A Brief Introduction to Fitting Models and Parameter Estimation

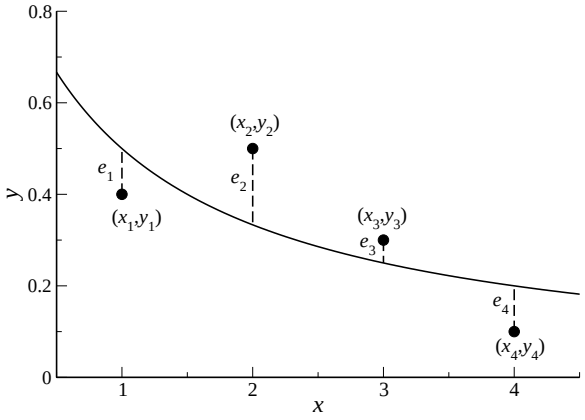
Least Squares Fitting

Suppose we have N pairs of observations, $(x_1, y_1), \ldots (x_N, y_N)$ and we expect x and y to be related according to some function, g, which itself involves one or more parameters, which we write as the vector $\mathbf{a}$.

$$y = g(x; \mathbf{a}).$$

Typically, the points (x_i, y_i) won't fall exactly on the curve. (Why? We might have **measurement error** or perhaps the relationship doesn't exactly follow the function given). We need to find the value of the parameter vector $\mathbf{a}$ that makes the function $g(x; \mathbf{a})$ provide the best fit to the data points.

One measure of how well the data fits is to take the difference between the predicted value of y and the observed value of y (we call this the error), square it (to get the squared error), and add up all these quantities for each of the data points (the sum of squared errors, or the error sum of squares, often abbreviated as the **sum of squares**).



(Why don't we just work with the unsquared errors? Positive and negative differences can cancel. Why not the absolute values of the errors? It turns out the math is easier.)

For our data point (x_i, y_i), the predicted value of y is $g(x_i; \mathbf{a})$ so our sum of squares is

$$L(\mathbf{a}) = \sum_{i=1}^{N} \left(y_i - g(x_i; \mathbf{a})\right)^2 .$$

Obviously, the sum of squares depends on the parameter vector $\mathbf{a}$, so we write L as a function of $\mathbf{a}$.

Our task is now to find the minimum of $L(\mathbf{a})$ over different values of $\mathbf{a}$. In general, this is not a straightforward task. If the function g is **linear** in its parameters then we can make analytic progress. One example of this is simple linear regression.

Linear Regression

Suppose we have N pairs of observations, $(x_1, y_1), \ldots (x_N, y_N)$ and we expect x and y to be linearly related:

$$y = ax + b.$$

We want to minimize the sum of squares

$$L(a, b) = \sum_{i=1}^{N} \left(y_i - (ax_i + b)\right)^2,$$

over the parameters a and b. This is a simple problem, often seen in multivariable calculus courses. We solve

$$\frac{\partial L}{\partial a} = 0 \quad \text{and} \quad \frac{\partial L}{\partial b} = 0.$$

Let's first look at $\partial L/\partial b$.

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \sum_{i=1}^{N} 2(-1)\left(y_i - (ax_i + b)\right) \\
&= -2\sum_{i=1}^{N} \left(y_i - ax_i - b\right) \\
&= -2\left(\sum_{i=1}^{N} y_i - a\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} b\right) \\
&= -2\left(\sum_{i=1}^{N} y_i - a\sum_{i=1}^{N} x_i - bN\right).
\end{aligned}$$

Setting the partial derivative equal to zero we get

$$\begin{aligned}
0 &= \sum_{i=1}^{N} y_i - a\sum_{i=1}^{N} x_i - bN \\
\Rightarrow \sum_{i=1}^{N} y_i &= a\sum_{i=1}^{N} x_i + bN \qquad (1) \\
\Rightarrow \frac{1}{N}\sum_{i=1}^{N} y_i &= \frac{a}{N}\sum_{i=1}^{N} x_i + b \\
\Rightarrow \overline{y} &= a\overline{x} + b. \qquad (2)
\end{aligned}$$

Here, $\overline{y}$ denotes the average of the y_i and $\overline{x}$ denotes the average of the x_i.

We see that the best fitting straight line passes through the "average" of the data points, $(\overline{x}, \overline{y})$.

