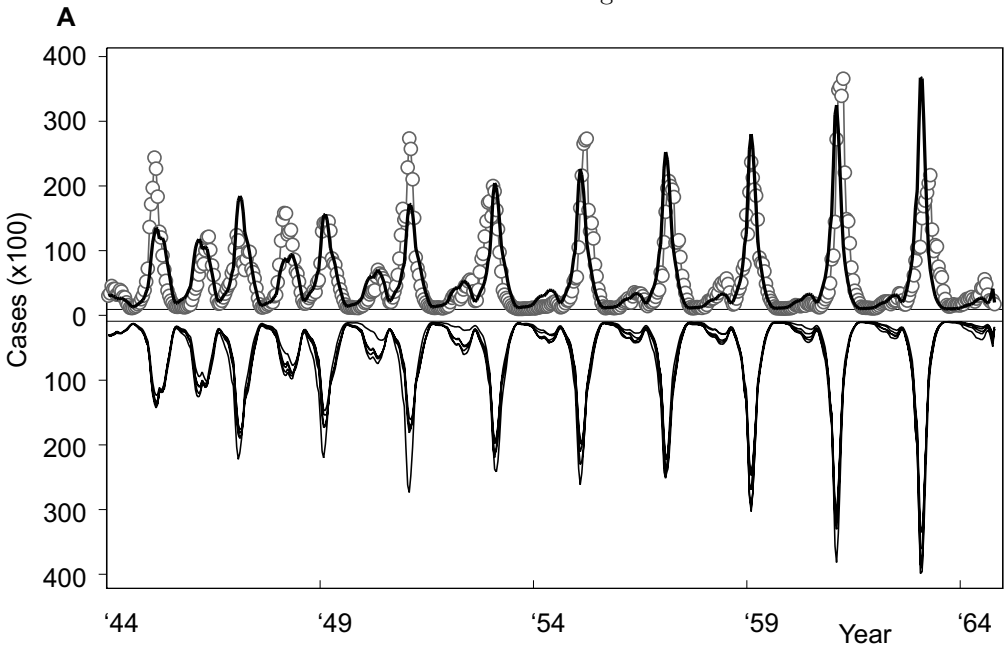
Measles

Ottar Bjørnstad

May 30, 2005

1 Preamle

In 2002 (Ecological monographs), Grenfell Finkenstadt and myself fit a mechanistic (discrete time) S-I-R-like model to measles incidence data to obtain the following fit:



The circles are the data, the top line is the determinsitic forecast (slightly tweaking the unknown initial coditions), and the bottom (mirrored lines) represents forecast from the stochastic (chain binomial-like) model.

We were quite chuffed with the ability of the model to forecast the dynamics for essentially 500 pathogen generations (20+year) from Jan 1, 1944 initial conditions. If anybody is interested, this handout details how we estimated the parameters.

2 Measles

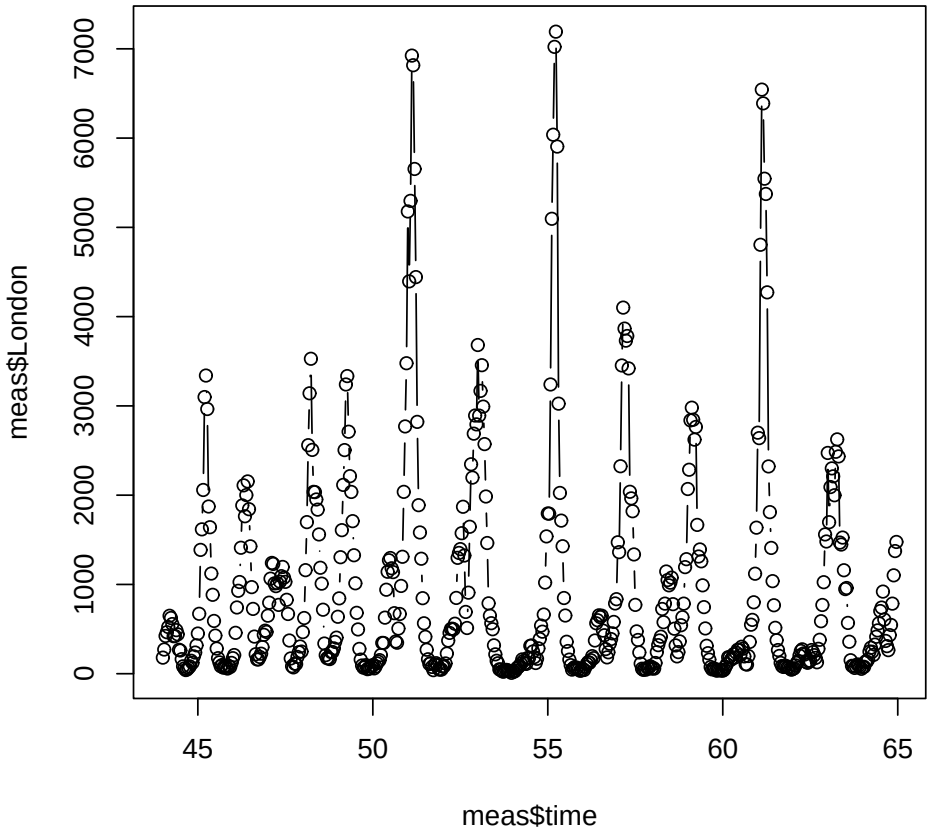
The biweekly incidence (number of cases for each two-week period) of measles has a long history in the study of infectious disease dynamics. The data set `meas.csv` contains the records from London between 1944 and 1966:

```
> meas = read.table("meas.csv", sep = ",", header = TRUE)
> names(meas)

[1] "year"   "week"   "time"   "London" "B"
```

