Estimation of infectious disease parameters: basics

Ben Bolker

June 2, 2008

1 Preliminaries:

1.1 Data types

- presence/absence
- incidence
- prevalence
- mortality
- seroprevalence
- time to mortality/infection/etc. for observed individuals

Other issues:

- misreporting/underreporting
- single value or by category (age, species, etc.)
- time-series, spatial (reporting frequency, spatial sampling design, etc etc etc.)

1.2 Things to estimate

- R_0 (intrinsic reproductive number)
- r (intrinsic rate of increase)
- λ (force of infection)

- generation time G, infectious period ($1/\gamma$), latent period, incubation period, existence of prodromal period (i.e. infectiousness before systems)

For the simple SIR, $R_0 \approx \exp(rG)$; $R_0 = \beta N/G$; $\lambda = \beta N$

2 Estimators for scalar data

From [1]:

- $R_0 \approx N/S^*$ (fraction susceptible): assumes endemic, equilibrium, homogeneous mixing, etc. e.g. have non-age-specific seroprevalence data (seroprevalence $\approx 1 - S^*/N$)
- $R_0 \approx L/A$ where A is average age at infection, L is average lifespan. Also assumes endemic, equilibrium, etc., "sudden death" mortality schedule

Final size: in principle, the fraction F who remain *uninfected* in an SIR epidemic is

$$F = e^{-R_0(1-F)};$$

for $R_0 > 2$, $F \approx \exp(-R_0)$ [12]. (So $R_0 \approx -\log F$.) Not terribly practical because outcome is highly variable (e.g. [3]), although there's lots of work on the distribution.

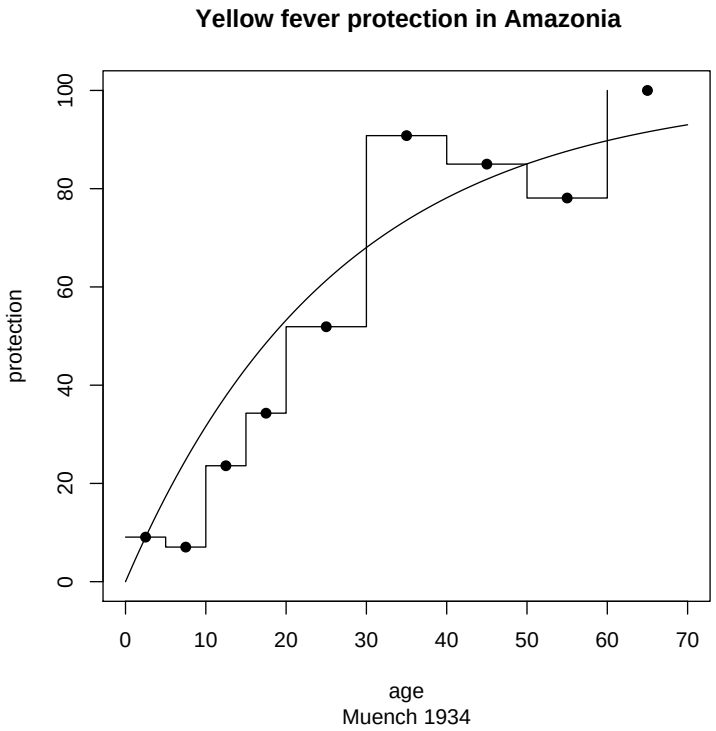
3 Estimators for age-structured data

The *catalytic curve* [10] is a simple exponential model for constant exposure with age. Can also handle reversion (loss of immunity etc.); non-constant *force of infection*; non-constant rates of disease-induced mortality with age [2, 8] — what if force of infection or rate of mortality varies with age?

The basic form of the catalytic curve is $P(a) = 1 - e^{-\lambda a}$, where P is the fraction protected/seropositive (i.e. $P(a) = 1 - S(a)$). We can then say $S = e^{-\lambda a}$ or $\log S = -\lambda a$, which we can fit with a log-linear regression without an intercept (in R, `lm(y~x-1)` fits a regression of y on x with the intercept forced to 0).

```
> x = read.table("yellowfever1.dat",header=TRUE)
> a = (x$age1+x$age2)/2  ## average age of category
> plot(prot~age1,data=x,type="s",  ## "stair-step" plot
```

```
+      xlim=c(0,70),ylim=c(0,100),
+      main="Yellow fever protection in Amazonia",
+      sub="Muench 1934",
+      xlab="age",ylab="protection")
> points(prot~a,data=x,pch=16)     ## midpoints
> unprot = (100-x$prot)/100        ## unprotected
> fit1 = lm(log(unprot)~a-1,data=x,subset=prot<100) ## fit
> curve(100*(1-exp(coef(fit1)*x)),add=TRUE)  ## add the curve
```



`coef(fit1)` and `confint(fit1)` tell us that $\lambda = 0.038$ with confidence intervals $(0.025, 0.051)$.

