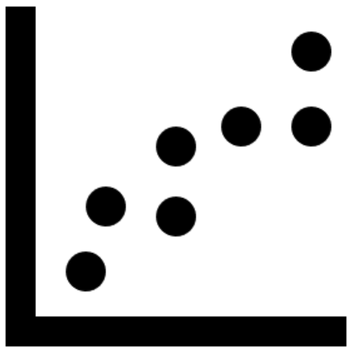
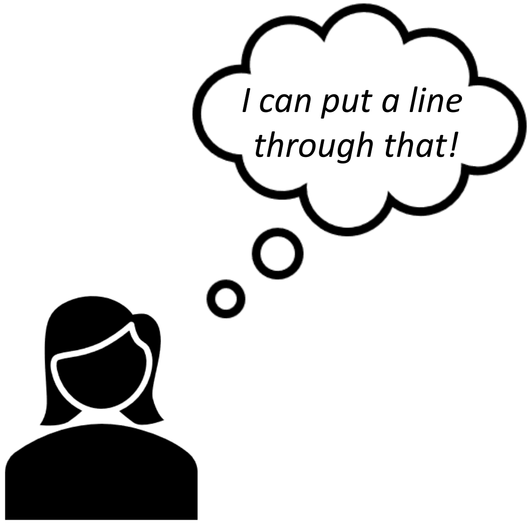
Simple Linear Regression
(Part 1)

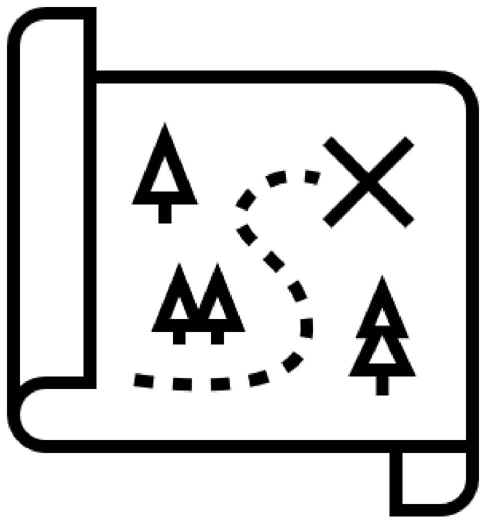




Guest Lecture By:
Brian Gallagher
PhD Candidate
Department of Biology

Goals for today

- Review basic statistical analyses and the types of variables used
- Introduce linear regression and its applications
- Learn about how regression models are ‘fitted’ to real data
- Interpreting coefficients from regression models



Where are we going?

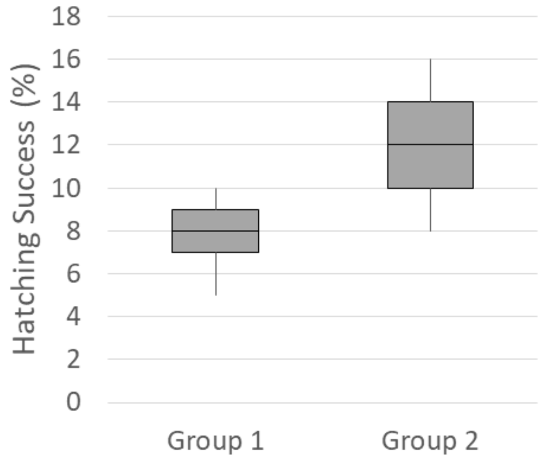
Statistical analysis

- Science is hard, but statistics help us make inferences based on data

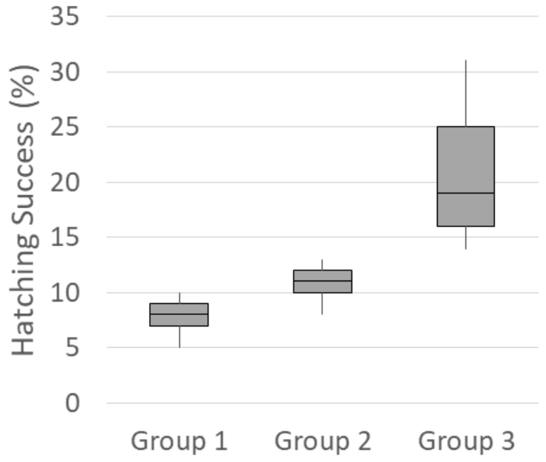
- Many analyses are based on detecting **differences in means**
  - Do means differ in two groups? (t-test)
  - Do means differ across multiple groups? (ANOVA)

- Regression builds directly off these previous analyses
  - Does the mean value of one variable change based on another variable?
  - Uses **continuous variables**

Analysis	Data type	N groups
t-test	Categorical	2



Analysis	Data type	N groups
ANOVA	Categorical	>2



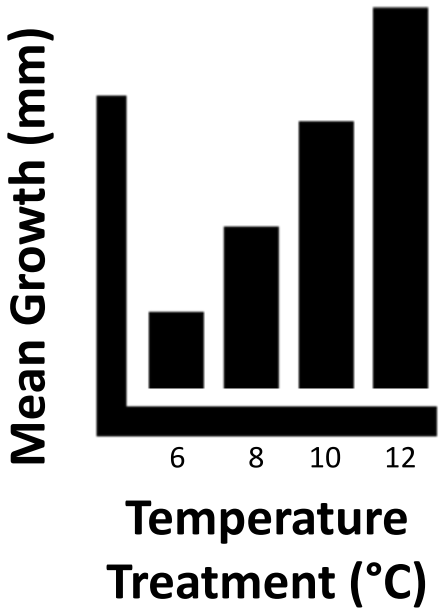
Categorical vs continuous variables

- Question: how does fish growth rate respond to increasing temperature?
  - Two possible ways to answer this

- <u>Approach 1:</u> conduct experiments where fish grow in tanks at different temperatures that are precisely controlled
  - Temperature is **categorical** (ANOVA)

- <u>Approach 2:</u> measure growth rates of fish that experience different temperatures in the wild
  - Temperature is **continuous** (regression)

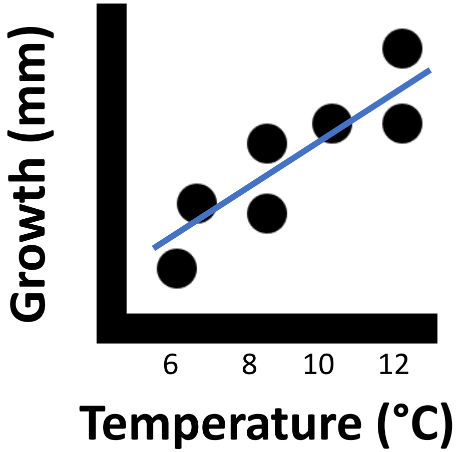
- Very similar qualitatively, but regression provides **different quantitative results**



