Model-Data Interface

Parameter estimation and statistical inference

Parameter estimation

- We've seen that basic reproductive ratio, R_0, is a very important quantity

- How do we calculate it?

- In general, we might not know (many) model parameters. How do we achieve parameter estimation from epidemiological data?

- Review some simple methods

1a. Final outbreak size

- From lecture 1, we recall that at end of epidemic:
    - $S(\infty) = 1 – R(\infty) = S(0)\ e^{–R(\infty)\ R_0}$

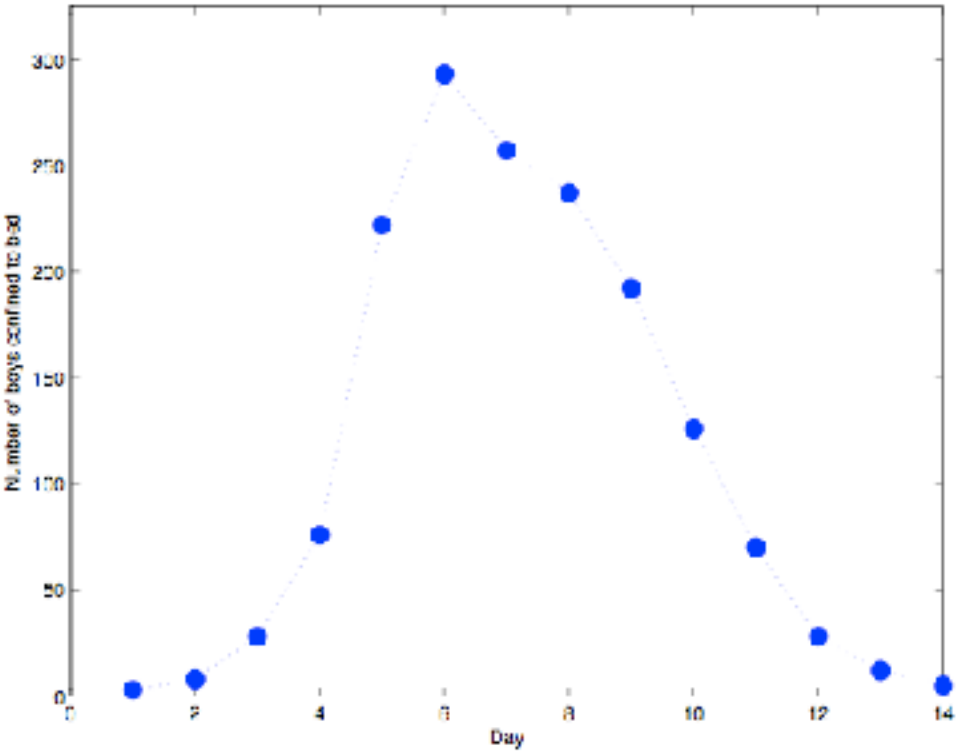
- So, if we know population size (N) , initial susceptibles (to get S(0)), and total number infected (to get R(∞)), we can calculate R_0

$$R_0 = -\frac{\log(1 - R(\infty))}{R(\infty)}$$

Note: Ma & Earn (2006) showed this formula is valid even when numerous assumptions underlying simple SIR are relaxed

1. Final outbreak size

- Worked example:



Influenza epidemic in a British boarding school in 1978

$N = 764$
$X(0) = 763$
$Z(\infty) \sim 512$

$R_0 \sim 1.65$

1b. Final outbreak size

- Becker showed that with more information, we can also estimate R_0 from

$$R_0 = \frac{(N-1)}{C}\ln\left\{\frac{X_0+\frac{1}{2}}{X_f-\frac{1}{2}}\right\} \qquad (\sim 1.66)$$

- Again, we need to know population size (N) , initial susceptibles (X_0), total number infected (C)
- Usefully, standard error for this formula has also been derived

$$SE(R_0) = \frac{(N-1)}{C}\sqrt{\sum_{j=X_f+1}^{X_0}\frac{1}{j^2}+\frac{CR_0{}^2}{(N-1)^2}}$$

Small aside: mean age at infection

- An epidemiologically interesting quantity is mean age at infection – how do we calculate it in simple models?
- From first principles, it's mean time spent in susceptible class
- At equilibrium, this is given by 1/(βI*), which leads to

$$A = \left(\frac{1}{\mu(R_0 - 1)} \right)$$

- This can be written as R_0-1 ≈ L/A (L= life expectancy)
- Historically, this equation's been an important link between epidemiological estimates of A and deriving estimates of R_0

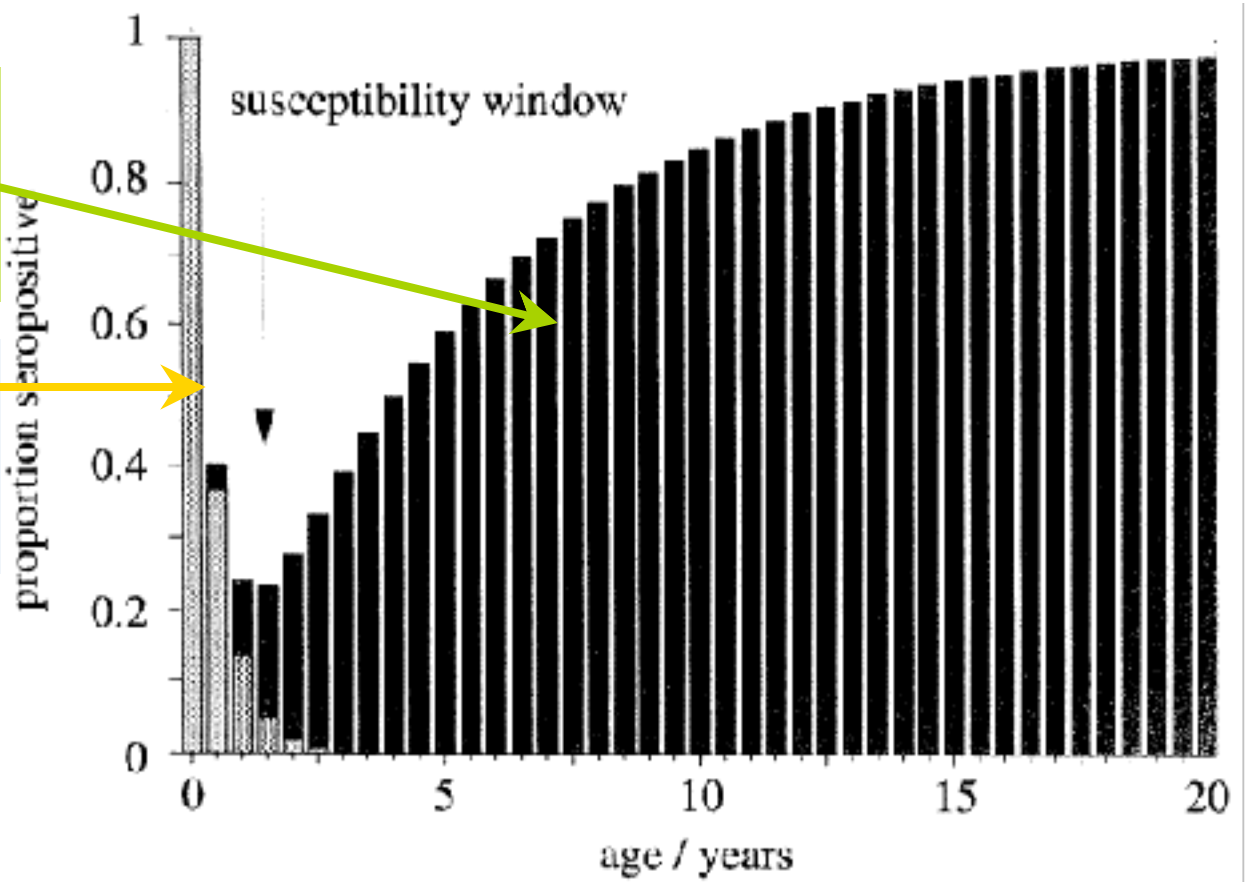
2. Independent data

- For *S(E)IR* model, we can calculate average length of time it takes for an individual to acquire infection (assuming born susceptible)
- Expression for *Mean Age at Infection* is

$$A \approx \frac{1}{\mu R_0} \qquad \Rightarrow A \approx \frac{L}{R_0} \qquad \Rightarrow R_0 \approx \frac{L}{A}$$

R_0 is mean life expectancy (L) divided by mean age at infection (A)

Measles Age-Stratified Seroprevalence



Infection-derived immunity

Maternally-derived antibodies

Mean age at infection (A) is ~4.5 years

Assume L~75, so R_0 ~ 16.6

Historical significance

Anderson & May (1982; *Science*)

