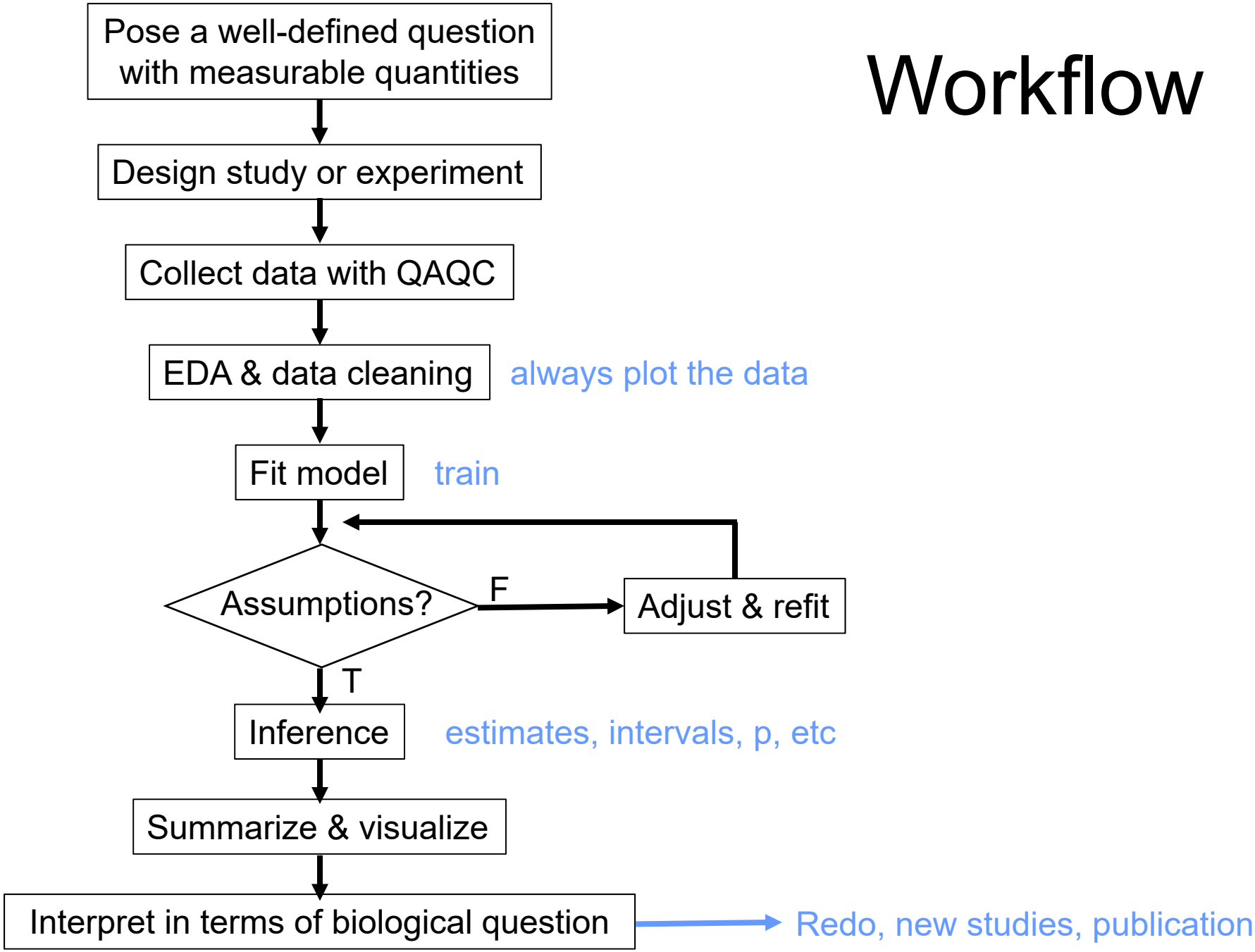
Today

- Collective results from Monday
  - lm goals
  - lm workflow
- Recap & questions from homework
- Coding the grid search algorithm

Goals of your lm analyses

- Estimation (various quantities) 7
  - what is the state of the world?
- Is there an effect? (H test) 5
  - ... but there is always an effect, just small
  - causation (mechanism) vs association
- Prediction 1
  - what will be the state of the world given some other information we know?

Workflow


Pose a well-defined question with measurable quantities
Design study or experiment
Collect data with QAQC
EDA & data cleaning
always plot the data
Fit model
train
Assumptions?
F
Adjust & refit
T
Inference
estimates, intervals, p, etc
Summarize & visualize
Interpret in terms of biological question
Redo, new studies, publication

Model algorithm

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

deterministic component

stochastic component

$y \sim x$ ← R shorthand "formula notation"

Equivalent to:

$$y \sim 1 + x$$

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 9 \\ 11 \end{pmatrix} = \begin{pmatrix} 1 & 40 \\ 1 & 45 \\ 1 & 39 \\ 1 & 50 \\ 1 & 52 \\ 1 & 57 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

design matrix

parameter vector

Take aways from Kery

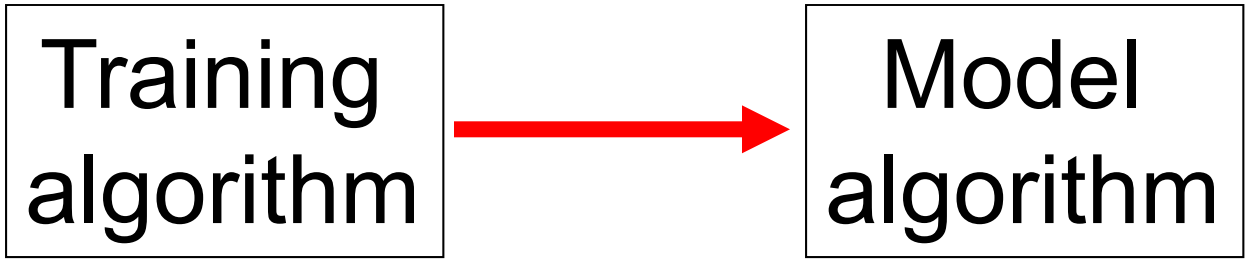
"Sometimes one may get away ... with not exactly knowing what parameterization of a model is fit by the software and what the parameters effectively mean"