More complex examples where pathology or detection occurs on reinfection — malaria, filariasis, dengue, etc. [11].

"Who acquires infection from whom" ("WAIFW") matrices – converting from FOI estimates to contact rates for an age-structured contact model [1].

4 Epidemic data

4.1 Initial epidemic curves

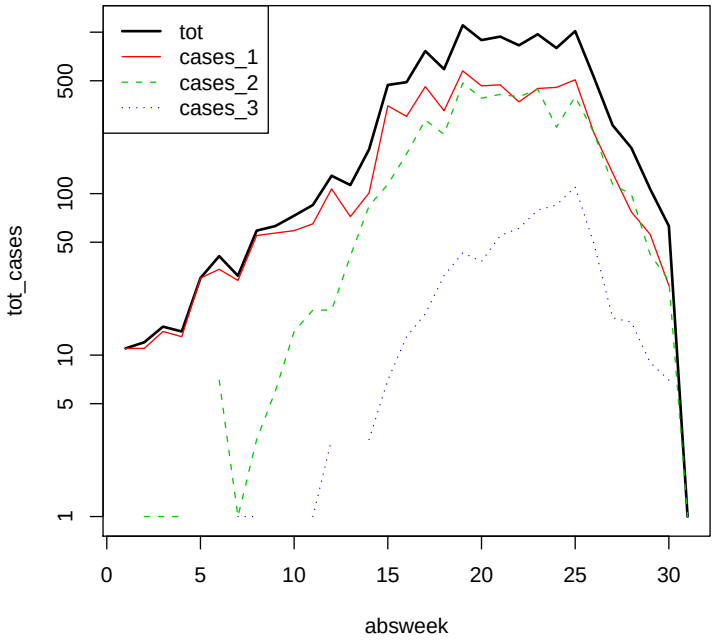
Estimate r from initial epidemic curve — put can be hard to decide when to cut off the series. Could try to fit the entire curve (dynamic SIR model), but epidemics tend to be much simpler at the beginning (heterogeneity less important, pre-intervention, etc etc)

Converting from r to R_0: for the standard SIR (exponential infectious period, no latency), $R_0 = \exp(rG)$. More complex with different latent and infectious periods [13, 14]. Doesn't work for mortality series (although maybe well enough for wildlife disease?)

Plot data:

```
> ndata = read.csv("niamey_weekly.csv")

> plot(tot_cases ~ absweek, data = ndata, log = "y", lwd = 2, type = "l")
> matlines(ndata$absweek, ndata[c("cases_1", "cases_2", "cases_3")],
+     col = 2:4)
> legend("topleft", c("tot", "cases_1", "cases_2", "cases_3"),
+     col = 1:4, lty = c(1, 1:3), lwd = c(2, rep(1, 3)))
```



Now fit linear regressions to carefully chosen subsets:

```
> fit1 = lm(log(tot_cases)~absweek,data=ndata,subset=absweek<=17)
> lmc1 = coef(fit1)
> curve(exp(lmc1[1]+lmc1[2]*x),add=TRUE)
> abline(v=17,lty=2) ## show the cutoff
```

(Normally we could use `abline(fit1)` to add the results of a linear regression to an existing plot, but the log scale on this plot gets R all confused ...) We can get crude confidence intervals on R_0 in this case by looking at `summary(fit1)` to get confidence intervals on the slope and filling in the estimate $\pm 2\sigma$ in the $R_0 = \exp(rG)$ formula.

(Producing the actual figure is left as an exercise. `summary(fit1)` shows that R^2 (no relationship to R_0!) is 0.959, and so the picture looks pretty nice.)

Epidemics in the other two areas (other reporting district in the same city) are later, and appear to start off faster. I'll leave calculating their r and R_0 as an exercise too.