<u>Approach 1:</u>

Qualitative: Fish growth increases with increasing temperature

Quantitative: Mean growth differs significantly between temperature treatments



<u>Approach 2:</u>

Qualitative: Fish growth increases with increasing temperature

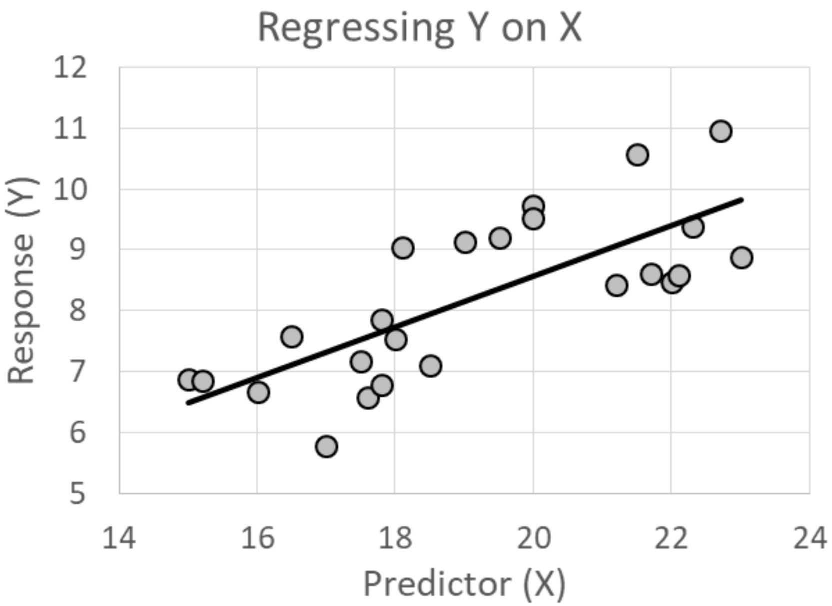
Quantitative: Mean growth *increases by 2 units for every 1°C increase in temperature*

Regression

- Regression estimates the relationship between **two continuous variables**
  - Response (Y), or dependent variable
  - Predictor (X), or independent variable

- Y is always **regressed on** X, and relationship is expressed as a linear model
  - Intercept (a)
  - Slope (b)

- Slope and intercept must be **estimated** from the data



$$Y = a + b(X)$$

<u>Intercept:</u>
Value of Y when X=0

<u>Slope:</u>
How Y changes in response to changes in X

**** Note: regression notation can vary A LOT depending on the source!***

Fitting regression models

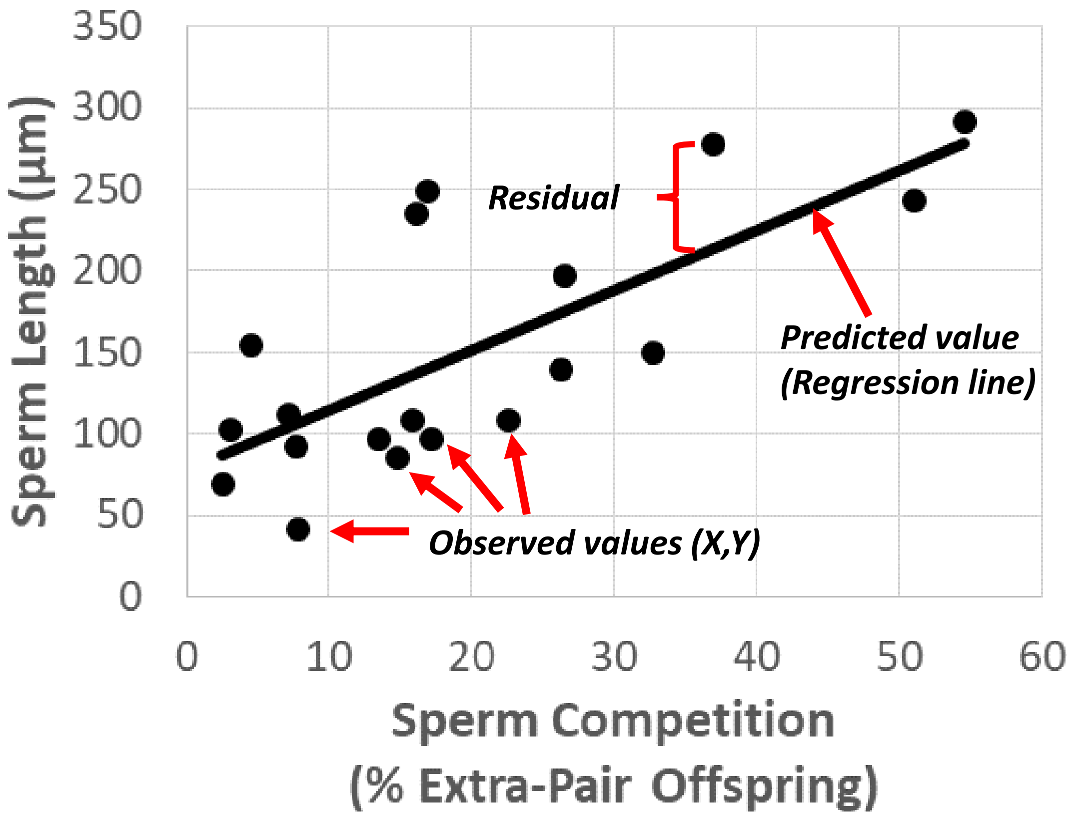
- Three main components

- Observed values: raw data points with Y and X coordinates

- Predicted values: line predicting the **average value** of Y at each value of X

- Residuals: **differences** between observed and predicted values (on Y-axis)
  - Also called "deviations"

- Ideally, observed and predicted values should be **similar**



Process:
In birds, males and females will often form pairs to raise young together. However, sometimes females will mate with other males (extra-pair reproduction), driving sperm competition.

Qualitative summary:
Sperm length tends to increase on average with increasing sperm competition.

