The Mathematics and Statistics of Infectious Disease Outbreaks

Michael Höhle[1]

[1]Department of Mathematics
Stockholm University, Sweden

L2: Simulation and Fitting of Epidemic Models[1]

Stockholm University

[1] LaMo: 2020-06-16 @ 22:13:45

Overview

1. Reed-Frost model
2. Deterministic SIR model
3. Stochastic SIR model in continuous time

Outline

1. Reed-Frost model
2. Deterministic SIR model
3. Stochastic SIR model in continuous time

The Reed-Frost epidemic model

- Discrete-time SIR model, where individuals are either
  - *S*usceptible,
  - *I*nfectious or
  - *R*ecovered / *R*emoved (dead, isolated or immune)
- Closed population with initially
  - $x_0 = n$ susceptible and
  - $y_0 = m$ infectious individuals
- Dynamics are described by a discrete-time Markov chain

$$\begin{aligned} Y_{t+1} | x_t, y_t &\sim \text{Bin}(x_t, 1 - (1-w)^{y_t}), \\ X_{t+1} &= x_t - Y_{t+1}, \end{aligned}$$

where w is the probability of an infectious contact between an infectious and a susceptible during one unit of time.

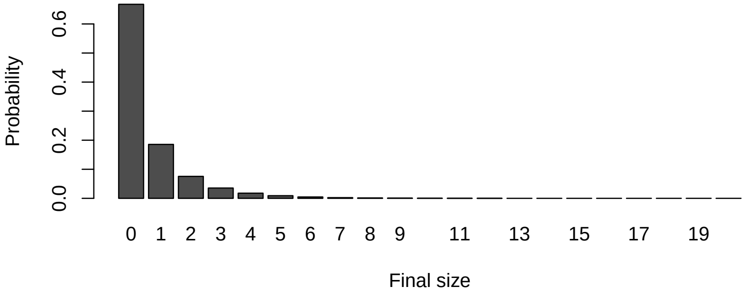
The Reed-Frost epidemic model (2)

Exercise 1

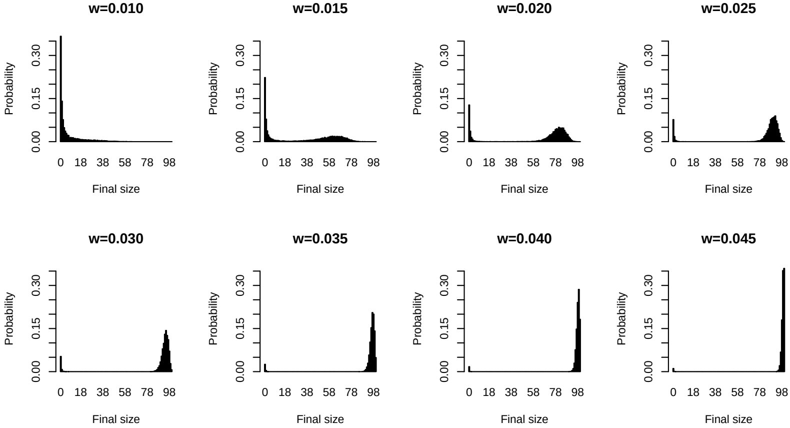
Show that the one-step success probability in the Markov chain equation on the previous slide is $1-(1-w)^{y_t}$.

Final size distribution

- The *final size* of the epidemic is $Z = Y_1 + Y_2 + Y_3 + \dots$
- Final size distribution can be computed exactly for small n, say $n \leq 30$.
- Final size distribution for $n = 20$, $m = 1$ and $w = 0.02$:



Simulated final size distribution for $n = 100$ and $m = 1$



R Code for Simulation of the Reed-Frost Model

```r
fsize.RF <-
function(n, m, w, samples) {
  #Initial susceptible
  xj <- matrix(data=n,nrow=samples,ncol=1)
  #Initial infectives
  yj <- matrix(data=m,nrow=samples,ncol=1)

  #Loop over all (samples) simulations until they all are ceased.
  while (sum(yj>0) & sum(xj>0)) {
    #Sample from all processes concurrently
    yj <- ifelse(xj > 0, rbinom(samples, xj, 1-(1-w)^yj), 0)
    #Update all xj
    xj <- xj - yj
  }
  #Done
  return(n-xj)
}
```

