Department of Mechanical Engineering, University of Bath

Mathematics 2 ME10305

Problem sheet — Least Squares Fitting of data

1. The Maths 2 webpage has links to the last five years of exam questions. Bizarrely, I think that the Least Squares questions on those papers are the best place to start this problem sheet! So do have a go at these before any of the other questions here, for they will, at least, inform you of precisely what you should expect in terms of the length of a question and how much calculator work I will expect.

ANSWER: The outline solutions are posted in the same place.

---

2. In this question we are going to play a different game in the sense that the data to be fitted has a different type of randomness. I would like to see how much accuracy one can gain from a set of data which has been rounded quite severely. We'll use the conversion from miles to kilometres for this purpose. The following is a Table of data which is subject to different degrees of rounding off, namely to zero, 1, 2 and 3 decimal places; x denotes miles and yn denotes kilometres with n decimal places. For each of these sets of data, fit a straight line through the origin to determine the least squares version of the conversion factor between the units. How accurate are they? You may compare with the exact value, 1.609344.

x :	1	2	3	4	5	6	7	8	9	10
$y0$:	2	3	5	6	8	10	11	13	14	16
$y1$:	1.6	3.2	4.8	6.4	8.0	9.7	11.3	12.9	14.5	16.1
$y2$:	1.61	3.22	4.83	6.44	8.05	9.66	11.27	12.87	14.48	16.09
$y3$:	1.609	3.219	4.828	6.437	8.047	9.656	11.265	12.875	14.484	16.093

If this has piqued your interest, then try the same with the conversion between ounces and grams, for which 1 ounce is equal to 28.349523 grams. Use 16 sets of data from 1oz to 16oz, and use the same number of decimal places as in the above miles/kilometres example. The raw data may be found at

`http://staff.bath.ac.uk/ensdasr/ME10305.bho/ounces-to-grams.txt`

although there is also a link to it at the unit webpage. Given that there are 16 data points, you might be able to coerce Excel into doing your calculations for you.

ANSWER: We are fitting the slope of the line $y = mx$, where $x =$ miles and $y =$ kilometres. The formula is

$$m = \frac{\sum_{i=1,N} x_i y_i}{\sum_{i=1,N} x_i}.$$

It shouldn't take too long to find $\sum_i x_i y_i$, while $\sum_i x_i^2 = 385$. We get the following results:

DPs	$\sum_i x_i y_i$	m	|error|
0	614.000	1.594805	0.014539
1	620.000	1.610390	0.001046
2	619.580	1.609299	0.000045
3	619.591	1.609327	0.000017

Quite frankly, I am amazed that we should get an error as small as about 1.5% for the zero decimal places case.

As the number of decimal places in the raw data increases, there is a clear trend towards having a greatly improved accuracy. However, I have to warn you that this is not at all straightforward: if the number which is being sought has only a few decimal places itself, then it is quite possible for the round-off error in the raw data to be biased towards one side of the correct value. In other words, errors incurred by the rounding off might not have a zero mean and if that is true, then the accuracy of this approach is impaired.

With regard to the ounces/grams example, we get the following Table of results:

DPs	m	|error|
0	28.350267	0.000744
1	28.348195	0.001328
2	28.349532	0.000009
3	28.349521	0.000002

So again we get some pretty spectacular accuracy, although there is small aberration in that the zero-DP case is slightly more accurate than the 1-DP case.

I have to admit that this question was motivated by a question I asked myself as a child with nothing much to do on a cold winter's evening, without access to wikipedia and before calculators could be bought. I had noticed that the conversion factor corresponding to the numbers of ounces and numbers of grams was not the same between a 4oz slab of butter and a 12oz jar of jam. Both were quoted in whole numbers and the factors are 28.25 and 28.333..., respectively. By assuming that the ounces were precise and that the grams had been rounded, I tried to work out from the various foodstuffs in the house the range of possible values that the conversion factor could take. Each quoted number of grams is equivalent to a range of values which are rounded the same the way, and therefore one is able to get a range of possible conversion factors for each case. I eventually settled on 28.35, which wasn't too bad, and which is why I have always remembered the number.

