Description of the likelihood function used and the corresponding fits

Petr Kouba

December 17, 2018

This document summarizes the final version of the fitting to the survival data.

1 Likelihood function

For our fits, we used the Maximum Likelihood Estimate. First we derive the likelihood function for our case. For this summary, we have closely followed [1].

Each time of death t_i , we have in our data, contributes to the likelihood with the following factor:

$$L_i = f(t_i) \tag{1}$$

Where L_i is the contribution of that particular survival time to the likelihood and $f(t_i)$ is the probability density of death in time t_i (in our discrete case, this is in fact not just probability density, but a probability death occured in the t_i^{th} time step).

Having these contribution to the likelihood, we can express the overall likelihood as follows:

$$L = \prod_i L_i \tag{2}$$

So we are only left with finding the expression for $f(t_i)$. The probability density function $f(t_i)$ corresponds to the cummulative distribution function:

$$F(t) = P\{T \leq t\} = \int_0^t f(T)dT \tag{3}$$

Where T is a non-negative random variable representing the waiting time till death occurs.

The above cummulative distribution function is by definition complementary to the survival function $S(t)$, which describes the probability od being alive at time t. We can therefore write:

$$S(t) = 1 - F(t) = P\{T \geq t\} = \int_t^{\inf} f(T)dT \tag{4}$$

The survival function can be written as:

$$S(t) = \exp(-\int_0^t \delta(T)dT) \tag{5}$$

Where $\delta(T)$ is the *hazard function*, that is the instantaneous rate of occurence of death at time T.

By differentiating equation 4 with respect to t, we obtain:

$$\frac{dS(t)}{dt} = -f(t) \tag{6}$$

For $S(t)$ we plug in the expression from equation 5, and obtain the realation between the probability density of death and the hazard function:

$$f(t) = \delta(t) \cdot \exp(-\int_0^t \delta(T)dT) = \delta(t) \cdot S(t) \tag{7}$$

Now we can express the overall likelihood:

$$L = \prod_i L_i = \prod_i f(t_i) = \prod_i \delta(t_i) \cdot S(t_i) \tag{8}$$

And taking the logarithm:

$$\log L = \sum_i [\delta(t_i) + \log S(t_i)] = \sum_i [\log \delta(t_i) - \int_0^{t_i} \delta(T) dT] \tag{9}$$

2 Fitting the survival population

We used the likelihood function derived in chapter 1 and the survival data on Daphnia, to obtain fits for the hazard functions.

2.1 Estimation of natural death rate

Since before we have already compared the exposed and the control population and have not noticed strong differences (Kolmogorov - Smirnov test yielded a difference between the distributions for age at infection = 5, but at that group we only had little sample of exposed and the test is too harsh anyway). We have therefore defined a compartment of uninfected, which includes both control population as well as the population of individuals who got exposed, but did not get infected. We used this largest possible pool of uninfected individuals, to estimate the natural death rate. We tried Ansatzes from 1st order to 4th order polynomial for our hazard function and we obtained the plot in Figure 1

