Linear Regression

In HW problem, wrote routine to fit a parabola to 3 points. In this case there was only one answer.

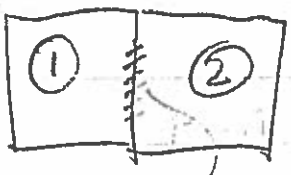
Often in data analysis you want to fit a curve to many data points

How do we do this? Use (usually) linear regression

This does not mean that we fit data with a line, rather that we use an algorithm which reduces the problem to a set of linear equations!

Let's look at a simple problem: measure mass transfer coefficients

Suppose we have a membrane between two reservoirs



membrane area = A

We start with some concentration in reservoir ① of C_0 and we assume that reservoir ② is maintained at a constant C_{eq}. If the volume of reservoir ① is V, then:

$$V \frac{\partial C}{\partial t} = -Ah(C - C_{eq})$$

where $C\big|_{t=0} = C_0$

We can solve this equation:

$$C = C_{eq} + (C_0 - C_{eq})\, e^{-\left(\frac{htA}{V}\right)}$$

We wish to determine h by measuring C as a function of time

In order for us to use linear regression, the model must be linear in the modelling parameter,

In this case, the modelling parameter is h. We can rewrite the eqn:

$$(C - C_{eq}) = (C_0 - C_{eq})\, e^{-\frac{hA}{V}t}$$

$$\ln(C - C_{eq}) = \ln(C_0 - C_{eq}) - \frac{hA}{V}t$$

So if we plot $\ln(C - C_{eq})$ vs $\frac{At}{V}$

we get a line w/ intercept of $\ln(C_0 - C_{eq})$ and a slope of h!

How do we get the best fit line?

We look at the deviation of the points from the line, and try to minimize this in some way.

Let's take:

$$y \equiv \ln(C - C_{eq}), \quad a \equiv \frac{Ah}{V}.$$

$$b \equiv \ln(C_0 - C_{eq})$$

We want to fit a series of points (t_i, y_i) by the model $y = at + b$

We shall use the method of least squares. We form the sum:

$$\text{Sum} = \sum_{i=1}^{N} (y_i - (at_i + b))^2$$

which is the sum of the squared distance between data pt. and model in the y direction.

We pick a & b so that this is a minimum

We let:

$$\left.\begin{aligned} \frac{\partial \text{sum}}{\partial a} &= 0 \\ \frac{\partial \text{sum}}{\partial b} &= 0 \end{aligned}\right\} \text{2 eqns for } a, b!$$

$$\frac{\partial \text{sum}}{\partial a} = \sum_{i=1}^{N} -2(y_i - (at_i + b))t_i$$

$$= \sum_{i=1}^{N} -2y_i t_i + 2a\sum_{i=1}^{N} t_i^2 + 2b\sum_{i=1}^{N} t_i = 0$$

and:

$$\frac{\partial \text{sum}}{\partial b} = \sum_{i=1}^{N} -2\left(y_i - (at_i + b)\right)$$

$$= \sum_{i=1}^{N} -2y_i + 2a\sum_{i=1}^{N} t_i + 2Nb = 0$$

This is a pair of linear equations for a & b!

We have the solution:

$$a\overline{t^2} + b\bar{t} = \overline{yt} \qquad \left(\text{let } \bar{x} \equiv \frac{1}{N}\sum_{i=1}^{N} x_i\right)$$

$$a\bar{t} + b = \bar{y}$$

$$\therefore\ a = \frac{\overline{yt} - \bar{y}\,\bar{t}}{\overline{t^2} - \bar{t}^2}$$

and $b = \bar{y} - a\bar{t}$

which can be used to get the mass transfer coefficient!

This is not the only way to approach the problem. Let's look at it as a system of equations!

We are trying to satisfy:

$$y_1 = a t_1 + b$$
$$y_2 = a t_2 + b$$
$$\vdots$$
$$y_N = a t_N + b$$

Let's write this in <u>matrix form</u>:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \approx \begin{pmatrix} t_1 & 1 \\ t_2 & 1 \\ \vdots & \vdots \\ t_N & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

we shall use $\tilde{X}$ to be the vector containing the modelling parameters:

$$X_1 \equiv a \quad X_2 \equiv b$$

If we let the matrix of modelling functions be $\underset{\approx}{A}$ (e.g. $A_{11} = t_1$, etc.)

then we want to satisfy:

$$\underset{\sim}{y} \approx \underset{\approx}{A}\,\underset{\sim}{x}$$

Actually, we want to minimize the residual:

$$\underset{\sim}{r} = \underset{\sim}{y} - \underset{\approx}{A}\,\underset{\sim}{x}$$

Or, for least squares, adjust $\underset{\sim}{x}$ s.t.:

$\|\underset{\sim}{r}\|_2 \equiv \|\underset{\sim}{y} - \underset{\approx}{A}\,\underset{\sim}{x}\|_2$ is as small as possible

↑ Euclidean norm

We want to pick $\underset{\sim}{x}$ s.t.

$$\underset{\sim}{r}^T\underset{\sim}{r} \equiv (\underset{\sim}{y} - \underset{\approx}{A}\,\underset{\sim}{x})^T(\underset{\sim}{y} - \underset{\approx}{A}\,\underset{\sim}{x})$$

is a minimum.