Aside: Inference by Maximum Likelihood Estimation

- Maximum Likelihood Estimation is a method in statistics to estimate the parameters of a statistical model
- The statistical model leads to a probability distribution for the observed data, i.e. in the discrete case $f_{Model}(\boldsymbol{y};\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y})$.
- Considering data as being fixed we can formulate the likelihood function as $L(\boldsymbol{\theta};\boldsymbol{y}) = f_{Model}(\boldsymbol{y};\boldsymbol{\theta})$.
- The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.

Statistical inference

- Estimation of w from time series data $\mathbf{y} = (y_0, y_1, y_2, \ldots, y_K)$ using the binomial likelihood

$$L(w) \propto \prod_{t=0}^{K-1} p_t^{y_{t+1}} (1 - p_t)^{x_t - y_{t+1}},$$

  here $p_t = 1 - (1 - w)^{y_t}$. $\rightarrow$ Knowledge of x_0 is required.

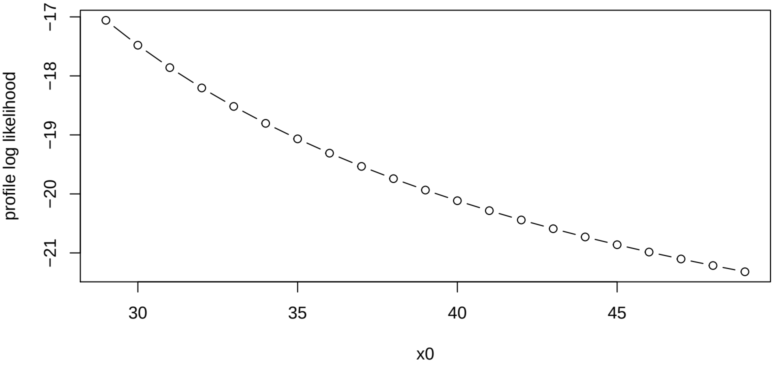
- Uncertainty of $\hat{w}$ can be quantified with a 95% confidence interval.
- Example: Generation sizes of a measles epidemic in St. Petersburg (from Table 4.1 in Daley and Gani, 1999): $\mathbf{y} = (1, 4, 14, 10, 1, 0)$
- Assume all susceptibles got infected: $x_0 = 4 + 14 + 10 + 1 = 29$

Example

```
######################################################################
# Likelihood function for the Reed-Frost model
#
# Parameters:
#  w.logit - logit(w) to have unrestricted parameter space
#  x       - vector containing the number of susceptibles at each time
#  y       - vector containing the number of infectious   at each time
#
######################################################################

l <- function(w.logit,x,y) {
  if (length(x) != length(y)) { stop("x and y need to be the same length") }
  K <- length(x)
  w <- plogis(w.logit)
  p <- 1 - (1-w)^y
  return(sum(dbinom( y[-1], size=x[-K], prob=p[-K],log=TRUE)))
}

# Epidemic D in Table 4.1 of Daley and Gani (1999), assuming all susceptibles got infected
y <- c(1, 4, 14, 10, 1, 0)
x <- numeric(length(y))
x[1] <- sum(y[-1])
x[2:length(x)] <- x[1]-cumsum(y[2:length(y)])

mle <- optim(par=0,fn=l,method="BFGS",x=x,y=y,control=list(fnscale=-1),hessian=TRUE)
# Maximum likelihood estimator
(w.hat <- plogis(mle$par))
## [1] 0.1700922
```

