Likelihood fitting and dynamic models II

Approaches to fitting
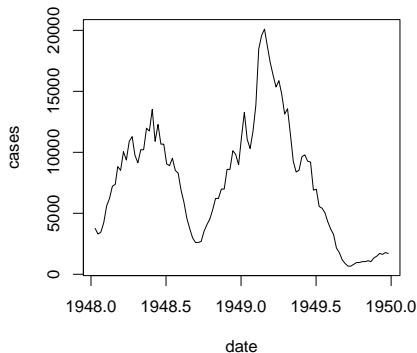
Jonathan Dushoff, McMaster University

MMED 2017

Summary

- Discuss theories and approaches behind fitting models to data

- Explain relationship between least squares and likelihood

- Discuss approaches to dealing with process error and observation error
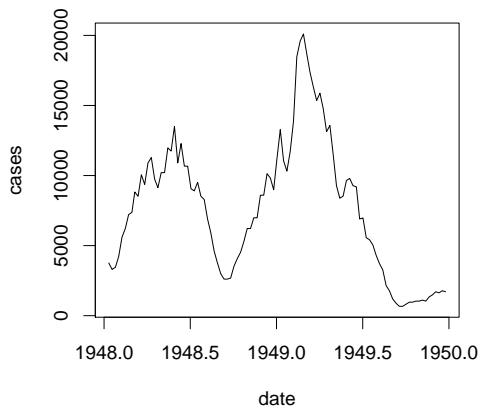
ICI3D

Measles data

▶

Measles reports from England and Wales

▶ Reconstruct the number of susceptibles

▶ Divide the data into generations

▶ Fit $\mathcal{R}_0$

▶ Predict

ICI3D

Why did I get the wrong answer?

Measles reports from England and Wales

ICI3D

Why did I get the wrong answer?

- Model structure may be wrong

- Population structure may be wrong

- Stochasticity in disease observation and recording

- Stochasticity in transmission

- Multi-parameter estimation
  - There may be different parameter combinations that work equally well
  - * Parameters may not be "identifiable"

ICI3D

Outline

ICI3D

Conceptual framework

- ▶ How do we assume our data relate to our model world?
  - ▶ **No error:** We could attempt to model everything we see, in exact detail
    - ▶ Impractical
  - ▶ **Observation error:** we could assume that the world is perfectly deterministic, but our *observations* are imperfect
    - ▶ Shooting
  - ▶ **Process error:** we could assume that we observe perfectly, but that the world is stochastic
    - ▶ Stepping
  - ▶ **Both kinds of error:** the world is stochastic, and our observations are imperfect
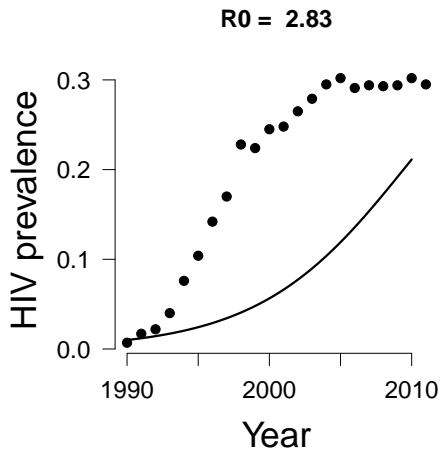   - ▶ Modern methods

ICI3D

Observation error only

- ▶ Point your model at the target

- ▶ Give it starting conditions and parameters

- ▶ Let it go
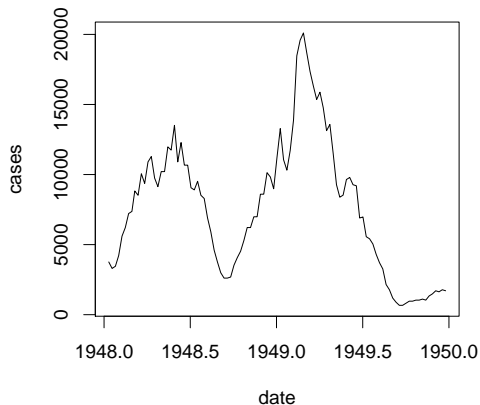
- ▶ Compare final results to observations