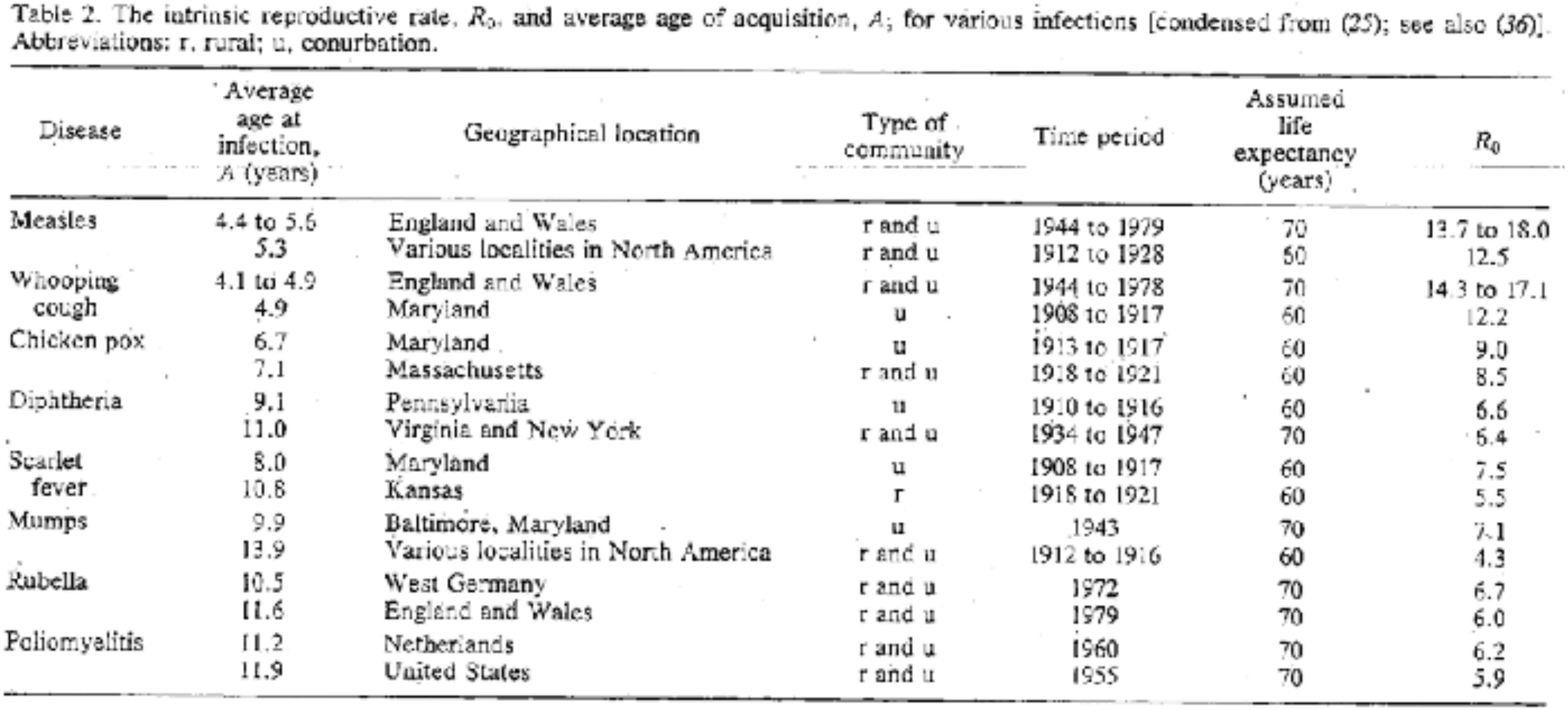


Table 2. The intrinsic reproductive rate, R_0, and average age of acquisition, A, for various infections [condensed from (25); see also (36)]. Abbreviations: r, rural; u, conurbation.

Disease	Average age at infection, A (years)	Geographical location	Type of community	Time period	Assumed life expectancy (years)	R_0
Measles	4.4 to 5.6	England and Wales	r and u	1944 to 1979	70	13.7 to 18.0
	5.3	Various localities in North America	r and u	1912 to 1928	60	12.5
Whooping cough	4.1 to 4.9	England and Wales	r and u	1944 to 1978	70	14.3 to 17.1
	4.9	Maryland	u	1908 to 1917	60	12.2
Chicken pox	6.7	Maryland	u	1913 to 1917	60	9.0
	7.1	Massachusetts	r and u	1918 to 1921	60	8.5
Diphtheria	9.1	Pennsylvania	u	1910 to 1916	60	6.6
	11.0	Virginia and New York	r and u	1934 to 1947	70	6.4
Scarlet fever	8.0	Maryland	u	1908 to 1917	60	7.5
	10.8	Kansas	r	1918 to 1921	60	5.5
Mumps	9.9	Baltimore, Maryland	u	1943	70	7.1
	13.9	Various localities in North America	r and u	1912 to 1916	60	4.3
Rubella	10.5	West Germany	r and u	1972	70	6.7
	11.6	England and Wales	r and u	1979	70	6.0
Poliomyelitis	11.2	Netherlands	r and u	1960	70	6.2
	11.9	United States	r and u	1955	70	5.9

3. Epidemic Take-off

A slightly more common approach is to study the epidemic take off

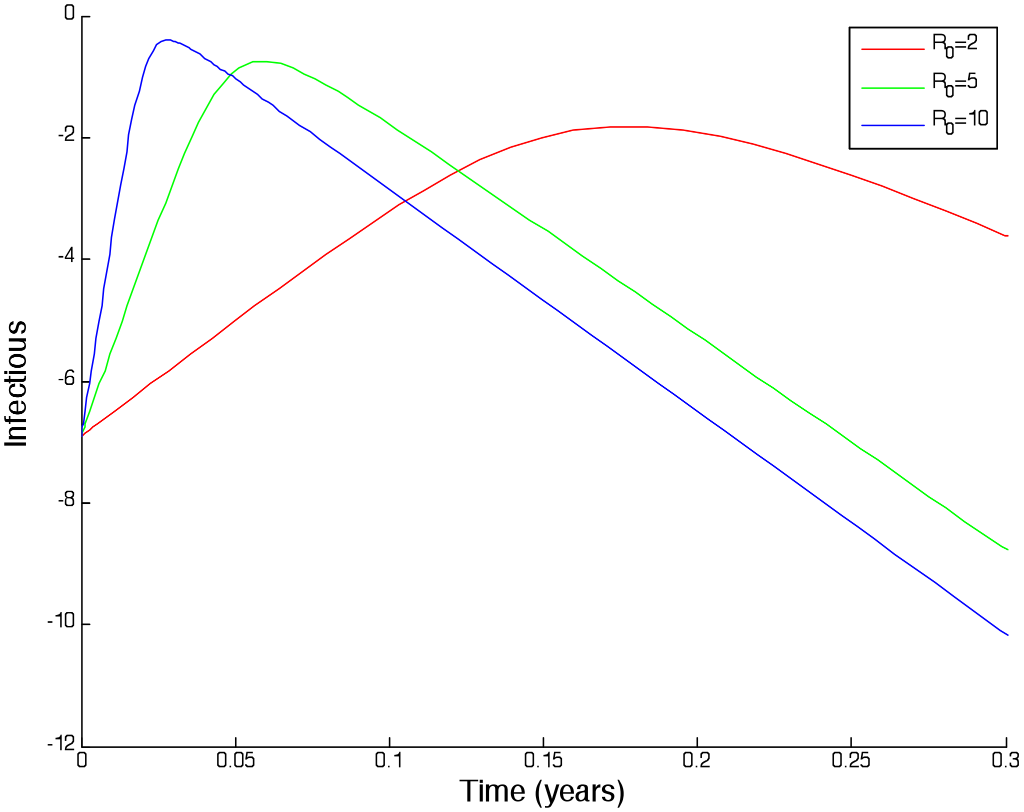
Recall from linear stability analysis that