Inference for x_0

Maximize log likelihood for $x_0 = 29, 30, 31, \ldots$



Outline

1. Reed-Frost model
2. Deterministic SIR model
3. Stochastic SIR model in continuous time

The SIR model

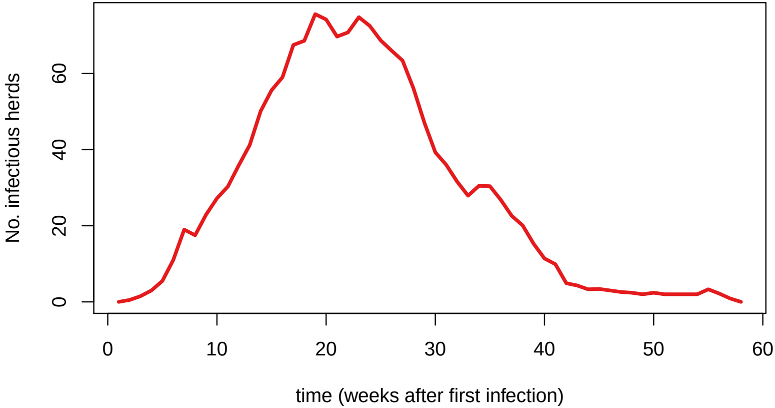
- When the population considered is large, it can be sufficient to disregard the stochasticity of the epidemic process and use deterministic models.
- Can formulate a continuous-time deterministic *SIR model* by using ordinary differential equations (ODEs).
- The deterministic system intends to model the mean behaviour of the underlying stochastic system.
- We assume a closed population (i.e. no demographics turnover) of size N.

Example: CSFV in The Netherlands (1)

- Classical swine fever virus (CSFV) is a highly contagious disease of pigs and wild boar.
- Characteristics of the disease are
  - Symptoms after infection: dullness and anorexia.
  - Acute form: rapid mortality often without clinical symptoms.
  - Secondary symptoms: diarrhea or respiratory problems.
- A huge outbreak in the Netherlands took place between February 1997 and May 1998.
  - 429 infected herds detected and stamped out ($\sim$ 700,000 pigs)
  - 1286 herds pre-emptively slaughtered ($\sim$ 1.1 million pigs)
  - Note: Netherlands has approximately 21,500 pig herds

Example: CSFV in the Netherlands (2)

- Stegeman et al. (1999) provide estimates on the weekly number of infectious herds from contact tracing and serological analysis:



SIR differential equation system

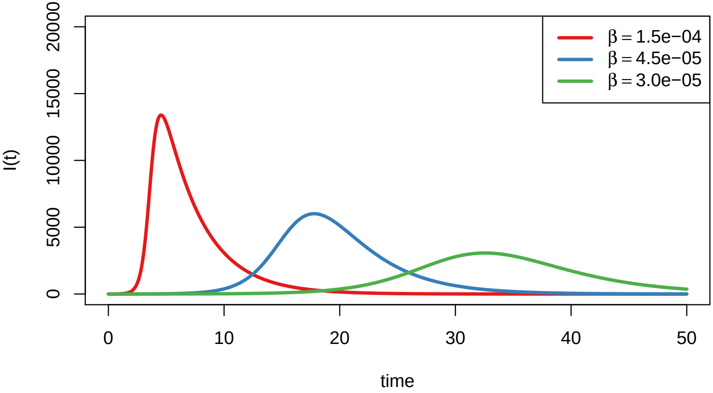
- As before, divide population into three groups
  - *S*usceptible,
  - *I*nfectious or
  - *R*ecovered / *R*emoved
- At all times $S(t) + I(t) + R(t) = N$, so $S(0) + I(0) = N$.
- Describe dynamics using an ordinary differential equation system

$$
\begin{aligned}
\frac{dS(t)}{dt} &= -\beta S(t) I(t) \\
\frac{dI(t)}{dt} &= \beta S(t) I(t) - \gamma I(t) \\
\frac{dR(t)}{dt} &= \gamma I(t)
\end{aligned}
$$

  where β, $\gamma > 0$.
- Solve ODE with initial condition $(S(0), I(0), 0)$ using numerical routines (`R` package `deSolve`).

Example

- Number of infected $I(t)$ for $\gamma = 0.3$, $N = 2 \times 10^4$, $I(0) = 1$ and different values of β.



