10 Parameters from Data

How do we learn the parameters of our models from data? This week, we'll explore strategies by which a model's parameters can be learned. Indeed, the fundamental difference between simple models for influenza, SARS-CoV-2, and measles simply comes down to our choice of parameters!

10.1 Estimating R_0 from epidemic growth data

The first method we will examine to estimate R_0 is difficult but perhaps the most intuitive. We have a notion that R_0 tells us something about the growth rate of an epidemic, and thus, if we have epidemic growth data available, we can utilize it to estimate R_0.

Recall that under the SIR model with natural death rate μ, we have

$$\dot{I} = \beta \frac{SI}{N} - (\gamma + \mu)I$$

In the early stages of the epidemic, when $S \approx N$, we get

$$\dot{I} \approx (\beta - (\gamma + \mu))I \qquad \rightarrow \qquad I(t) \approx I(0)e^{(\beta-(\gamma+\mu))t} = I(0)e^{(R_0-1)(\gamma+\mu)t} \, .$$

Thus, we expect exponential growth at rate $(R_0 - 1)(\gamma + \mu)$. As a consequence, were we to plot $\log[I(t)]$ we should get a straight line with slope $m = (R_0 - 1)(\gamma + \mu)$. Thus, if we can estimate that slope $\hat{m}$, we have

$$R_0 \approx 1 + \frac{\hat{m}}{\gamma + \mu} \, .$$

Note that to use this approximation, we need to have high accuracy observations of unhindered epidemic growth for a substantial amount of time—enough to estimate $\hat{m}$. We also need to have good data on μ and γ, or inversely, on the typical life expectancy and the typical duration of infection.

10.2 Estimating R_0 from equilibrium prevalence

If we cannot observe epidemic growth (Method 1), perhaps we can instead observe the equilibrium number of infections in the population. Pulling Eq. (10) from the Week 3 Lecture Notes,

$$i_{\text{eq}} = \left(1 - \frac{1}{R_0}\right) \frac{\mu}{\gamma + \mu} \, ,$$

which would provide an alternative estimate, by rearranging to find that

$$R_0 = \frac{1}{1 - i_{\text{eq}} \left(\frac{\gamma}{\mu} + 1\right)}$$

Table 2. The intrinsic reproductive rate, R_0, and average age of acquisition, A, for various infections [condensed from (*25*); see also (*36*)]. Abbreviations: r, rural; u, conurbation.

Disease	Average age at infection, A (years)	Geographical location	Type of community	Time period	Assumed life expectancy (years)	R_0
Measles	4.4 to 5.6	England and Wales	r and u	1944 to 1979	70	13.7 to 18.0
	5.3	Various localities in North America	r and u	1912 to 1928	60	12.5
Whooping cough	4.1 to 4.9	England and Wales	r and u	1944 to 1978	70	14.3 to 17.1
	4.9	Maryland	u	1908 to 1917	60	12.2
Chicken pox	6.7	Maryland	u	1913 to 1917	60	9.0
	7.1	Massachusetts	r and u	1918 to 1921	60	8.5
Diphtheria	9.1	Pennsylvania	u	1910 to 1916	60	6.6
	11.0	Virginia and New York	r and u	1934 to 1947	70	6.4
Scarlet fever	8.0	Maryland	u	1908 to 1917	60	7.5
	10.8	Kansas	r	1918 to 1921	60	5.5
Mumps	9.9	Baltimore, Maryland	u	1943	70	7.1
	13.9	Various localities in North America	r and u	1912 to 1916	60	4.3
Rubella	10.5	West Germany	r and u	1972	70	6.7
	11.6	England and Wales	r and u	1979	70	6.0
Poliomyelitis	11.2	Netherlands	r and u	1960	70	6.2
	11.9	United States	r and u	1955	70	5.9

Figure 1: Reproduction of Table 2 from Directly Transmitted Infectious Diseases: Control by Vaccination.

Note that this estimate, too, relies on knowing three other values from the data: i_{eq}, which may be rather small and γ/μ, which may be rather large[1] Consequently, this method may also be difficult to use.