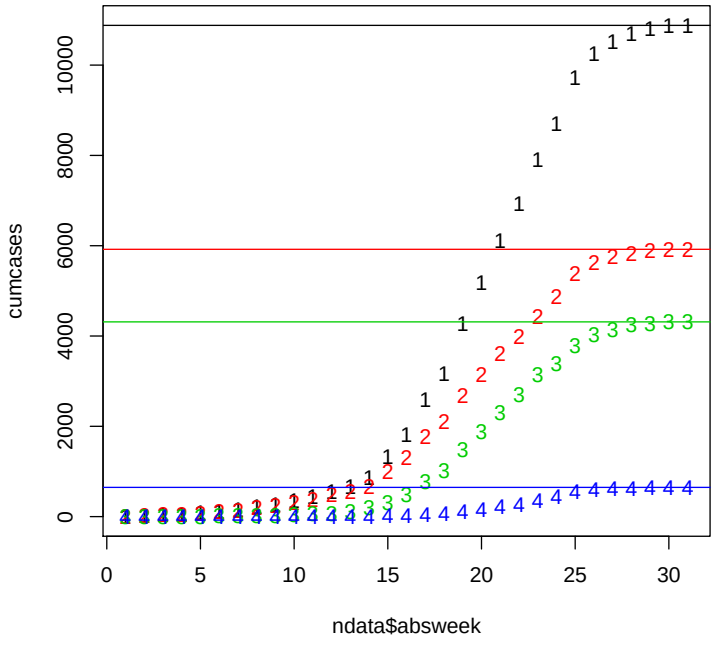
4.2 Complete epidemic curves

The cumulative epidemic curve (i.e. cumulative fraction infected, $C(t) = \sum_{\tau=0}^{t} I(\tau)/N$ for an SIR model is logistic, with "carrying capacity" (asymptote) equal to the epidemic final size, and growth rate r. The quick-and-dirty way to fit this is with a *logit transformation*, $y = \log(C)/(\log(F - C))$ which linearizes a logistic function — if you can eyeball the final size (or just set it to the final value), then you can just do linear regression.

```
> casedata = ndata[c("tot_cases","cases_1","cases_2","cases_3")]
> cumcases = apply(casedata,2,cumsum)  ## cumulative sum of each column
> finalsize = cumcases[nrow(cumcases),] ## last row
> cumpropcases = scale(cumcases,center=FALSE,scale=finalsize)
> logitcases = qlogis(cumpropcases)
```

It does *look* as though these epidemics have reached their final size ...

```
> matplot(ndata$absweek, cumcases)
> abline(h = finalsize, col = 1:4)
```

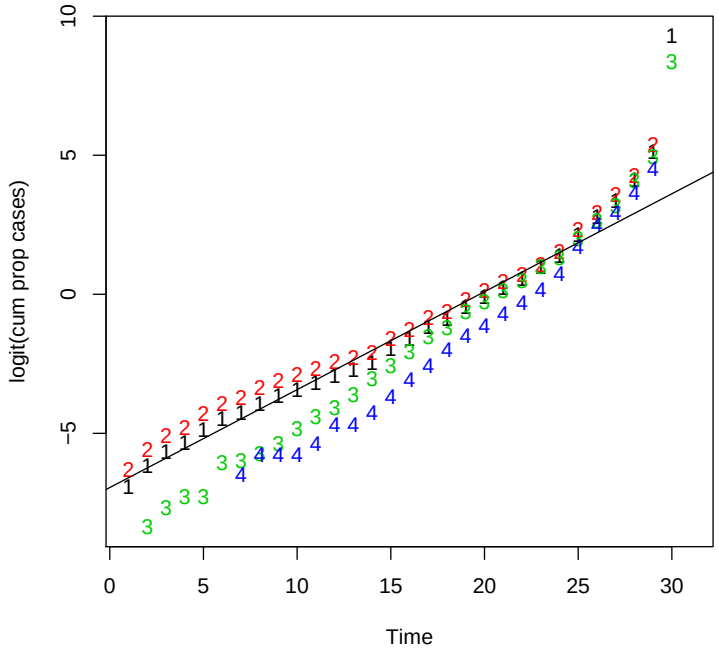


```
> matplot(ndata$absweek, logitcases, xlab = "Time", ylab = "logit(cum prop cases)")
> logitcase2 = data.frame(absweek = ndata$absweek, logitcases)
> (logitfit1 = lm(tot_cases ~ absweek, data = logitcase2, subset = absweek <=
+     27))

Call:
lm(formula = tot_cases ~ absweek, data = logitcase2, subset = absweek <=     27)

Coefficients:
(Intercept)      absweek
    -6.9490       0.3521

> abline(logitfit1)
```



I left off the last few weeks because (a) the very last week is infinite ($\text{logit}(1.0) = \infty$) and (b) the last few points look a bit wonky anyway. The r estimate here (0.352) is at least of the same order of magnitude as the initial-epidemic fit. Feel free to estimate the size of the error, and the values for each area.

Fitting the whole time-series: easiest procedure is assuming continuous-time deterministic ODE, assuming normally distributed measurement error only (this is called *trajectory matching*: e.g. [7]), fitting to the case reporting data. Often assume other parameters are known from other sources (because otherwise estimation may be nearly impossible: but see [4]). Relaxing the assumptions of either all-process or all-measurement error is hairier (ask Aaron King).

