Differential Equations and Likelihood Functions, a refresher

Tjalling Jager*

April 6, 2016

About this document

This document provides a refresher for those who have been exposed to mathematics and statistics but don't feel too confident in working with differential equations and likelihood functions. I assume that you are familiar with mathematical functions, powers and logarithms, derivatives and integration, and basic statistical distributions such as the binomial and the normal. This document is by no means complete, but should sufficiently prepare you to get started on toxicokinetic (TK) and toxicodynamic (TD) modelling and data fitting.

For the course, it is important that Section 1 and 2 of this refresher are clear to you (perhaps with the exception of Section 2.5, which is not so important). This is the level of detail in mathematics that we need for this course.

There is a good chance that Section 3 will be largely new to you, or at least that it presents several statistical concepts in an unfamiliar way. At least, try to understand the basic concepts behind fitting, likelihood and profiling. No problem if the detail in Section 3.4 and 3.5 is too much for you now; you can always return to it during or after the course. During the course, you will be using statistics in a passive way only, so focus on understanding the concepts and not the details.

A good entry-level book on modelling is the textbook by Doucet and Sloep [2], though it may not be easy to find. If you want more information about useful mathematical techniques for biologists (and are not afraid of math), download the document 'Basic methods in Theoretical Biology' from http://www.bio.vu.nl/thb/course/tb/tb.pdf.

*DEBtox Research, De Bilt, The Netherlands. Email: tjalling@debtox.nl, http://www.debtox.nl/

Contents

1 Starting from the basics

In this section, I will gently introduce the concept of differential equations.

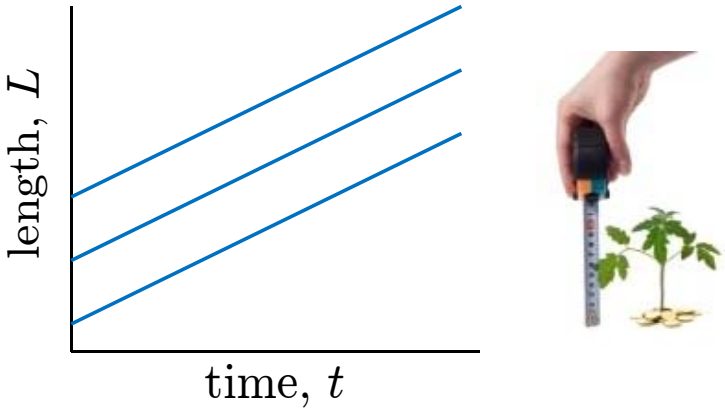

Figure 1: Plant length as a function of time. All three lines follow the growht equation that $dL(t)/dt = 2$.

1.1 Linear growth

Suppose that a plant grows by 2 cm every day. How can we describe the growth of the plant in a mathematical way? Growth is the change in size over time. If we plot length of the plant over time (Fig. 1), the growth rate is the slope (the derivative) of that curve. In short:

$$\frac{dL(t)}{dt} = 2 \tag{1}$$

L has a unit of cm, so $dL(t)/dt$ has a unit of cm/day. You can see that the units at both sides of the equal sign are the same. Integrating this equation, we get a description of the length of the plant as a function of time:

$$L(t) = 2t + C \tag{2}$$

The integration constant C reflects that Eq. 1 only specifies the growth *rate*. To fully specify the length as function of time, we need a boundary condition (the length at one time point). If I know that at $t = 0$ the plant is 5 cm tall, I can calculate the appropriate value of C:

$$L(0) = 5 = 2 \times 0 + C \tag{3}$$

from which it follows that $C = 5$. The equation for the length of plant as a function of time is thus fully described by:

$$L(t) = 2t + 5 \tag{4}$$

This is a rather trivial example, but it illustrates a fundamental approach: we start with how a system property (in this case the length of the plant) changes over time, and then derive an equation for the system property as a function of time (using a boundary condition). Checking the units at both sides of the equation is always a good way to identify errors in the derivation.

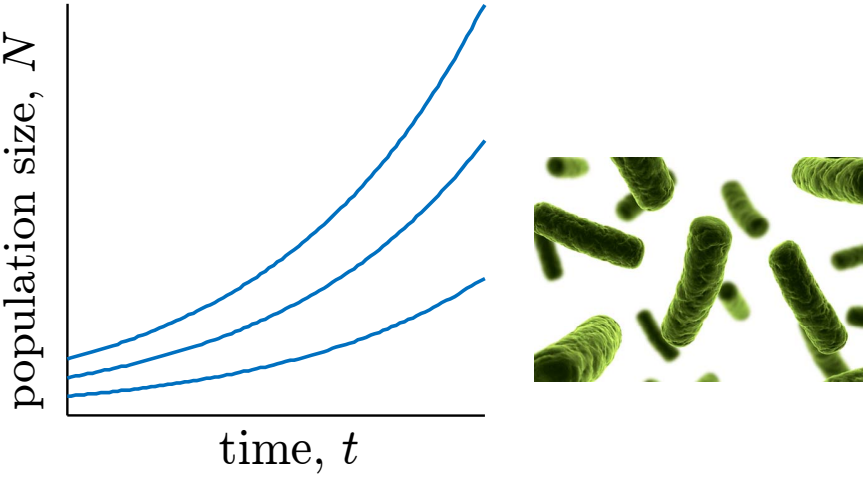

Figure 2: Bacterial population size as a function of time. All three lines follow the growth equation that $dN(t)/dt = rN(t)$, with the same value for r but a different initial population size.

1.2 Exponential growth

Under optimal conditions, bacteria divide after a constant period of time; the doubling time. Starting with one cell, we get two cells after one doubling time, four cells after two doubling times, eight cells after three doubling times, etcetera. The growth rate of the population thus increases with the size of the population. Starting with 100 cells, the growth rate is much larger than starting with a single cell. How much larger? Exactly a hundred times larger. The hundred cells all divide after one doubling time to yield two hundred cells; in the same time period, a single cell has yielded two cells. We can write this as an equation for the growth rate, which is the derivative of the population size N:

$$\frac{dN(t)}{dt} = rN(t) \tag{5}$$

The change in the population size (the growth rate) is thus proportional to the size of the population. The proportionality constant is here called r. What is the unit of r? At the left side of the equal sign, $dN(t)/dt$ has the unit of cells/day. At the right side of the equal sign, we have N in cells, so r must have the unit 1/day or 'per day' (more strictly: cells per cell per day).