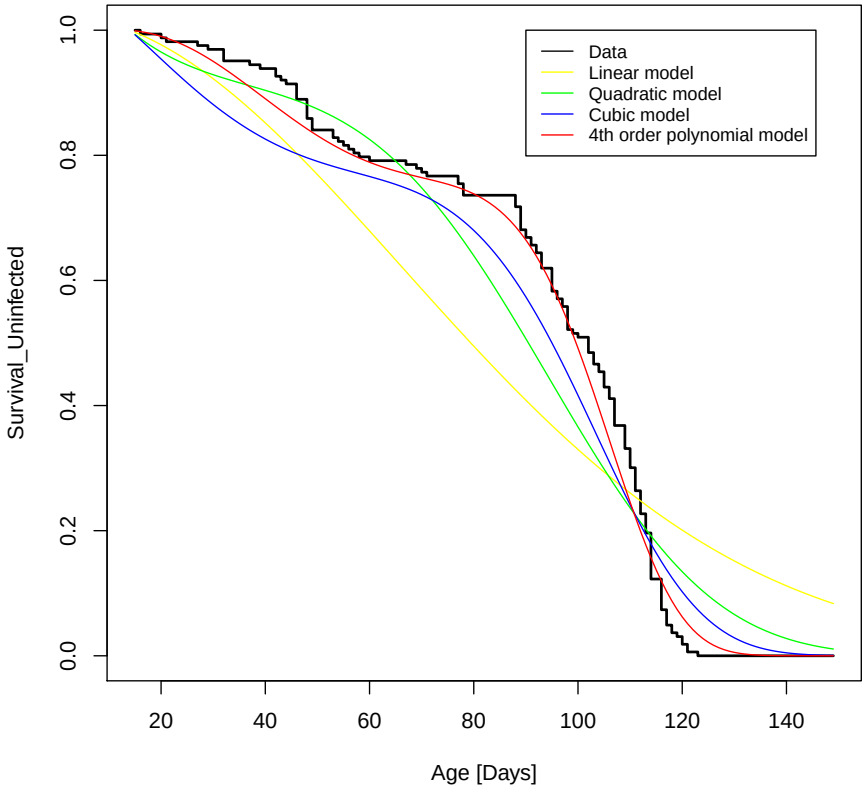


Figure 1: The survival curve of all of the uninfected individuals, with our polynomial fits, fitted since age = 16, since there were no deaths before

Using the likelihood ratio test, we have confirmed that the 4th order polynomial is the closest fit from those we have tried. We therefore now have a good estimate of the hazard function corresponding to the natural death rate.

In order to have the reference natural moratality rate for each of the age-at-infection compartment, we have fitted the uninfected survival two more times (age at infection = 15 overlapped, with the fit we made in Figure 1), each time we fitted starting from the age at infection of the desired compartment. The rates obtained like this should not however differ much from the deathrate described as obtained from Figure 1, we compare them in the Figure 2

Even in these other cases, we concluded (using likelihood ratio test), that the best fit is the one assuming quartic polynomial for the hazard function.s

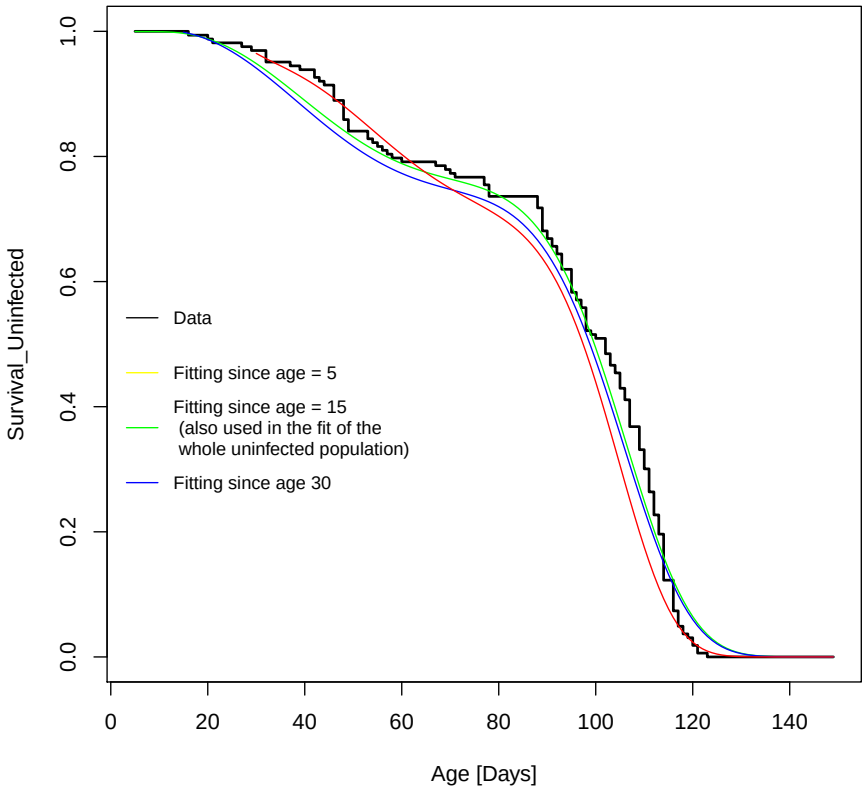
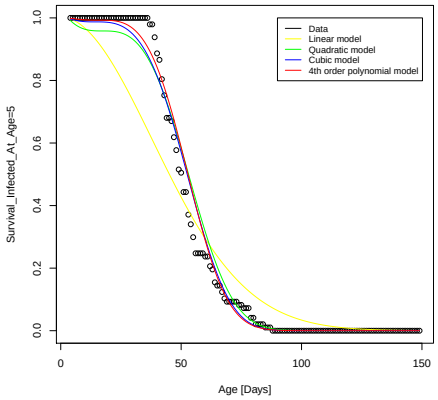