Numerical Solution of the SIR ODE

- Defining the vector of derivatives for the SIR ODE

```
##############################################################################
# Function to compute the derivative of the ODE system
#
#  t - time
#  y - current state vector of the ODE at time t
#  parms - Parameter vector used by the ODE system
#
# Returns:
#  list containing dS(t)/dt and dI(t)/dt
##############################################################################

sir <- function(t,y, parms) {
  beta <- parms[1]
  gamma <- parms[2]
  S <- y[1]
  I <- y[2]
  return(list(c(S=-beta*S*I,I=beta*S*I-gamma*I)))
}
```

- Use `deSolve::lsoda`

```
sim <- lsoda(y=c(N-1,1), times=times, func=sir,parms=c(beta.grid[1],gamma))
head(sim, n=3)
##            time        1        2
## [1,] 0.00000000 19999.00 1.000000
## [2,] 0.05005005 19998.84 1.144682
## [3,] 0.10010010 19998.66 1.310297
```

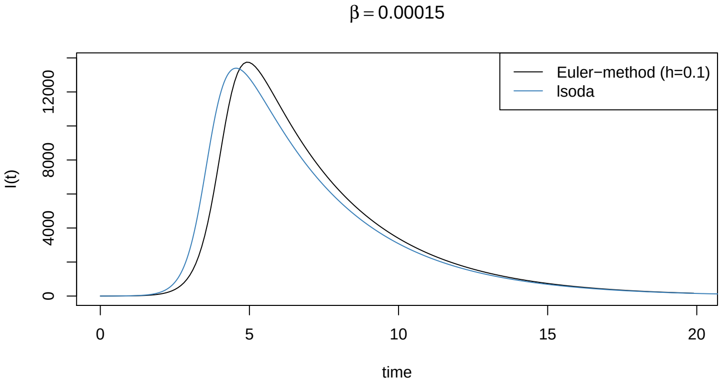
Aside: Numerical Solution of ODEs (1)

- Simple method to solve an ODE system numerically given the initial condition: Euler-Method
- Example in R: Stepwidth h and initial value is $S(0) = N - 1$ and $I(0) = 1$

```
# Step width of the Euler method
h <- 0.1
y <- matrix(NA, nrow=ceiling(20/h), ncol=3, dimnames=list(NULL, c("t","S","I")))
# Initial value
y[1,] <- c(0,N-1,1)
# Loop
for (i in 2:nrow(y)){
  y[i,] <- c(y[i-1,"t"]+h, y[i-1,c("S","I")] +
         h * sir(y[i-1,"t"], y[i-1,c("S","I")], parms=c(beta.grid[1],gamma))[[1]])
}
```

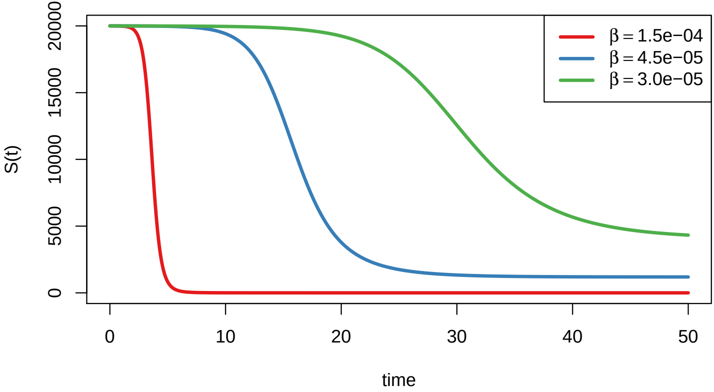
Aside: Numerical Solution of ODEs (2)

- Plot for Euler solve for SIR system



Example cont.

- Number of susceptibles $S(t)$ for $\gamma = 0.3$, $N = 2 \times 10^4$, $I(0) = 1$ and different values of β.



Estimating parameters (1) – Gaussian observations

- We have k observations $\boldsymbol{y}_i = (S(t_i), I(t_i))'$ at times $t_1, \ldots, t_k$ with mean $\mathrm{E}(\boldsymbol{y}_i; \boldsymbol{\theta})$, determined by the SIR ODE.
- Least squares estimates $\boldsymbol{\theta} = (\beta, \gamma)'$ minimizes the function

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{k} ||\boldsymbol{y}_i - \mathrm{E}(\boldsymbol{y}_i; \boldsymbol{\theta})||_2,$$

- Solution $\hat{\boldsymbol{\theta}}$ is found using numerical optimizing routines.
- Often only $I(t)$ is available, but not $S(t)$. Then least squares corresponds to MLE for Gaussian observations with

$$I(t_i) \sim N(\mathrm{E}(I(t_i); \boldsymbol{\theta}), \sigma^2).$$

  where σ^2 is variance of the observation noise (kept fixed).
- Square-root transform of $I(t_i)$ and $\mathrm{E}(I(t_i); \boldsymbol{\theta})$ might be useful.

