2.1 Fitting to Data (or empirical modelling)

Given a set of data, useful formulae can be reached simply by finding relationships which fit the data. (This what Kepler did, for example). Simple relations often exist where measurements on a set of data are governed by the same physical processes. The most common are as follows:

- **Linear relations** have the form
  
  \[ y = A + Bx \]
  
  for “constants” A and B with “variables” x and y. In a graph of y against x, data points are approximated by a straight line with intercept at y=A and slope B.
  
  Linear Relations are typical in “cause and effect” situations, such as stretching an elastic band, force and acceleration, etc. (where else may they arise?)

- **Exponential relations** have the form
  
  \[ y = Ce^{dx} \]
  
  for ”constsnts” c and d, with ”variables” x and y (or alternatively \( x = C' e^{dy} \) with y and x transposed). Note that we could also write the exponential relation as
  
  \[ y = C \times D^x \quad \text{for} \quad D = e^{\frac{d}{x}} \quad \text{or} \quad d = \ln DTaking \logarithms:lny = \ln C + dx \]
  
  In a graph of ln y against x, data points are approximately by a straight line with intercept \( \ln y = \ln c \) and slope d.

  Exponential Relations are typical on cases of growth (eg population growth) or decay or relaxation to a “stable” state, such as radioactive decay, terminal velocity on free-fall, etc (where else might they arise?)

- **Power-laws** have the form \( y = Fx^g \)
  
  (for “constants” F and g and “variables” x and y). Taking logarithms \( \ln y = \ln F + g \ln x \)
  
  In a graph of ln y against ln x, data points are approximated by a straight line with intercept \( \ln y = \ln C \) and slope g.

  Power Laws are typical in many physical situations where two (or more) aspects of the same phenomenon are being compared. For example period and radius of planetary orbits, etc. (where else might they arise?)
- **Mixed relationships**

  Mixed relationships, such as

  \[ y = A + Bx^c \quad \text{or even} \quad y = Ae^{bx} + Cx^d + E \]

  are less easy to identify. An example is given in Question 4 of problem sheet 1, where successively improved approximations are used to arrive at "Bode’s Law".

  **Example (Kepler’s third law, 1619)**

  Find a relationship between \( T \) and \( R \) that fits the data

<table>
<thead>
<tr>
<th>Planet</th>
<th>Mean distance from sun ( k(10^9 m) )</th>
<th>Period of orbit ( T ) (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>57.9</td>
<td>88</td>
</tr>
<tr>
<td>Venus</td>
<td>108.2</td>
<td>225</td>
</tr>
<tr>
<td>Earth</td>
<td>149.6</td>
<td>365</td>
</tr>
<tr>
<td>Mars</td>
<td>227.9</td>
<td>687</td>
</tr>
<tr>
<td>Jupiter</td>
<td>778.3</td>
<td>4329</td>
</tr>
<tr>
<td>Saturn</td>
<td>1427</td>
<td>10753</td>
</tr>
<tr>
<td>Uranus</td>
<td>2870</td>
<td>30660</td>
</tr>
<tr>
<td>Neptune</td>
<td>4497</td>
<td>60150</td>
</tr>
<tr>
<td>Pluto</td>
<td>5907</td>
<td>90670</td>
</tr>
</tbody>
</table>

  (a) Try linear \( T = A + BR \)

  (b) Try exponentials \( T = Ce^{dR} \quad \text{or} \quad R = C' e^{dT} \)

  (c) Try power-law \( T = FR^8 \)

  Only the latter gives a good straight-line fit with

  \[ \ln T = 41.69 + 1.50 \ln k \quad [\text{exercise}] \]

  or

  \[ T = 0.192 \times R^3 \quad [\text{exercise}] \]
showing that the period of the orbit varies with the mean radius for the power of 3/2. (what are the units of the constant, 0.192, above?)

**Note:** In physically based systems, power-laws commonly have a simple exponents (such as 3/2 in Kepler’s 3rd Law). Questions 2 and 3 of problem-sheet 1 give further examples.

**Fitting model parameters to data**

Drawing a graph from some data and checking that it is a straight is fairly straightforward. Its much easier to recognise a straight line than and other curve! Doing it ‘by hand’ enables us to recognise if any points don’t fit the general pattern, and we might check the data.

