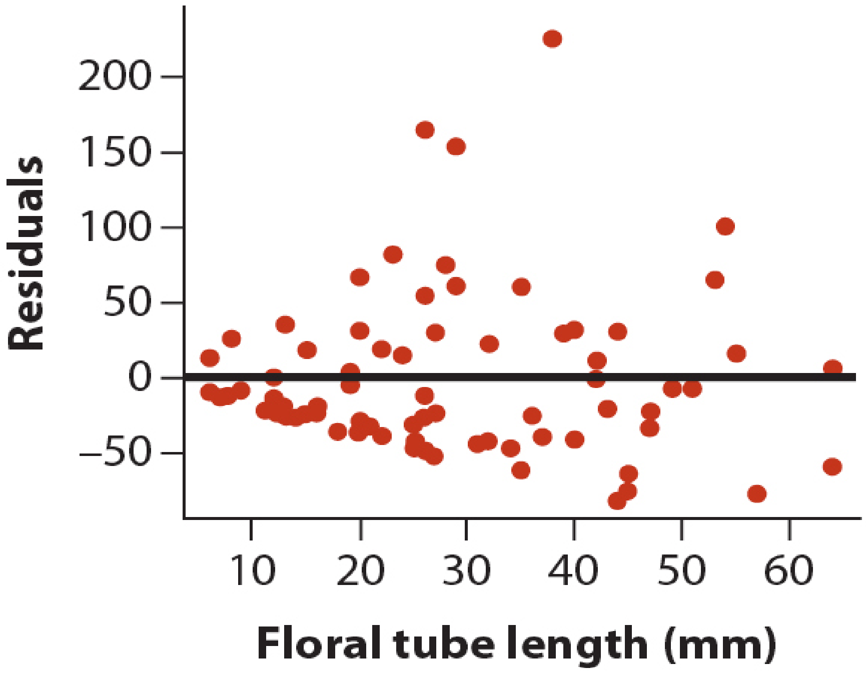
Precision of the predicted **mean**

Precision of the predicted **value**

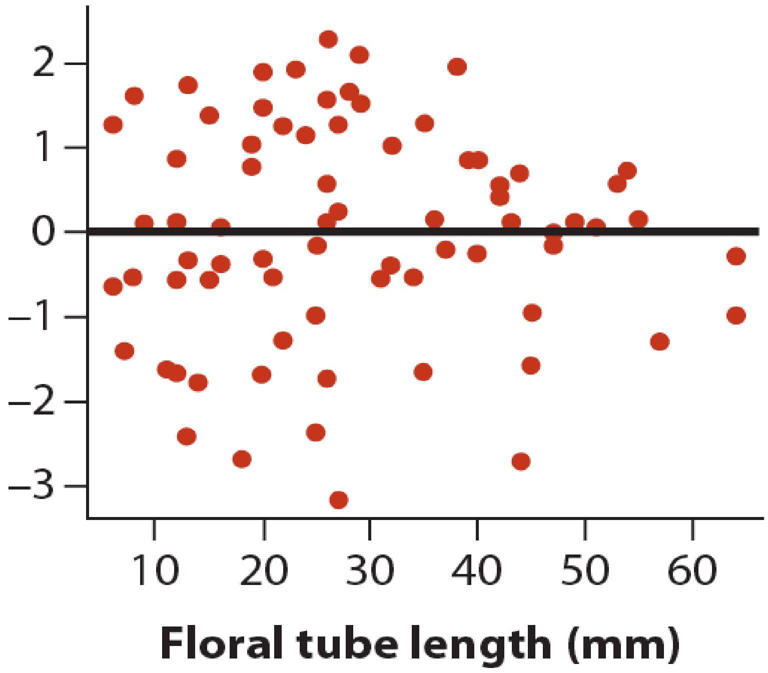
Residual plots are essential



Inspecting residuals

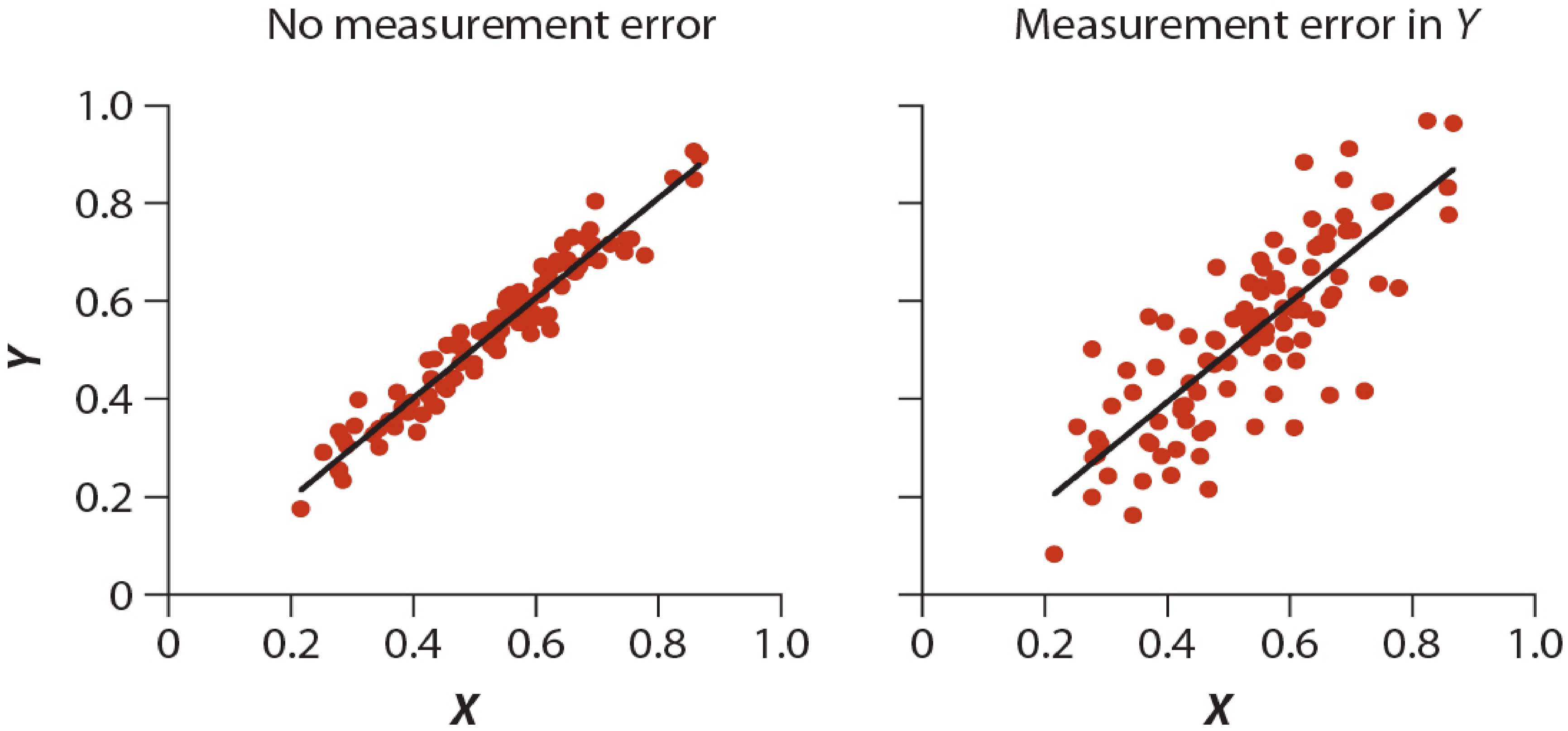


y = a + bx

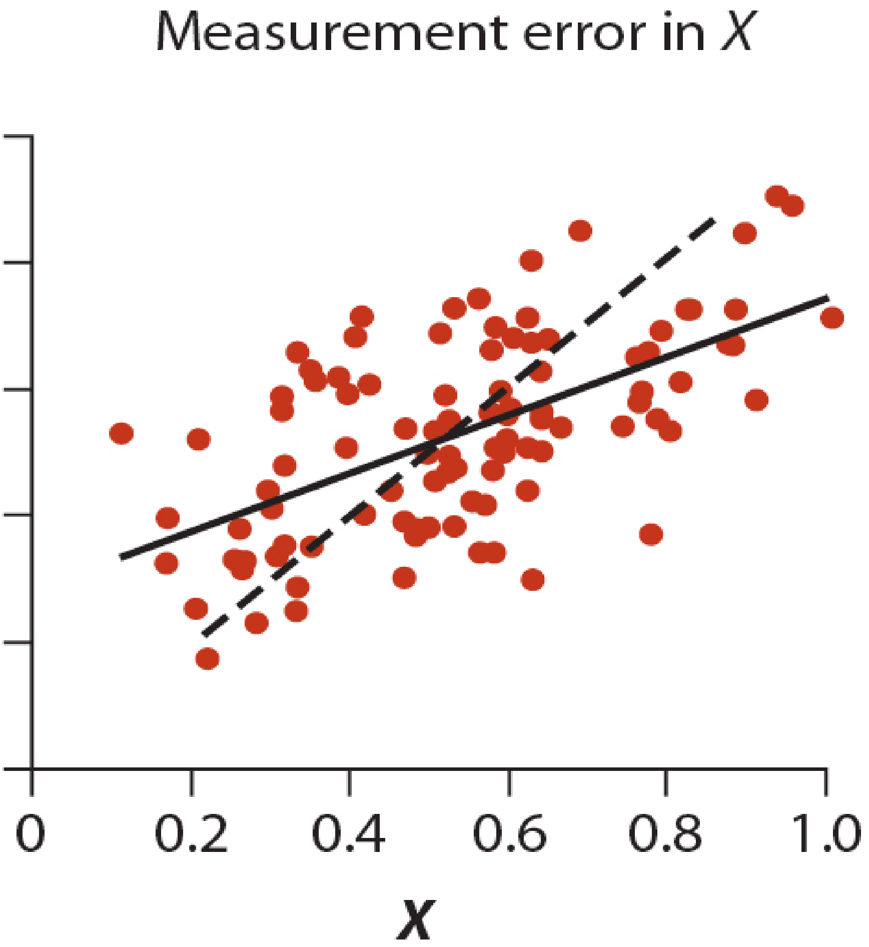


sqrt(y) = a + bx