The incidence are accessed as `meas$London`. In addition, the data set contain columns reporting `meas$year`, `meas$month`, the two combined into `meas$time`, the incidence (`meas$London`), and biweekly number of births (`meas$B`).

```
> plot(meas$time, meas$London, type = "b")
```



Ideally we should be able to use this data to estimate key epidemiological parameters ... However, we rarely have detailed information on all state variables (time series of susceptibles AND infecteds). Some simple tricks have recently been suggested. We will illustrate this using estimation with the so-called TSIR (time-series S-I-R) model. The model (approximation) is as follows: If we use a discrete time step equal to the generation time of the pathogen (2 weeks in the case of measles). We can then write (very crudely) the model (ignoring a latent period) as:

$$S_{t+1} = S_t + B_t - I_t, \lambda_{t+1} = \beta S_t I_t^{\alpha}, \tag{1}$$

where S_t and I_t are the numbers of susceptibles and infecteds in (pathogen) generation time, t. B_t is the number of births (into the susceptible class) in the time step, β – of course – is the transmission rate, and the mysterious α is a fudge exponent that partially accounts for discretizing the underlying contiuous process (Glass et al.) and partially for certain heterogenietes in mixing (Liu et al.).

The final variable, λ_{t+1} represents the *expectation* for the new number of infecteds in the next generation. Evidently, the actual number of infecteds that will appear in generation $t+1$ will follow some stochastic distribution around λ_{t+1}. For example I_{t+1} $Po(\lambda_{t+1})$ (Miramontes and Rohani 1998) or I_{t+1} $NegBin(\lambda, I_t)$.

2.1 inference (hypothetical)

Given time series I and S the candidate for estimation is obvious:

$$log(I_{t+1}) = log(\beta) + log(S_t) + \alpha log(I_t). \tag{2}$$

We can estimate the unknown parameters β and α by a regression of $log(I_{t+1})$ on $log(I_t)$ with $log(S_t)$ as an *offset* (that means the slope for variable is fixed at unity). The intercept of this regression would be the estimate of $log(\beta)$ and the slope against $log(I_t)$ would be the estimate α.

In R:

```
N = length(meas$London)

Inow = log(meas$London[2:N])
Ilag = log(meas$London[1:(N-1)])

Slag = log(S[1:(N-1)]) #NB! This variable does not exist in the data

#now the regression
glm(Inow ~ Ilag + offset(Slag))
```

3 inference (the real example)

The challenge is that most real data sets do not contain perfect records on all state variables. For example, the `meas` data does not contain information on S, and I is under-reported. However, we do have information about births. Another challenge is the strong seasonality in transmission rates that result from aggregation of children during school term.

We clearly need some more elaborate (*ad hoc?*) scheme – In this case tailored to the biology and data on measles....