---

3. Suppose that you were given a set of experimental data where it is suspected that the data should satisfy an equation of the form

$$y = a + b/x.$$

How could the data be manipulated in order to use standard Least Squares theory?

[Note: I can think of at least two different ways of doing this.]

What about the equation,

$$y = \frac{a}{x+b}?$$

ANSWER: The first way would be to multiply both sides of the given equation by x to get

$$xy = ax + b.$$

If now we define, $Y = xy$ and $X = x$, we can do a standard linear Least Squares fit to $Y = aX + b$ to find a and b, from which we can then find y as a function of x.

The second way would be to let $X = x^{-1}$ and $Y = y$, to obtain $Y = a + bX$, and hence we would find a and b using standard Least Squares theory..

I will leave it to you to decide whether these two ways would yield the same coefficients.

The equation, $y = a/(x + b)$ may be rearranged into the form,

$$x = \frac{a}{y} - b,$$

and then we may use one of the two methods above to find a and b.

---

4. In many experimental situations the observable, y, is a power law function of the parameter, x. In other words it takes the form,

$$y = ax^b,$$

where a and b need to be found. [For example, the rate of heat transfer from a hot vertical surface is proportional to the $\frac{1}{4}$ power of the temperature difference between the heated surface and the ambient conditions.] How would you convert this power-law relationship into a straight line relationship?

ANSWER: If one takes natural logarithms of the given equation, then it becomes,

$$\ln y = \ln a + b \ln x.$$

It is very very likely that both x and y will be single-signed for the type of experiments which satisfy such relationships, such as the quoted example where the both the rate of heat transfer and the temperature difference are positive. However, should x be negative, then we can always replace it by $x^* = -x$, and use x^* instead.

Therefore we may define $Y = \ln y$, $X = \ln x$, $A = \ln a$ and $B = b$, and use standard Least Squares theory to fit $Y = A + BX$.

---

5. Experimental measurements have been taken of z, which is a function of both x and y. It is suspected that z is a linear function of x and y, and therefore it represents a plane in 3D space. Use least squares theory to determine the three unknown coefficients in the following equation for the plane,

$$z = ax + by + c.$$

ANSWER: The answer for this question is no more difficult than the derivation for fitting a quadratic line. We need to minimise the sum of the squares of the residuals for $z = ax + by + c$:

$$S = \sum_{i=1}^{n} (z_i - ax_i - by_i - c)^2.$$

We need to set to zero in turn each of the three first partial derivatives of S with respect to a, b and c. This gives the following equation,

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} y_i^2 & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} y_i & N \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i z_i \\ \sum_{i=1}^{N} y_i z_i \\ \sum_{i=1}^{N} z_i \end{pmatrix}.$$

---

6. An obsessive cyclist has a comprehensive set of data for his/her ride-times over the same route for a period of several years. Naturally the cyclist's journey times are slower when the weather is colder, and faster when it is warmer. The cyclist wishes to determine (i) what the long term general trend is in terms of speed, (ii) what seasonal effect should be expected given the time of year. To this end, the cyclist proposes a least squares fit of the form,

$$T = a + bt + c\cos(2\pi t) + d\sin(2\pi t)$$

where T is the ride-time and t is time measured in years. Develop the least squares theory which will allow the cyclist to achieve his/her twin objectives.

ANSWER: The appropriate equation to solve is

$$\begin{pmatrix} N & \sum_{i=1}^{N} t_i & \sum_{i=1}^{N} \cos 2\pi t_i & \sum_{i=1}^{N} \sin 2\pi t_i \\ \sum_{i=1}^{N} t_i & \sum_{i=1}^{N} t_i^2 & \sum_{i=1}^{N} t_i \cos 2\pi t_i & \sum_{i=1}^{N} t_i \sin 2\pi t_i \\ \sum_{i=1}^{N} \cos 2\pi t_i & \sum_{i=1}^{N} t_i \cos 2\pi t_i & \sum_{i=1}^{N} \cos^2 2\pi t_i & \sum_{i=1}^{N} \sin 2\pi t_i \cos 2\pi t_i \\ \sum_{i=1}^{N} \sin 2\pi t_i & \sum_{i=1}^{N} t_i \sin 2\pi t_i & \sum_{i=1}^{N} \sin 2\pi t_i \cos 2\pi t_i & \sum_{i=1}^{N} \sin^2 2\pi t_i \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} T_i \\ \sum_{i=1}^{N} T_i t_i \\ \sum_{i=1}^{N} T_i \cos 2\pi t_i \\ \sum_{i=1}^{N} T_i \sin 2\pi t_i \end{pmatrix}.$$

Clearly the cyclist will wish to see that b is negative to indicate that the journey time is decreasing overall, and hence that the speed is improving.

