Model assessment

Jonathan Dushoff, McMaster University

Goals

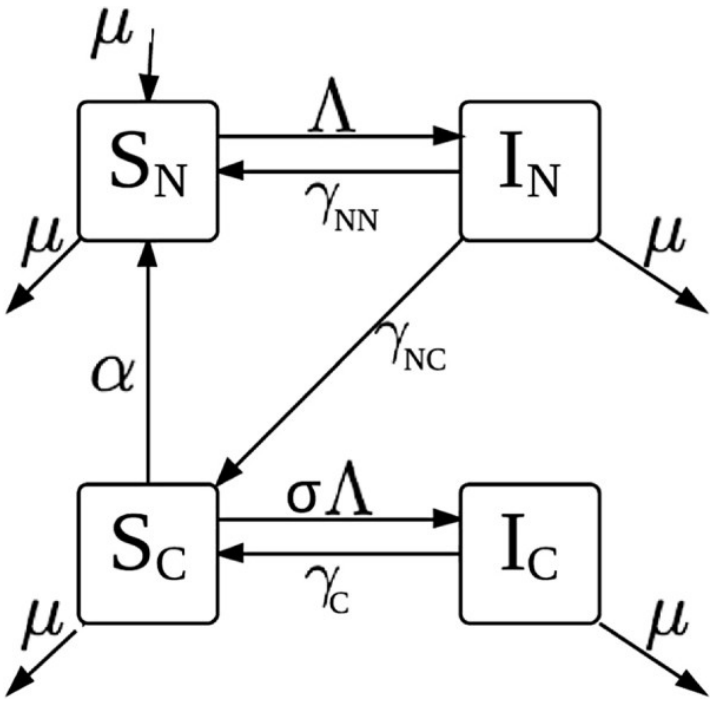
- Discuss model types and model goals
- Discuss the value of simulation for validating models
- Discuss metrics for evaluating fit
  - Put the Goodness of fit test in its place

Do I have a good model?

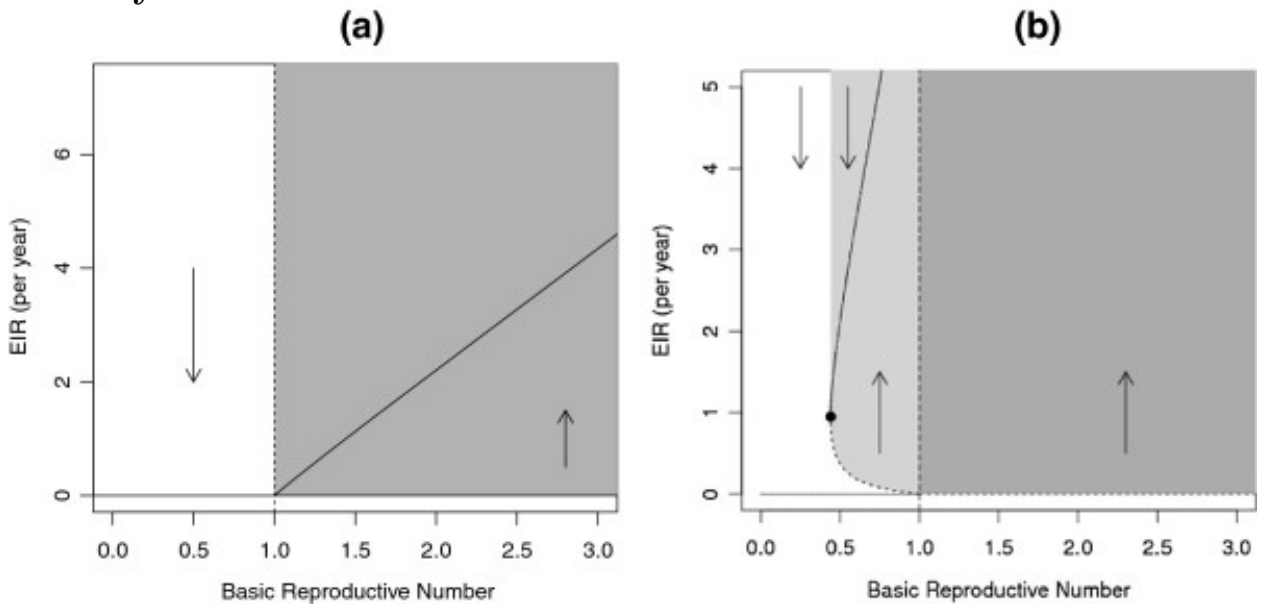
- What is my model trying to accomplish?
  - Generating hypotheses
  - Evaluating plausibility
  - Prediction
  - Mechanistic understanding
  - Evaluating scenarios

1 Conceptual models

Disease thresholds
Effects of clinical immunity



Bistability



2 Prediction

Ptolemy v. Copernicus
Where will we see cholera cases?

