L02 Estimation in Compartmental Models[1]

Michael Höhle[1]

[1]Department of Mathematics, Stockholm University, Sweden
m_hoehle

STA427 FS2021
Statistical Methods in Infectious Disease Epidemiology
Epidemiology, Biostatistics and Prevention Institute
University of Zurich, Switzerland

University of Zurich[UZH]

[1]LaMo: 2021-02-27 @ 21:50:05

Outline

Overview

Outline

Statistical challenges

Statistics in a nutshell:

Stochastic model + data $\rightarrow$
Parameter estimation + quantification of uncertainty

- Only one realization of the epidemic is observed.
- The data used for estimation can contain serious problems, e.g. under-reporting, changes in the test behaviour.
- The analysis is conducted using all available covariates, but important risk covariates might be missing in the analysis.

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.

Outline

Statistical inference

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

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

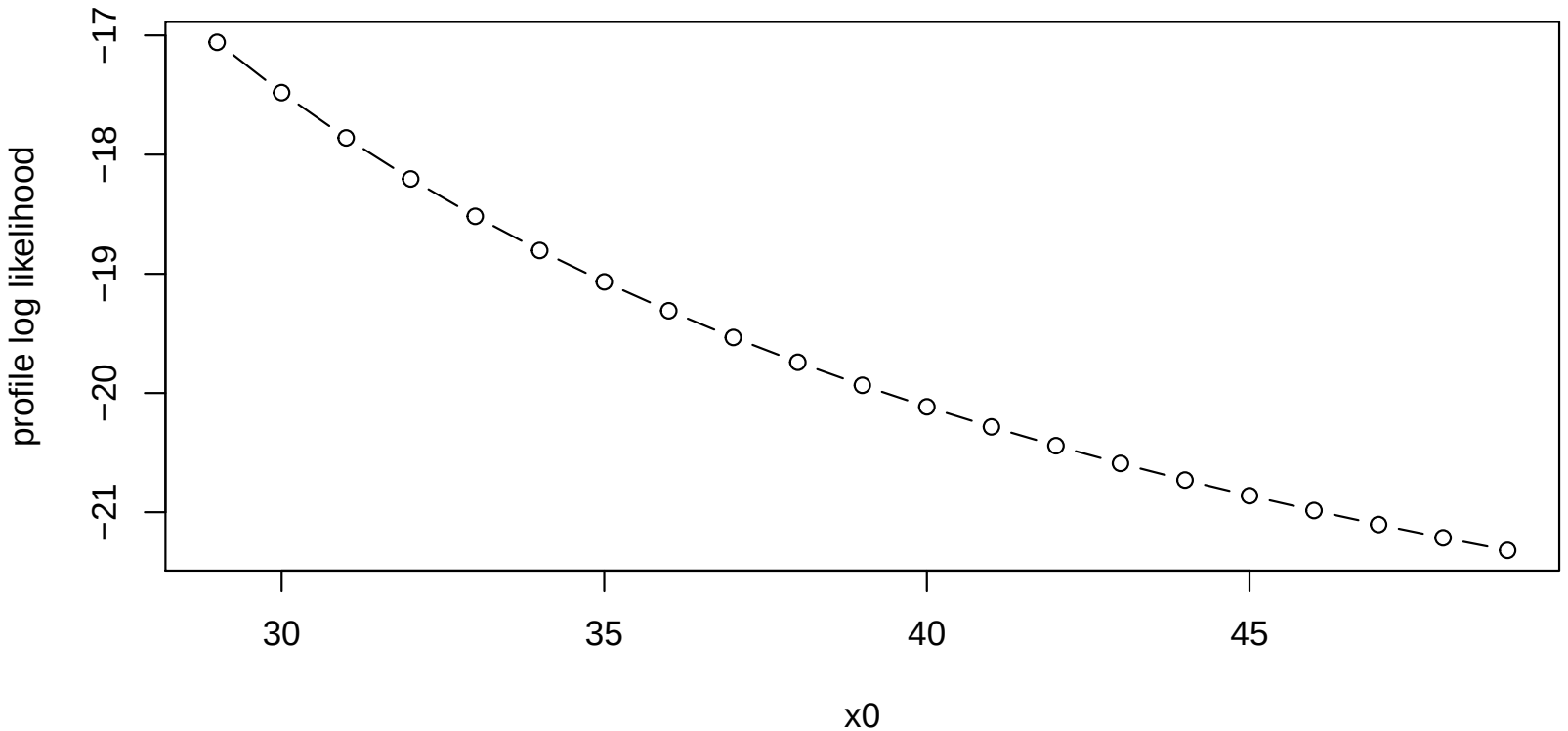
  here $p_t = 1 - (1 - w)^{y_t}$. $\rightarrow$ Knowledge of x_0 is required.
- Uncertainty of $\hat{w}$ can be quantified with a 95% confidence interval.
- Example: Generation sizes of a measles epidemic in St. Petersburg (from Table 4.1 in Daley and Gani, 1999): $\mathbf{y} = (1, 4, 14, 10, 1, 0)$
- 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