Estimating parameters (3) – MLE for CSFV Data

- Define the log-likelihood function

```
######################################################################
#Least-squares fit
######################################################################

ll.gauss <- function(theta, take.sqrt=FALSE) {
  #Solve ODE using the parameter vector theta
  res <- lsoda(y=c(N-1,1), times=csfv$t, func=sir, parms=exp(theta))
  #Squared difference?
  if (take.sqrt==FALSE) {
    return(sum(dnorm(csfv$I,mean=res[,3],sd=1,log=TRUE)))
  } else {
    return(sum(dnorm(sqrt(csfv$I),mean=sqrt(abs(res[,3])),sd=1,log=TRUE)))
  }
}
```

- Maximize the log-likelihood using optim and compute estimates

```
#Determine MLE
N <- 21500
mle <- optim(log(c(0.00002,3)), fn=ll.gauss,control=list(fnscale=-1))

#Show estimates and resulting R0 estimate
beta.hat <- exp(mle$par)[1]
gamma.hat <- exp(mle$par)[2]
R0.hat <- beta.hat*N/gamma.hat
```

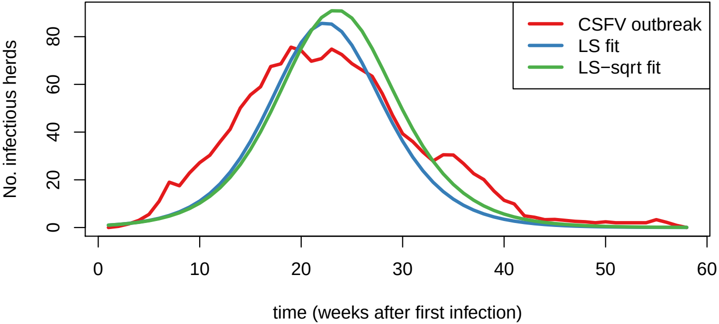
Estimating parameters (3) – MLE for CSFV Data

- Plug-in of the MLE to find solution of the ODE

```
mu <- lsoda(y=c(N-1,1), times=csfv$t, func=sir,parms=exp(mle$par))
head(mu, n=3)
##      time        1        2
## [1,]    1 21499.00 1.000000
## [2,]    2 21495.42 1.313401
## [3,]    3 21490.71 1.723989
```

Estimating parameters (3) – MLE for CSFV Data

- Example: SIR model fitted to CSFV curve by Gaussian likelihood



The MLEs are $\hat{\beta} = 0.00015$ (0.00014 for LS-sqrt), $\hat{\gamma} = 2.85$ (2.65) and $\hat{R}_0 = 1.10$ (1.10).

Estimating parameters (4) – Poisson observations

- Assuming Gaussian observation ignores the fact that we actually observe count data. For small counts this may become problematic.
- An alternative is to use a count data distribution, e.g.

$$y_i \sim \mathrm{Po}(I(t_i)).$$

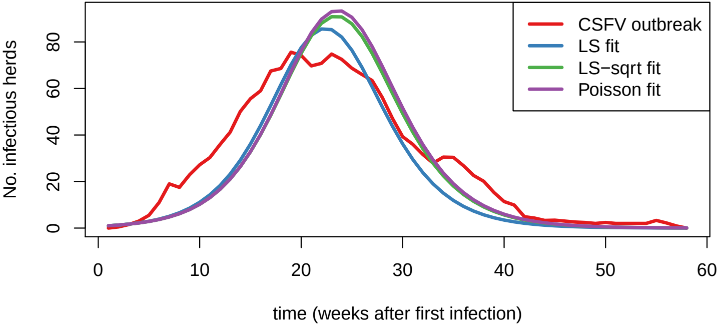
- As a consequence the log-likelihood is

$$\log(L(\boldsymbol{\theta})) = \sum_{i=1}^{k} y_i \log(I(t_i)) - I(t_i) + const.$$

- Since for the Poisson distribution $E(y_i) = Var(y_i)$, it might be necessary to address additional over-dispersion in the data using, e.g., a negative binomial distribution.