Data from: Birkhead and Montgomerie (2020)

Example: lion conservation

- Trophy hunting can provide revenue that helps fund conservation initiatives
  - Sustainability is crucial

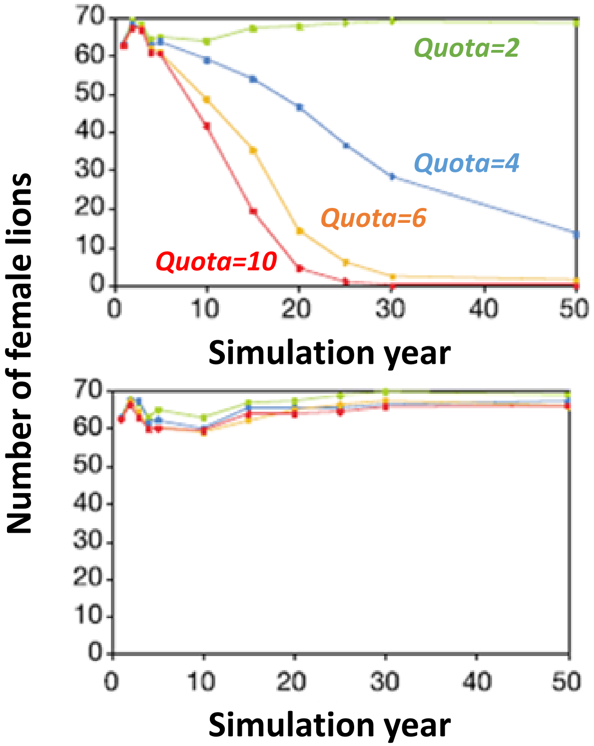
- Researchers found that hunting **male lions 6 years old or older** had negligible impacts on long-term abundance
  - Driven by social structure and infanticide

- Question: is there an easy way to estimate the age of individual lions?
  - Can help ensure sustainable hunting

Sustainable trophy hunting of African lions

Karyl Whitman, Anthony M. Starfield, Henley S. Quadling & Craig Packer

Department of Ecology, Evolution and Behavior, University of Minnesota, 1987 Upper Buford Circle, Saint Paul, Minnesota 55108, USA

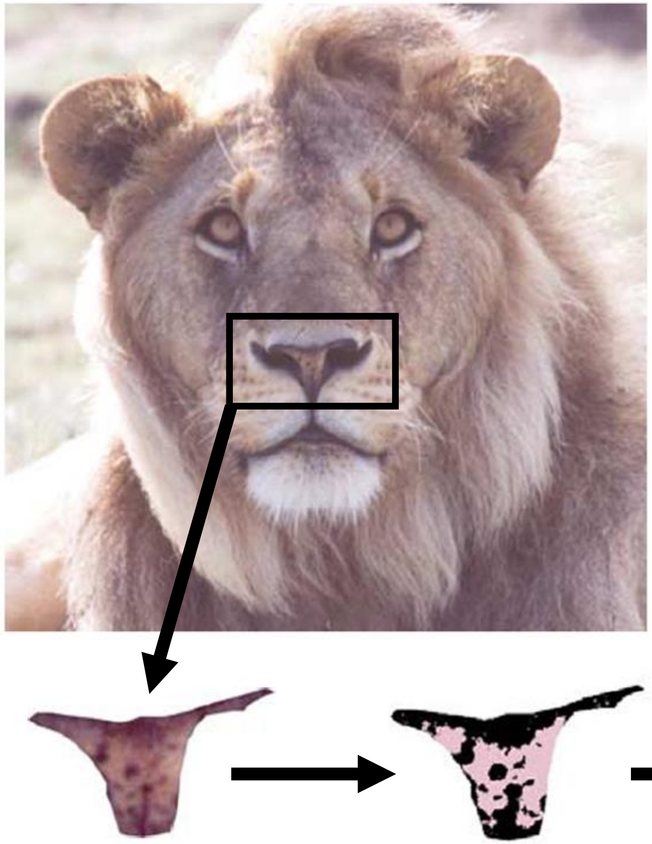


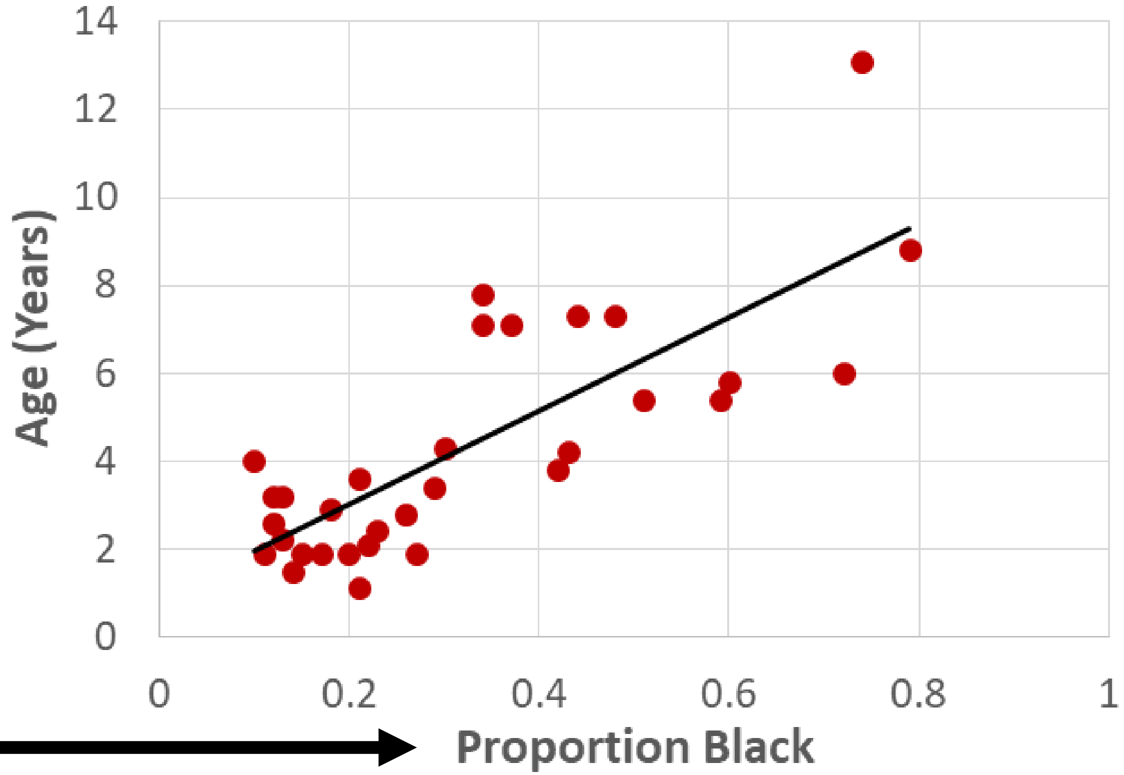
Hunting males 3 years old or older results in **declines** in abundance, especially when quotas are high