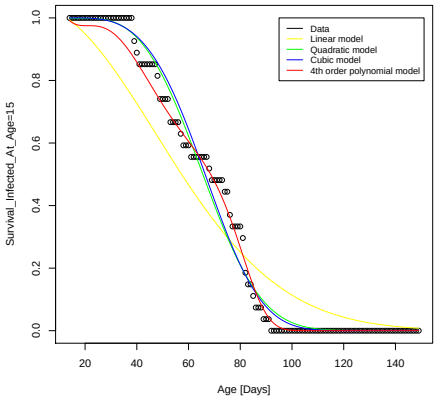
Figure 2: Comparing the fitting of the survival of the uninfected population, using different starting points

2.2 Estimation of death rates in the infected population

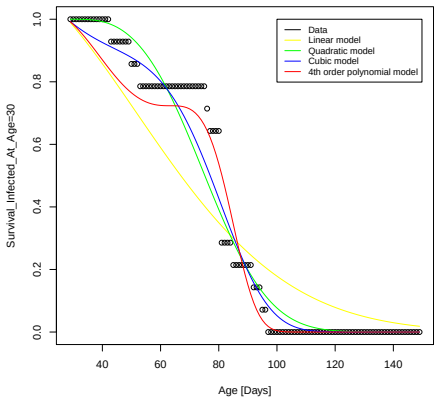
We then proceeded with estimating the death rate in populations with different age at infection. We obtained the plots in Figure 3



(a) Fitting the survival curve for age at infection = 5



(b) Fitting the survival curve for age at infection = 15



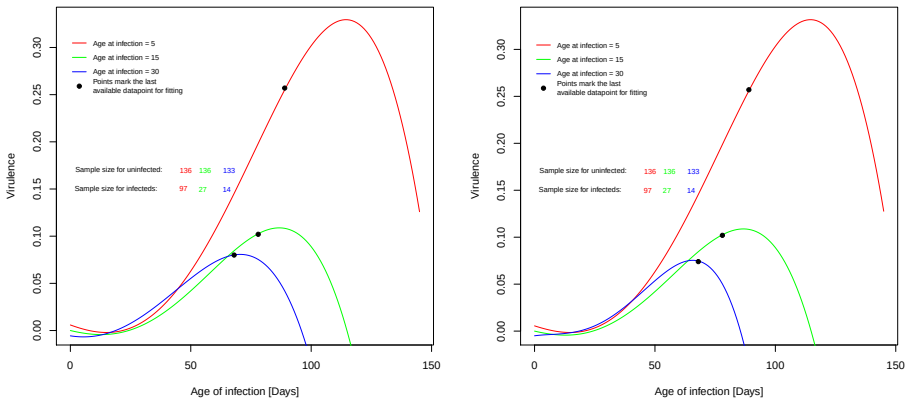
(c) Fitting the survival curve for age at infection = 30

Figure 3: Survival curves in infected populations with different ages at infection, fitted everytime since the respective age at infection

Using the likelihood ratio test, we have concluded that the quadratic fit was the best one in each of the cases.

3 Virulence

Using the hazard functions obtained in Section 2.1, we then constructed the virulence function for each age at infection. We constructed two versions, one using the identical natural death rate and one reflecting the different starting age for the fitting. We provide the plots corresponding to both of these versions in the Figure 4



(a) Virulence computed using the same death rate in all age-at-infection compartments

(b) Virulence computed using the death rate fitted in each case since the respective age at infection

Figure 4: Two versions of virulence

We can see, that the virulence for the age at infection $= 5$ seems much higher then for the other ages, we would like to confirm this statistically now.

TODO: Test the hypothesis of the age-at-infection effect being significant
Question: What kind of test to use?

References

[1] German Rodriguez. Generalized linear models. URL `https://data.princeton.edu/wws509/notes/c7.pdf`.