The values of the coefficients, c and d, will allow the cyclist to determine at which point of the year that one might expect the journey time to be at its greatest. Given that,

$$c\cos(2\pi t) + d\sin(2\pi t) = \sqrt{c^2 + d^2}\,\cos(2\pi t - \phi),$$

where ϕ is a phase which satisfies both

$$\cos\phi = \frac{c}{\sqrt{c^2 + d^2}} \qquad \text{and} \qquad \sin\phi = \frac{d}{\sqrt{c^2 + d^2}},$$

it is clear that $\sqrt{c^2 + d^2}$ will be the expected variation in journey time from the mean during the year. In addition, the value of t which satisfies, $2\pi t = \phi$, will give that point in the year when the journey time is expected to be at its maximum. I suspect that that might very well be in January or February.

---

7. An experiment has two measurables, $y(x)$ and $z(x)$, as the control parameter, x, is varied. Both y and z should be linear with different slopes, but should have the same intercept on the vertical axis. That is, we wish to fit the following to the data, where there are three constants to find:

$$y = ax + c, \qquad z = bx + c.$$

How is this done? [This is a simplified version of a problem a couple of third year students brought to me where they had to fit a straight line to five measureables all of which had the same intercept. So this question isn't a product of my wild imaginings! What was the answer for their problem?]

ANSWER: The sums of the squares of the residuals may now be written in the form,

$$S = \sum_{i=1}^{N} \Big[\big(y_i - ax_i - c\big)^2 + \big(z_i - bx_i - c\big)^2 \Big]. \tag{1}$$

There are three constants to find, a, b and c, and therefore we will need to set all three first partial derivatives to zero. Here's the final answer:

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & 0 & \sum_{i=1}^{N} x_i \\ 0 & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i & 2N \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i x_i \\ \sum_{i=1}^{N} z_i x_i \\ \sum_{i=1}^{N} (y_i + z_i) \end{pmatrix}.$$

Note the presence of the $2N$ in the matrix.

Note: this question arose from a couple of third year Group Business and Design students' need to fit five lines with a common intercept. I hadn't seen such a type of least squares analysis before, but the extension to the above theory from two to five lines apparently worked perfectly for their data. If we were to write their five equations as $y^{(1)} = m^{(1)}x + c$, $y^{(2)} = m^{(2)}x + c$, and so on, then the required matrix vector system for the six constants would be:

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & 0 & 0 & 0 & 0 & \sum_{i=1}^{N} x_i \\ 0 & \sum_{i=1}^{N} x_i^2 & 0 & 0 & 0 & \sum_{i=1}^{N} x_i \\ 0 & 0 & \sum_{i=1}^{N} x_i^2 & 0 & 0 & \sum_{i=1}^{N} x_i \\ 0 & 0 & 0 & \sum_{i=1}^{N} x_i^2 & 0 & \sum_{i=1}^{N} x_i \\ 0 & 0 & 0 & 0 & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i & 5N \end{pmatrix} \begin{pmatrix} m^{(1)} \\ m^{(2)} \\ m^{(3)} \\ m^{(4)} \\ m^{(5)} \\ c \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i^{(1)} x_i \\ \sum_{i=1}^{N} y_i^{(2)} x_i \\ \sum_{i=1}^{N} y_i^{(3)} x_i \\ \sum_{i=1}^{N} y_i^{(4)} x_i \\ \sum_{i=1}^{N} y_i^{(5)} x_i \\ \sum_{j=1}^{5} \Big[\sum_{i=1}^{N} y_i^{(j)} \Big] \end{pmatrix}.$$

D.A.S.R. 25/03/2021