This is exactly equivalent to what we did before!

Let's multiply this out:

$$r^2 \equiv \underset{\sim}{y}^T\underset{\sim}{y} - \underset{\sim}{y}^T\underset{\approx}{A}\underset{\sim}{x} - \underset{\sim}{x}^T\underset{\approx}{A}^T\underset{\sim}{y} + \underset{\sim}{x}^T\underset{\approx}{A}^T\underset{\approx}{A}\underset{\sim}{x}$$

We take the gradient of this w.r.t. $\underset{\sim}{x}$ and set it equal to zero!

$$\underset{\sim}{\nabla}_{\underset{\sim}{x}} r^2 = 0 = -2\underset{\approx}{A}^T\underset{\sim}{y} + 2\underset{\approx}{A}^T\underset{\approx}{A}\underset{\sim}{x} = 0$$

Thus we have the problem:

$$\underset{\approx}{A}^T\underset{\approx}{A}\underset{\sim}{x} = \underset{\approx}{A}^T\underset{\sim}{y} \quad !$$

This gives us a set of $\underline{\underline{n}}$ eq'ns for the n unknowns in $\underset{\sim}{x}$!

These are known as the normal equations.

Linear Regression Error

Let's apply all this back to linear regression. Suppose we have some data and we are trying to fit it, and calc. the uncertainty in the modelling parameters.

We measu… a lead weight … meas. its pos'n a…

We get

t_i (s…	
0	
1	1.8
2	20.7
3	15.9
4	2.0

Skip this (go straight to vector approach)

We fit this to the equation:

$$h = h_0 + V_0 t - \frac{1}{2} g t^2$$

We want to both calculate g, and determine the uncertainty in g.

First let's get g. We have the problem:

$$h_i = (1)\,h_0 + (t)\,V_0 + \left(-\tfrac{1}{2}t^2\right) g$$

$h_0 \to x_1$, $V_0 \to x_2$, $g \to x_3$

So:

$$\underset{\approx}{A}\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & -.5 \\ 1 & 2 & -2 \\ 1 & 3 & -9/2 \\ 1 & 4 & -8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \underset{\sim}{h}\begin{pmatrix} 0.3 \\ 14.8 \\ 20.7 \\ 15.9 \\ 2.0 \end{pmatrix}$$

Using the normal eq'ns we get:

$$\underset{\approx}{A}^T \underset{\approx}{A}\, \underset{\sim}{x} = \underset{\approx}{A}^T \underset{\sim}{h}$$

$$\underset{\sim}{x} = \left[\underset{\approx}{A}^T \underset{\approx}{A}\right]^{-1} \underset{\approx}{A}^T \underset{\sim}{h}$$

Or numerically:

$$\underset{\sim}{X} = \begin{pmatrix} .886 & .257 & -.086 & -.143 & .086 \\ -.771 & .186 & .571 & .386 & -.371 \\ -.286 & .143 & .286 & .143 & -.286 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \end{pmatrix}$$

This yields:

$$x_1 = 0.197, \quad x_2 = 19.7 \quad x_3 = 9.64$$

Now we want to examine the error in the position measurements h_i.

To do this we look at the deviation:

$$s_h^2 \approx \frac{1}{N-3} \sum_{i=1}^{N} \left(h_i - h(t_i) \right)^2$$

$\frac{1}{N-3}$: because 3 adjustable parameters det. from data

$h(t_i)$: using fitted parameters

The quantity $h_i - h(t_i)$ is just the residual;

$$\underset{\sim}{r} = \underset{\sim}{h} - \underset{\approx}{A}\,\underset{\sim}{x} \quad !$$

$$\therefore\ s_h^2 = \frac{1}{N-3}\,\|\underset{\sim}{r}\|_2^2 = 0.110$$

↳ you should really have more pts for accuracy!

Ok, how do we use this to get the error in $g(x_3)$?

We had:

$$\underset{\sim}{x} = \left\{\left[\underset{\approx}{A}^T\underset{\approx}{A}\right]^{-1}\underset{\approx}{A}^T\right\}\underset{\sim}{h}$$

$$\uparrow$$

$$\equiv \underset{\approx}{K}$$

Thus $x_i \equiv \sum_{j=1}^{N} K_{ij}\, h_j$

(K_{ij}: i'th row of $\underset{\approx}{K}$!)

We can thus use the formula for a linear combination of random variables!

$$S_{x_i}^2 = \sum_{j=1}^{N} K_{ij}^2 S_{h_i}^2 = 0.0315$$

these are all the same in this problem!

Thus $g = 9.64 \pm 0.18$

to the 69% confidence level ($1\bar{\sigma}$)

Often we want to develop correlations for their predictive value, for example what is the expected ht. of the lead wt. at some time t?

$$h_m = X_1 + X_2 t - \tfrac{1}{2} t^2 X_3 \; !$$

But what is the error? We have a linear combination of X_1, X_2, X_3 each of which have some error.