How can we use this equation to go to population size as a function of time? We cannot simply integrate this function, as we did for the plant example, because $N(t)$ is on both

sides of this equation. We have an equation that species the derivative of a function, as a function of that function itself. This is a differential equation (DE; in this case an 'ordinary' differential equation or ODE). Solving differential equations is problematic, and there are only a few methods available (which rapidly fail when the differential equation becomes biologically interesting). In this simple case, however, we *can* find the solution. We are looking for a function whose derivative is the same function times a constant. Exponential functions fit this bill:

$$N(t) = Ce^{rt} \tag{6}$$

We can easily prove that this is the function we are looking for by taking the derivative:

$$\frac{dN(t)}{dt} = rCe^{rt} = rN(t) \tag{7}$$

The derivative is r times the original function. Note that I can select any value for the constant C that I like. Again, writing down the *change* in population size does not fully specify the population size over time, and again we require a boundary condition. Suppose that I know the population size at $t = 0$ and call it N_0:

$$N(0) = N_0 = Ce^{r\times 0} \tag{8}$$

It follows that the constant C in this case equals the initial population size (please note that this is *not* necessarily true for other differential equations). The solution is thus:

$$N(t) = N_0 e^{rt} \tag{9}$$

The units on both sides of the equality also match. Because r was 1/day, rt is dimensionless (which is required in the exponent). Both $N(t)$ and N_0 have the unit of cells.

How can we prove that there is indeed a constant doubling time T_2 from our solution, and how does the doubling time relate to the parameter r?

$$N(t+T_2) = 2 \times N(t) \qquad \text{fill in the equations for } N: \tag{10a}$$
$$N_0 e^{r(t+T_2)} = 2 \times N_0 e^{rt} \qquad \text{divide both sides by } N_0: \tag{10b}$$
$$e^{r(t+T_2)} = 2 \times e^{rt} \qquad \text{remove brackets in the exponent:} \tag{10c}$$
$$e^{rt+rT_2} = 2 \times e^{rt} \qquad \text{write as two exponents:} \tag{10d}$$
$$e^{rt} \times e^{rT_2} = 2 \times e^{rt} \qquad \text{divide both sides by } e^{rt}: \tag{10e}$$
$$e^{rT_2} = 2 \qquad \text{take natural logarithm on both sides:} \tag{10f}$$
$$rT_2 = \ln 2 \qquad \text{and rearrange:} \tag{10g}$$
$$T_2 = \frac{\ln 2}{r} \tag{10h}$$

This shows that there is a constant doubling time, which does not depend on the initial population size N_0, nor on the time t. It only depends on the rate constant r.

In this example, we could have written down a function for the population size immediately, using the constant doubling time:

$$N(t) = N_0 2^{t/T_2} \tag{11}$$

However, this is mathematically equivalent to the equation we derived by solving the differential equation (the proof is left as an exercise for the reader).

1.3 What if we cannot find a solution?

In the previous section, we could solve the differential equation and come up with a 'normal' (algebraic) function for the system property that we were interested in (population size). However, for most of the biologically-interesting problems, the ODEs cannot be solved analytically anymore. Two options remain: inspection of the ODE or simulation.

An analysis of the ODE itself can for example provide information about the existence of equilibria and their stability. As an example, consider a population of bacteria with a constant removal of bacteria. The removal rate is independent of the population size, b cells/day. The differential equation then is:

$$\frac{dN(t)}{dt} = rN(t) - b \tag{12}$$

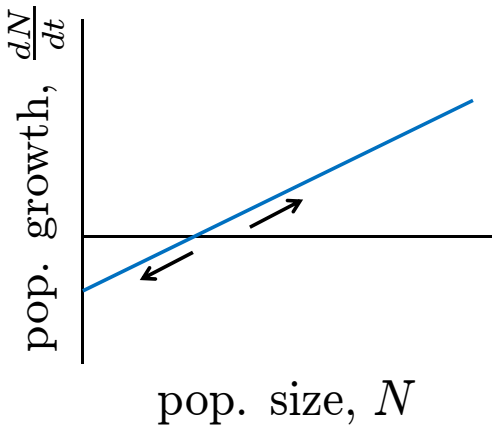

Figure 3: Bacterial growth rate versus actual population size for the example with constant removal: $dN(t)/dt = rN(t) - b$.

This equation can still be solved analytically, but we can also obtain quite a lot of information from the ODE itself. An equilibrium exists where the growth rate is zero, and thus $dN(t)/dt = 0$. We can see from the differential equation that this occurs when $N = b/r$. When the population size is exactly b/r, the population does not grow or shrink. We call this equilibrium 'stable' when a small deviation from this equilibrium produces a derivative that leads the population back to the equilibrium. Below the stable equilibrium population size, the growth rate should be positive, and above this size it should be negative (so that

the population shrinks back to the equilibrium size). So what is the verdict in our example? If N is below the equilibrium size b/r, the derivative is negative, so the population would shrink to zero. Above this point, the derivative is positive, so the population would increase to infinity (in the model at least). Clearly, this equilibrium is unstable.

The other option to say something about the solution of an ODE, without analytically solving it, is to simulate it. For most of the interesting models, this is the only way to say something about the solution. To start, we need the initial condition of the system, for example the population size at $t = 0$, N_0. At $t = 0$, we can calculate the derivative $dN(t)/dt$, so we know how N will change in the close neighbourhood of N_0. We can predict where the population will be after a small amount of time Δt by taking a linear approximation:

$$N(0 + \Delta t) = N(0) + \left.\frac{dN(t)}{dt}\right|_{t=0} \times \Delta t \tag{13}$$

The vertical line after the derivative should be read as: "the value of the derivative when evaluated in the point $t = 0$." Starting in N_0 at $t = 0$, we can thus calculate a new population size at $t = \Delta t$, from there one at $t = 2 \times \Delta t$ etc. (see Fig. 4). In general, starting from a time point $t = T$, we calculate a new value at $t = T + \Delta t$ as:

$$N(T + \Delta t) = N(T) + \left.\frac{dN(t)}{dt}\right|_{t=T} \times \Delta t \tag{14}$$

Of course, this is an approximation of the real function for $N(t)$, but the approximation improves the smaller we take Δt. This is not something to do with a pocket calculator; computers are perfectly suited to perform this tedious task.

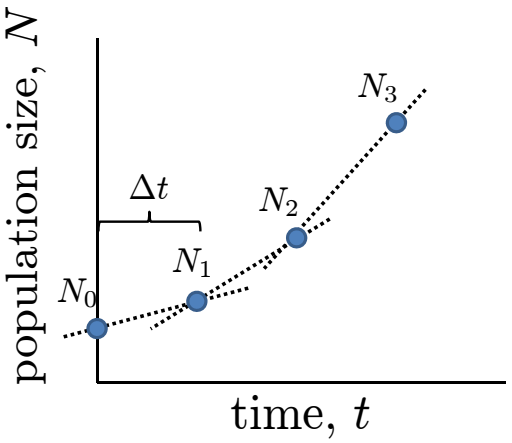

Figure 4: Numerical approximation of a differential equation using Euler's method. The slope in point N_0 is used to calculate point N_1 after a small time step Δt, etcetera.