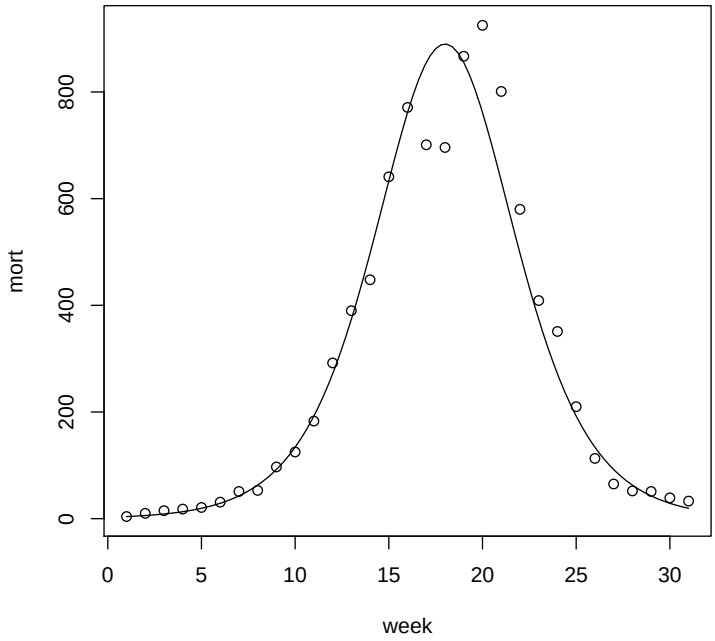
Discrete-time models use the *chain binomial*; may use *susceptible reconstruction* to try to come up with the number of S at time t. We usually say something like $S(t) = S_0 - \sum I(t)$ (that is, every case that occurs is subtracted from the number of susceptibles). Then, roughly speaking, we can say $\beta(t-1) = I(t)/(S(t-1) \cdot I(t-1))$. This assumes a closed population, homogeneous mixing, etc.; one can try to account for births, etc. [5, 6].

Some (not all) of these fits have a likelihood basis, so can be compared between populations/ extended to vary across space, etc. etc. etc. (although not necessarily easily!)

5 Exercise

From [9] (via Steve Ellner, probably (??) from `http://math.arizona.edu/~dsl/bbombay.htm`):

```
> bombay = read.csv("bombayplague.csv")
> plot(mort ~ week, data = bombay)
> curve(890/cosh(0.2 * (x - 1) - 3.4)^2, add = TRUE)
```



Given that the generation time for pneumonic plague is approx. 3 days (and for flea-borne is ≈ 7, but this epidemic is more likely to have been pneumonic), what can you say about these data?

References

[1] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford Science Publications, Oxford, 1991.

[2] P. Caley and J. Hone. Estimating the force of infection; *Mycobacterium bovis* infection in feral ferrets *Mustela furo* in New Zealand. *Journal of Animal Ecology*, 71:44–54, 2002.

[3] J. M. Drake. Limits to forecasting precision for outbreaks of directly transmitted diseases. *PLoS Medicine*, 3, 2006. PMC1288026.

[4] B. D. Elderd, V. M. Dukic, and G. Dwyer. Uncertainty in predictions of disease spread and public health responses to bioterrorism and emerging

diseases. *Proceedings of the National Academy of Sciences*, 103:15693–15697, Oct. 2006.

[5] P. E. M. Fine and J. A. Clarkson. Measles in England and Wales-I: an analysis of factors underlying seasonal patterns. *International Journal of Epidemiology*, 11:5–15, 1982.

[6] B. F. Finkenstädt and B. T. Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 49:187–205, 2000.

[7] R. Gani and S. Leach. Transmission potential of smallpox in contemporary populations. *Nature*, 414:748–751, 2001.

[8] D. M. Heisey, D. O. Joly, and F. Messier. The fitting of general force-of-infection models to wildlife disease prevalence data. *Ecology*, 87:2356, 2006.

[9] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115:700–721, Aug. 1927. ArticleType: primary_article / Full publication date: Aug. 1, 1927 / Copyright 1927 The Royal Society.

[10] H. Muench. Derivation of rates from summation data by the catalytic curve. *Journal of the American Statistical Association*, 29:25–38, Mar. 1934.

[11] A. Srividya, P. K. Das, S. Subramanian, K. D. Ramaiah, B. T. Grenfell, E. Michael, and D. A. P. Bundy. Past exposure and the dynamics of lymphatic filariasis infection in young children. *Epidemiology and Infection*, 117:195–201, Aug. 1996.

[12] J. Swinton. Extinction times and phase transitions for spatially structured closed epidemics. *Bulletin of Mathematical Biology*, 60:215–230, Mar. 1998.

[13] J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings. Biological sciences / The Royal Society*, 274:599–604, Feb. 2007. PMID: 17476782.

[14] J. Wallinga and P. Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160:509–16, Sept. 2004. PMID: 15353409.