3.1 susceptible reconstruction

The idea of susceptible reconstruction was layed out Bobashev et al. and Finkenstadt and Grenfell; Consider the reccursive equation 1 which can be rewritten as:

$$S_t = \overline{S} + D_0 + \sum_{k=0}^{t} B_k - \sum_{k=0}^{t} I_k/\rho, \tag{3}$$

where $\overline{S}$ is the mean number of susceptibles, D_0 is the unknown deviations around the mean at time 0, and ρ is the (known or unknown) reporting rate. We can reconstruct the time series D_t of how the susceptible numbers deviate from the mean value, $D_t = S_t - \overline{S}$, by rewriting (3) as,

$$\sum_{k=0}^{t} B_k = \overline{S} + D_0 + 1/\rho \sum_{k=0}^{t} I_k + D_t, \tag{4}$$

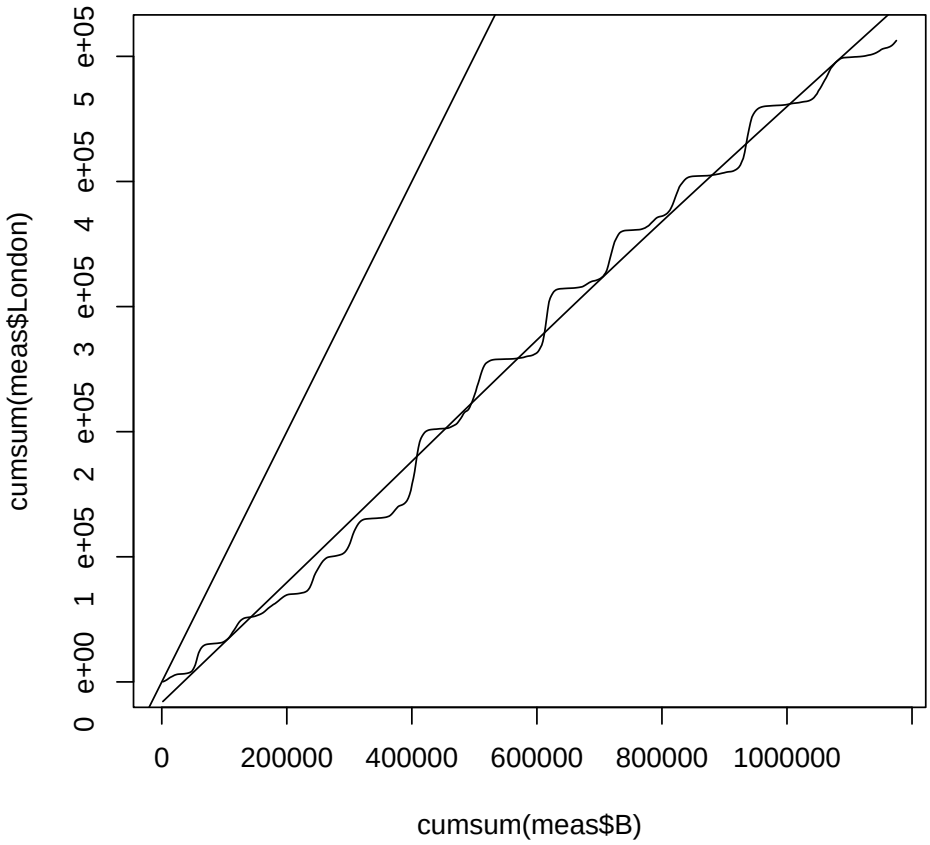
from which it is clear that D_t is the residual from the regression of cumulative number of births on the cumulative number of cases. Note, that this reconstruction still works when D_0, $\overline{S}$ and the reporting rate ρ is unknown because these are accommodated by the intercept and slope of the cumulative-cumulative regression. The method does not allow the independent estimation of the mean number of susceptibles.

As it turns out, reporting rates sometimes varies subtly through time so it is good to use a slightly more flexible regression than linear regression – a smoothing spline (with 2.5 degrees-of-freedom) for example.

In R:

```
> cum.reg = smooth.spline(cumsum(meas$B), cumsum(meas$London),
+     df = 2.5)
```

```
> D = predict(cum.reg)$y - cumsum(meas$London)
> plot(cumsum(meas$B), cumsum(meas$London), type = "l")
> lines(cum.reg)
> abline(a = 0, b = 1)
```



The 1-to-1 line generated by the `abline`-command shows that the cumulative number of cases are less than the cumulative number of births. However, this discrepancy is very informative for measles because we know that more than 95% of children were infected with measles before the age of 20 (in the pre-vaccination era); The slope of the cumulative regression, therefore, is an estimate of the under-reporting rate in this system. We can get these estimated under-reporting rates for each time step. from the cumulative regression as follows:

```
> ur = predict(cum.reg, deriv = 1)$y
> summary(ur)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4525  0.4620  0.4792  0.4734  0.4833  0.4866
```

The reporting rate is clearly almost constant across the 20-years at around 52%. As a simple hack, we can correct the time series data for the under-reporting:

```
> Ic = meas$London/ur
```