This particular method to approximate an ODE is called Euler's method. More sophisticated methods exist, and software packages such as Matlab include various solver algorithms to choose from.

1.4 Notation issues

Different notations are used for derivatives, which are actually equivalent, such as:

$$\frac{dN(t)}{dt} \equiv \frac{dN}{dt} \equiv N' \equiv \dot{N} \tag{15}$$

Removing the (t) after the function name (in this case N) is often used to enhance readability of differential equations. The notations with the accent or dot are also frequently used, but have the disadvantage that it is not clear how the derivative is taken. In biology, derivatives will generally be taken with respect to time, because we are interested in processes over time.

1.5 Dimensions and units

In the previous, I talked explicitly about the units of the parameters and variables in the equations. Units are an expression for the 'dimension' of a parameter. For example, grammes is a unit for the dimension mass, and day is a unit for time. Dimension analysis is a very handy tool in building and testing of models. We use the following simple rules:

1. You can only add or subtract terms with the same dimension (or units). Example: you cannot add apples and pears unless you put both in the 'unit' of 'pieces of fruit.'

2. In a multiplication of two terms, the dimension (or units) multiply. In a division the dimensions (or units) are divided. When you divide apples by pears you get a property with the unit apples/pear. If both are expressed in the same unit ('pieces of fruit'), the result is a dimensionless fraction.

3. Terms in an exponential or in a logarithm must be dimensionless. Example: the term e^{rt} can only be correct of the units of r and t cancel (e.g., when they are expressed as per day and day, respectively).

4. At both sides of an equal sign there should be the same dimension (or unit). Example: $dN/dt = rN$ can only be a meaningful expression when r has a dimension 'per time' (with time in the same units as t).

A model expression that is not dimensionally correct can never be a meaningful representation of the real world.

1.6 Why are differential equations so commonly used?

An enormous number of practical mathematical models take the form of differential equations. This is not by accident: the behaviour of many systems depends on the state of the system. The example I started with (the continuously growing plant, growing with 2 cm/day) was a rather artificial example. The growth of a plant can not be constant (at least not for long). The size of the plant must influence the growth rate in some way, e.g., because the size of the plant determines the area for capturing sunlight and relates to

the size of the root system for uptake of nutrients. When the change of a system's state depends on its current state, we are immediately in the realm of differential equations.

2 Compartment models

Compartment models form a particularly insightful example of the application of ODEs to solve practical problems. In these models, there is a 'carrier medium' which holds and transports a 'tracer.' In this example, we follow a pesticide that is dumped into a lake (example modified from [2]). This type of modelling follows naturally in chemical transport problems in the environment (fate modelling) or in organisms (toxicokinetics or pharmacokinetics).

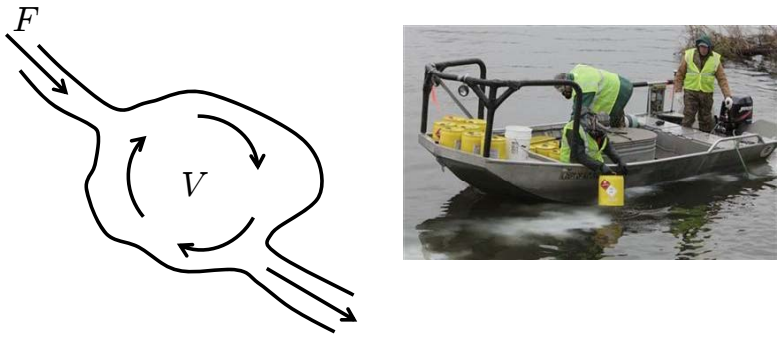

Figure 5: A well-mixed lake with volume V, with a river of flow rate F running through it. These men are dumping rotenone to kill asian carp; an invasive species in the US.

2.1 Unilake

Consider a lake with a volume V m^3 with a river running through it at a flow rate F m^3/s. At $t = 0$, someone dumps a pesticide into the lake. What is the time course of the pesticide's concentration in the lake? This is not a simple question at all. The change in the lake's concentration depends on the lake's morphology, the chemical's dispersion from the point of entry, etcetera. Let's start by making a huge simplifying assumption: assume that the lake is turbulently mixed, to such an extent that the pesticide is homogeneously distributed instantaneously. Therefore, the concentration of the pesticide is the same throughout the lake. With this assumption, we can treat the lake as a compartment with respect to its chemical content. After the initial dumping of the pesticide into the lake, the concentration will decrease because of the river flowing through the lake. Per second, the river takes F m^3 water out of the lake, that has the associated pesticide concentration. Clearly, the *change* in the lakes pesticide content depends on its *actual* content. We can put this information into a differential equation for the amount of chemical in the lake Q (in mg):

$$\frac{dQ(t)}{dt} = -FC(t) = -\frac{F}{V}Q(t) \tag{16}$$

Here, $Q(t)/V$ is the concentration of the pesticide in the lake, and the river flow F takes it out (so the derivative is always negative). Dimension analysis shows that this equation can make sense. Is there an equilibrium in this model? The derivative $dQ(t)/dt$ is only

zero when $Q(t)$ is zero. Thus, the only stable situation is the situation where there is no pesticide in the lake.

We can easily translate this equation for the amount of pesticide into an equation for the concentration by dividing both sides by V (this is only allowed when the volume is constant):

$$\frac{dC(t)}{dt} = -\frac{F}{V}C(t) \tag{17}$$

Why did I start with amounts of chemical and not directly with the concentration? In this case, it does not matter. However, it will in the next example, where the chemical from one lake enters another. The combination F/V is also known as the 'dilution rate.'

With the knowledge of the previous section, we can solve this differential equation, which yields an exponential decay (as long as the parameters F and V are constant over time):

$$C(t) = C(0)e^{-\frac{F}{V}t} \tag{18}$$

In the solution, the initial concentration in the lake $C(0)$ appears. The differential equation is only concerned with the *change* in concentration, so we need an additional piece of information to fix the concentration's absolute time course.

When t becomes very large, the concentration C becomes very small. Does it ever become zero? No, it will approach zero, but never reaches it exactly. This is one point where model and reality do not match. In reality, there will be a point where the last molecule of the pesticide has left the lake. In the model, there will always be 'fractions of molecules' that can be divided *ad infinitum*. In practice, this usually does not bother us, unless we are interested in the detailed behaviour in situations where there are only a few molecules.