10.3 Estimating R_0 from age of first infection

A third method is to conceive of an entirely different source of data: the age at which individuals in the population are first infected. The logic behind this method is that older individuals, who have been around longer, are more likely to have been infected, while newborns are by definition susceptible. Thus there must be some typical or average age at which infections occur. Such information could, in principle, come from public health records on childhood diseases.

How quickly do people leave the susceptible compartment after being born, due to infection? The answer is the *force of infection* βi. Thus the typical time spent susceptible is $\frac{1}{\beta i}$. When the system is at equilibrium, $i = i_{\text{eq}}$, and thus, the typical age at infection a is given by

$$a = \frac{1}{\beta\left(1 - \frac{1}{R_0}\right)\frac{\mu}{\gamma+\mu}} = \frac{1}{\mu(R_0 - 1)}$$

[1] Another way to think about $\frac{\gamma}{\mu}$ is that it is the ratio of life expectancy to infection duration. Thus, for long lives and short infections, this ratio may be large.

and so

$$R_0 = 1 + \frac{\frac{1}{\mu}}{a} = 1 + \frac{L}{a} \, .$$

In other words, if life expectancy is $L = \mu^{-1}$, then R_0 is simply the ratio of life expectancy to age of first infection plus one.

Does this sound crazy? It's not! In fact, Anderson and May provided estimates of R_0 from data using this method in 1982 (Fig. 1). Similarly, but from an angle of vaccination, a study from Panagiotopoulos, Antoniadou, and Valassi-Adam [2] of Greek rubella cases and vaccination over the second half of the 20th century showed that imperfect rubella vaccine coverage had the effect of pushing up the typical age at first infection. This led to an increase in infections during pregnancies, which can result in miscarriage (more likely if the infection occurs early during pregnancy) and congenital rubella syndrome (in which newborns suffer birth defects including cataracts, deafness, heart problems, and neurological problems, or death). In other words, by decreasing R_0 via vaccination, the typical age of first infection was increased, but R_0 was not sufficiently decreased to eliminate infections altogether.

10.4 Estimating R_0 from seroprevalence

Yet another approach to estimating R_0 is to use serology. The roots of the word **serology** indicate that it is the study of the serum, the part of the blood that contains antibodies. Thus, when people refer to serological data, they typically mean data regarding the prevalence of antibodies in the population. In fact, **serosurveys** are studies that try to investigate whose antibodies indicate a prior infection with a particular pathogen, and **seroprevalence** refers to the prevalence of said antibodies.

If seroprevalence tells us the fraction of the population that *does* have antibodies, then one minus seroprevalence tells us the fraction of the population that is susceptible. Once more, noting that at steady state we expect $s_{\text{eq}} = \frac{1}{R_0}$, a model-based estimate of R_0 would suggest that

$$R_0 = \frac{1}{1 - \text{seroprevalence}} \, .$$

Here, again, data quality may be difficult to work around. First of all, **seroconversion** refers to the process by which an individual goes from being seronegative to seropositive, but **seroreversion** is its opposite. The fact that we have this vocabulary should be a warning that not everyone lacking antibodies is necessarily in the never-infected category. Second, as we will explore in the coming weeks, serology studies require calibration and thus will typically have an imperfect sensitivity and specificity.

[2] https://www.bmj.com/content/319/7223/1462

10.5 Estimating R_0 using the generation interval distribution

From time to time, a method is developed that becomes associated with the names of its authors. So it is with the method of Wallinga and Teunis, who advanced a new idea after SARS which was published in 2004.

One of the critiques of the methods above is that they all rely on assumptions about models, life expectancy, or other interactions between measurable parameters and model structure. A way around this, typically, would be to take a likelihood-based approach, in which we compute the "who-infected-whom" likelihood over all possible transmission trees, given measurements of cases over time. This approach, of course, has issues, including the fact that, among N cases there are [[FILE NOT FOUND]] unique trees.[3] This is why directly observed transmission trees are powerful, relative to simply counting cases over time.