Polite way of saying you should understand how the model is parameterized and what the parameters mean

Take aways from Kery

- Linear models can be parameterized in different ways
  - e.g. effects versus means parameterization
- 2 sample t-test & ANOVA are special cases of simple linear model
  - Categorical variables coded 0,1 for each category
  - aka dummy variable coding

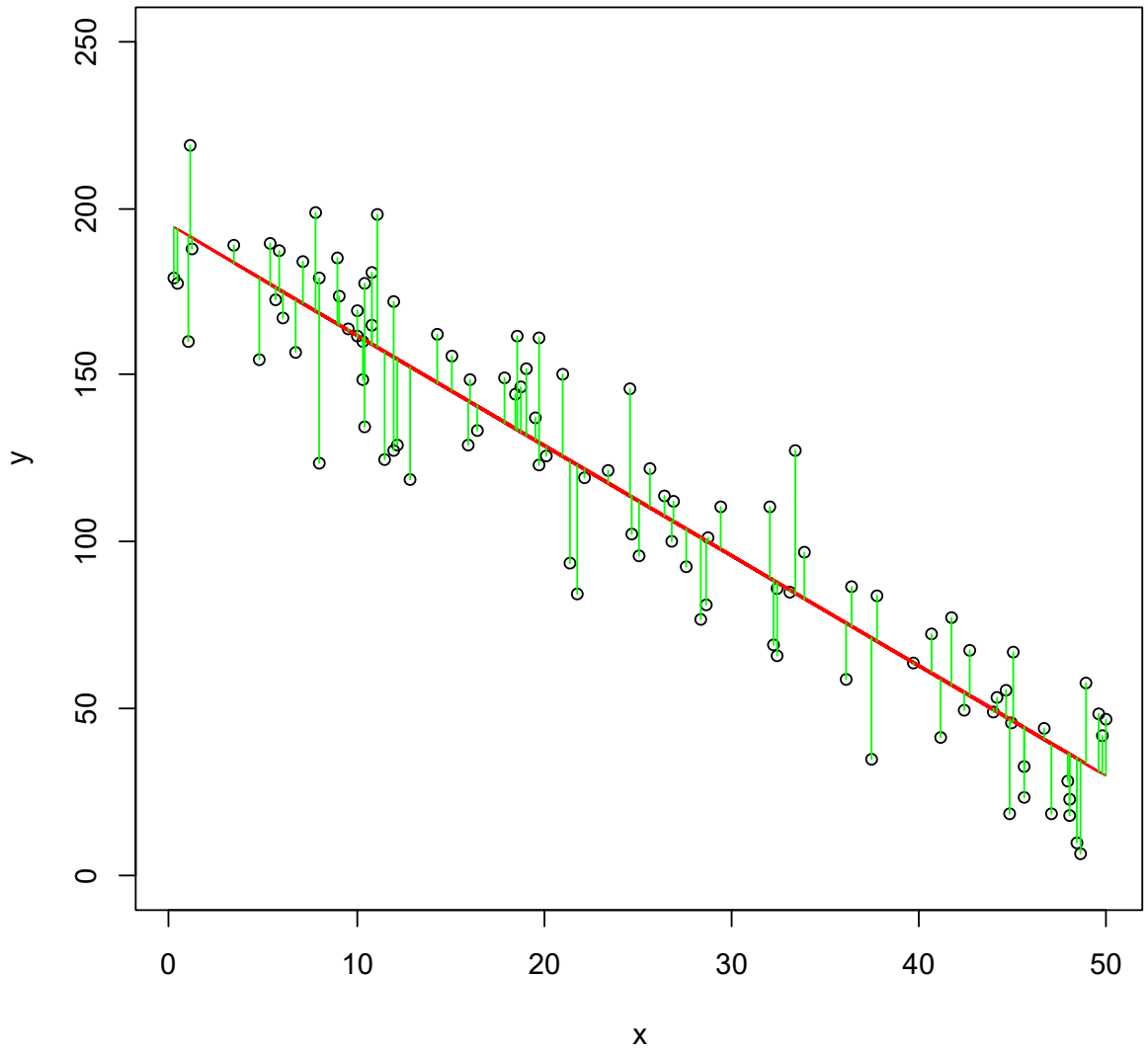
Training algorithm



Big idea in data science

Legendre 1805: comet orbits, SSQ.

Least squares algorithms



Vary model parameters

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

Minimize distance from data

$$ssq = \sum_i e_i^2$$

Grid search algorithm Pseudocode

Read in data
Set up values of β_0 and β_1 to try
Set up storage for ssq, β_0, β_1
For each value of β_0
    For each value of β_1
        Calculate model predictions
        Calculate deviations
        Sum squared deviations
        Store ssq, β_0, β_1
Plot sum of squares profiles (ssq vs β_0, ssq vs β_1)
Report best ssq, β_0, β_1
Plot fitted model with the data

Translate this to R code and use it to train the model with your data