So, the concentration in the lake approaches zero after a long period of time. To express this mathematically, we use a limit:

$$\lim_{t\to\infty} C(t) = 0 \tag{19}$$

This expression tells us that if t goes in the direction of infinity, we can get $C(t)$ as close to zero as we like (but never reach it). You should note that infinity is not a number; it is best to use it only in the context of a limit, where it indicates that a parameter *goes to* infinity.

2.2 Twin lakes

We take the lake example one step further by adding a second lake through which the river flows after the first lake. For the first lake, the differential equation for the amount of pesticide is exactly the same as in the single lake example:

$$\frac{dQ_1(t)}{dt} = -\frac{F}{V_1}Q_1(t) \tag{20}$$

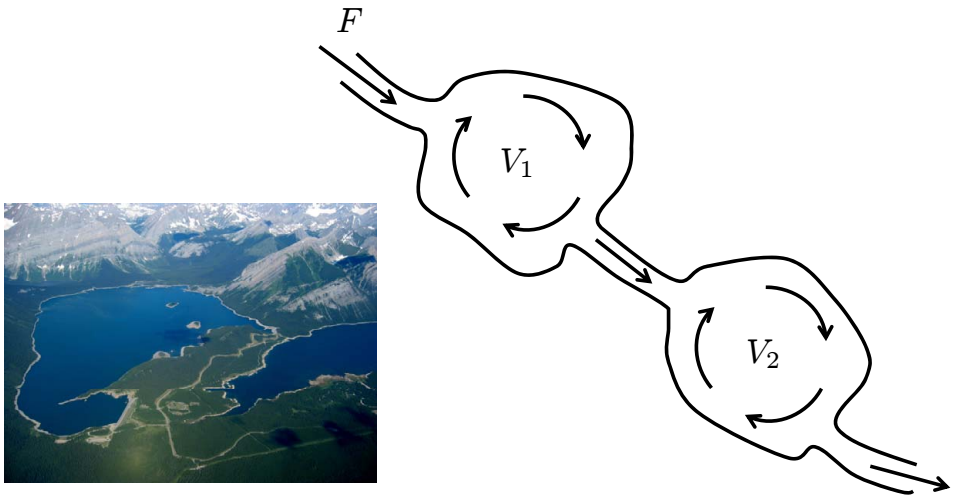

Figure 6: Two lakes with a river of flow rate F running through both.

The first lake is not influenced by what happens in the second lake. The second lake receives the river flow that comes out of the first lake. Clearly, all of the pesticide that leaves the first lake will arrive as input in the second lake (we are making a mass balance for the pesticide here). Thus, the negative term in the ODE for lake 1 (Eq. 20) enters as a positive term in lake 2:

$$\frac{dQ_2(t)}{dt} = \frac{F}{V_1}Q_1(t) - \frac{F}{V_2}Q_2(t) \tag{21}$$

The second lake also has a negative term, which is what the river takes with it when it leaves the lake system. This negative term looks very much like that of the first lake, but now on the basis of Q_2 and V_2.

Does the second lake have an equilibrium? $dQ_2(t)/dt$ can be zero, so there is a point where the Q_2 does not change in time. However, this point also depends on the value of Q_1, which continuously decreases in time. From looking at the differential equation, we can see that Q_2 initially increases, until it reaches a certain point, after which it decreases again. Thus, this point is not an equilibrium.

Now convert the equations for amount of pesticides to concentrations. Equation 20 for lake 1 is divided by V_1:

$$\frac{dC_1(t)}{dt} = -\frac{F}{V_1}C_1(t) \tag{22}$$

and for lake 2, Equation 21 is divided by V_2:

$$\frac{dC_2(t)}{dt} = \frac{F}{V_2}C_1(t) - \frac{F}{V_2}C_2(t) \tag{23}$$

The interesting thing is now that the positive term for lake 2 is *not* the same as the negative term for lake one, unless the volumes of both lakes are identical. This example shows that

mass is conserved but not concentrations. When 5 mg of pesticides leaves lake 1, there must be 5 mg entering lake 2. However, when 5 mg/L leaves lake 1 this does not mean that the concentration in lake 2 increases by 5 mg/L. For this reason, it is a good strategy to start building your model in terms of quantities that are preserved (masses, numbers, energy, etc.) and not densities, concentrations or fractions. When the model is finished, and mass and/or energy balances are checked, the equations can subsequently be converted to densities or concentrations.

2.3 Unilake with a continuous input

Now suppose that instead of a single pulse of pesticide there is a continuous input into the lake at a rate of E mg pesticide per day. The differential equation for the lake is now:

$$\frac{dC(t)}{dt} = E - \frac{F}{V}C(t) \tag{24}$$

In this case, there is an equilibrium situation. The derivative is zero when $C = (EV)/F$ mg/m^3. Again, this equilibrium is never reached exactly (unless we are able to start with the lake exactly at the equilibrium concentration), but will be approached ever closer over time. We can thus write:

$$\lim_{t\to\infty} C(t) = \frac{EV}{F} \tag{25}$$

Is this equilibrium stable? Yes. When the concentration in the lake is below the equilibrium concentration, the derivative is positive, so the concentration increases over time (Fig. 7). If the concentration exceeds the equilibrium concentration, the concentration will decrease over time.

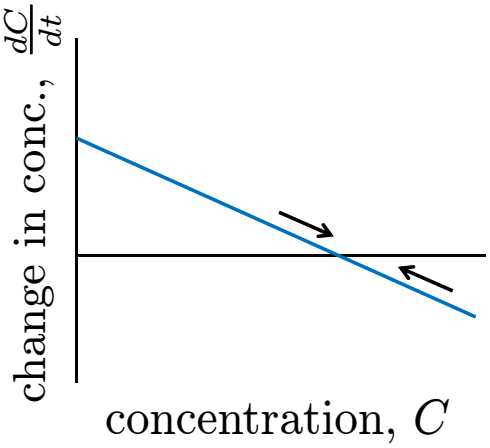

Figure 7: Change in the lake's concentration rate versus actual concentration for the example with constant emission: $dC(t)/dt = E - F/VC(t)$.

2.4 Adding processes

Modern-day pesticides are degradable, so it is unlikely that they will persist in the lake. For degradation, the simplest approach is to assume a first-order decay (because the rate with which the chemical is degraded depends linearly on the concentration of one compound). In a closed bottle, this leads to an exponential decrease of the concentration. We can add a first-order rate constant k_d to the differential equation for concentration (Eq. 17) as follows:

$$\frac{dC(t)}{dt} = -\frac{F}{V}C(t) - k_d C(t) = -\left(\frac{F}{V} + k_d\right) C(t) \tag{26}$$

The first-order constant k_d is of the same type as the dilution rate of the lake F/V, so both terms can be added.