$$I_{SIR} \approx I(0) \times e^{(R_0-1)\gamma t}$$

Take logarithms

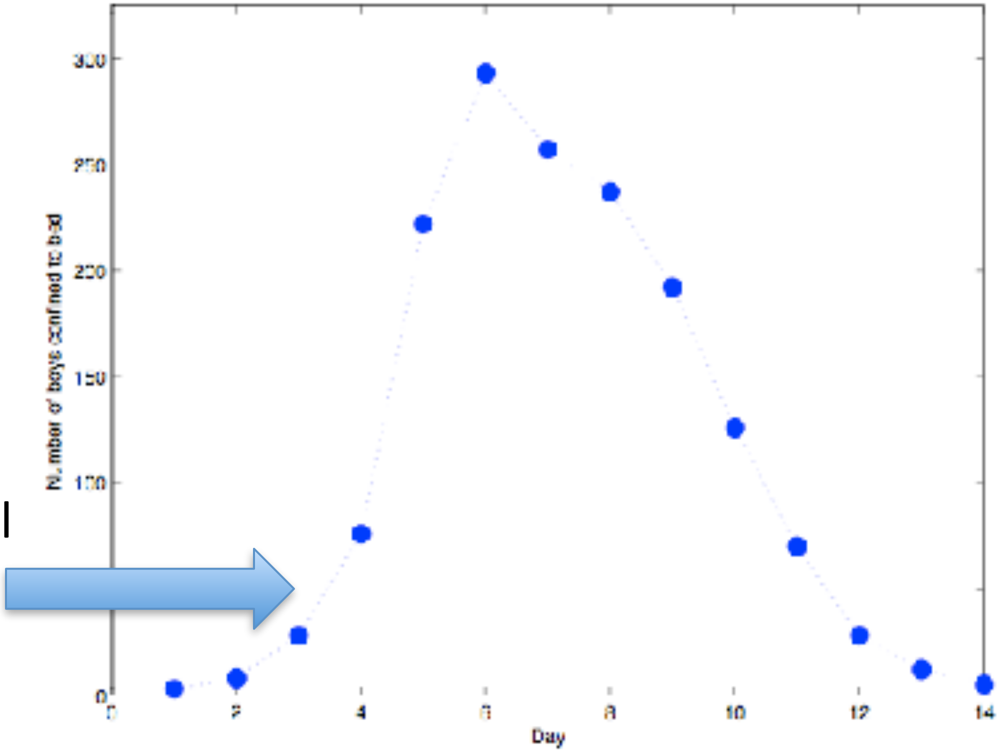
$$\log(I_{SIR}) = \log(I(0)) + (R_0 - 1)\gamma t$$