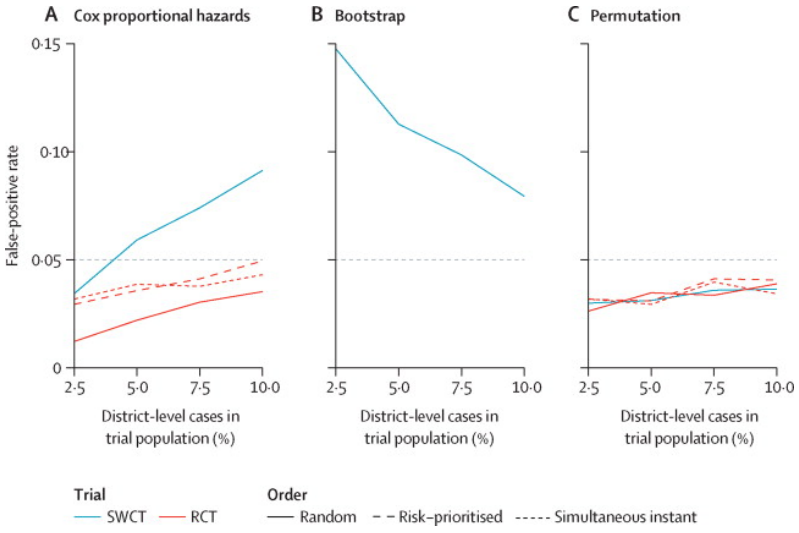
3 Model Validation

- Does your fitting algorithm match your *model world*?
- If you use your fitting algorithm on simulations from your model world, then you *know the right answer*!

Validation measures

- Coverage
- Precision
- Bias?
- Accuracy?

Coverage



- The right answer should be inside your 95% confidence interval 95% of the time
  - If more, your model is *too conservative*
  - If less, your model is *invalid*
- In many cases it's good to look at the two tails separately:
  - How often do you overestimate? Underestimate?

Precision

- A good model tries to provide a precise answer
  - Confidence intervals should be narrow, if possible
  - But not at the price of overconfidence (invalidity)
- As data increases, your precision should increase

  - CIs should approach zero width
  - . . . as long as you have data about *everything*
- Conversely, CIs should reflect a variety of sources of uncertainty

Bias and accuracy

- Good coverage and high precision should ensure high accuracy and low bias
- Don't worry about "unbiased estimators"
  - Your estimator doesn't need to be absolutely unbiased
  - Your reasonable estimator will be asymptotically unbiased

4 Model Evaluation

- Does your model match the *real world*?
  -
- How well does your model match the real world?

4.1 Goodness of fit

- Goodness of fit *statistics* describe how well a model prediction matches observed data
- Goodness of fit *tests* attempt to determine whether the observed difference between model and data is statistically significant

Your model is false!
. . . or at least, incomplete

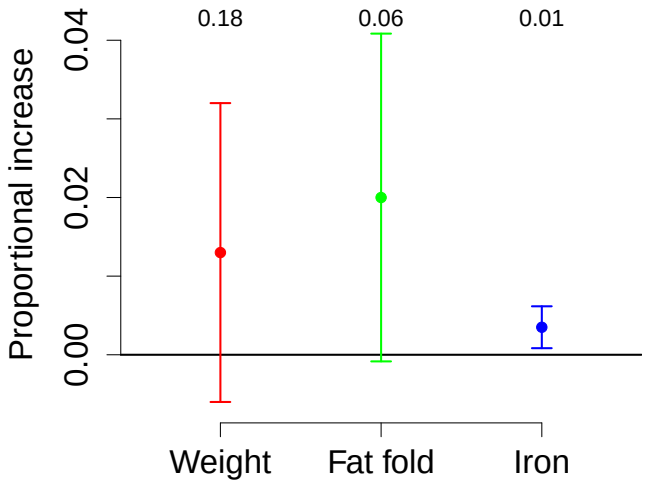
- A goodness of fit test won't make it true
- You can "pass" a goodness of fit test by:
  - having a good model
  - making very broad predictions
  - having bad data
  - choosing an inappropriate way to compare
- So why would we do this?
- For that matter, why do we use P values at all in biology?