There are automated ways of fitting a straight line to data. It all depends on what you consider the best fit. The most straight forward way to fit the line \( y = mx + c \) to several points \((x_i, y_i)\) is to minimise the sum of the squared error in the \( y \) values. That is find \( m \) and \( c \) which minimises

\[
(mx_1 + c - y_1)^2 + (mx_2 + c - y_2)^2 + \cdots + (mx_n + c - y_n)^2.
\]

(Why square the error?) This has the advantage of being reasonably easy to do. The linear regression function on your calculator (read the manual) or spreadsheet does it this way. You should never use this blindly, at least plot the data and see if it looks like a straight line and if there are any suspect points.

There is a statistical theory of fitting a straight line (or other models) to data which takes account of the distribution of errors (see statistics courses eg 154,258), and gives a measure of how well the model fits.

Just one more point to consider, minimising the error in the predicted \( y \) values seems a good idea if we want to use known \( x \) values to predict \( y \) values, but we would get a different answer if we minimised the error in the \( x \) values, or the distance between the straight line and the data points.

**2.2 Modelling from Theoretical Principles**

Arguments based on reasonable assumptions about the factors that contribute to a situation can often be used to deduce the basic form that a model should take.

We will use some examples to illustrate this. Recall that the following steps are involved

- identifying the problem
- simplification and
- mathematical form and solution assumptions
- validation
Let us obtain a model for the number of stray dogs a park and then try to improve it:

- Park has a fixed size with entrances and exits.
- Dogs enter and leave the park.
- Dogs might die and puppies might be born in the park.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of dogs at time + $n(t)$ dogs</td>
<td></td>
</tr>
<tr>
<td>rate at which dogs enter park $e(t)$ dogs/sec</td>
<td></td>
</tr>
<tr>
<td>rate at which dogs leave park $l(t)$ dogs/sec</td>
<td></td>
</tr>
<tr>
<td>rate at which puppies are born $b(t)$ dogs/sec</td>
<td></td>
</tr>
<tr>
<td>rate at which dogs die $d(t)$ dogs/sec</td>
<td></td>
</tr>
<tr>
<td>size of park (constant) $A$ $m^2$</td>
<td></td>
</tr>
</tbody>
</table>

**Parameters**:

- birth and death negligible: $b = d = 0$

**Assumptions**:

- We must relate $n$ to $e$ and $l$ and $A$
- rate of increase of $n$ must be $e - l$ ($e$ increases $n$, $l$ decreases $n$) so:

$$\frac{dn}{dt} = e(t) - l(t)$$

Note 1: This is an example of a “conservation law”: dogs do not spontaneously materialise or disappear so that the number in a given space only changes as dogs enter or leave the space through its boundaries.

Note 2: The “dimensions” of each of the terms in this equation are the same, namely

$$[\frac{dn}{dt}] = \frac{[n]}{[t]} = \frac{dogs}{time}$$

$$[e] and [l] = \frac{dogs}{time}$$

(Square brackets mean ‘the dimensions of’). The equation is “dimensionally consistent”. Mathematical models do not make sense unless they are dimensionally consistent.
To solve this equation we would need to know how $e(t)$ and $l(t)$ vary with time and some “starting” value of $n(t)$ (e.g. $n(t_0) = 0$ at opening time) so that (exercise):

$$n(t^0) = n(t_0) + \int_{t_0}^{t} (e(t) - l(t))dt$$

– this formula would require further data or assumptions to carry further (Why?) (We will leave it in this form for now).

**What might be wrong with the assumptions made?**

(a) If $A$ was very big, say the size of Britain, then birth and death would be important: $b \neq 0$ and $d \neq 0$.

We must then have:

$$\frac{dn}{dt} = e(t) - l(t) + b(t) - d(t)$$

**Notes:** This is an example of a more general conservation law, that includes production and destruction. It is also dim. constant solving would need integrating:

$$n(t) = n(t_0) + \int_{t_0}^{t} (e(t) - l(t) + b(t) - d(t))dt$$

**Note:** That the most appropriate model probably depends very much on the size of $A$. For Britain, $e$ and $l$ are probably negligible so that we might simplify further by assuming $e = l = 0$ making the park a “closed system”.

(b) If $e$, $l$, $b$ and $d$ are finite then $n(t)$ is a “continuous” function of time. for example:

That is $n$ varies through the real numbers, not the natural numbers (ie 0,1,2,3...). What would the model mean if it predicted $n = 17.5392$ dogs at some moment?