Looking at the other partial derivative:

$$\frac{\partial L}{\partial a} = \sum_{i=1}^{N} -2x_i\left(y_i - (ax_i + b)\right)$$

$$
\begin{aligned}
&= -2\sum_{i=1}^{N}\left(x_i y_i - a x_i^2 - b x_i\right) \\
&= -2\left(\sum_{i=1}^{N} x_i y_i - a\sum_{i=1}^{N} x_i^2 - b\sum_{i=1}^{N} x_i\right).
\end{aligned}
$$

Setting the partial derivative equal to zero gives

$$
\begin{aligned}
\sum_{i=1}^{N} x_i y_i &= a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i \qquad (3) \\
\frac{1}{N}\sum_{i=1}^{N} x_i y_i &= \frac{a}{N}\sum_{i=1}^{N} x_i^2 + \frac{b}{N}\sum_{i=1}^{N} x_i \\
\Rightarrow \overline{xy} &= a\overline{x^2} + b\overline{x}. \qquad (4)
\end{aligned}
$$

Here, $\overline{xy}$ means the average value of $x_i y_i$ and $\overline{x^2}$ is the average value of x_i^2. Notice that in general, $\overline{xy} \neq \overline{x} \cdot \overline{y}$ and $\overline{x^2} \neq (\overline{x})^2$.

Equations (1) and (3) (or the pair 2 and 4) are a pair of simultaneous linear equations for a and b and can be written as the matrix equation

$$
\begin{pmatrix} \sum x_i & N \\ \sum x_i^2 & \sum x_i \end{pmatrix}
\begin{pmatrix} a \\ b \end{pmatrix}
= \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.
$$

In this form, the equations are usually referred to as the **normal equations**. (Here, all sums are taken from $i = 1$ to $i = N$.) Provided that the matrix on the left hand side is invertible, these are easily solved for a and b.

Equivalently, we can work with equations 2 and 4 and get

$$
\begin{pmatrix} \overline{x} & 1 \\ \overline{x^2} & \overline{x} \end{pmatrix}
\begin{pmatrix} a \\ b \end{pmatrix}
= \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix}.
$$

These are just the normal equations divided by N. Either set can be solved to give

$$
\begin{aligned}
a &= \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \\
b &= \frac{\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy}}{\overline{x^2} - \overline{x}^2} \\
&= \overline{y} - a\overline{x}.
\end{aligned}
$$

Thinking back to our problem of minimizing $L(a, b)$, we should also check that the point we get is a minimum (how do you do this?).

We now have our best fitting straight line.

***More General Linear Regression**

(A starred section is one that you would not be expected to know the details of for the test.)

It turns out that we can apply this procedure to fit a general linear model of the form

$$y_i = \sum_{j=1}^{k} b_j g_j(x_i) + e_i,$$

where the b_j are constants and the g_j are any functions (not necessarily linear) of x alone. By "linear" here we mean that the model is linear in the parameters b_j, not (necessarily) in x. We write the number of parameters as k.

Our linear regression model can be written in this form by taking $b_1 = b$, $b_2 = a$, $g_1(x) = 1$, $g_2(x) = x$. We could fit a quadratic model by adding $b_3 = c$ and $g_3(x) = x^2$, giving $y = b + ax + cx^2$.

The solution for the "best fitting" set of parameters b_i is usually obtained via a matrix calculation. Various similar quantities are grouped together into matrices and vectors. The N observations get stacked up in the vector $\mathbf{y}$ and the k parameters in the vector $\mathbf{b}$. The (i, j) entry of the $N \times k$ matrix X is set equal to $g_j(x_i)$. This allows the model to be written as

$$\mathbf{y} = X\mathbf{b} + \mathbf{e},$$

where $\mathbf{e}$ is a vector containing the N "errors".

As in the linear regression analysis, we can find the best fitting value of $\mathbf{b}$, often written as $\hat{\mathbf{b}}$ by calculating the partial derivatives of the function $L(\mathbf{b})$ with respect to each of the b_j, setting the derivatives equal to 0 and solving. It turns out that this gives the following set of normal equations

$$X^{\mathrm{T}} X \hat{\mathbf{b}} = X^{\mathrm{T}} \mathbf{y}.$$

Notice that the normal equations are **linear** for the general linear model.

If the matrix $X^T X$ is invertible, the normal equations can be solved to give

$$\hat{\mathbf{b}} = \left(X^T X\right)^{-1} X^T \mathbf{y}.$$

The matrix $X^T X$ has (i, j) entry equal to $\sum_k g_i(x_k) g_j(x_k)$ and the matrix $X^T \mathbf{y}$ has entries equal to $\sum_k g_i(x_k) y_k$.

Fitting Different Models

(not starred!)

Imagine fitting the straight line $y = a + bx$ and the quadratic $y = a + bx + cx^2$ to the same set of data. Which would fit better? (In other words, have a smaller minimum sum of squares?)