Hunting males 6 years old or older results in **stable** abundance, regardless of quota

Whitman et al. (2005)

Example: lion conservation





Best-fit line:

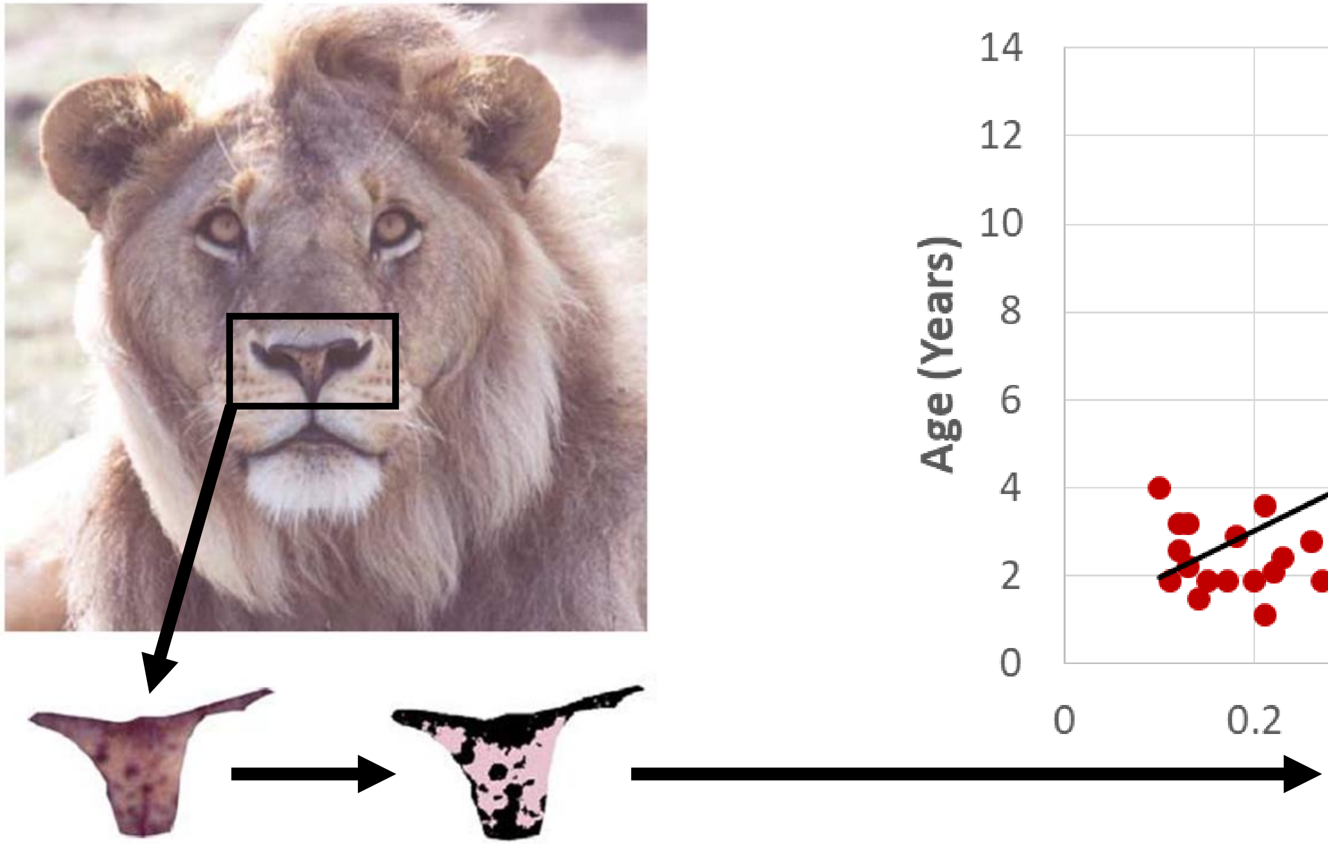
Y = 0.88 + 10.65(X)

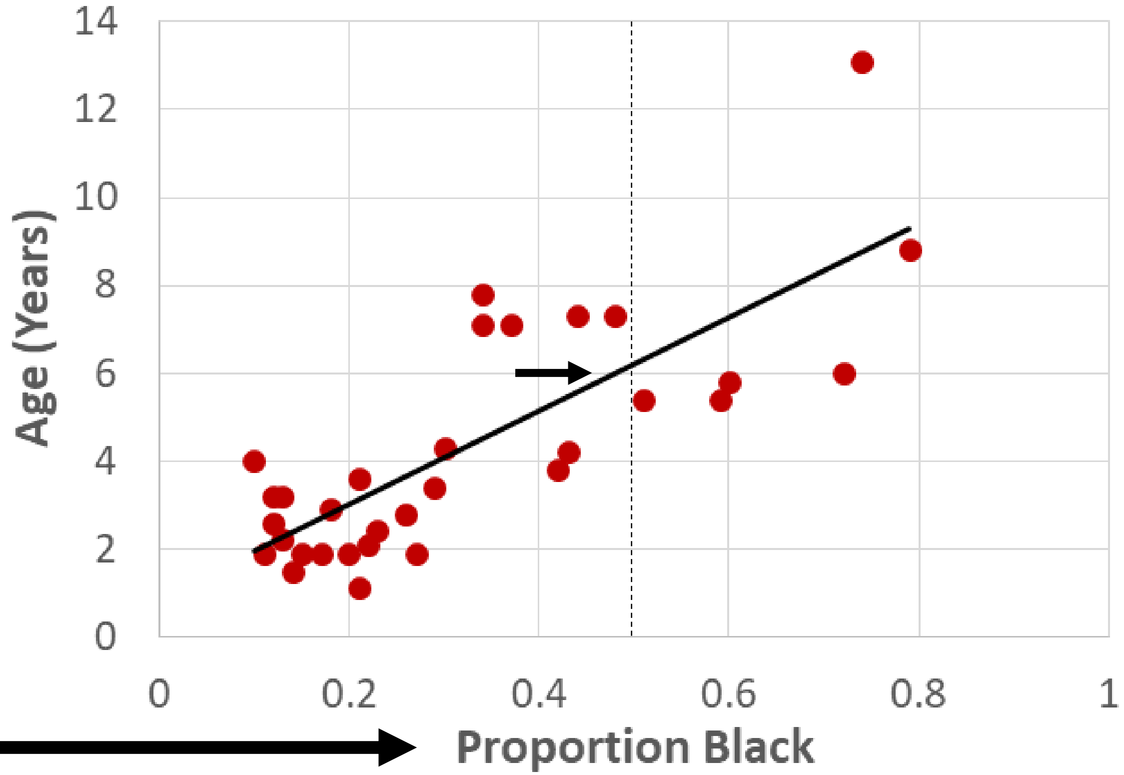
Intercept:
Age=0.88 years when proportion black=0