So, regression slope will give R_0



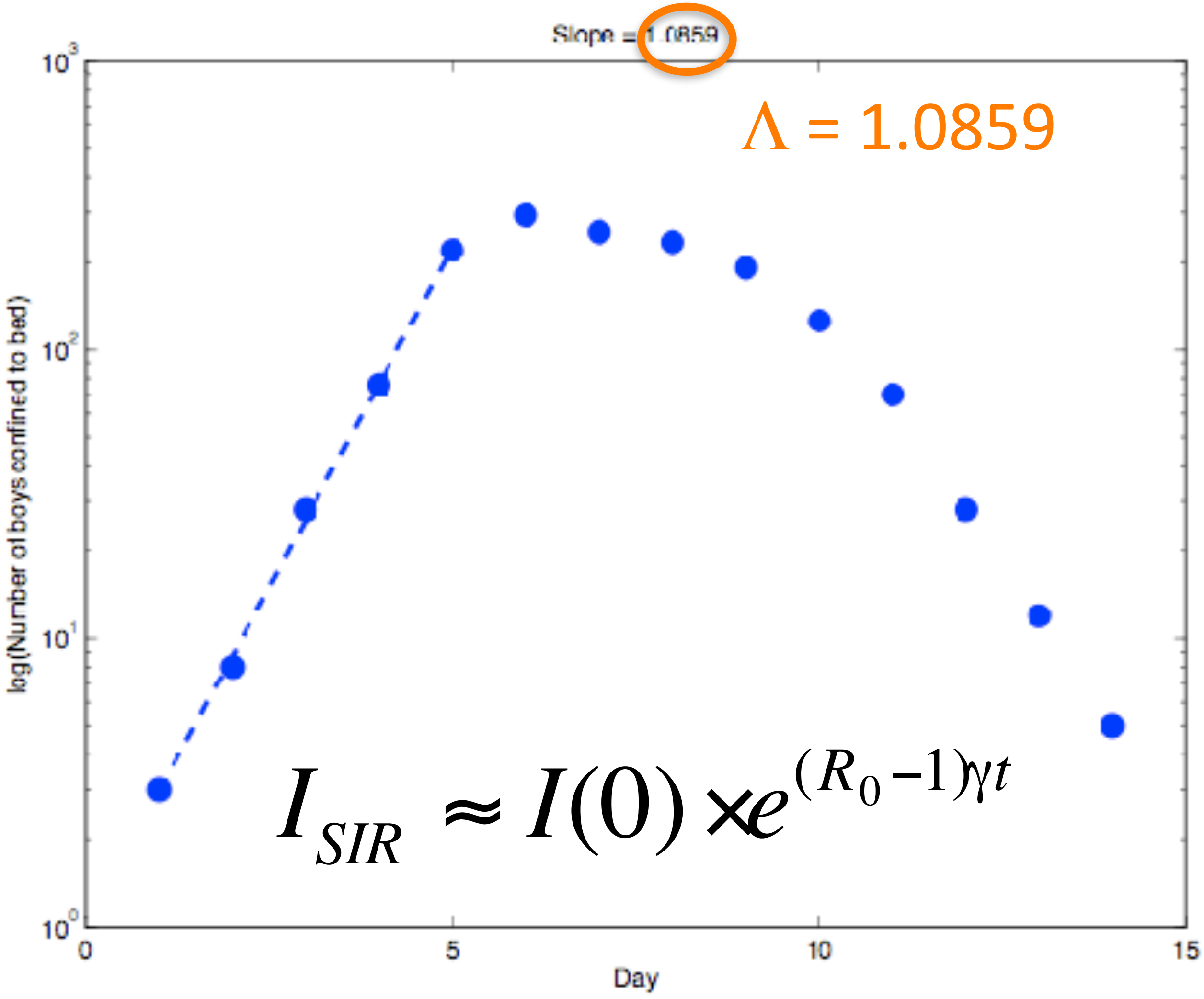
3. Epidemic take-off

- Back to school boys



Looks like classic exponential take-off

Epidemic take-off



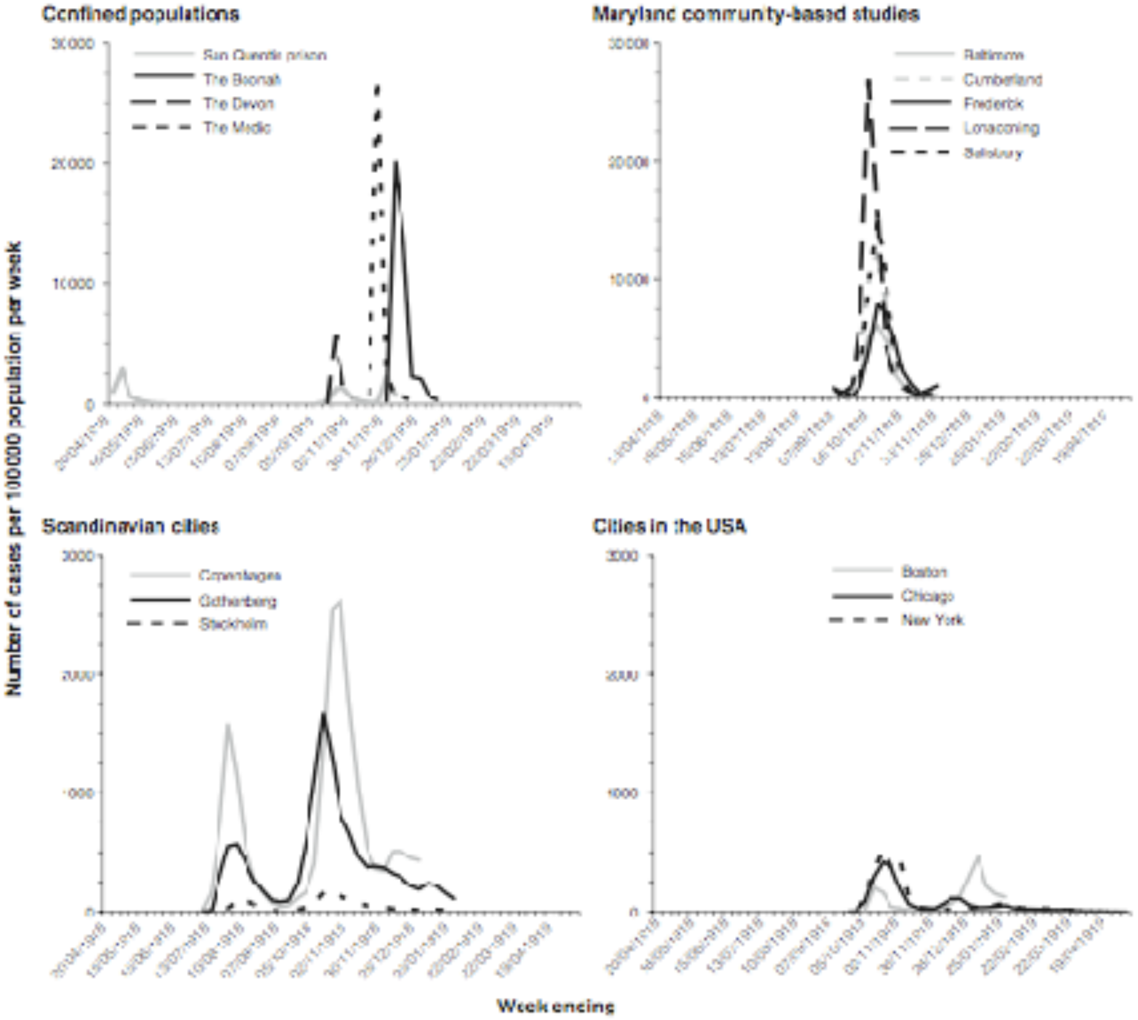
$\Lambda = 1.0859$

$$I_{SIR} \approx I(0) \times e^{(R_0 - 1)\gamma t}$$

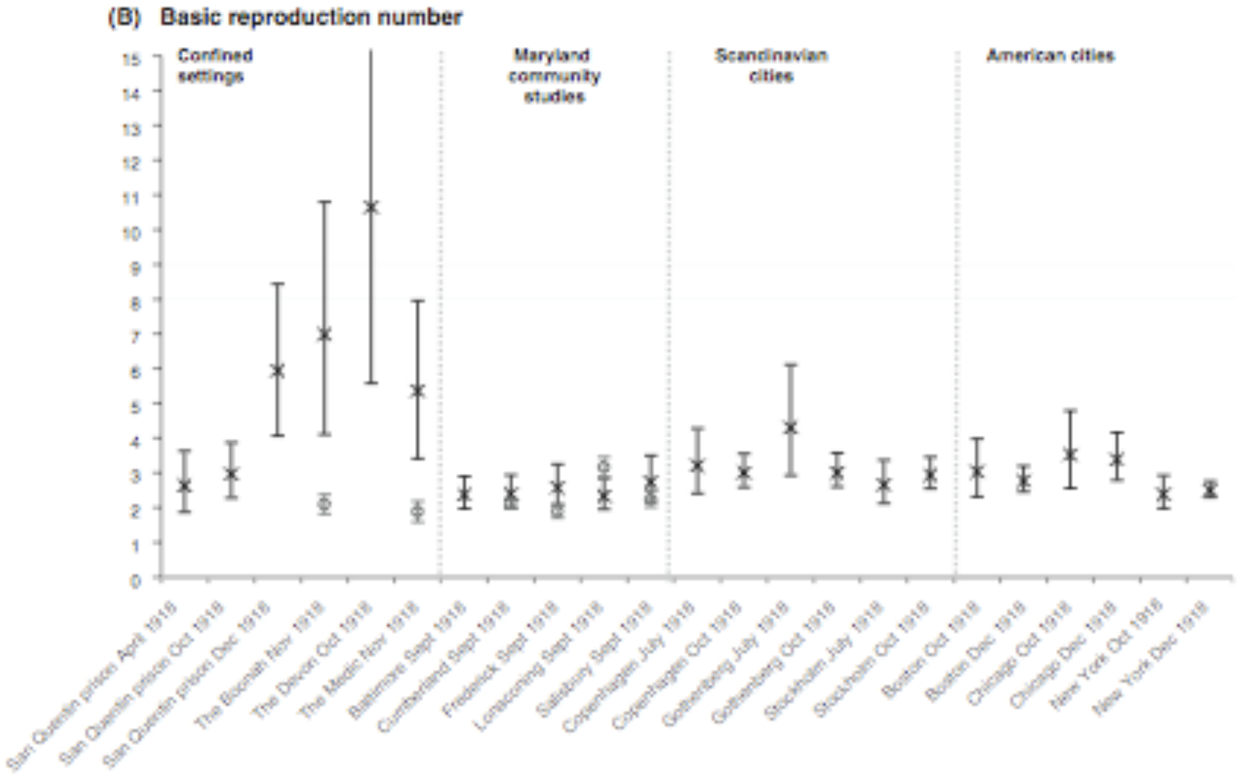
Our value for 'flu incubation period

So,
$R_0 = 1.0859 * 2.5 + 1$
$= 3.7$

Vynnycky *et al.* (2007)



Vynnycky *et al.* (2007)



Variants on this theme

- Recall

$$\log(I_{SIR}) = \log(I(0)) + (R_0 - 1)\gamma t$$

- Let T_d be ‘doubling time’ of outbreak
- Then,
  - ⋆ **$R_0 = \log(2) / T_d\gamma + 1$**

4. Likelihood & inference

- We focus on random process that (putatively) generated data
- A model is explicit, mathematical description of this random process
- "The likelihood" is probability that data were produced given model and its parameters:

$$L(\text{model} \mid \text{data}) = \Pr(\text{data} \mid \text{model})$$

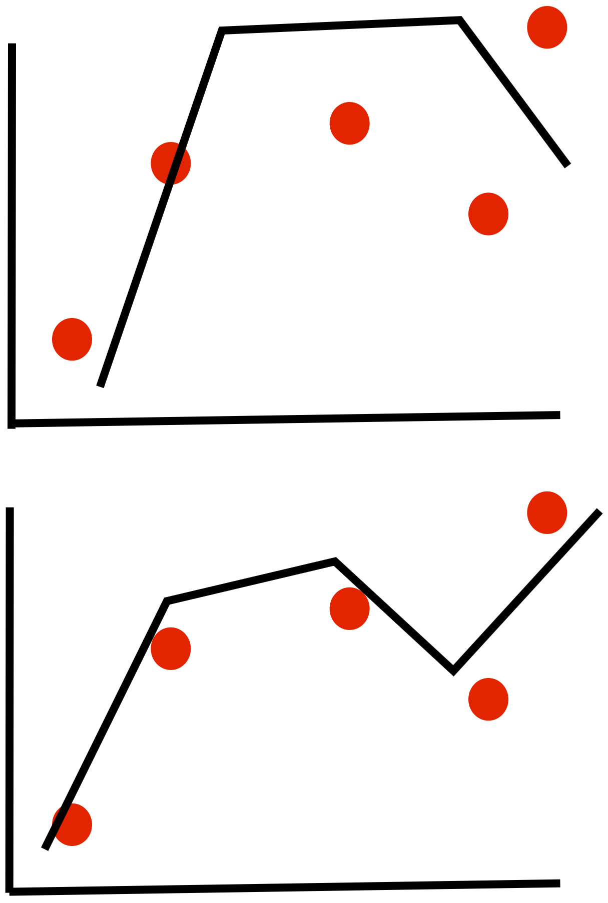
- Likelihood quantifies (in some sense optimally) model goodness of fit

4. Likelihood & estimation

- Assume we have data, D, and model output, M (both are vectors containing state variables). Model predictions generated using set of parameters, θ
- Transmission dynamics subject to
  - "process noise": heterogeneity among individuals, random differences in timing of discrete events (environmental and demographic stochasticity)
  - "observation noise": random errors made in measurement process itself

4. Likelihood & estimation

- If we ignore process noise, then model is deterministic and all variability attributed to measurement error
- Observation errors assumed to be sequentially independent
- Maximizing likelihood in this context is called 'trajectory matching'



4. Likelihood & estimation

- Data, D
- Model output, M
- Parameters, θ

- If we assume measurement errors are normally distributed, with mean μ and variance σ^2 then

$$L(M(\theta) \mid D) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(D_i - M_i)^2}{2\sigma^2}}$$