Shooting

Shooting

R0 = 2.83

HIV prevalence vs Year
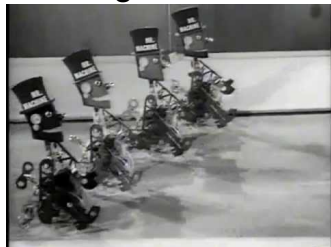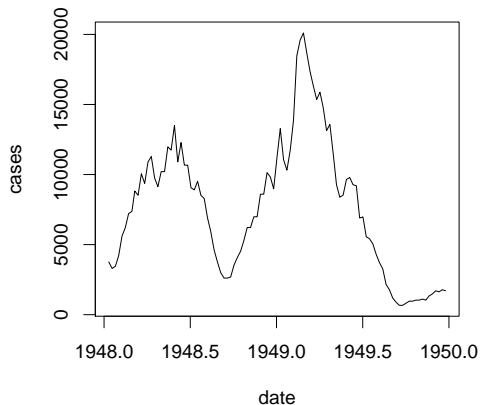
Shooting

Measles reports from England and Wales

Process error only

- Look at each step separately.

- See how the model is doing for that step.
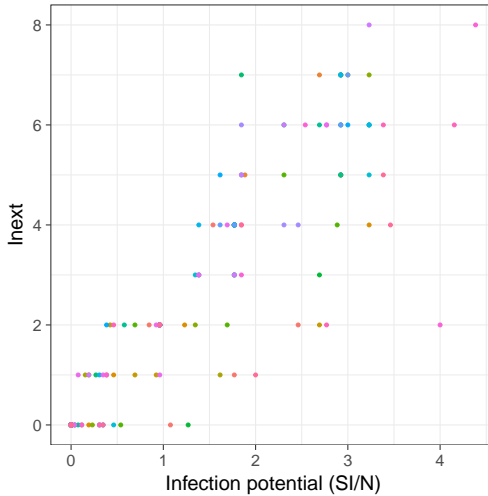
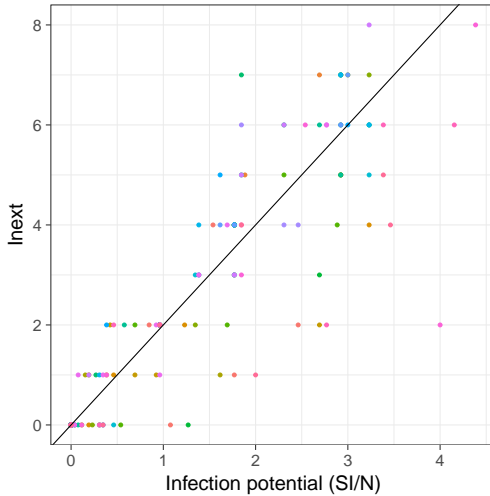- Reset based on observed data before taking the next step

Stepping

ICI3D

Stepping

Measles reports from England and Wales

Stepping

Stepping

Modern methods

- Is it better to ignore process error, or observation error?

- What if we have a small number of cases, and good reporting (Ebola in small villages)
  - * Ignore observation error?

- What if we have a large number of cases, and poor reporting (HIV in Harare)?
  - * Ignore process error?

- What if we have a new epidemic and poor reporting (West African Ebola)?
  - * Can't really ignore either kind

ICI3D

Outline

ICI3D
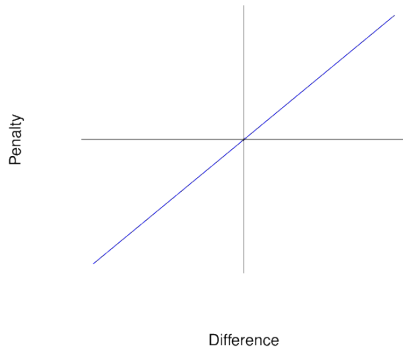
How to fit?