Slope:
Age increases by 10.65 years per 1 unit increase in proportion black

Data from: Whitman et al. (2005)

Example: lion conservation





<u>Best-fit line:</u>
Y = 0.88 + 10.65(X)

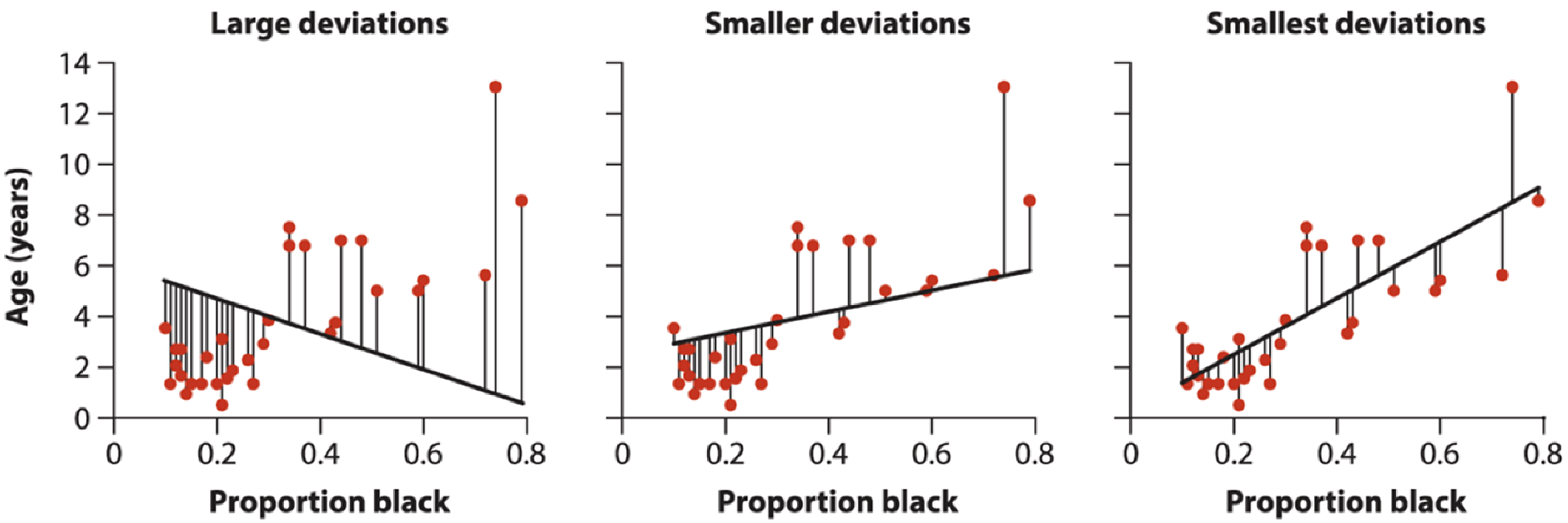
Based on this regression line, it is likely that an individual is 6 or older if his nose is **more than 50% black!**

Lion nose photos can help manage hunting!

Data from: Whitman et al. (2005)

How to fit a regression model?

A regression model uses an algorithm called **ordinary least squares (OLS)** to find the intercept and slope that best fit the data. OLS is designed to **minimize** residuals (deviations) so they are as small as possible. When residuals are minimized, this should result in predicted values that are, on average, as close as possible to observed values.



Residuals (vertical black lines) are shown for three possible regression lines. The left panel has a poor fit with many large residuals, while the right panel shows the best-fit line where residuals are as small as possible.

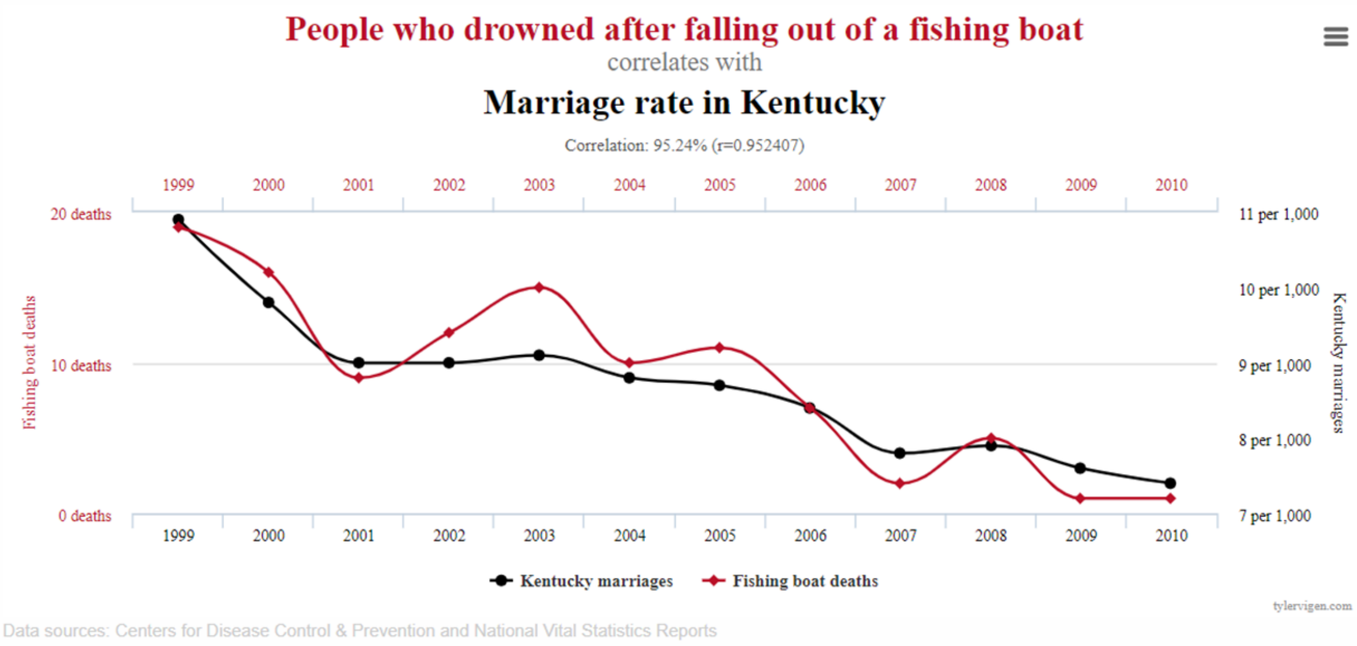
Figure from: Whitlock and Schluter (2020)