4.2 Digression

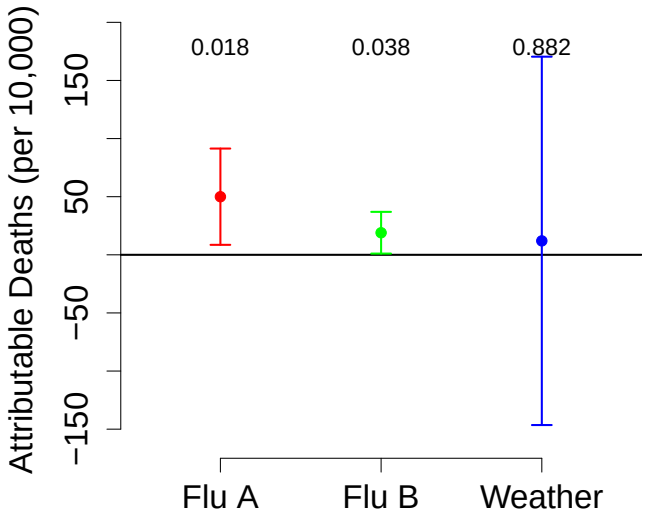
Passing goodness of fit tests

- I can make any model pass a goodness of fit test by broadening the uncertainty
- That doesn't make it a good model

Vitamin study



... with confidence intervals



- **Never** say: A is significant and B isn't, so $A > B$
- **Instead:** Construct a statistic for the hypothesis $A > B$
  - May be difficult

What does the P value mean?

- Low: you are seeing something clearly
- High: you are seeing something unclearly

Goodness of fit test

- Your model is *not* reality (null hypothesis is false)
- Can we see the difference clearly?
  - If *no*, model may be *good* or *bad*.
    - We probably can't add any more complexity based on current data
  - If *yes*, model may be *good* or *bad*.
    - We *may* be able to add more complexity based on current data
    - But we may not need to

Capturing patterns

- You can ask:
  - Does your model do a reasonable job of capturing the data?
    - You can use a goodness of fit *statistic* for this, and not worry about the P value
  - Does your model capture patterns and relationships that you (or other experts) think are important?

4.3 Going beyond

Out-of-sample validation

- Does your model make predictions *outside* the range on which you calibrated it?
  - Predicting gravitational shifts in star positions from measurements in Earth laboratories
  - Predicting cholera outbreaks in Bangladesh from a model calibrated to Haiti
  - Predicting influenza patterns in 2010 from a model calibrated from 2000–2009

Predicting way out of sample

Saturn's shepherd moons were predicted before they were seen!

Essentially, all models are wrong, but some are useful.

– Box and Draper (1987), *Empirical Model Building ...*

Test sets

- What is **test set** spelled backwards?
- Hold some data out while fitting your model
- Or just *pretend* to do this as an evaluation method
  - In other words, test what would happen under various withholding scenarios
  - This can get very elaborate, and we should probably do it more

Other model worlds

- The model you're *fitting* is probably pretty simple
- But you can *simulate* very complicated models, indeed
- How well can you do? Which details are important?

Generating hypotheses

For example:

- Safe burial is key to interrupting Ebola transmission
- Vaccinating domestic dogs can eliminate transmission of canine rabies

Testing hypotheses

- Both the Farr model and the Snow model made testable predictions about cholera
- Snow tested his hypotheses by removing the pump handle

Hard questions

Answers are not always easy

5 Conclusion

Summary
Dynamic models

- Clarify thinking
  - What are our assumptions, what else do we need to know?
- Understand outcomes
  - Can heterogeneity explain the time course of HIV epidemics?
  - Is it possible that MDA could break the cycle of malaria transmission in some areas?
- Predict outcomes
  - What is the potential for a hepatitis A outbreak in Cape Town?
  - What might happen if I improve testing-and-treatment outreach in Jamaica?
- Find new mechanisms
  - Why can't I explain my data? What haven't I thought of?

Summary
Evaluation

- Validation (inside your model world)
  - Does my fitting method work (assuming my model is right)?
- Inspection (compare patterns)
- Prediction (and other out-of-sample comparison)
  - Can my model predict things I haven't told it yet?
- Generate and test mechanistic hypotheses