4. Likelihood & estimation

- Data, D
- Model output, M
- Parameters, θ

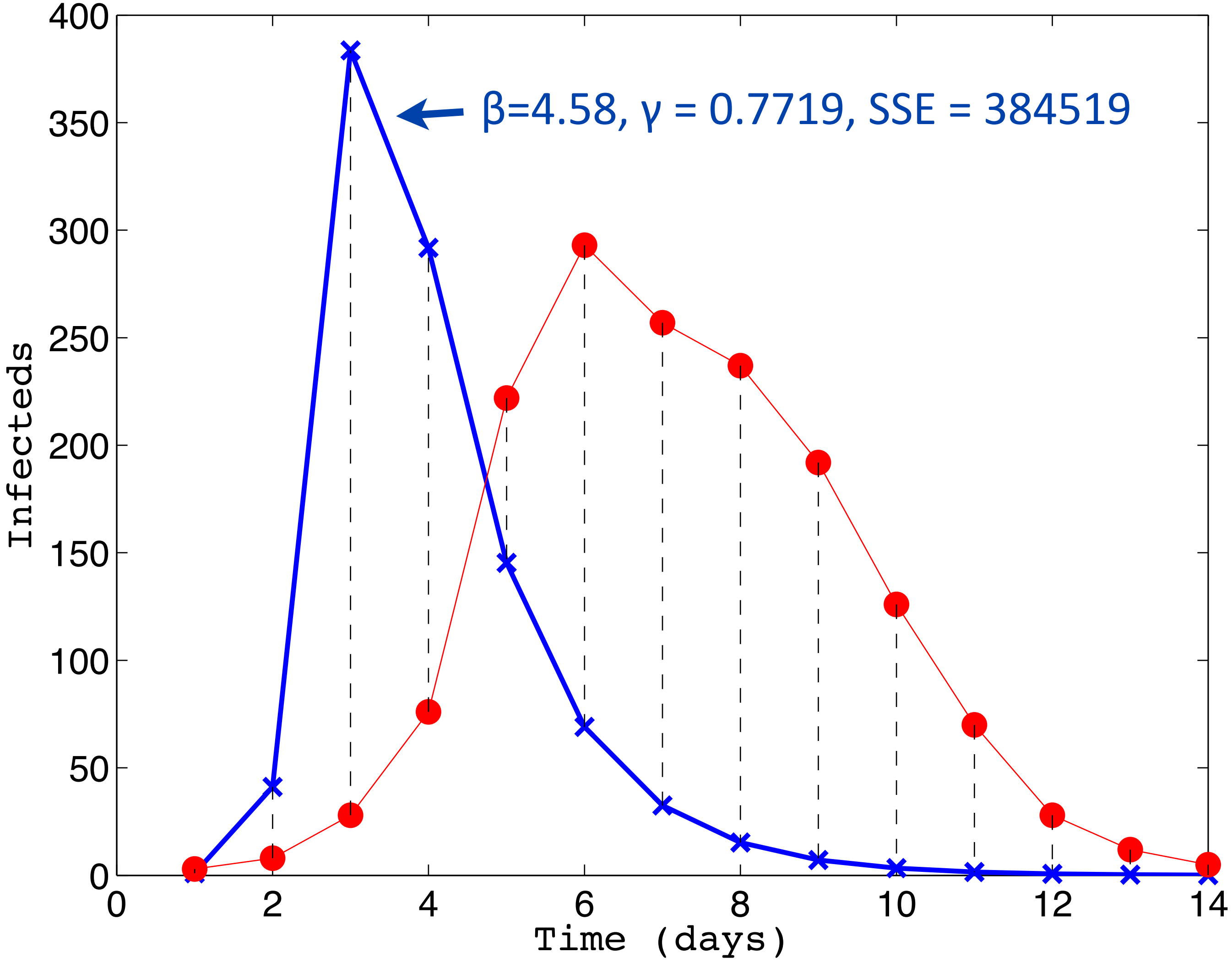
- Often easier to deal with Log-likelihoods:

$$\log\left(L(M(\theta)\,|\,D)\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_i (D_i - M_i)^2$$

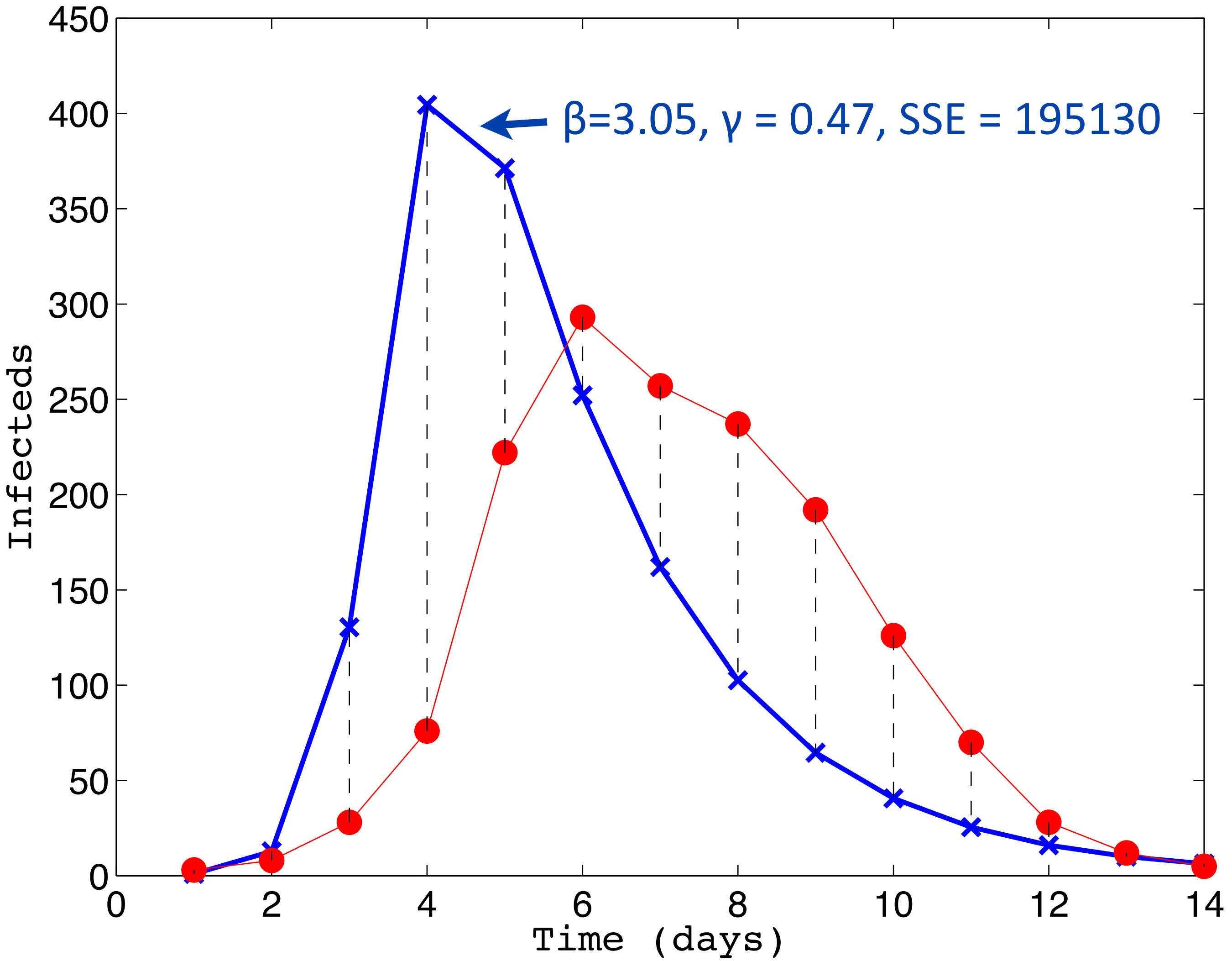
4. Likelihood & estimation

- Under such conditions, Maximum Likelihood Estimate, MLE, is simply parameter set with smallest deviation from data
- Equivalent to using least square errors, to decide on goodness of fit
  - Least Squares Statistic = SSE = $\Sigma(D_i - M_i)^2$
- Then, miminise SSE to arrive at MLE