What Wallinga and Teunis observed was that, for any pair of infections i and j, one can write down the relative probability that it was j who infected i, instead of one of the other k possible infections in our time series. Intuitively, for example, this same logic is what was implicitly used in the CU Boulder study of roommates: if Dan Pemic was symptomatic on 1 October, and his roommate Flynn Uenza was symptomatic on 30 November, we deem it highly unlikely that Dan infected Flynn. On the other hand, if their other roommate Mara Lia became symptomatic on 6 October, we deem it much more likely that Dan infected Mara.

With this in mine, let the generation interval for the infectious disease have distribution $w(\tau)$. Then the relative likelihood that case i was caused by case j is

$$p_{ij} = \frac{w(t_i - t_j)}{\sum_{k,\, k\neq i} w(t_i - t_k)} \tag{1}$$

where t_i, t_j, and t_k are the times of symptom onset for cases i, j, and k, respectively.[4] Then, the effective reproduction number for case j is the sum over all cases i, weighted by the relative likelihood that it was case j that caused i,

$$R_j = \sum_i p_{ij} \tag{2}$$

What can we say about this equation? First of all, it should be noted that any two cases that are reported on the same day will have the same R_j. That's because this method uses only the time of reporting to infer.

Second, this method will clearly produce estimates that are time varying, in the sense that for the same disease and the same population, our estimates of R may (and will) vary based on t.

[3] Uh oh! Looks like the notes were corrupted. How many unique directed trees are possible among N entities?

[4] As a refresher, the **latent period** is the time between infection and infectiousness. The **generation interval** or **generation time** is the time between the infection of a primary case and its secondary cases. Similarly, the **incubation period** is the time between infection and the manifestation of symptoms. The **serial interval** is the time between the symptoms of a primary case and the symptoms of a secondary case. The incubation period may be longer or shorter than the latent period, depending on the disease.

It is for this reason that the method of Wallinga and Teunis has demonstrated substantial value to the modeling community: not only can we *average* over R_j to get an estimate of R_0 during an epidemic outbreak, but we can also simply track R_t, the real-time reproductive number over time.

10.6 Estimating overdispersion parameters from transmission trees

Suppose that we have the data from a transmission tree, such that we know, with certainty, who infected whom, and thus how many secondary infections were produced for each initial infection. In other words, suppose that we have a set of secondary infection data, $\{z_i\}$ for $i = 1, 2, \ldots, N$ samples.

Last week, we learned the Negative Binomial model for secondary infections, which says that the likelihood of observing data point z_i is

$$Pr(z_i) = \frac{\Gamma(z_i + k)}{z_i!\Gamma(k)} \left(\frac{k}{k+\mu}\right)^k \left(1 - \frac{k}{k+\mu}\right)^{z_i} , \tag{3}$$

where k is the dispersion parameter and μ is the expected value—last week's lecture notes simply wrote this as R_0 straight away. Assuming that the data in our transmission tree are independent[5] of each other,

$$Pr(\{z_i\}) = \prod_{i=1}^{N} \frac{\Gamma(z_i + k)}{z_i!\Gamma(k)} \left(\frac{k}{k+\mu}\right)^k \left(1 - \frac{k}{k+\mu}\right)^{z_i} , \tag{4}$$

and taking a log, we arrive at the log-likelihood of

$$\mathcal{L} = \sum_{i=1}^{N} \log \frac{\Gamma(z_i + k)}{z_i!\Gamma(k)} + k \log \left[\frac{k}{k+\mu}\right] + z_i \log \left[1 - \frac{k}{k+\mu}\right] . \tag{5}$$

Finally, it turns out to be easier to estimate $\alpha = 1/k$ rather than k. As a result, we'll use the log-likelihood

$$\mathcal{L} = \sum_{i=1}^{N} \log \left[\frac{\Gamma(z_i + \alpha^{-1})}{\Gamma(\alpha^{-1})}\right] - \log[z_i!] + \alpha^{-1} \log \left[\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right] + z_i \log \left[1 - \frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right] . \tag{6}$$