A similar extension is to follow the transformation of the parent pesticide Q_p into a metabolite Q_m. This leads to a system that is rather similar to the twin lake example:

$$\frac{dQ_p(t)}{dt} = -\frac{F}{V}Q_p(t) - k_m Q_p(t) \tag{27}$$

$$\frac{dQ_m(t)}{dt} = -\frac{F}{V}Q_m(t) + k_m Q_p(t) \tag{28}$$

The second negative term for the parent compound (the loss due to transformation) enters as a positive term for the metabolite. We can easily convert this set of ODEs to concentrations by dividing both equations by V. In this case, there is only a single volume, so we could have immediately started with concentrations. However, starting with absolute amounts makes it easier to check that our mass balance is intact.

This example shows that the definition of compartments is flexible. We can use spatially separated states such as the concentration in lake 1 and lake 2, or two compounds in the same lake.

2.5 What if the volume of the lake varies?

Suppose that the water volume in the lake varies over time. The pesticide concentration in the lake at each time point is now given by:

$$C(t) = \frac{Q(t)}{V(t)} \tag{29}$$

But what about the derivative $dC(t)/dt$? What is the derivative of a quotient of two functions? We have the 'quotient rule' for derivation, but we can also make use of the 'product rule' as follows (after the first equation I do not write (t) behind the functions

anymore for readability):

$$\frac{dC(t)}{dt} = d\left(\frac{Q(t)}{V(t)}\right)/dt \qquad \text{rewrite quotient:} \tag{30a}$$

$$= \frac{d(QV^{-1})}{dt} \qquad \text{apply product rule:} \tag{30b}$$

$$= \frac{dQ}{dt}V^{-1} + Q\frac{d(V^{-1})}{dt} \qquad \text{apply chain rule for second term:} \tag{30c}$$

$$= \frac{dQ}{dt}V^{-1} - Q\frac{1}{V^2}\frac{dV}{dt} \qquad \text{rewrite with } Q/V = C\text{:} \tag{30d}$$

$$= \frac{dQ}{dt}V^{-1} - C\frac{1}{V}\frac{dV}{dt} \tag{30e}$$

In words, the derivative for the concentration (C) is the derivative for the amounts (Q) divided by the volume as function of time, *minus* the concentration times the relative change in volume. This makes sense because when the change in volume is zero, we have our original ODE back. This last term takes care of the dilution of the pesticide concentration when the lake increases in volume, and the fact that the pesticide is concentrated when lake volume decreases.

The volume V can be any function of time; for the ODE this does not matter. However, solving the ODE to an explicit algebraic expression is usually impossible. There is an interesting exception. Suppose the volume of the lake decreases exponentially. This implies that:

$$\frac{dV(t)}{dt} = -aV(t) \Rightarrow \frac{1}{V(t)}\frac{dV(t)}{dt} = -a \tag{31}$$

We can thus write the ODE for the pesticide concentration as:

$$\frac{dC(t)}{dt} = -\frac{F}{V(t)}C(t) + aC(t) = \left(a - \frac{F}{V(t)}\right)C(t) \tag{32}$$

The exponential decay rate a is of the same type as the lake's dilution rate $F/V(t)$, so we can subtract one from the other. At some point, we get an equilibrium for the concentration, where the removal of pesticide from the lake is exactly compensated by the decrease in volume. The amount of pesticide Q continuously decreases, but the concentration remains stable.

Of course, an exponential decay for the volume of a lake (without changes in the river flow F) is not a very realistic situation. However, some (populations of) organisms may grow exponentially, at least for some time. In toxicokinetic models for the concentration of a chemical in an organism, such a growth rate constant is sometimes used. Care should be taken that this simplification only holds for *exponential* growth or shrinking.

2.6 The role of assumptions

Assumptions play a pivotal role in modelling. Models are always a simplification of reality, and simplifications involve assumptions about the processes. In drawing up a model it is proper etiquette to explicitly state all assumptions that go into a model. The equations itself are then nothing more than a translation into mathematics; no additional assumptions should go into the translation. Unfortunately, most models will contain hidden assumptions, forcing the reader to scrutinise the equations to tease out which assumptions are made.

In our initial unilake example, the main assumption is that the lake is well mixed: the concentration should always be the same throughout the lake. There should be no losses of the chemical from the water of the lake, and thus no degradation, volatilisation, or sorption to sediment. The volume of the lake and the river flow rate should be constant. Furthermore, the concentration is treated as a continuous variable, even though it is made up of a discrete number of chemical molecules (this is no problem as long as the number of molecules is very large). For the twin lake example, we are adding the assumption that the chemical that leaves lake 1 immediately turns up in lake 2 (no chemical is 'in transit' in between the two lakes).

The assumptions are the crucial part of the thought process that takes us from the complex real world to the simplified model. But, how complex or simple should a model be? Clearly, there is no straightforward answer to this question; model complexity depends on the question that needs to be addressed and the information that is available. The more complex a model is, the better the fit on the data can be, but the less meaningful that fit becomes. Truly complex models cannot be 'falsified' anymore, and a meaningful parameterisation puts high demands on the data in terms of quality and quantity (or strong *a priori* information about the parameter values). Furthermore, complex models are unlikely to provide general insight into problems. it must be stressed that the purpose of a modelling exercise should not be to get a curve through a set of data; that works better without a model (take a pencil and connect the dots). Instead, models are simplifications that help us understand the system that we are interested in, and to make predictions for the system's behaviour under untested (or untestable) circumstances.

Model building is thus a balancing act. Complex models can include more realism and provide a better fit to data, at the cost of higher parameter uncertainty and lack of generality. In general, striving for simplicity is a good thing, but as Albert Einstein appears to have said: "Everything should be made as simple as possible, but no simpler." If a model is too simple, it will be totally useless because it cannot explain anything about the real system. The lake models I discussed in this Chapter are at the lowest end of the complexity scale: I doubt that you can dream up a simpler model that still bears any relevance for reality. However, it is easy to make these models more complex, and most of the published lake models that you will find are indeed more complex.

3 Fitting and likelihood

Interpreting a set of data requires two types of model: a model for the process (e.g., in terms of a set of ODE's) and a model for the deviations between model and data (a statistical model). It is obvious that we need a process model, but it is perhaps less obvious that we need to think about a statistical model. The statistical model is needed to help us judge what a good fit is, and thus what the best parameter values are, and how certain we are of them (confidence intervals). In this section, I will discuss the 'sum-of-squares' as the most popular criterion, and then introduce the far more general framework of likelihood functions.