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
```

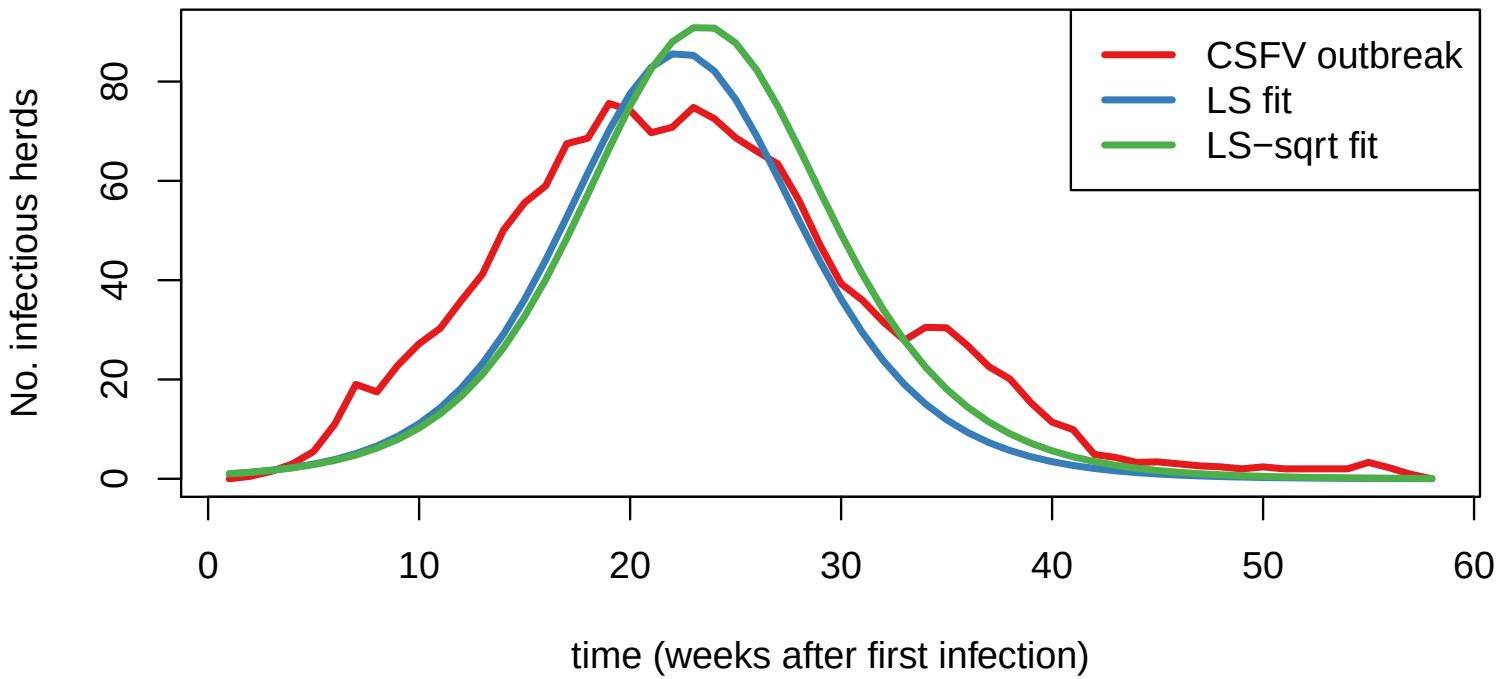
Estimating parameters (3) – MLE for CSFV Data

- 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.

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

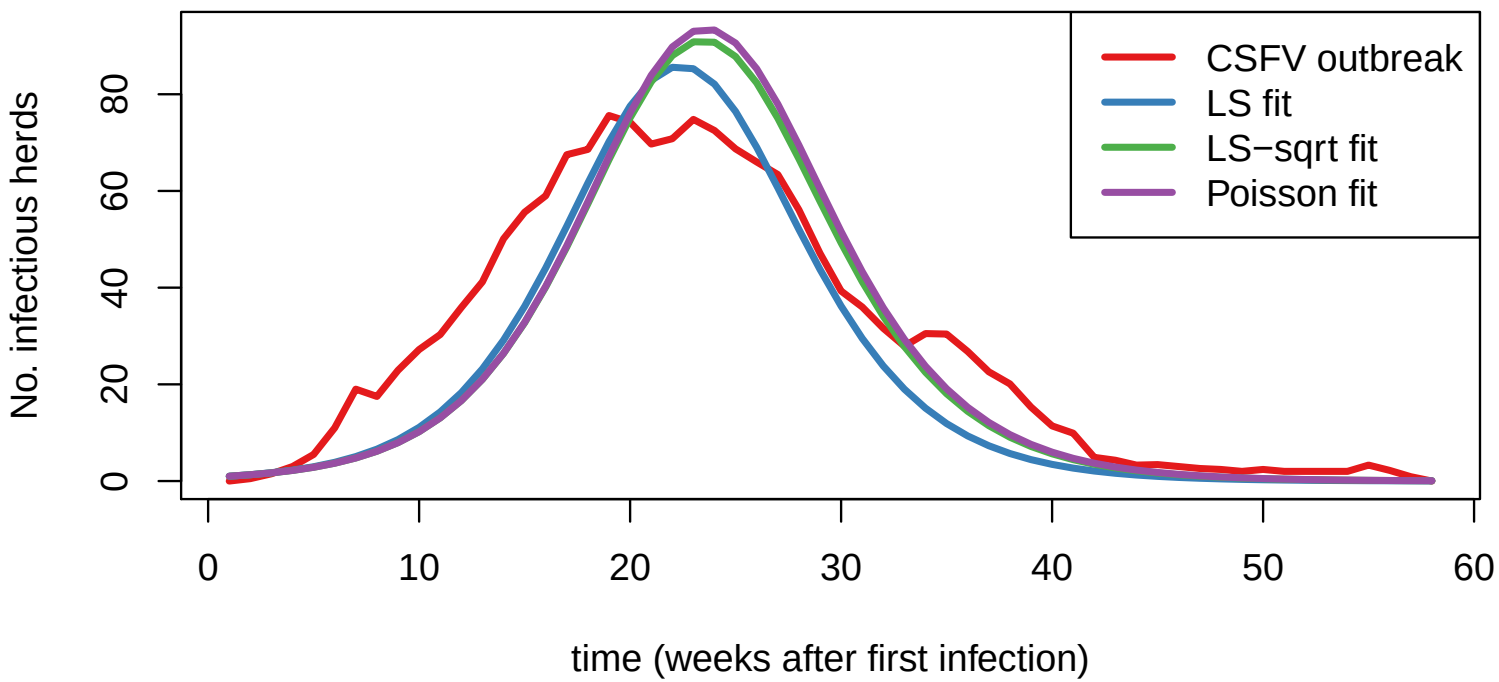
- As a consequence the log-likelihood is

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

- 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 $\hat{R}_0 = 1.10$.

Outline

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)

- The contact times $t_i, i = 1, \ldots, k$, are unlikely to be observed, i.e. the previous likelihood can not be constructed since $S(t)$ is unknown.
- To make inference tractable assume that the duration of the infectious period is a constant μ_y, say, (known or to be estimated). Let $u_1, \ldots, u_k$ denote the individual removal times.
- In this case $t_i + \mu_y = u_i$ and hence

$$S(t^-) = S(0) - \sum_{i=1}^{k} \mathbb{1}_{(u_i - \mu_y, \infty)}(t)$$

- The likelihood is now

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

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 $\rightarrow$ 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


Becker, N. G. 1989. Analysis of Infectious Disease Data. Chapman & Hall/CRC.