- ▶ Solving an equation

- ▶ By eye (fiddling with parameters)

- ▶ *Minimizing a distance function*

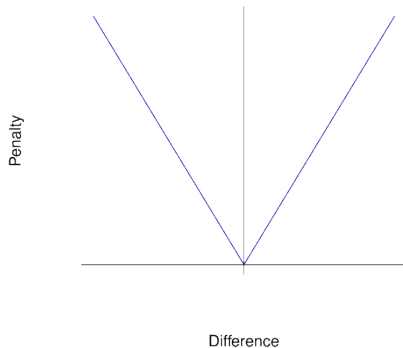- ▶ Likelihood

ICI3D

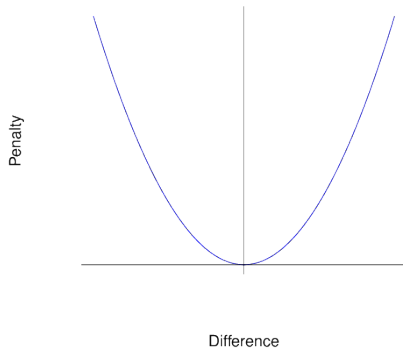# Distance functions

$$D = \sum_i y_i - \hat{y}_i$$

ICI3D

Distance functions

$$D = \sum_i |y_i - \hat{y}_i|$$

Penalty

Difference

Distance functions

$$D = \sum_i (y_i - \hat{y}_i)^2$$

Outline

Likelihoods

- Assume that the difference between the estimate $\hat{y}_i$ and the data point y_i is normally distributed. What is the log likelihood?

-
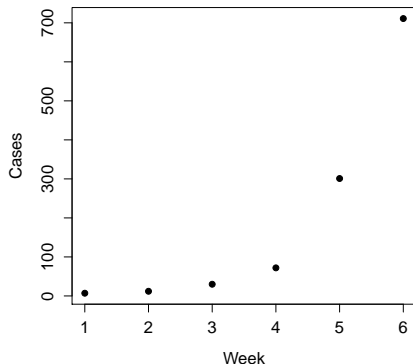$$L = \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(\hat{y}_i - y_i)^2}{2\sigma^2}\right)$$

-
$$\ell = \sum_i -\log(\sigma\sqrt{2\pi}) - \sum_i \frac{(\hat{y}_i - y_i)^2}{2\sigma^2}$$

- *We minimize the likelihood by minimizing the sum of squares*
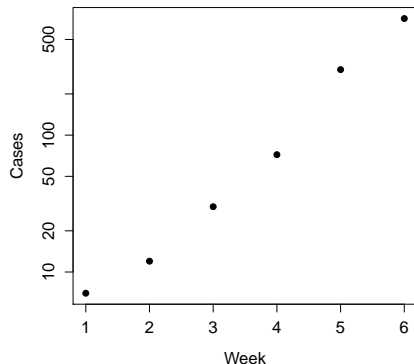 - and then solving for σ

Least squares → likelihood

- Attaching your least squares fit to a likelihood means:
  - You can *use it* for statistical inference (LRT)
  - You can *challenge* the assumptions