Estimating parameters (5) – MLE for CSFV Data

- Example: SIR model fitted to CSFV curve by Poisson likelihood



- The MLEs are $\hat{\beta} = 0.00013$, $\hat{\gamma} = 2.61$ and hence $\hat{R}_0 = 1.10$.

Exercise

Exercise 2

Read the blog post “Flatten the COVID-19 curve” and experiment with different containment strategies using the Shiny App. Discuss the pros and cons of different strategies. Discuss limitations of the model when used to evaluate strategies.

As background information you might want to read the blog post “Coronavirus: The Hammer and the Dance” by Tomas Pueyo.

Outline

1 Reed-Frost model

2 Deterministic SIR model

3 Stochastic SIR model in continuous time

Stochastic SIR model in continuous time (1)

- If the population under study is large enough, deterministic approximations are reasonably valid to obtain an understanding of the disease.
- In small populations, however, stochasticity plays an important role for extinction, which cannot be ignored.
- Stochastic epidemic modeling is described in Becker (1989), Daley and Gani (1999) and Andersson and Britton (2000), who all rely heavily on the theory of stochastic processes.

Stochastic SIR model in continuous time (2)

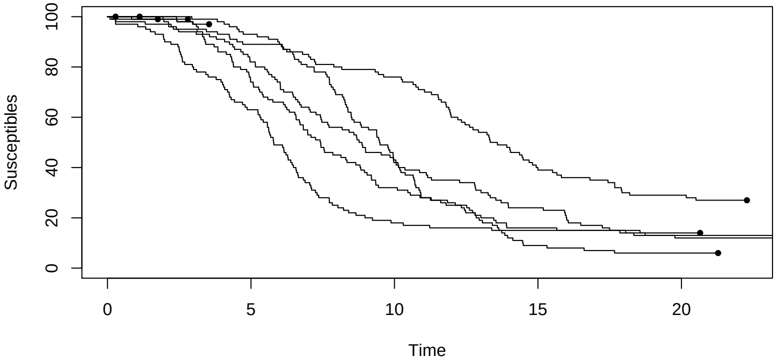
- The *stochastic SIR model* can be described as a continuous-time Markov process, where the event rates for infection and removal are:

Event	Transition		Rate
Infection:	$(S(t), I(t))$	$\rightarrow (S(t)-1, I(t)+1)$	$\beta \cdot S(t) \cdot I(t)$
Removal:		$\rightarrow (S(t), I(t)-1)$	$\gamma \cdot I(t)$

- Again, $R(t)$ is implicitly given, because a fixed population of size $S(0)+I(0)$ is assumed.
- The integer size of the population is now taken into account: Once $I(t)=0$, the epidemic ceases.
- Point process viewpoint: piecewise constant rates, while the length of the infective period is exponentially distributed.

Stochastic SIR model in continuous time (3)