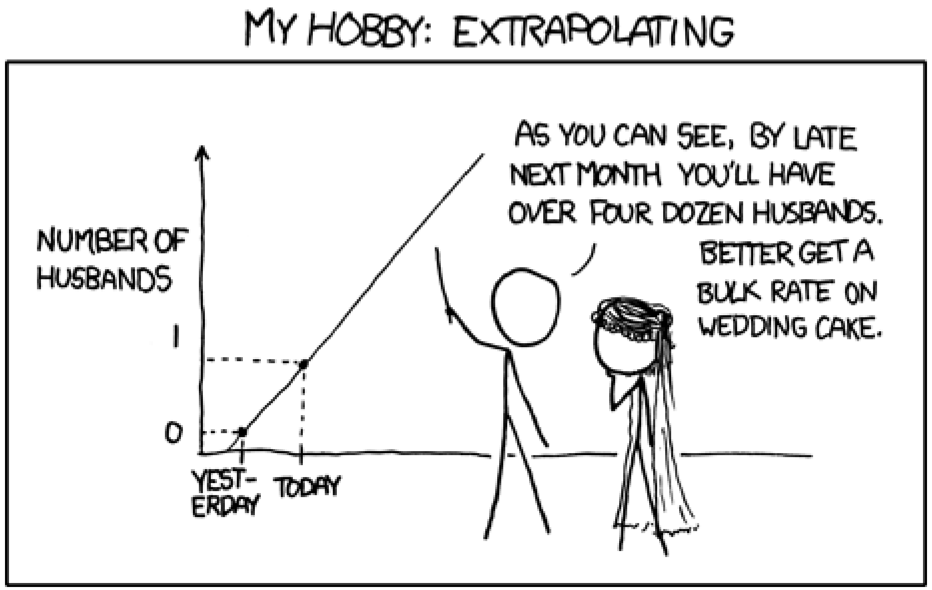
Note of caution

- Regression can be useful, but it has limitations!

- Implies a **causal relationship**
  - Need to be thoughtful when choosing X and Y variables

- **Avoid extrapolating** far beyond the range of X values that you have
  - If growth is regressed on temperature, which ranges from 5-15°C, do **NOT** use regression to estimate growth at 25°C





Images: Spurious correlations by Tyler Vigen, xkcd comics

Take a break…

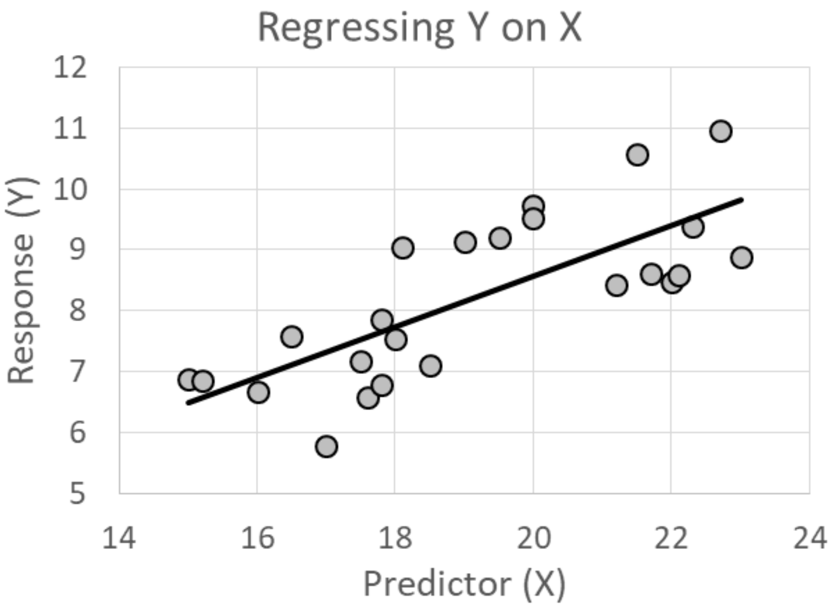
Image: Maggie (my dog)

Regression coefficients

- The intercept and slope are **coefficients** estimated in regression models

- Differences in intercepts and slopes can support inferences
  - Slopes inform the **magnitude and direction** of effects
  - Intercepts can (sometimes) provide a **useful baseline**

- Also important to think about **units**
  - Intercept = Y unit
  - Slope = Y unit/X unit



$$Y = a + b(X)$$

__Intercept:__
Value of Y when X=0

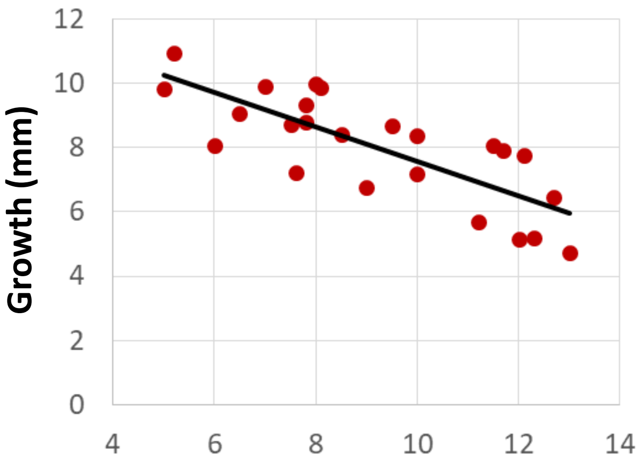
Units = Y

__Slope:__
How Y changes in response to changes in X

Units = Y per 1 unit increase in X

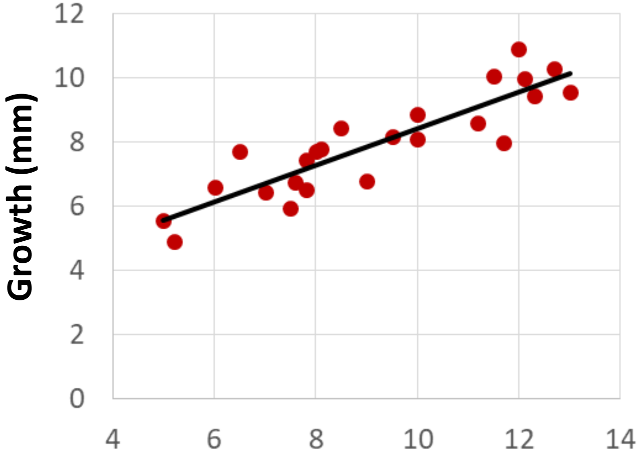
Slope direction

- Three types of possible relationships
  - Units = mm/°C
- Negative (-)
  - Slope < 0
- Positive (+)
  - Slope > 0
- None
  - Slope = 0