The quadratic must fit better. Why is that? Because the straight line is a special case of the quadratic (with $c = 0$), so the best fitting quadratic can never do any worse than the best fitting line. The quadratic is more flexible in this sense.

This is a general point: if we have a more flexible function, we will be better able to fit data. If we took higher and higher degree polynomials, we could improve further. (A polynomial of degree n means that there will be $n + 1$ simultaneous linear equations to solve.)

We can find a straight line that perfectly fits a data set consisting of any two points (unless they lie on a vertical line). We can find a quadratic equation that perfectly fits (almost any) data set consisting of three points. And so on...

By making our function sufficiently complex we can often make it pass through all of the points, even though the data contains error (e.g. measurement error). The point is that complex functions are prone to **overfitting** data. (Remember the point about us preferring simple models over complex ones?)

Statisticians have a methodology to answer the following questions raised by these considerations:

- Does one model fit "significantly" better than another?
- Does the improved fit justify the additional complexity of the model?

Nonlinear Least Squares: Moving Beyond Linear Models

What happens if we try to fit an exponential function directly to the population growth data? (Earlier on we talked about converting the exponential growth of the population data to linear growth by logging the data.)

To simplify, let's just fit the model $y = e^{ax}$ to some data. (Imagine we can set the initial value equal to one.) Then we have

$$\begin{aligned} L(a) &= \sum (y_i - e^{ax_i})^2 \\ \Rightarrow \frac{\partial L}{\partial a} &= -2 \sum x_i e^{ax_i} (y_i - e^{ax_i}) \\ &= -2 \left(\sum x_i y_i e^{ax_i} - \sum x_i e^{2ax_i} \right). \\ \frac{\partial L}{\partial a} = 0 &\Rightarrow \sum x_i y_i e^{ax_i} = \sum x_i e^{2ax_i}. \end{aligned}$$

Can we solve for a? Looks unlikely: the normal equations are **nonlinear**, with a appearing in the exponent of a sum of (up to) $2N$ exponential terms.

The key difference is that a polynomial is **linear** in its **parameters**, the exponential is **nonlinear** in its **parameter**. This leads to the normal equations being linear (easy to solve) or nonlinear (difficult to solve), respectively.

For a nonlinear model we typically cannot solve the least squares problem analytically. Instead we turn to numerical approaches.

Numerical optimization is, in general, a non-trivial problem. Finding a minimum is easy if a function has a single minimum— you just vary the parameter value so that you move "downhill" towards the minimum. More generally, functions have one or more local minima in addition to their global minimum. It is then not easy for an optimization routine to figure out whether a minimum is just a local one or is the global minimum.

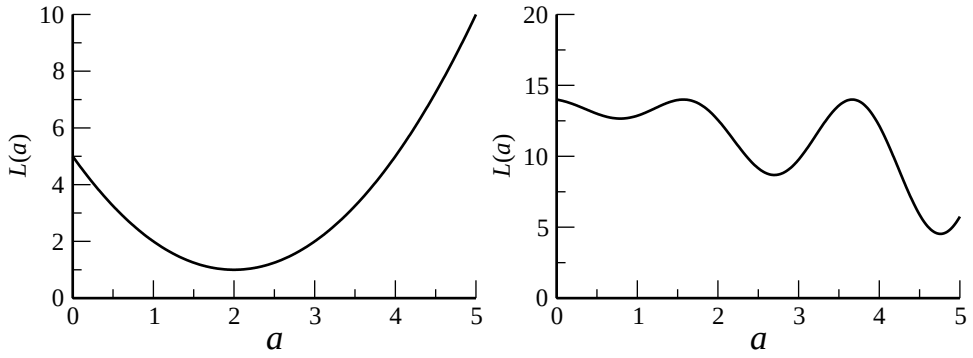


Figure 1: The minimum of the function in the left hand panel is easily found, but the local minima in the right hand panel make the process more difficult.

People spend their entire careers working on minimization problems, but this isn't a course on numerical analysis and optimization. We will follow the many people who typically (and often over-optimistically) don't spend too much time worrying about this. Usually, they employ a pre-packaged optimization routine and hope that everything will be OK. MATLAB provides several such routines, and `fminsearch` seems to be a popular choice. We will soon discuss using MATLAB to fit population growth models.

Other points:

- Finding a minimum may become more difficult if our model has more parameters since we have a larger "parameter space" to explore.
- What does the shape of $L(\mathbf{b})$ near its minimum tell us about how accurately we can find the values of parameters?