10 SIR simulations with $S(0) = 100$, $I(0) = 1$, $\beta = 0.01$ and $\gamma = 0.5$:

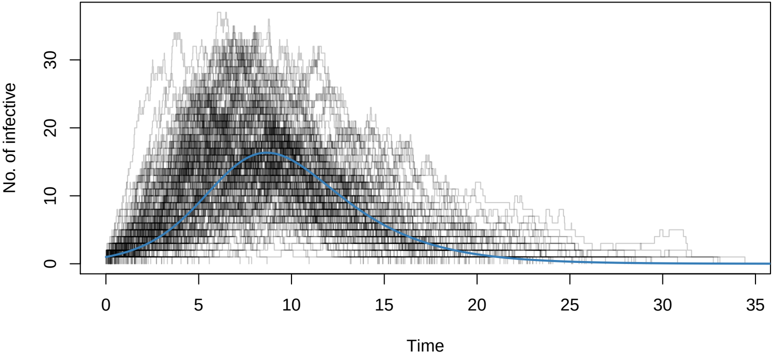


R Code for Simulation of the Stochastic SIR Model

```r
rSIR <-
function(T, beta, gamma, n, m) {
  #Initialize (x= number of susceptibles)
  t <- 0
  x <- n
  y <- m

  #Possible events
  eventLevels <- c("S->I","I->R")
  #Initialize result
  events <- data.frame(t=t,x=x,y=y,event=NA)
  #Loop until we are past time T
  while (t < T & (y>0)) {
      #Draw the waiting type for each possible event
      wait <-  rexp(2,c("S->I"=beta*x*y,"I->R"=gamma*y))
      #Which event occurs first
      i <- which.min(wait)
      #Advance Time
      t <- t+wait[i]
      #Update population according to the eventy type
      if (eventLevels[i]=="S->I") { x <- x-1 ; y <- y+1}
      if (eventLevels[i]=="I->R") { y <- y-1 }
      #Store result
      events <- rbind(events,c(t,x,y,i))
  }
  #Recode event type and return
  events$event <- factor(eventLevels[events$event], levels=eventLevels)
  return(events)
}
```

Stochastic SIR vs. determinstic SIR model

Same parameters as previously - 250 trajectories vs. the ODE solution:



Exercise

Exercise 3

Set $\beta = 0.01$ and $\gamma = 0.5$ and $S(0) = 100$ and $I(0) = 1$. Simulate 1000 instances of the final size of the epidemic using `rSIR` and make a histogram of the result.

Likelihood inference* (1)

- Assume that the epidemic process is completely observed over the interval $(0, \tau]$, where τ is the duration of the epidemic.
- Denote the successive times of the k infectious contacts by $T_1, \ldots, T_k$.
- Denote the PDF of the duration of the infectious period by $f_Y(y)$, e.g. exponentially distributed durations: $f_Y(y) = \gamma \exp(-\gamma y)$.
- Likelihood of the data $\{(t_i, y_i), i = 1, \ldots, k\}$ is

$$L = \left[\prod_{i=1}^{k} f_Y(y_i)\right] \left[\prod_{i=1}^{k} \lambda(t_i)\right] \exp\left(-\int_0^\tau \lambda(u) du\right),$$

where $\lambda(t) = \beta \cdot I(t^-) \cdot S(t^-)$ is the conditional intensity function (CIF) and t^- denotes the time just prior to t_i.

Likelihood inference* (2)

- A complication of the presented equations is that the CIF has to be integrated over time. However, for the SIR model the CIF is a piecewise constant function → integration is tractable.
- A binomial approximation exists for time series data, where $C(t)$ denotes the number of new cases in the interval $(t, t+1]$ (Becker, 1989):
  - The conditional probability of a given susceptible escaping infection during the interval $(t, t+1]$ is approximately $\pi_t = \exp\{-\lambda(t)\}$.
  - We then have

$$C(t) \quad \sim \quad \text{Bin}(S(t), 1 - \pi_t)$$

Likelihood inference* (3) – GLM's

- For the SIR model, $\lambda(t) = \beta \cdot I(t)$ and binomial regression with log link is applicable.
- If $\lambda(t)$ can be assumed to be small, we have

$$1 - \pi_t = 1 - \exp\{-\lambda(t)\} \approx \lambda(t), \text{ so}$$

$$C(t) \sim \text{Bin}(S(t), \lambda(t)) \approx \text{Poisson}(S(t) \cdot \lambda(t))$$

- For the linear formulation $\lambda(t) = \beta I(t)$, a Poisson regression with identity link can be used, with explanatory variables $(S(t))'$.

Literature I

- Andersson, H. and T. Britton (2000). *Stochastic Epidemic Models and their Statistical Analysis*. Vol. 151. Springer Lectures Notes in Statistics. Springer-Verlag.
- Becker, N. G. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall/CRC.
- Daley, D. J. and J. Gani (1999). *Epidemic Modelling: An introduction*. Cambridge University Press.
- Stegeman, A., A. R. W. Elbers, J. Smak, and M. C. M. de Jong (1999). "Quantification of the transmission of classical swine fever virus between herds during the 1997- 1998 epidemic in The Netherlands". In: *Preventive Veterinary Medicine* 42, pp. 219–234.