- perhaps a dog is 0.5392 of its way through the entrance, or a puppy is in the middle of being born, or an old dog is 0.4608 dead!

- either: a more ”discrete” model (not continuous) is needed.

- or: the function $n(t)$ still has meaning as representing an approximation for the number of dogs at time $t$.

**Continuum Hypothesis:** when $n$ is very large the approximation that $n(t)$ varies smoothly (the continuum hypothesis) is reasonable.

(c) We could be sophisticated and make reasonable models for $e,l,b,d$ . . . but let’s use the ideas developed here to look at something simpler . . .

This example has illustrated:

- Conservation laws
Example 2
An experiment shows that on average a bacterium (found on milk, say) grows from maturity to the point where it divides to form two new immature bacteria in $T$ seconds. Use this to model an infection of your pinta.

- Identify the problem
- List simplifications and assumptions
- State model on mathematical form

There are two basic models that could be used to represent the multiplication of the bacteria:

1. a difference equation $n(t) = 2n(t - T)$
2. a differential equation $\frac{dn}{dt} = dn$

- What different assumptions are contained in each model?
- What are the solutions like?
- What value must the constant take if the two models are to agree?

**Significant features:** A bacteria doubles every $T$ sec, on average

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Symbol</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of bacteria at time $t$</td>
<td>$n(t)$</td>
<td>bugs</td>
</tr>
<tr>
<td>mean doubling time</td>
<td>$T$</td>
<td>sec</td>
</tr>
</tbody>
</table>

**Simplification:** (seems simple enough)

**Assumptions:** Let’s suppose population always exactly doubles in $T$ sec (precisely $T$ sec after any moment in time every bacterium will have divided just once, exactly doubling $n$).

**Mathematical form:** $n(t) = 2n(t - T)$ Alternatively, writing $t = t_0 + kT$ we obtain: $n(t) = 2^{\frac{t-t_0}{T}}n(t_0)$

**Solution:** Given an initial number of bacteria at time $t_0$, $n(t_0)$ we can solve to find $n(t_0 + T)$, $n(t_0 + 2T)$ etc. $n(t_0 + T) = 2n(t_0)$

$n(t_0 + 2T) = 2n(t_0 + T) = 4n(t_0)$

$n(t_0 + kT) = 2^k n(t_0) \quad \text{for } k = 0, 1, 2, 3 \ldots$

**Some points to Note:**
1. This holds for whole-number values of $\frac{t-t_0}{T}$ only. (Why?)

2. With this model, $n(t)$ will always take integer values if $n(t_0)$ is an integer. (Why?)

3. We need to know $n(t_0)$ for all values of $t_0$ over one interval of time of length $T$ to predict $n(t)$ at all later times. That is, at any time $t_0$ we must know not only the number but the age-distribution of the bacteria to predict subsequent growth at all later times.

**Note:** This is also seen on the growth of the human population, eg. the baby booms of 40’s, 50’s 60’s 80’s . . .

**Criticism:** Individual bacteria would not on general divide every $T$ seconds precisely. They only do so on “average”. The fact that they don’t tends to “smooth out” the age distribution. Bacteria are so numerous and multiply so rapidly that there is not likely to be much variation in age distribution quite soon after they start dividing.

**New Assumptions:**

- $n(t)$ might not be the precise number of bugs but closely estimates the number (we hope).

- Age distribution does not vary, so that the growth-rate at any time is proportional to the total population at that same time.

**Mathematically:**

\[
\frac{dn}{dt} = \alpha n
\]

**Solution:** (Exercise) is

\[
n(t) = n(t_0)e^{\alpha(t-t_0)}
\]

given that the population is $n(t_0)$ at time $t_0$.

**Observations:**

1. This solution changes continuously with time (continuum hypothesis)

2. If $t$ increases by $T$ the value of $n$ must double and so we must have:

\[
n(t_0 + T) = 2n(t_0) = n(t_0)e^{\alpha T}
\]

so that

\[
e^{\alpha T} = 2 \quad \text{or} \quad \alpha = \frac{1}{T} \ln 2
\]

substituting for $\alpha$ thus gives [exercise]

\[
n(t) = n(t_0)2^{(t-t_0)/T}
\]