We'll also use the fact that $\frac{\alpha^{-1}}{\alpha^{-1}+\mu} = \frac{1}{1+\alpha\mu} = (1+\alpha\mu)^{-1}$ and thus $1 - \frac{\alpha^{-1}}{\alpha^{-1}+\mu} = \frac{\alpha\mu}{1+\alpha\mu}$ to get

$$\mathcal{L} = \sum_{i=1}^{N} \log \left[\frac{\Gamma(z_i + \alpha^{-1})}{\Gamma(\alpha^{-1})}\right] - \log[z_i!] - \alpha^{-1} \log[1 + \alpha\mu] + z_i \log \left[\frac{\alpha\mu}{1+\alpha\mu}\right] . \tag{7}$$

Rearranging terms, we get

$$\mathcal{L} = \sum_{i=1}^{N} \log \left[\frac{\Gamma(z_i + \alpha^{-1})}{\Gamma(\alpha^{-1})}\right] - \log[z_i!] - (z_i + \alpha^{-1}) \log[1 + \alpha\mu] + z_i \log \alpha + z_i \log \mu . \tag{8}$$

[5] When is this a good assumption? When is this a bad assumption?

And, we'll simplify the ratio of the Γ functions by using the trick that

$$\frac{\Gamma(I+k)}{\Gamma(k)} = (I-1+k)(I-2+k)\dots(I+k)k$$

for positive integer I, and thus

$$\log\left[\frac{\Gamma(z_i+\alpha^{-1})}{\Gamma(\alpha^{-1})}\right] = \sum_{x=0}^{z_i-1}\log\frac{1+\alpha x}{\alpha} = \sum_{x=0}^{z_i-1}\log\left[1+\alpha x\right] - z_i\log\alpha$$

so

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{x=0}^{z_i-1}\log\left[1+\alpha x\right] - \log\left[z_i!\right] - (z_i+\alpha^{-1})\log\left[1+\alpha\mu\right] + z_i\log\mu\,. \tag{9}$$

Because our goal is to find the value of k (now α) that maximizes the log likelihood, we are free to multiply $\mathcal{L}$ by a constant, or even add or subtract constants; doing so will not change the location of the maximum. Therefore, we will multiply through by N^{-1} and drop the $z!$ term to get

$$\mathcal{L} = \frac{1}{N}\left[\sum_{i=1}^{N}\sum_{x=0}^{z_i-1}\log\left[1+\alpha x\right]\right] - \left(\bar{z}+\frac{1}{\alpha}\right)\log\left[1+\alpha\mu\right] + \bar{z}\log\mu\,, \tag{10}$$

where we have used $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$ to represent the sample mean.

There are two parameters μ and α, and so we take the partial derivative with respect to each, and set each equation to zero, yielding

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mu} &= -\left(\bar{z}+\frac{1}{\alpha}\right)\frac{\alpha}{1+\alpha\mu} + \frac{\bar{z}}{\mu} = \frac{\bar{z}}{\mu} - \frac{1+\alpha\bar{z}}{1+\alpha\mu} = 0 \quad \rightarrow \quad \mu = \bar{z}\,. \\
\frac{\partial\mathcal{L}}{\partial\alpha} &= \frac{1}{N}\left[\sum_{i=1}^{N}\sum_{x=0}^{z_i-1}\frac{x}{1+\alpha x}\right] + \frac{1}{\alpha^2}\log\left[1+\alpha\mu\right] - \mu\frac{\bar{z}+\alpha^{-1}}{1+\alpha\mu} = 0 \quad \rightarrow \quad \text{numerical root-finding}\,.
\end{aligned} \tag{11}$$

Once one has estimated $\hat{\alpha}$, then $\hat{k} = \hat{\alpha}^{-1}$. For a discussion of why this is appropriate, see Lloyd-Smith 2007, a paper which shows that maximum likelihood estimates of k can be biased upward if sample sizes are small or when there is under-reporting of zeroes, but will not be biased downward by other common issues with data. This means that *overdispersion is unlikely to be overestimated.*