Mexican flu example
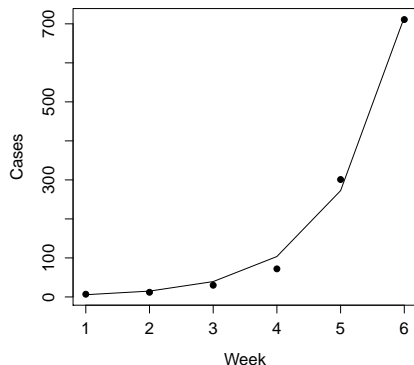

- How fast is it growing? *r*

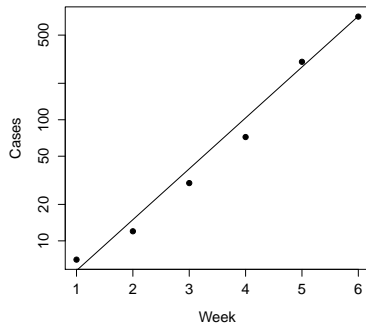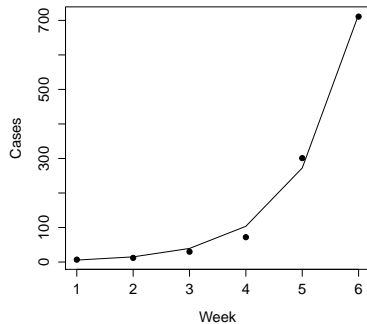- How hard will it be to control? $\mathcal{R}_0$

A different perspective

- ► We could make the normal assumption on either scale
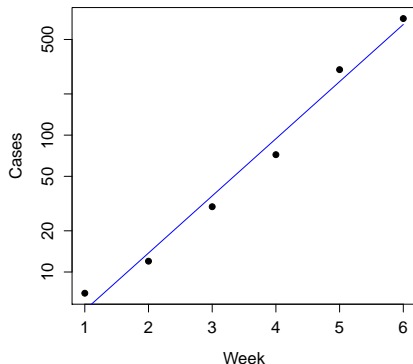
- ► How much does it matter?

Normal assumption

- Least squares on the linear scale

- 10:50 :: 980:1020

- Gives relatively too much weight to large observations

Normal assumption

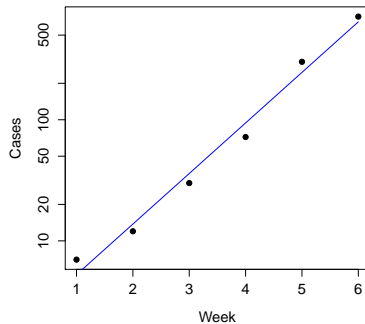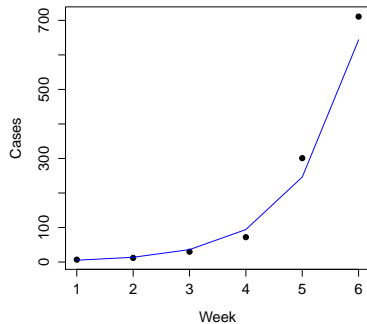# Lognormal assumption

- Least squares on the log scale

- 3:5 :: 300:500

- Gives relatively too much weight to small observations

Lognormal assumption

A more realistic error distribution

- My case counts are *individuals*

- What distributions can I use to reflect that?

- * Poisson or binomial
  - * WRONG!
  - * *Sorry:*
    - * OK, technically it's right, but you shouldn't do it.
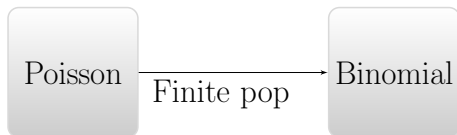
ICI3D

Reality is complicated

- Poisson and binomial reflect *only* individual-level variation
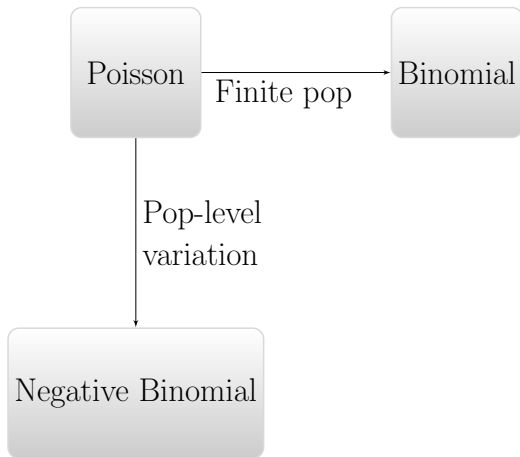   - No temporal variation
    - No clustered sampling
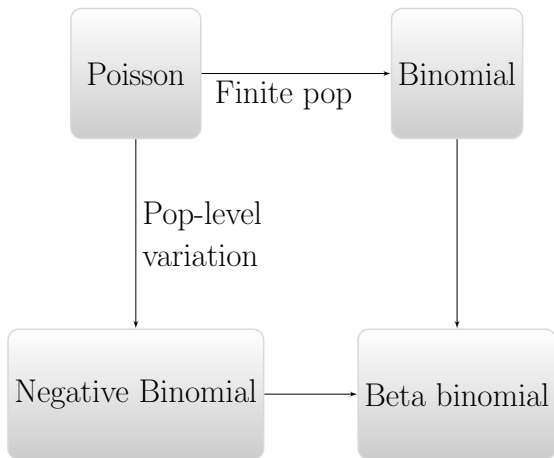   - . . .