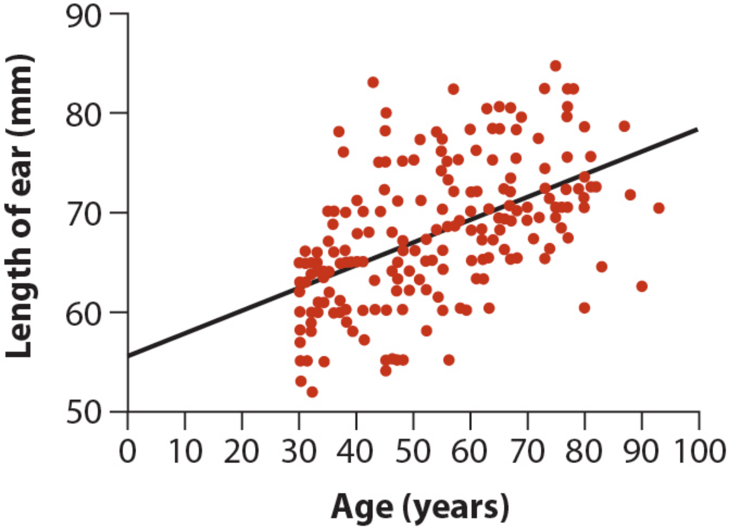
Measurement errors on Y…



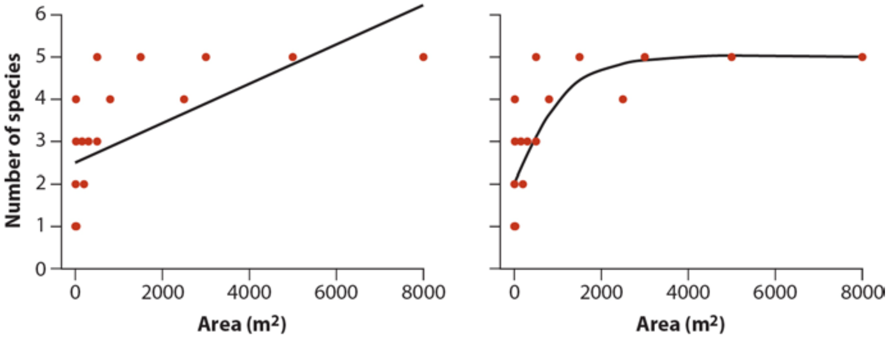
Measurement errors on X

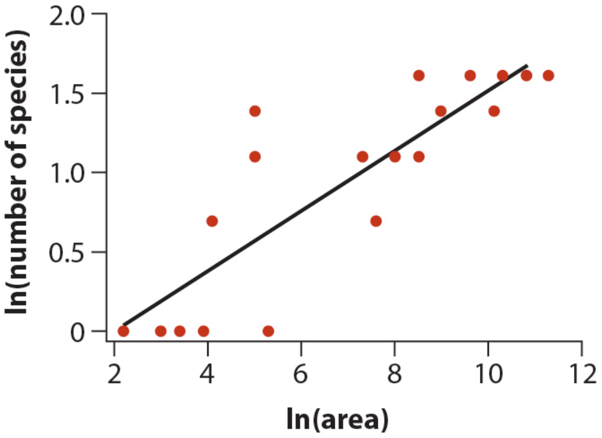


Extrapolation



Transformation



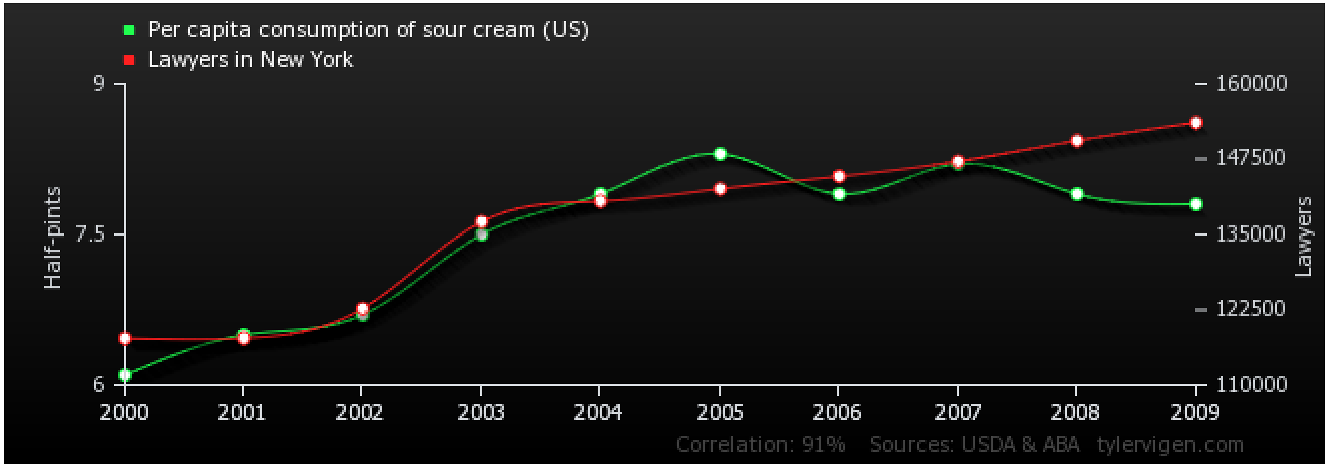


How are they connected?

- What is the difference between correlation and regression?
- What are similarities?
- Sanity checks?
- Transformations?

What does it mean?

- Correlation != causation



Correlation: 91% Sources: USDA & ABA tylervigen.com

Exercises

- Distribution of levels of gene expression in Orangutans
- Is evolution (dN/dS) related to gene expression?
- We will summarize all genes pr. Chromosome
- Chromosomes with high gene expression
- Chromosome with high dN/dS
- What is dN/dS?
- What do you expect? Why?

Pet project – guns and USA

- I have taken data on gun ownership pr. State and combined with death rate (by firearms)
- Also data on countries in the world

- Are these two variables correlated?
- Is it significant?
- How strong is the effect? (slope)
- What is R^2 ?