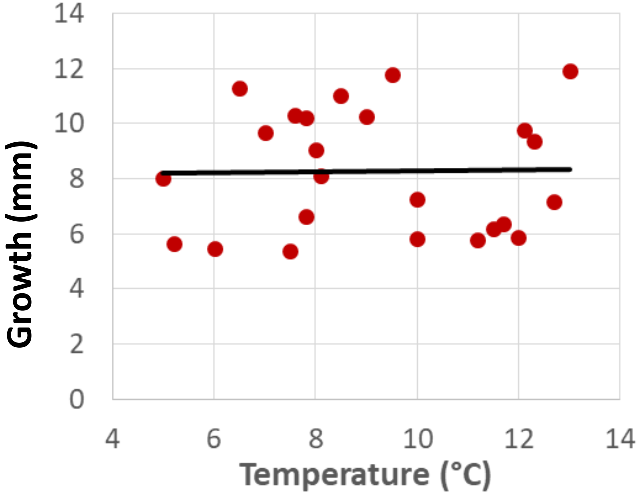
Negative relationship

<u>Quantitative:</u> as temperature increases by 1°C, predicted growth **decreases** by…



Positive relationship

<u>Quantitative:</u> as temperature increases by 1°C, predicted growth **increases** by…

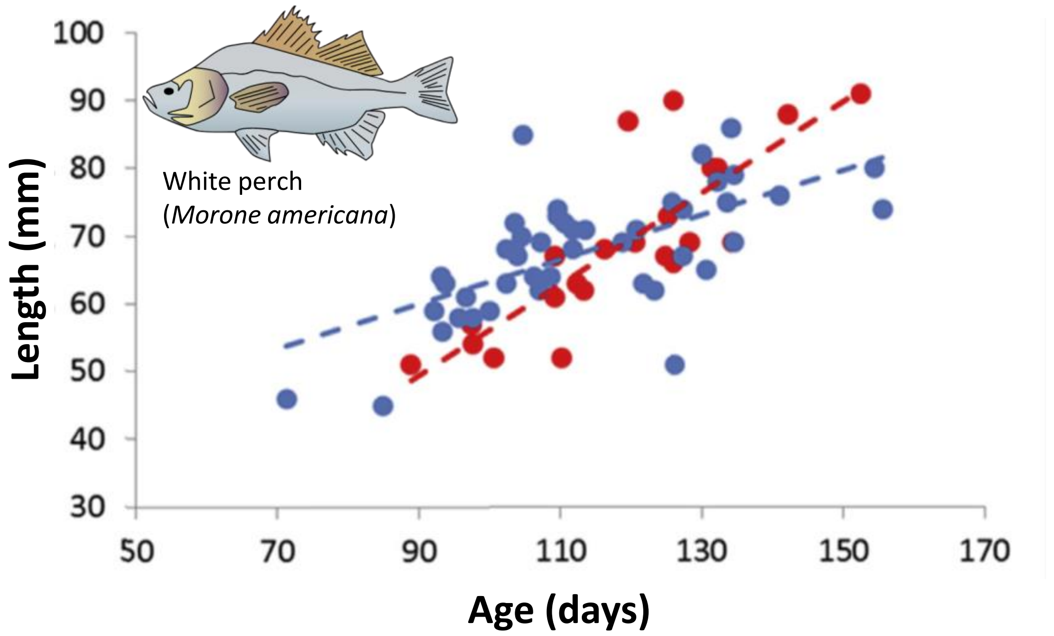


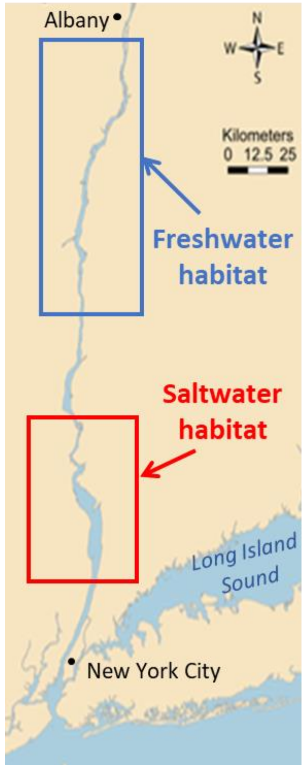
No relationship

<u>Quantitative:</u> as temperature increases by 1°C, predicted growth **remains constant**

Slope steepness

- Comparing regression slopes can reveal different **strengths of relationships**
- Example: white perch in the Hudson River
  - Regressed length on age for individuals caught in freshwater (blue) or saltwater (red) habitats
  - **Different slopes**
  - Units = mm/day (growth rate)
- Inference: average growth rates are **faster** in saltwater than freshwater
  - Slopes tell us *how much* faster (2x)





Coefficients:
Slope = 0.33 (freshwater)
Slope = 0.67 (saltwater)

Qualitative:
White perch in saltwater grow faster than in freshwater

Quantitative:
White perch in saltwater grow 0.67 mm/day on average, compared to 0.33 mm/day in freshwater

Data: Gallagher et al. (2018), Images: IAN

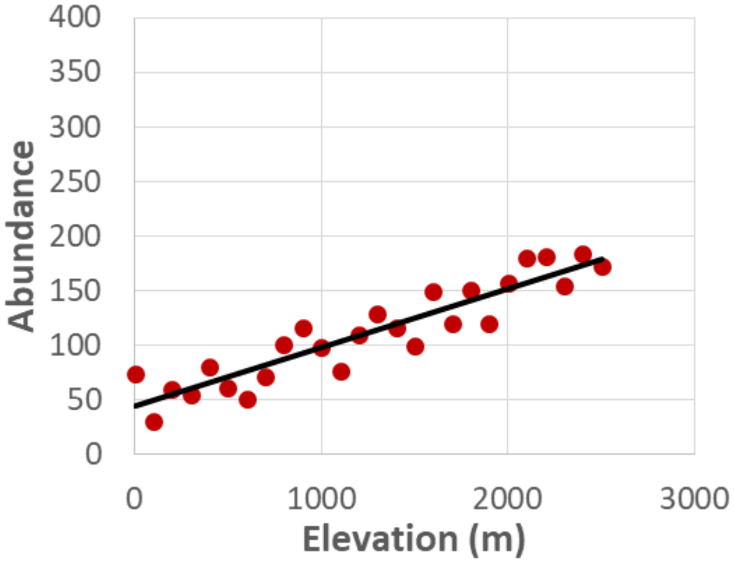
Intercept differences

- Comparing regression intercepts can inform **baseline responses**

- <u>Example:</u> elevation gradients in abundance
  - Same slope, **different intercepts**
  - Units = abundance (# of individuals)