Trajectory matching

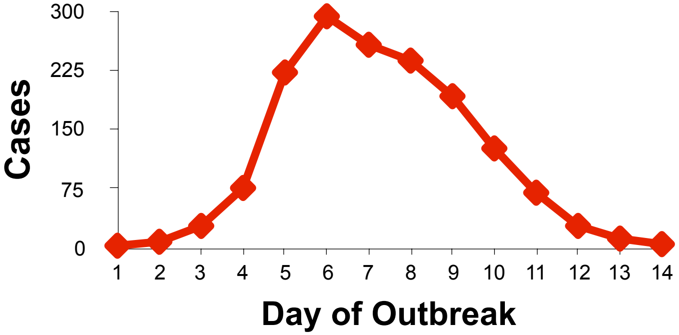


Trajectory matching



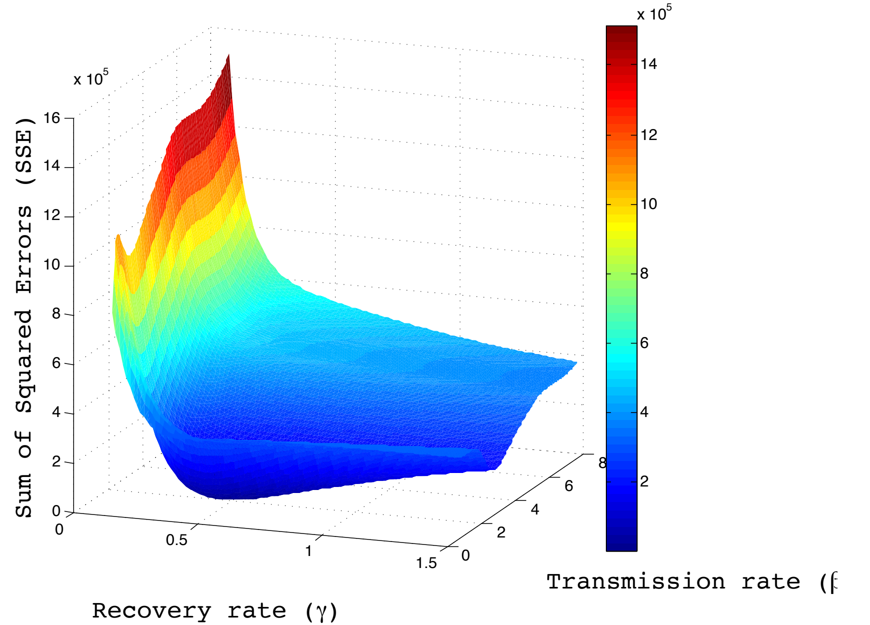
β=3.05, γ = 0.47, SSE = 195130

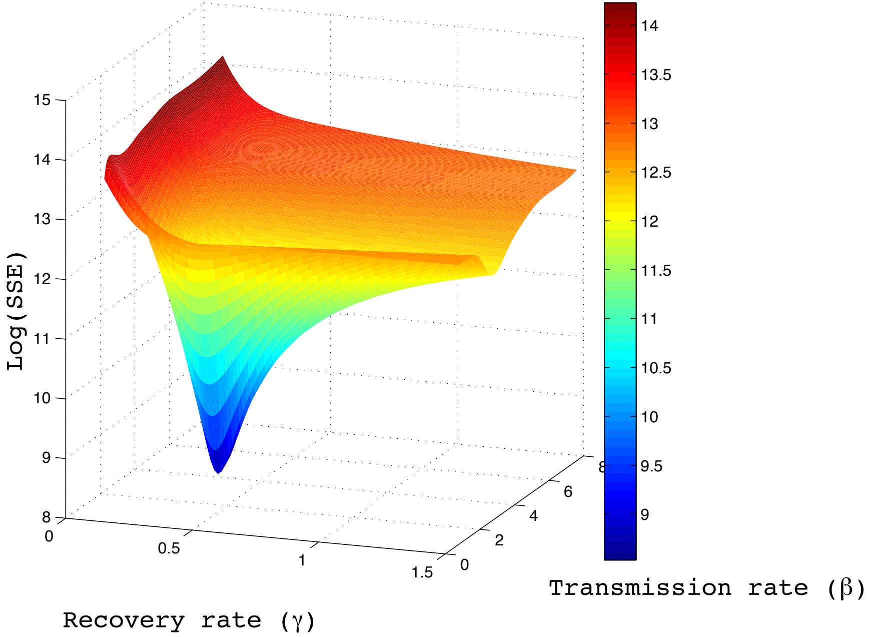
Model estimation: Influenza outbreak



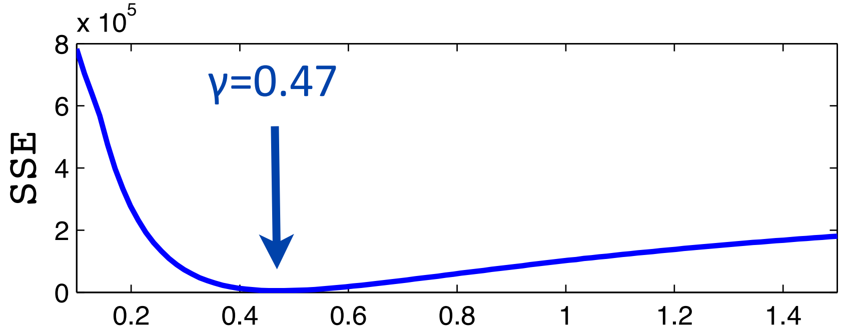
- Systematically vary β and γ, calculate SSE

- Parameter combination with lowest SSE is 'best fit'



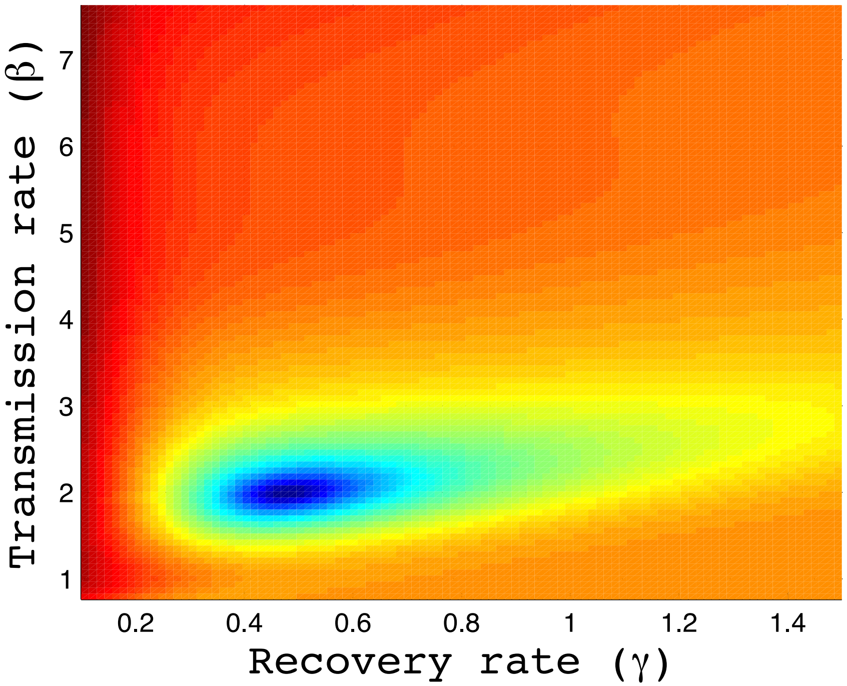


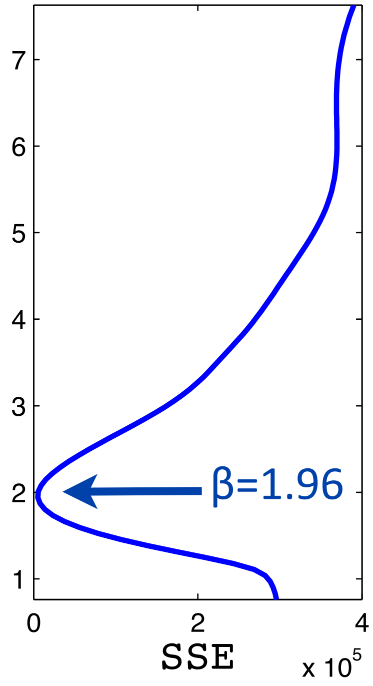
Model estimation: Influenza outbreak



Best fit parameter values:

1. β = 1.96 (per day)
2. 1/γ = 2.1 days
3. R_0 ~ 4.15





Generally, may have more parameters to fit, so grid search not efficient

Nonlinear optimization algorithms (eg Nelder-Mead) would be used

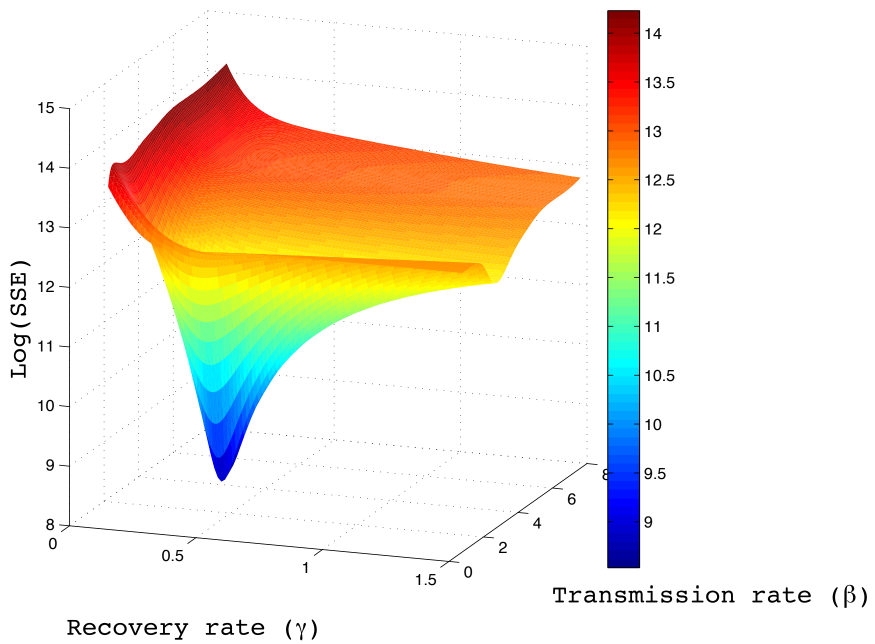
4. Likelihood & estimation

- How do we relate SSE to logLik?