3.2 estimation

To estimate parameters we rewrite the model (1) in terms of the data and unknown parameters on a log-scale as (recall that λ_{t+1} is the expected number of cases in time $t+1$):

$$log(\lambda_{t+1}) = log(\beta_u) + log(D_t + \overline{S}) + \alpha log(I_t)$$

which obviously is an almost (but not quite) a linear regression with unknown parameters β_u, α and $\overline{S}$. Before we are ready to estimate the parameters, however, we need to consider the fact that β varies seasonally (because of the school year); thus the subset u. The most flexible model is to assume the each of the 26 biweeks of the year has its own transmisson rate. Under that assumption we have 28 parameters to estimate. Let us define a vector that flags these across the 21 years, and create the three vectors of current and lagged infecteds and lagged 'residual susceptibles':

```
> seas = rep(1:26, 21)[1:545]
> lInew = log(Ic[2:546])
> lIold = log(Ic[1:545])
> Dold = D[1:545]
```

A simple trick is to realize that given a vaule for $\overline{S}$, the models falls neatly within the linear regression framework (thought had it not, we can alwasy write out the likelihood, and use `optim` to find the MLEs). We can therefore use `glm` to find a profile likelihood estimate of $\overline{S}$. We know from serology that the average proportion of susceptibles in measles is somewhere in the 5%-10% range and – given the size of London at the time (3.3M) – we can postulate a reasonable range of candidate values:

```
> Smean = seq(0.01, 0.2, by = 0.001) * 3300000
```

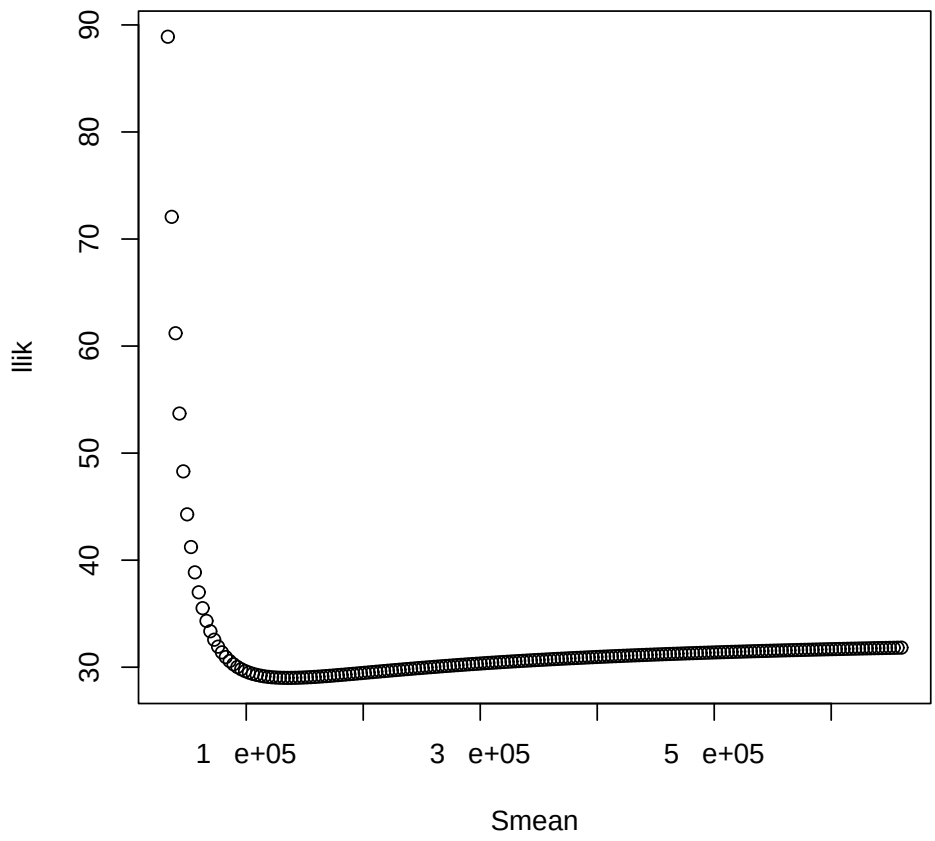
We then set up a vector to store the log-likelihood values corresponding to each candidate:

```
> llik = rep(NA, length(Smean))
```

We then loop over all the values. Note the `-1` in the regression formula removes the intercept, so that `as.factor(seas)` becomes the estimates of the log-beta's. Note further that `glmfit$deviance` holds -2*log likelihood.

```
> for (i in 1:length(Smean)) {
+     lSold = log(Smean[i] + Dold)
+     glmfit = glm(lInew ~ -1 + as.factor(seas) + lIold + offset(lSold))
+     llik[i] = glmfit$deviance
+ }
> plot(Smean, llik)
> Smean[which(llik == min(llik))]

[1] 135300
```