You might expect that:

$$S_{h_m}^2 = S_{x_1}^2 + (t)^2 S_{x_2}^2 + \left(-\tfrac{1}{2}t^2\right)^2 S_{x_3}^2$$

But this is incorrect!!

The problem is that x_1, x_2, x_3 are not independent! Their covariance is non-zero!

Thus we should have:

$$S_{h_m}^2 = S_{x_1}^2 + (t)^2 S_{x_2}^2 + \left(-\tfrac{1}{2}t^2\right)^2 S_{x_3}^2$$
$$+ 2\,(1)(t)\, S_{x_1 x_2}^2 + 2\,(1)\left(-\tfrac{1}{2}t^2\right) S_{x_1 x_3}^2$$
$$+ 2\,(t)\left(-\tfrac{1}{2}t^2\right) S_{x_2 x_3}^2$$

How do we calculate the covariance?

Recall:

$$x_i \equiv \sum_{j=1}^{N} K_{ij}\, h_j$$

where $\underset{\approx}{K} \equiv \left[\underset{\approx}{A}^T \underset{\approx}{A}\right]^{-1} \underset{\approx}{A}^T$

Thus we let:

$$x_1 = \sum_{j=1}^{N} \alpha_j h_j$$ where α is 1st row of $\underset{\approx}{K}$

and

$$x_2 = \sum_{j=1}^{N} \beta_j h_j$$ where $\underset{\sim}{\beta}$ is the 2nd row of $\underset{\approx}{K}$

Now we need to determine the covariance

$$\sigma^2_{x_1 x_2} = E\left((x_1 - \mu_{x_1})(x_2 - \mu_{x_2})\right)$$

↑ value for x_1 & x_2 expected for an infinite sample

$$\mu_{x_1} = \sum_{j=1}^{N} \alpha_j \mu_{h_j}$$

↑ from infinite sample

$$\mu_{x_2} = \sum_{j=1}^{N} \beta_j \mu_{h_j}$$

So:

$$\sigma^2_{x_1 x_2} \equiv E\left(\left(\sum_{j=1}^{N} \alpha_j (h_j - \mu_{h_j})\right)\left(\sum_{j=1}^{N} \beta_j (h_j - \mu_{h_j})\right)\right)$$

Now if all of the h_j are *independent* then

$$E\left((h_j - \mu_{h_j})(h_i - \mu_{h_i})\right) = 0$$

for $i \neq j$

$$\therefore \sigma^2_{x_1 x_2} = E\left(\sum_{j=1}^{N} \alpha_j \beta_j (h_j - \mu_{h_j})^2\right)$$

$$= \sum_{j=1}^{N} \alpha_j \beta_j \sigma_h^2$$

↑ provided all the h_j have the same uncertainty

$$\therefore \sigma^2_{x_1 x_2} = \underset{\sim}{\alpha}\, \underset{\sim}{\beta}^T \sigma_h^2$$

So for this problem if we let

$\underset{\sim}{\gamma} \equiv 3^{rd}$ row of $\underset{\approx}{K}$

$$x_1 = \underset{\sim}{\alpha}\,\underset{\sim}{h}\,, \quad x_2 = \underset{\sim}{\beta}\,\underset{\sim}{h}\,, \quad x_3 = \underset{\sim}{\gamma}\,\underset{\sim}{h}$$

$$s_h^2 = \frac{1}{N-3}\,\|\underset{\sim}{r}\|_2^2\,, \quad \underset{\sim}{r} = \underset{\sim}{h} - \underset{\approx}{A}\,\underset{\sim}{x}$$

$$s_{x_1}^2 = (\underset{\sim}{\alpha}\,\underset{\sim}{\alpha}^T)\, s_h^2$$

$$s_{x_2}^2 = (\underset{\sim}{\beta}\,\underset{\sim}{\beta}^T)\, s_h^2 \quad \text{etc.}$$

and:

$$h_m = (1)\,x_1 + (t)\,x_2 + (\tfrac{1}{2}t^2)\,x_3$$

$$\sigma_{h_m}^2 = (1)^2\sigma_{x_1}^2 + (t)^2\sigma_{x_2}^2 + (\tfrac{1}{2}t^2)^2\sigma_{x_3}^2$$
$$+ 2(1)(t)\,\underset{\sim}{\alpha}\,\underset{\sim}{\beta}^T\sigma_h^2 + 2(1)(\tfrac{1}{2}t^2)\,\underset{\sim}{\alpha}\,\underset{\sim}{\gamma}^T\sigma_h^2$$
$$+ 2(t)(\tfrac{1}{2}t^2)\,\underset{\sim}{\beta}\,\underset{\sim}{\gamma}^T\sigma_h^2$$
$$= \left\| \left[(1)\,\underset{\sim}{\alpha} + (t)\,\underset{\sim}{\beta} + (\tfrac{1}{2}t^2)\,\underset{\sim}{\gamma} \right] \right\|_2^2 \sigma_h^2$$