ICI3D

Distribution diagram

Poisson

ICI3D

Distribution diagram

Poisson — Finite pop → Binomial
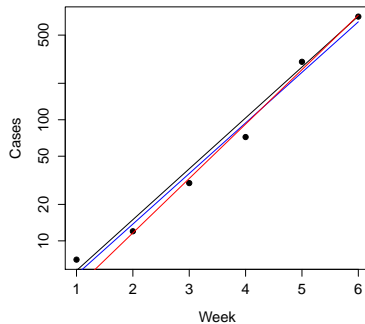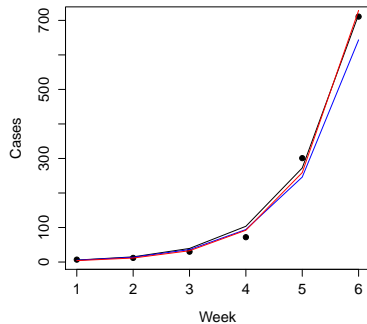
ICI3D

Distribution diagram

ICI3D

Distribution diagram

Discrete distributions

- Negative binomial is a good general-purpose discrete distribution
  - Individual- and population-level variability

- Beta binomial takes size into account
  - Good when denominator is clear and important
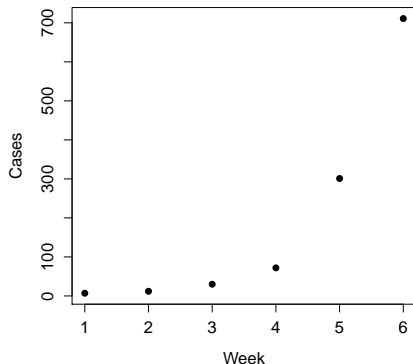 - For exampling, when sampling

Negative binomial fits

Comparison

- Realistic error distribution provides (apparently) better fits

- Confidence intervals
  - Normal: r = 0.96–0.97/wk
  - Lognormal: r = 0.64–1.29/wk
  - Negative binomial: r = 0.90–1.14/wk

- How would you test these methods?
  - * Validation: use simulated data to see if your method is reliable

ICI3D

Identifiability

- What if we tried to estimate $\mathcal{R}_0$ instead of r from Mexican flu data?
  - * Disease could be fast with low $\mathcal{R}_0$ or slow with high $\mathcal{R}_0$.

Outline

Modern approaches

- ► Why are people using model worlds with no observation error?
  - ► or no process error?

- ► Sometimes they are good enough (model validation)

- ► Combining both is *hard*

Filtering

- Filtering is a little like shooting
  - Simulate from beginning to end, but use *stochastic* simulations

- You need a lot of simulations, and often ways of selecting and refining them

- A popular, state-of-the-art method is implemented in the R package pomp

## *Filtering*

Measles reports from England and Wales
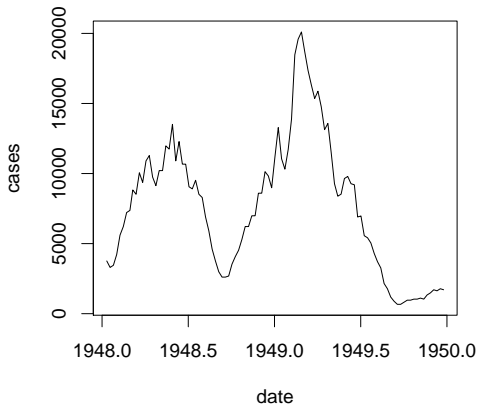
Latent variable methods