Our best estimates then is:

```
> lSold = log(Smean[which(llik == min(llik))] + Dold)
> glmfit = glm(lInew ~ -1 + as.factor(seas) + lIold + offset(lSold))
> summary(glmfit)

Call:
glm(formula = lInew ~ -1 + as.factor(seas) + lIold + offset(lSold))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.060377  -0.142732  -0.003244   0.136641   0.756131

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
as.factor(seas)1  -11.574672   0.076903  -150.5   <2e-16 ***
as.factor(seas)2  -11.131830   0.077004  -144.6   <2e-16 ***
as.factor(seas)3  -11.266591   0.079988  -140.9   <2e-16 ***
as.factor(seas)4  -11.387843   0.082013  -138.9   <2e-16 ***
as.factor(seas)5  -11.458274   0.083131  -137.8   <2e-16 ***
as.factor(seas)6  -11.452393   0.083687  -136.8   <2e-16 ***
as.factor(seas)7  -11.524541   0.084211  -136.9   <2e-16 ***
as.factor(seas)8  -11.733467   0.084169  -139.4   <2e-16 ***
```

```
as.factor(seas)9  -11.536306   0.082668  -139.6   <2e-16 ***
as.factor(seas)10 -11.495535   0.082542  -139.3   <2e-16 ***
as.factor(seas)11 -11.537266   0.082690  -139.5   <2e-16 ***
as.factor(seas)12 -11.691471   0.082547  -141.6   <2e-16 ***
as.factor(seas)13 -11.670301   0.081375  -143.4   <2e-16 ***
as.factor(seas)14 -11.702528   0.080404  -145.5   <2e-16 ***
as.factor(seas)15 -11.801345   0.079279  -148.9   <2e-16 ***
as.factor(seas)16 -11.986965   0.077588  -154.5   <2e-16 ***
as.factor(seas)17 -12.074714   0.074840  -161.3   <2e-16 ***
as.factor(seas)18 -11.910984   0.071779  -165.9   <2e-16 ***
as.factor(seas)19 -11.526471   0.069937  -164.8   <2e-16 ***
as.factor(seas)20 -11.340105   0.070469  -160.9   <2e-16 ***
as.factor(seas)21 -11.428370   0.072138  -158.4   <2e-16 ***
as.factor(seas)22 -11.565334   0.073278  -157.8   <2e-16 ***
as.factor(seas)23 -11.468243   0.073575  -155.9   <2e-16 ***
as.factor(seas)24 -11.504891   0.074493  -154.4   <2e-16 ***
as.factor(seas)25 -11.608515   0.075180  -154.4   <2e-16 ***
as.factor(seas)26 -11.222460   0.075546  -148.6   <2e-16 ***
lIold               0.963917   0.008642   111.5   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for gaussian family taken to be 0.05598971)

    Null deviance: 14858.805  on 545  degrees of freedom
Residual deviance:    29.003  on 518  degrees of freedom
AIC: 3.9412

Number of Fisher Scoring iterations: 2
```

That is, $\overline{S}$ is

```
> Smean[which(llik == min(llik))]

[1] 135300
```

α is

```
> glmfit$coef[27]

    lIold
0.9639174
```

and the $log - \beta's$ are

```
> glmfit$coef[1:26]

 as.factor(seas)1  as.factor(seas)2  as.factor(seas)3  as.factor(seas)4
        -11.57467         -11.13183         -11.26659         -11.38784
 as.factor(seas)5  as.factor(seas)6  as.factor(seas)7  as.factor(seas)8
        -11.45827         -11.45239         -11.52454         -11.73347
 as.factor(seas)9 as.factor(seas)10 as.factor(seas)11 as.factor(seas)12
        -11.53631         -11.49554         -11.53727         -11.69147
as.factor(seas)13 as.factor(seas)14 as.factor(seas)15 as.factor(seas)16
        -11.67030         -11.70253         -11.80135         -11.98696
as.factor(seas)17 as.factor(seas)18 as.factor(seas)19 as.factor(seas)20
```

```
        -12.07471         -11.91098         -11.52647         -11.34011
as.factor(seas)21 as.factor(seas)22 as.factor(seas)23 as.factor(seas)24
        -11.42837         -11.56533         -11.46824         -11.50489
as.factor(seas)25 as.factor(seas)26
        -11.60851         -11.22246
```