- <u>Inference:</u> species 2 is more tolerant of low elevations than species 1

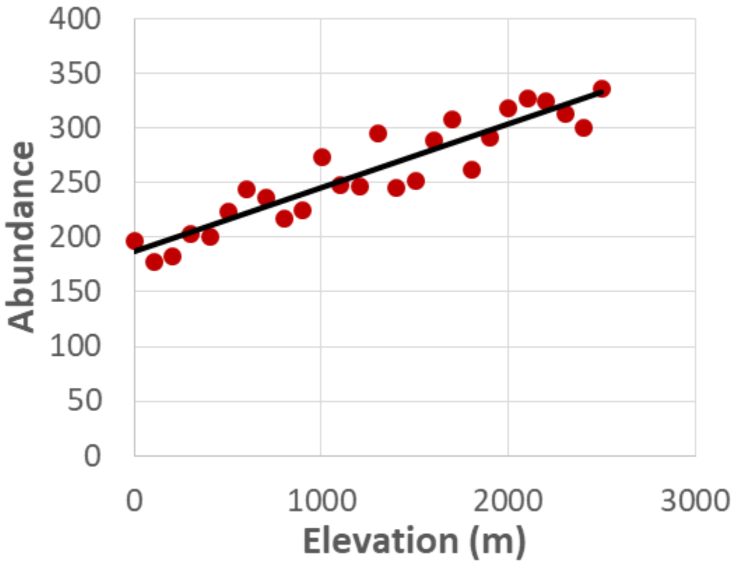
- Not always ecologically meaningful
  - Depends on X and its units (e.g. a species cannot have negative abundance)
  - Range of X values should include zero



<u>Species 1:</u>

Slope = 0.05
Intercept = 50

<u>Quantitative:</u> when elevation is zero (sea level), abundance is predicted to be **~50 individuals**



<u>Species 2:</u>

Slope = 0.05
Intercept = 200

<u>Quantitative:</u> when elevation is zero (sea level), abundance is predicted to be **~200 individuals**

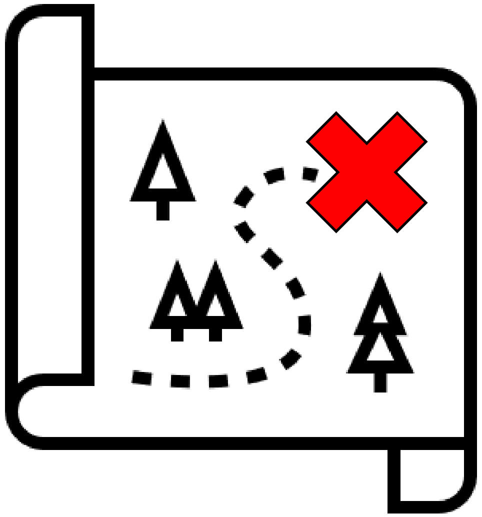
Overview

- Review basic statistical analyses and the types of variables used
  - *Inferences from categorical vs. continuous data*

- Introduce linear regression and its applications
  - *Estimates the relationship between two continuous variables*

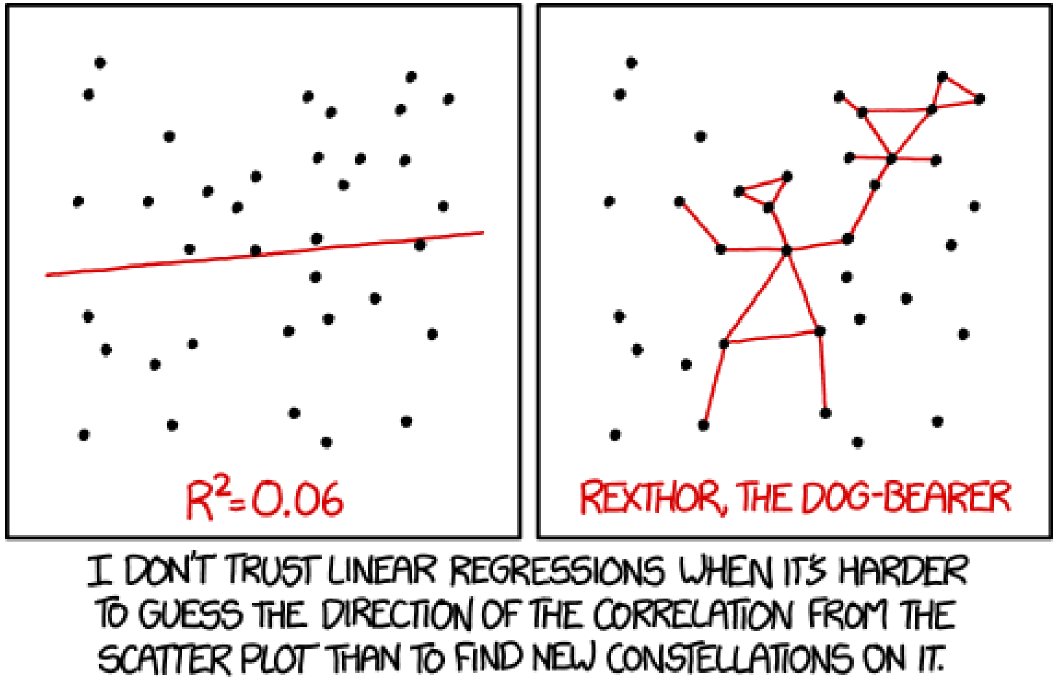
- Learn about how regression models are 'fitted' to real data
  - *OLS minimizes residuals to find best slope and intercept*

- Interpreting coefficients from regression models
  - *Intercepts and slopes can support inferences about ecological processes*



Where have we been?

Thanks!
Any questions?



Feel free to e-mail me:

brian.kenneth.gallagher@gmail.com

Image: xkcd comics