The Method of Least Squares

Steven J. Miller*

Department of Mathematics and Statistics
Williams College
Williamstown, MA 01267

Abstract

The Method of Least Squares is a procedure to determine the best fit line to data; the proof uses calculus and linear algebra. The basic problem is to find the best fit straight line $y = ax + b$ given that, for $n \in \{1, \ldots, N\}$, the pairs (x_n, y_n) are observed. The method easily generalizes to finding the best fit of the form

$$y \ = \ a_1 f_1(x) + \cdots + c_K f_K(x); \tag{0.1}$$

it is not necessary for the functions f_k to be linearly in x – all that is needed is that y is to be a linear combination of these functions.

Contents

1 Description of the Problem

Often in the real world one expects to find linear relationships between variables. For example, the force of a spring linearly depends on the displacement of the spring: $y = kx$ (here y is the force, x is the displacement of the spring from rest, and k is the spring constant). To test the proposed relationship, researchers go to the lab and measure what the force is for various displacements. Thus they assemble data of the form (x_n, y_n) for $n \in \{1, \ldots, N\}$; here y_n is the observed force in Newtons when the spring is displaced x_n meters.

*E-mail: sjm1@williams.edu

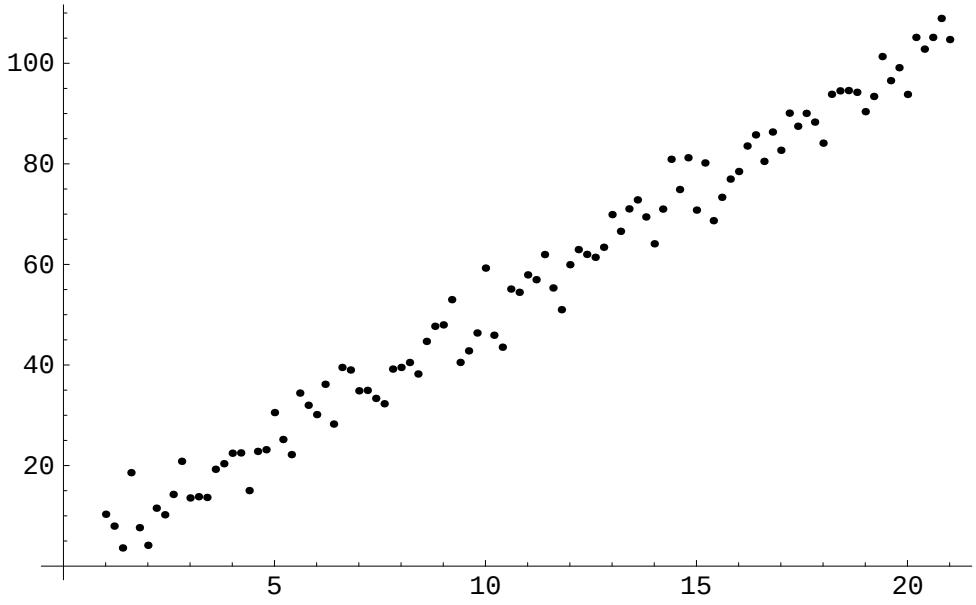

Figure 1: 100 "simulated" observations of displacement and force ($k = 5$).

Unfortunately, it is extremely unlikely that we will observe a perfect linear relationship. There are two reasons for this. The first is experimental error; the second is that the underlying relationship may not be exactly linear, but rather only approximately linear. (A standard example is the force felt on a falling body. We initially approximate the force as $F = mg$ with g the acceleration due to gravity; however, this is not quite right as there is a resistive force which depends on the velocity.) See Figure 1 for a simulated data set of displacements and forces for a spring with spring constant equal to 5.

The Method of Least Squares is a procedure, requiring just some calculus and linear algebra, to determine what the "best fit" line is to the data. Of course, we need to quantify what we mean by "best fit", which will require a brief review of some probability and statistics.

A careful analysis of the proof will show that the method is capable of great generalizations. Instead of finding the best fit line, we could find the best fit given by *any* finite linear combinations of specified functions. Thus the general problem is given functions $f_1, \ldots, f_K$, find values of coefficients $a_1, \ldots, a_K$ such that the *linear* combination

$$y \;=\; a_1 f_1(x) + \cdots + a_K f_K(x) \tag{1.1}$$

is the best approximation to the data.

2 Probability and Statistics Review

We give a quick introduction to the basic elements of probability and statistics which we need for the Method of Least Squares; for more details see [BD, CaBe, Du, Fe, Kel, LF, MoMc].

Given a sequence of data $x_1, \dots, x_N$, we define the **mean** (or the **expected value**) to be $(x_1 + \cdots + x_N)/N$. We denote this by writing a line above x: thus

$$\overline{x} \ = \ \frac{1}{N} \sum_{n=1}^{N} x_n. \tag{2.2}$$

The mean is the average value of the data.