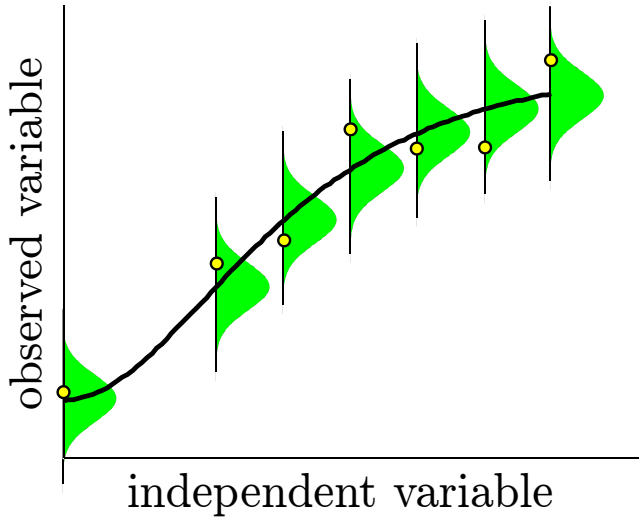

Figure 8: The principle underlying least squares: normal independent distributions for the error, with constant variance. The mean of the distributions is given by the process model. For the residuals (the difference between observation and process model), the mean of the distribution is zero.

3.1 The sum-of-squares

Probably the best known and most popular approach used to fit models to data is the least-squares method. The residuals (the difference between model prediction and observed value) are squared and summed. Suppose we have a data set Y (consisting of n values) and predicted values $\hat{Y}$ (depending on the parameter set θ), then the sum-of-squares (SSQ) is:

$$\mathrm{SSQ}(\theta; Y) = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i(\theta) \right)^2 \tag{33}$$

This criterion, however, follows from assumptions about the nature of the deviations between model and data. These assumptions are:

1. The residuals are random trials from a normal distribution with mean zero and some unknown standard deviation.
2. This standard deviation is the same for all data points (homoscedasticity).
3. These normal distributions are uncorrelated (independent trials).
4. The values on the x-axis have no associated error.

From this set of assumptions, the SSQ can be derived as the logical consequence. What kind of process did we specify with these assumptions? It most closely represent the situation where the deviations between model and data are caused by random measurement error. Measurement error is, however, not our main problem in biology or ecotoxicology. In practice, the deviations result from the model being wrong, and because of biological variation (between and within individuals). Often, our observations are not independent, for example when we follow the body size of the same group of individuals over time. Homoscedasticity is often compromised because large observed values tend to have a higher variation than small ones. In other cases, a normal distribution does not apply at all because we follow discrete responses such as the number of surviving individuals.

Unfortunately, not all of these issues have a satisfying solution (at least not a workable one in practice). Therefore, we have to be sloppy to proceed. This sloppiness means that we should not put too much emphasis on the exact values of parameters or their confidence intervals. Estimates and intervals are only representative when both the process model *and* the statistical model are 'true'; a situation that is rarely approached in biology.

3.2 Defining the likelihood

The likelihood (L) of a set of parameters (θ), given a data set (Y), is the probability (P) to obtain the data set if that parameter set would have been the correct one (and the model is true). In math:

$$L(\theta|Y) = P(Y|\theta) \tag{34}$$

In 'frequentist' statistics, the data set represents a random trial from a stochastic process, so we can talk about the probability of the data. The parameter set however has no probability; each parameter has a fixed but unknown value. Therefore we speak about the likelihood of the parameters, and not their probability (Bayesians do, by the way; they define probability in a more general fashion). In a coin-tossing experiment, we can talk about the *probability* of finding 20 times heads in 50 trials, but we talk about the *likelihood* of this coin being fair (i.e., that $p = 0.5$).

How do we calculate this probability? We can calculate the probability of the data when we specify the distribution that they follow. Let's start with an extremely simple example. Suppose I throw a coin 50 times and observe 20 times heads. The number of successes (here defined as throwing heads), follows from the bionomial distribution:

$$L(p|Y) = P(Y|p) = \binom{50}{20} p^{20}(1-p)^{30} \tag{35}$$

To find the maximum of the likelihood function, we can plot the function in Equation 35 and see where it reaches its maximum (Fig. 9). Alternatively, we can use a numerical optimisation routine. However, in this case, we can also do it analytically. How can we find the maximum or minimum of a function? At these points, the derivative is zero. Therefore, we can take the derivative of the likelihood to the parameter p to find out for which value of p the function has a maximum. In this simple case, this is indeed possible (this is a nice exercise to test your math skills), and the result is rather unsurprising: $\hat{p} = \frac{20}{50}$. The maximum likelihood estimate for p is thus the frequency of heads found in the experiment.

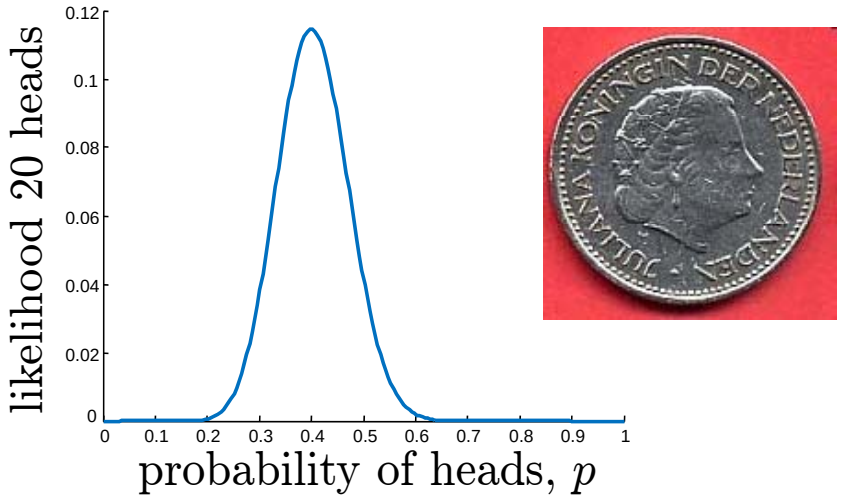

Figure 9: The likelihood of 20 times heads in 50 throws with a coin, as function of the probability of success. This is the function specified in Eq. 35.

3.3 Properties of likelihood functions

The big advantage of likelihoods is that we can combine different sources of information. When we have a model that predicts both survival and growth of an organism, we need to make a simultaneous fit on both types of data set. We cannot add sums-of-squares, because the SSQ depends on the unit of the observations. The likelihood is, however, dimensionless, and independent likelihood functions can be multiplied to yield one combined likelihood function.