- Latent variable methods are a little like stepping
  - But we step to and from unknown values (our latent variables), so we need a way of exploring many possibilities

- Popular, state-of-the-art methods are available in the R packages rjags and rstan

ICI3D

Latent-variable methods

Measles reports from England and Wales
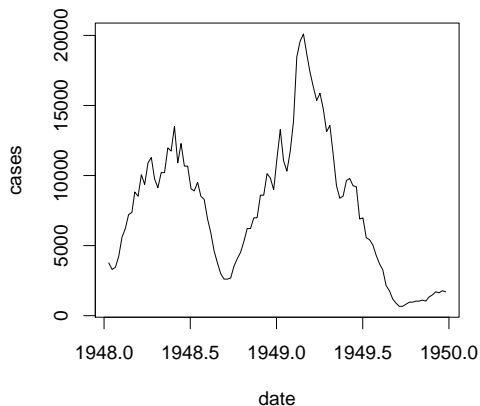
Latent variables

- We model *observed* variables:
  - e.g., reported cases, estimated prevalence

- Using *unobserved* latent variables
  - true cases, true prevalence, true number of susceptibles

Reauires a framework that can address the fact that our latent
variables have many possible values

ICI3D

Multi-parameter inference

- ► Modern methods are already hard, and when you consider various sources of uncertainty, you're really on the bleeding edge

- ► Many high-profile models for Ebola, for example failed to consider process error.

- ► The biggest paper talking about process error neglected uncertainty in generation intervals

- ► Once you do multi-parameter inference, you may find that confidence intervals are very large – this may reflect the reality of knowledge, but may not make you look good
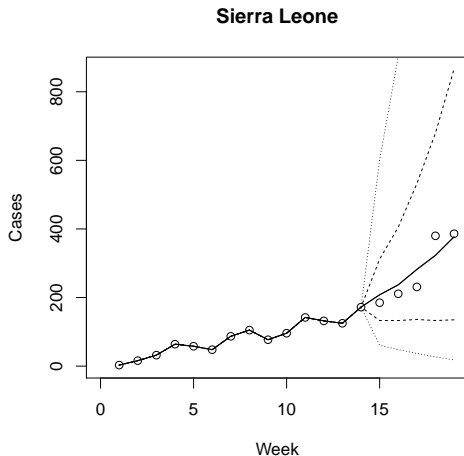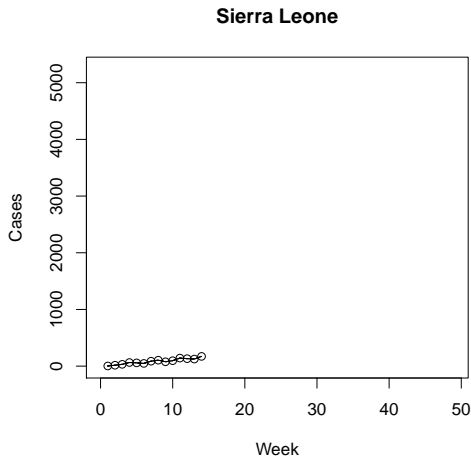
ICI3D