Consider the following two sequences of data: $\{10, 20, 30, 40, 50\}$ and $\{30, 30, 30, 30, 30\}$. Both sets have the same mean; however, the first data set has greater variation about the mean. This leads to the concept of variance, which is a useful tool to quantify how much a set of data fluctuates about its mean. The **variance**[1] of $\{x_1, \dots, x_N\}$, denoted by σ_x^2, is

$$\sigma_x^2 \ = \ \frac{1}{N} \sum_{n=1}^{N} (x_i - \overline{x})^2; \tag{2.3}$$

the **standard deviation** σ_x is the square root of the variance:

$$\sigma_x \ = \ \sqrt{\frac{1}{N} \sum_{n=1}^{N} (x_i - \overline{x})^2}. \tag{2.4}$$

Note that if the x's have units of meters then the variance σ_x^2 has units of meters^2, and the standard deviation σ_x and the mean $\overline{x}$ have units of meters. Thus it is the standard deviation that gives a good measure of the deviations of the x's around their mean, as it has the same units as our quantity of interest.

There are, of course, alternate measures one can use. For example, one could consider

$$\frac{1}{N} \sum_{n=1}^{N} (x_n - \overline{x}). \tag{2.5}$$

Unfortunately this is a signed quantity, and large positive deviations can cancel with large negatives. In fact, the definition of the mean immediately implies the above is zero! This, then, would be a terrible measure of the variability in data, as it is zero regardless of what the values of the data are.

We can rectify this problem by using absolute values. This leads us to consider

$$\frac{1}{N} \sum_{n=1}^{N} |x_n - \overline{x}|. \tag{2.6}$$

[1]For those who know more advanced statistics, for technical reasons the correct definition of the sample variance is to divide by $N - 1$ and not N.

While this has the advantage of avoiding cancellation of errors (as well as having the same units as the x's), the absolute value function is not a good function analytically. It is not differentiable. This is primarily why we consider the standard deviation (the square root of the variance) – this will allow us to use the tools from calculus.

We can now quantify what we mean by "best fit". If we believe $y = ax+b$, then $y-(ax+b)$ should be zero. Thus given observations

$$\{(x_1, y_1),\ \ldots,\ (x_N, y_N)\}, \tag{2.7}$$

we look at

$$\{y_1 - (ax_1 + b),\ \ldots,\ y_N - (ax_N + b)\}. \tag{2.8}$$

The mean should be small (if it is a good fit), and the sum of squares of the terms will measure how good of a fit we have.

We define

$$E(a, b)\ :=\ \sum_{n=1}^{N} \left(y_n - (ax_n + b)\right)^2. \tag{2.9}$$

Large errors are given a higher weight than smaller errors (due to the squaring). Thus our procedure favors many medium sized errors over a few large errors. If we used absolute values to measure the error (see equation (2.6)), then all errors are weighted equally; however, the absolute value function is not differentiable, and thus the tools of calculus become inaccessible.

Remark 2.1 (Choice of how to measure errors)**.** *As the point is so important, it is worth looking at one more time. There are three natural candidates to use in measuring the error between theory and observation:*

$$E_1(a, b)\ =\ \sum_{n=1}^{N} \left(y_i - (ax_i + b)\right), \tag{2.10}$$

$$E_2(a, b)\ =\ \sum_{n=1}^{N} |y_i - (ax_i + b)| \tag{2.11}$$

and

$$E_3(a, b)\ =\ \sum_{n=1}^{N} \left(y_i - (ax_i + b)\right)^2. \tag{2.12}$$

The problem with (2.10) *is that the errors are signed quantities, and positive errors can cancel with negative errors. The problem with* (2.11) *is that the absolute value function is not differentiable, and thus the tools and results of calculus are unavailable. The problem with* (2.12) *is that errors are not weighted equally: large errors are given significantly more weight than smaller errors. There are thus problems with all three. That said, the problems with* (2.12) *is not so bad when compared to its advantages, namely that errors cannot cancel and that calculus is available. Thus, most people typically use* (2.12) *and measure errors by sums of squares.*

3 The Method of Least Squares

Given data $\{(x_1, y_1), \dots, (x_N, y_N)\}$, we defined the error associated to saying $y = ax + b$ by

$$E(a, b) \ := \ \sum_{n=1}^{N} \left(y_n - (ax_n + b)\right)^2 . \tag{3.13}$$

Note that the error is a function of two variables, the unknown parameters a and b.

The goal is to find values of a and b that minimize the error. In multivariable calculus we learn that this requires us to find the values of (a, b) such that the gradient of E with respect to our variables (which are a and b) vanishes; thus we require

$$\nabla E \ = \ \left(\frac{\partial E}{\partial a}, \frac{\partial E}{\partial b}\right) \ = \ (0, 0), \tag{3.14}$$

or

$$\frac{\partial E}{\partial a} \ = \ 0, \quad \frac{\partial E}{\partial b} \ = \ 0. \tag{3.15}$$

Note we do not have to worry about boundary points: as $|a|$ and $|b|$ become large, the fit will clearly get worse and worse. Thus we do not need to check on the boundary.