The likelihood-ratio principle can be used to compare two model fits. The two models we compare should be nested; that is, one model is a reduced version of the other (reduced by fixing one or more parameters to a certain value). The ratio of two likelihoods (and thus the difference in two log likelihoods) from nested models can be tested for significance; whether the fits of the models are different enough to yield a low probability of this happening purely by chance. To do this, we can use an 'asymptotic property' of the likelihood ratio

(which means that it becomes a better approximation of reality, the larger the number of observations): two times the difference in log-likelihood will follow a chi-square distribution with as degrees of freedom (v) the number of parameters in which the two models differ. Using the natural logarithm of the likelihood ($\ell = \ln L$):

$$2\ (\ell(\theta|Y) - \ell(\theta_1|Y)) \sim \chi^2_{v,1-\alpha} \tag{36}$$

We compare the full the parameter set θ, and a reduced parameter set θ_1. When two times the difference between the two likelihoods exceeds the critical value of the chi-square distribution, the zero hypothesis ('the fits are equally good') is rejected.

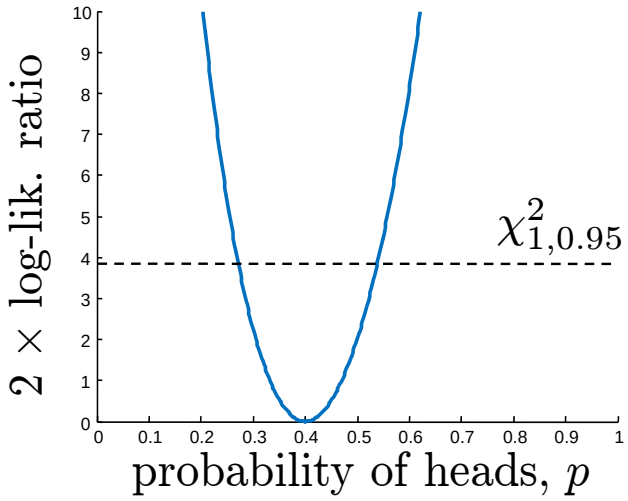

Figure 10: Likelihood-ratio criterion for the coin-tossing example. Where the ratio exceeds the 95% critical value for the chi-square distribution (one degree of freedom), the parameter values are rejected. The borders of the confidence interval are $0.272-0.538$ (for comparison: using a normal approximation for the distribution of $\hat{p}$ leads to $0.264 - 0.536$).

The likelihood function can also be used to generate confidence intervals for parameter estimates. There are several ways to do this, probably the most popular one is by using the curvature of the likelihood around the best value. These procedures yield so-called asymptotic standard errors (asymptotic because they are accurate for large data sets). Confidence intervals from this approach are always symmetric around the best estimate.

A more robust approach (accurate for smaller data sets and nasty models) is 'profiling the likelihood'. In profiling a single parameter, one parameter is fixed to a value and the other parameters estimated. The likelihood is calculated for a range of values for the fixed parameter. The result is a plot like Figure 10 (although there is only one parameter in that case, so no optimisations are needed). We can then use the chi-square criterion to find the borders of the confidence interval: the interval is defined as all those values of the parameter that are not rejected in a likelihood-ratio test at confidence level α. The resulting confidence intervals can be strangely shaped, or even discontinuous. More information about calculation of confidence intervals can be found from textbooks [5] or the open literature [4].

The likelihood function can also be used in a Bayesian statistical framework, which apart from the data, also relies on prior information to generate a probability distribution for a parameter set. Bayesian statistic is largely outside the scope of this text, you might consult textbooks [1], but I want to give you a hint about its concepts. Bayesians can speak about the 'probability of parameters' because they apply a broader definition of 'probability'; the probability distribution of a parameter relates to our degree of belief about its value. The probability of a set of parameters θ, given the available data, is given by Bayes' theorem:

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} \propto L(\theta|Y)P(\theta) \tag{37}$$

where $P(\theta)$ is the prior probability of the parameter set (before looking at the data), and $P(Y)$ the probability of the data (which is generally irrelevant because it does not depend on the parameters, and we only need to know the probability up to a proportionality). The resulting probability $P(\theta|Y)$ is our 'posterior' probability distribution: how our *a priori* belief in the parameter's value ($P(\theta)$) is modified by the data (through the likelihood $L(\theta|Y)$).

3.4 Bi- and multinomial likelihood

In general, the probability to find x successes in n bionomial trials, when the probability of success is p, can be calculated as:

$$P(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{38}$$

In ecotoxicology, the problem of coin tossing is not a very common one. Nevertheless, we can make use of this result when we look at survival (or any other all-or-nothing response). Suppose we set up a test with n animals per container, and k treatments (e.g., different concentrations of a toxicant). After a fixed exposure period we look at the number of survivors in each container x_i. These values x_i can be viewed as the outcome of k coin-tossing experiments with an unknown probability of 'success' p. When I have a model to predict p as a function of a parameter set θ, we can calculate the likelihood as the product of k independent trials:

$$L(\theta|Y,n) = P(Y|\theta,n) = \prod_{i=1}^{k} \binom{n_i}{x_i} p(\theta)^{x_i} \left(1-p(\theta)\right)^{n_i - x_i} \tag{39}$$

Fitting a dose-response curve in this manner thus assumes a binomial 'error model' for the data (Fig. 11).

In the binomial case, we only have two options: either the animal is dead when we end the test or it is alive. However, suppose we make observations on the status of the animals at more points in time. Suppose that we count the survivors after 1 day and after 2 days (a situation that occurs in the standard acute toxicity tests with the water flea *Daphnia magna*). An individual animal might have died between the start of the experiment and day

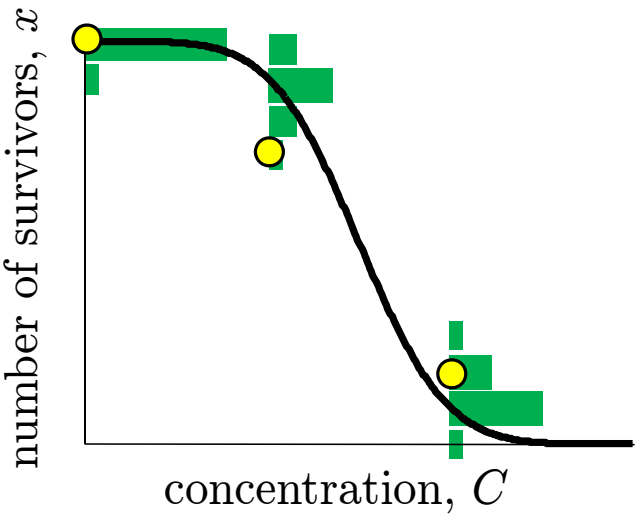

Figure 11: Assuming a binomial distribution for the error. The probability of success is given by the model curve.

1, or between day 1 and day 2, or it might still be alive at the end of the test. Clearly, there are 3 possible outcomes, each has a probability, and these probabilities add up to 1. For this situation, we can use a generalisation of the bionomial distribution: the multinomial. If we follow n individuals, and they can end up in three categories, the probability of a certain outcome x is (see Fig. 12):

$$P(x_1, x_2, x_3|n, p_1, p_2, p_3) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3} \quad \text{with } \sum p_i = 1 \text{ and } \sum x_i = n \tag{40}$$

If we have a model for p with parameters (θ), we can easily draw up a likelihood function for these parameters (see [3] for details).