Assessing and reporting uncertainty
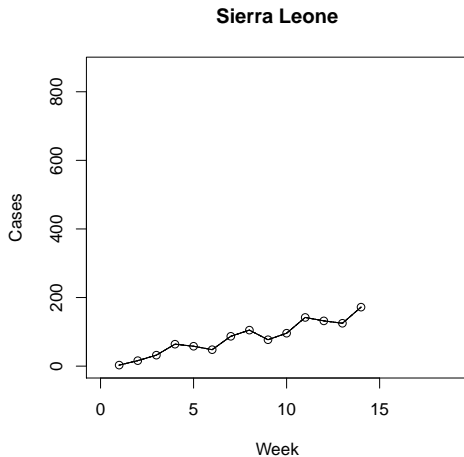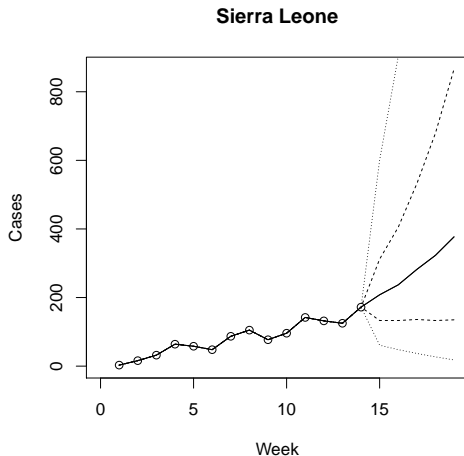
Sierra Leone

Assessing and reporting uncertainty

Sierra Leone

Assessing and reporting uncertainty

Sierra Leone

Assessing and reporting uncertainty

Sierra Leone

Assessing and reporting uncertainty

Sierra Leone

Assessing and reporting uncertainty

Sierra Leone

Outline

ICI3D
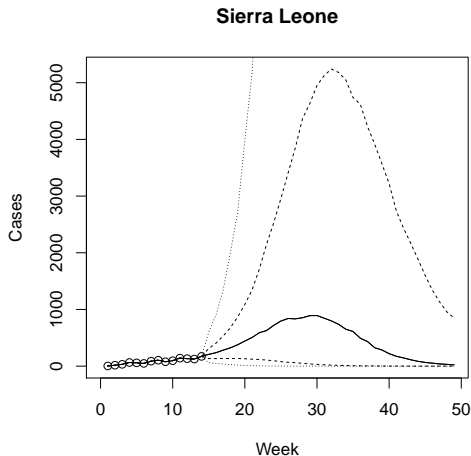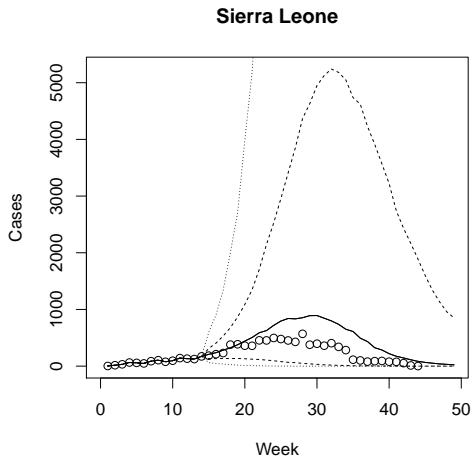
Likelihood

- Maximum likelihood and likelihood are not the same thing

- Bayesian approaches and frequentist approaches (including maximum likelihood) *both* depend on calculating (or approximating) likelihood

ICI3D

Frequentist inference

- To do frequentist inference on these complicated likelihoods, we need to:
  - estimate likelihoods
  - find the maximum likelihood
  - use the likelihood ratio test to find confidence intervals

- This is hard

Bayesian inference

- To do Bayesian inference on these complicated likelihoods, we need to:
  - construct prior distributions
  - estimate likelihoods
  - estimate the posterior

- Usually *a little* less hard
  - But still requires more assumptions

ICI3D

Summary

- We need **dynamics** to understand links between processes and outcomes
  - How do things work?

- We need **statistics** to understand uncertainty
  - What can we learn from *data*

- Combining these two is difficult, but progress is being made.

ICI3D

Summary

- Making your fit into a likelihood clarifies assumptions and creates a foundation for statistical inference

- Accounting for both process and observation error is hard
  - and not always necessary

- Stepping methods don't allow for observation error
  - filtering methods can address this

- Shooting methods don't allow for process error
  - latent variable methods can address this

Title: Likelihood fitting and dynamic models II
Attribution: Jonathan Dushoff, McMaster University, MMED 2017

Source URL: http://tinyurl.com/ici3d-figshare
For further information please contact admin@ici3d.org.