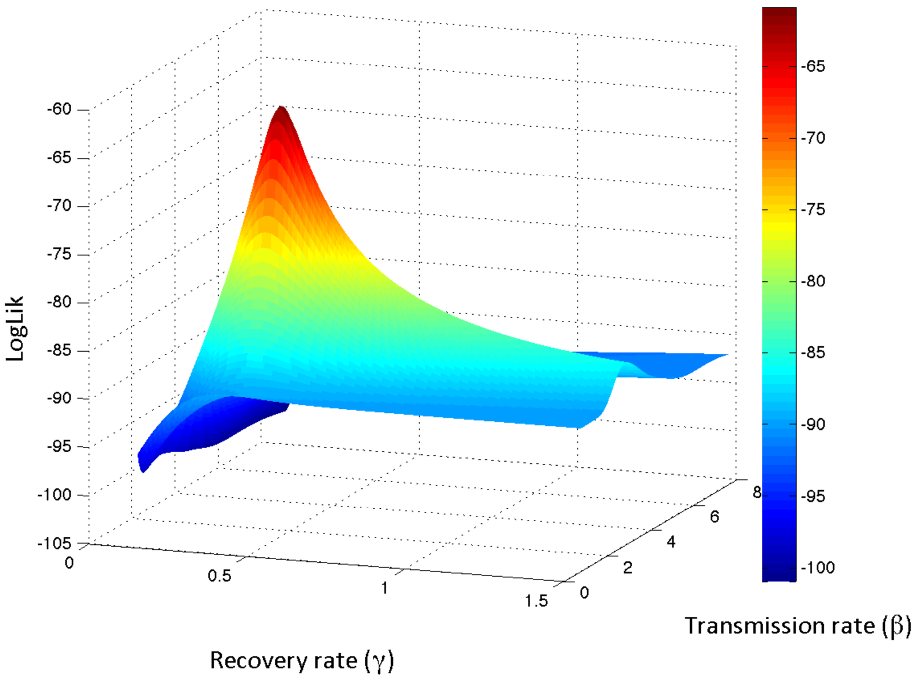
$$\log\left(L(M(\theta)\mid D)\right) = -\frac{\overset{\text{=length of data}}{n}}{2}\log\left(2\pi\underbrace{\sigma^2}_{\text{=SSE/n}}\right) - \frac{1}{2\sigma^2}\underbrace{\sum_i (D_i - M_i)^2}_{\text{=SSE}}$$

Model estimation: Influenza outbreak

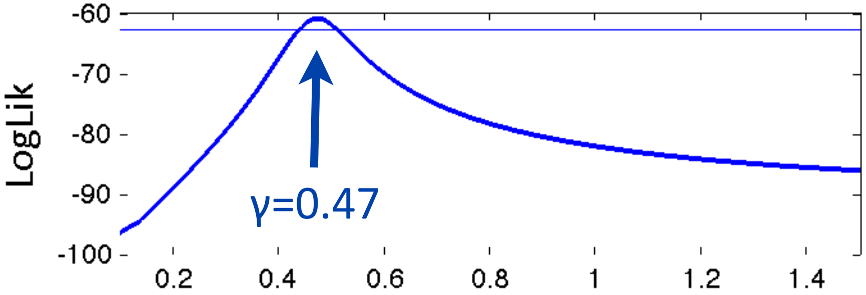
SSE



LogLik

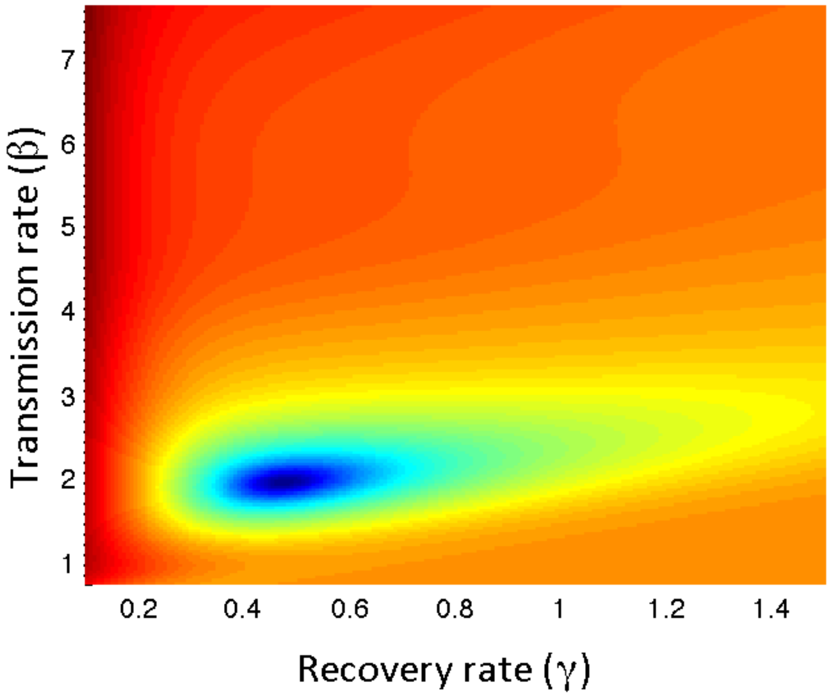


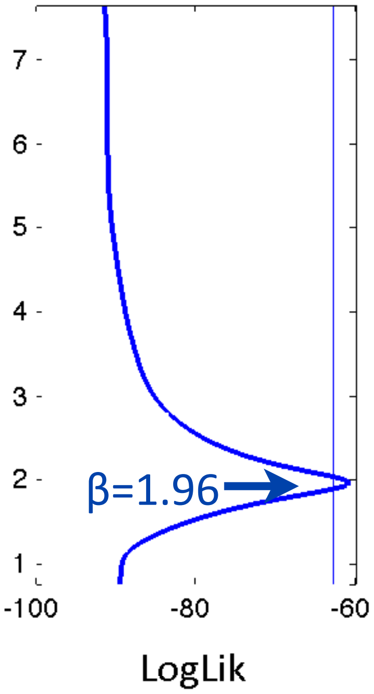
Model estimation: Influenza outbreak



Maximum Likelihood Estimates:

1. β = 1.96 (per day)
2. $1/\gamma$ = 2.1 days
3. R_0 ~ 4.15





Recall 2 log-likelihood units indicate significant difference

Can use likelihood profiles to put confidence intervals on estimates

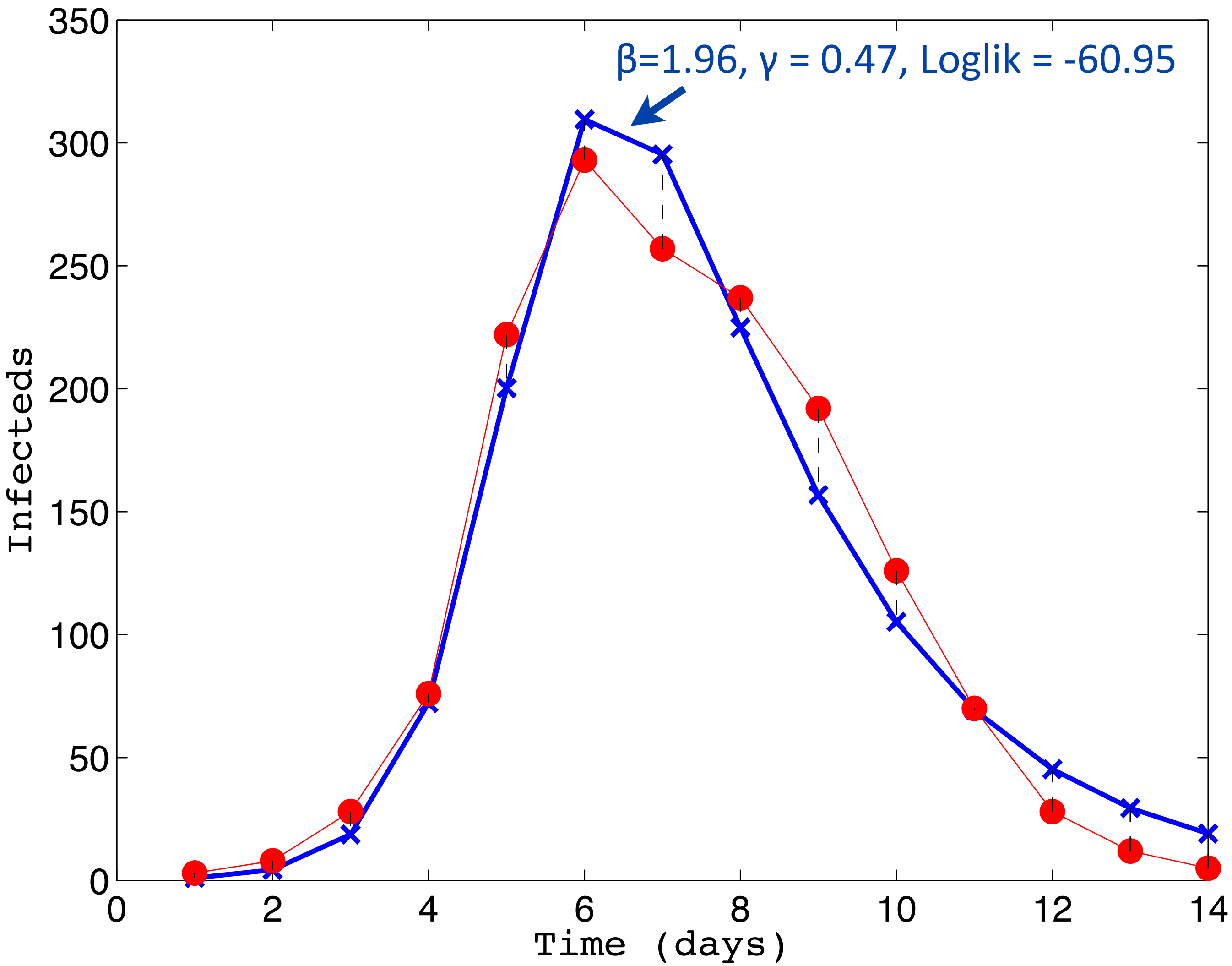
β=1.96 (1.90,2.04)
γ=0.47 (0.43,0.50)

Model comparison

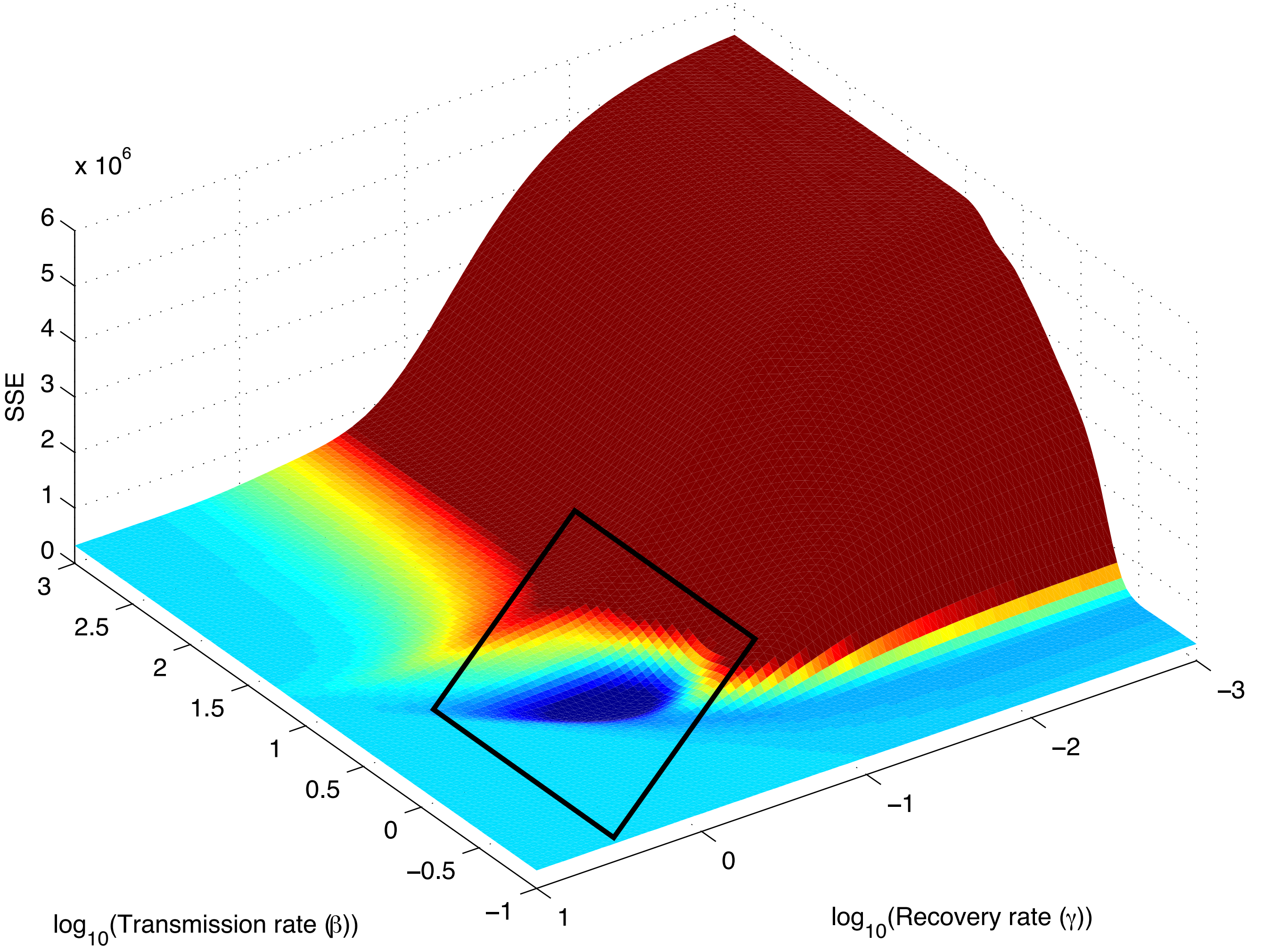
- How to compare models with different number of estimated parameters?
- Commonly use Akaike's Information Criterion
- AIC = 2 *p* - 2 logLik, where *p* is number of estimated parameters for model
- rule-of-thumb: if AIC difference < 2, models indistinguishable

	SIR	Model 2
β	1.96 (1.90,2.04)	
γ	0.47 (0.43,0.50)	
logLik	-60.95	
AIC	125.9	

Likelihood estimation


β=1.96, γ = 0.47, Loglik = -60.95

Likelihood surface



When likelihood surface is somewhat complex, success of estimation using gradient-based optimization algorithms (eg Nelder-Mead) will depend on providing a good initial guess

Caveat

- In boarding school example, data represent number of boys sick ~ Y(t)
- Typically, data are 'incidence' (newly detected or reported infections)
- Don't correspond to any model variables
- May need to 'construct' new information:
  - $dC/dt = \gamma Y$ diagnosis at end of infectiousness
  - $dC/dt = \beta XY/N$
- Set $C(t+\Delta t) = 0$ where Δt is sampling interval of data

Lecture Summary …

- R_0 can be estimated from epidemiological data in a variety of ways
  - Final epidemic size
  - Mean age at infection
  - Outbreak exponential growth rate
  - Curve Fitting
- In principle, variety of unknown parameters may be estimated from data

Further, ...

1. Include uncertainty in initial conditions
   - We took I(0) = 1. Instead could estimate I(0) together with β and γ (now have 1 fewer data points)
2. Explicit observation model
   - Implicitly assumed measurement errors normally distributed with fixed variance, but can relax this assumption
3. What is appropriate model?
   - SEIR model? (latent period before becoming infectious)
   - SEICR model? (“confinement to bed”)
   - Time varying parameters? (e.g. action taken to control spread)

Further, ...

4. Assumed model deterministic -- how do we fit a stochastic model?
   - Use a 'particle filter' to calculate likelihood
5. Can we simultaneously estimate numerous parameters?
   - More complex models have more parameters... estimate all from 14 data points? ⇒ identifiability
6. More complex models are more flexible, so tend to fit better
   - How do we determine if increased fit justifies increased complexity? ⇒ information criteria