3.5 Normal likelihood

The normal distribution is (as its name suggests) the most common distribution in science. Its popularity comes from the central limit theorem, which shows that the sum of a large number of independent random variables (which each can have a different distribution) will follow a normal distribution. How large 'large' is depends on the shapes of the distributions (the sum of 2 normal distributions is already perfectly normal, but for uniform distributions we need quite a few more). The probability density function f for the normal distribution is (note that $\exp(x)$ is another way to write e^x):

$$f(x|\mu;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{41}$$

Suppose we have a set of data Y with n data points. Further suppose that the observations come from a normal distribution with variance σ^2 around the 'true' value $\hat{Y}$ that is predicted

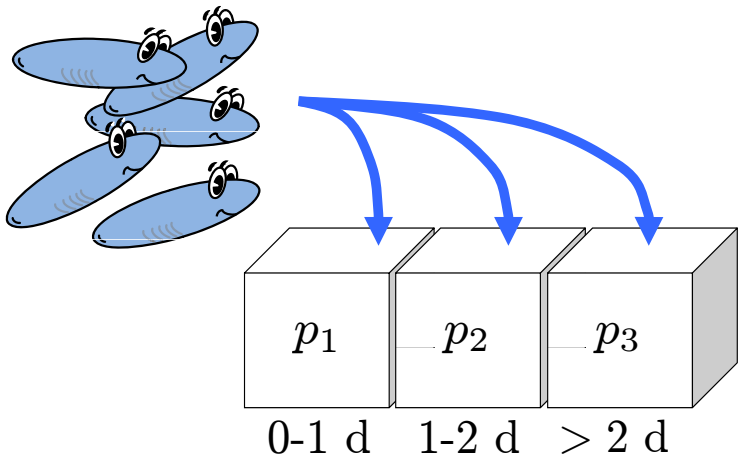

Figure 12: Illustrating the multinomial distribution; each animal will end up in one of the three categories (the time interval in which it dies) with a certain probability p_i. These probabilities sum to one over the three categories.

by a model, which has a set of parameters θ. The probability density for a single observation Y_i is:

$$f(Y_i|\theta;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \hat{Y}_i(\theta))^2}{2\sigma^2}\right) \tag{42}$$

The likelihood of the complete data set is the product of the probability densities for each Y_i (note that we have seamlessly gone from probabilities to probability densities; this should not concern us here):

$$L(\theta|Y;\sigma) = \prod_{i=1}^{n} f(Y_i|\theta;\sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \hat{Y}_i(\theta))^2}{2\sigma^2}\right) \tag{43}$$

Which we can rewrite to:

$$L(\theta|Y;\sigma) = \frac{n}{\sqrt{2\pi\sigma^2}} \exp\left(-\sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i(\theta))^2}{2\sigma^2}\right) \tag{44}$$

$$= \frac{n}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\mathrm{SSQ}(\theta;Y)\right) \tag{45}$$

The value of σ (the residual s.d.) does not depend on i, nor on the parameter set θ. So, you can see, even without differentiation, that the value of L is maximum when the sum-of-squares is minimal. Least squares thus follows from assuming normal, independent, distributions for the error, with a constant variance (whose value we don't need to know).

In practice, we usually use the natural logarithm of the likelihood ℓ (the maximum of a function is at the same location as the maximum of the ln of the function):

$$\ell(\theta|Y;\sigma) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\mathrm{SSQ}(\theta;Y)}{2\sigma^2} \tag{46}$$

Using the normal likelihood to fit model parameters to data thus requires an estimate for the residual variance σ^2. We can base this variance on previous experience, estimate it from the data set, or 'profile it out.' The profiling strategy rests on the idea of replacing σ^2 by its maximum likelihood estimate estimate $\mathrm{SSQ}(\theta;Y)/n$. Entering this replacement into the log-likelihood function leads to a very simple result:

$$\ell(\theta|Y) = -\frac{n}{2}\ln \mathrm{SSQ}(\theta;Y) + C \tag{47}$$

The C absorbs all the constant terms (that do not depend on the parameters), which do not concern us in finding the maximum of the likelihood function. The advantage is that the error variance σ^2 falls out of the equation, and does not need to be estimated or guessed.

If we know the variance σ^2 (or can take a good guess), we can also simplify the log-likelihood function:

$$\ell(\theta|Y;\sigma) = -\frac{\mathrm{SSQ}(\theta;Y)}{2\sigma^2} + C \tag{48}$$

This is a rather interesting result: we cannot add SSQs for different types of data, but we *can* add them after dividing them by their respective error variances.

3.6 Evaluation

The maximum likelihood estimates for parameters, and their confidence intervals, are only representative when both the process model and the statistical model are 'true.' So if either of these models is poorly supported, the estimates and especially their intervals should be considered as 'approximate', at best.

In the situation of dynamic modelling, the assumptions of independence and randomness of the errors probably hurt most. In a biological setting, we can measure body size, reproductive output and survival with good accuracy. Measurement errors are not the main cause of the deviations between model and data. Far more important will be that our animals are different, and that there is a degree of stochasticity in their behaviour. When we follow the same group of individuals over time, independence is obviously severely compromised.

For the endpoint survival, these problems are not very important. Even though we follow the same group of individuals over time, we only take one observation for each individual: the interval in which it dies. As long as we can assume that the probability to die is independent for each interval, we are fine. Exception would be the situation where the death of one individual affects the probability to die for the remaining individuals in the same container.

At this moment, I am not aware of good (practical) solutions to the problems of dependence and non-randomness of the errors. A proper approach probably involves the definition of stochastic models for the individual, quantifying the differences between individuals, and large numbers of Monte Carlo simulations of possible outputs. This is clearly

an area that needs more work. In the mean time, we have to make do with 'approximations'.

References

[1] G. E. P. Box and G. C. Tiao. *Bayesian inference in statistical analysis*. John Wiley & Sons, Inc, New York, US, 1992.

[2] P. Doucet and P. B. Sloep. *Mathematical modeling in the life sciences*. Ellis Horwood Limited, London, UK, 1992.

[3] T. Jager, C. Albert, T. G. Preuss, and R. Ashauer. General Unified Threshold model of Survival - a toxicokinetic-toxicodynamic framework for ecotoxicology. *Environmental Science & Technology*, 45:2529–2540, 2011.

[4] W. Q. Meeker and L. A. Escobar. Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, 49(1):48–53, 1995.

[5] Y. Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, Oxford, UK, 2001.