Differentiating $E(a, b)$ yields

$$\begin{aligned} \frac{\partial E}{\partial a} &= \sum_{n=1}^{N} 2\left(y_n - (ax_n + b)\right) \cdot (-x_n) \\ \frac{\partial E}{\partial b} &= \sum_{n=1}^{N} 2\left(y_n - (ax_n + b)\right) \cdot (-1). \end{aligned} \tag{3.16}$$

Setting $\partial E/\partial a = \partial E/\partial b = 0$ (and dividing by -2) yields

$$\begin{aligned} \sum_{n=1}^{N} \left(y_n - (ax_n + b)\right) \cdot x_n &= 0 \\ \sum_{n=1}^{N} \left(y_n - (ax_n + b)\right) &= 0. \end{aligned} \tag{3.17}$$

Note we can divide both sides by -2 as it is just a constant; we cannot divide by x_i as that varies with i.

We may rewrite these equations as

$$\begin{aligned} \left(\sum_{n=1}^{N} x_n^2\right) a + \left(\sum_{n=1}^{N} x_n\right) b &= \sum_{n=1}^{N} x_n y_n \\ \left(\sum_{n=1}^{N} x_n\right) a + \left(\sum_{n=1}^{N} 1\right) b &= \sum_{n=1}^{N} y_n. \end{aligned} \tag{3.18}$$

We have obtained that the values of a and b which minimize the error (defined in (3.13)) satisfy the following matrix equation:

$$\begin{pmatrix} \sum_{n=1}^N x_n^2 & \sum_{n=1}^N x_n \\ \sum_{n=1}^N x_n & \sum_{n=1}^N 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^N x_n y_n \\ \sum_{n=1}^N y_n \end{pmatrix}. \tag{3.19}$$

We need a fact from linear algebra. Recall the inverse of a matrix A is a matrix B such that $AB = BA = I$, where I is the identity matrix. If $A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ is a 2×2 matrix where $\det A = \alpha\delta - \beta\gamma \neq 0$, then A is invertible and

$$A^{-1} = \frac{1}{\alpha\delta - \beta\gamma} \begin{pmatrix} \delta & -\gamma \\ -\beta & \alpha \end{pmatrix}. \tag{3.20}$$

In other words, $AA^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ here. For example, if $A = \begin{pmatrix} 1 & 2 \\ 3 & 7 \end{pmatrix}$ then $\det A = 1$ and $A^{-1} = \begin{pmatrix} 7 & -2 \\ -3 & 1 \end{pmatrix}$; we can check this by noting (through matrix multiplication) that

$$\begin{pmatrix} 1 & 2 \\ 3 & 7 \end{pmatrix} \begin{pmatrix} 7 & -2 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.21}$$

We can show the matrix in (3.19) is invertible (so long as at least two of the x_n's are distinct), which implies

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^N x_n^2 & \sum_{n=1}^N x_n \\ \sum_{n=1}^N x_n & \sum_{n=1}^N 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{n=1}^N x_n y_n \\ \sum_{n=1}^N y_n \end{pmatrix}. \tag{3.22}$$

Denote the matrix from (3.19) by M. The determinant of M is

$$\det M = \sum_{n=1}^N x_n^2 \cdot \sum_{n=1}^N 1 - \sum_{n=1}^N x_n \cdot \sum_{n=1}^N x_n. \tag{3.23}$$

As

$$\overline{x} = \frac{1}{N} \sum_{n=1}^N x_n, \tag{3.24}$$

we find that

$$\begin{aligned} \det M &= N \sum_{n=1}^N x_n^2 - (N\overline{x})^2 \\ &= N^2 \left(\frac{1}{N} \sum_{n=1}^N x_n^2 - \overline{x}^2 \right) \\ &= N^2 \cdot \frac{1}{N} \sum_{n=1}^N (x_n - \overline{x})^2, \end{aligned} \tag{3.25}$$

where the last equality follows from simple algebra. Thus, as long as all the x_n are not equal, $\det M$ will be non-zero and M will be invertible. Using the definition of variance, we notice the above could also be written as

$$\det M \;=\; N^2\sigma_x^2. \tag{3.26}$$

Thus we find that, so long as the x's are not all equal, the best fit values of a and b are obtained by solving a linear system of equations; the solution is given in (3.22)**.**

We rewrite (3.22) in a simpler form. Using the inverse of the matrix and the definition of the mean and variance, we find

$$\begin{pmatrix} a \\ \\ b \end{pmatrix} \;=\; \frac{1}{N^2\sigma_x^2}\begin{pmatrix} N & -N\overline{x} \\ -N\overline{x} & \sum_{n=1}^N x_n^2 \end{pmatrix}\begin{pmatrix} \sum_{n=1}^N x_n y_n \\ \\ \sum_{n=1}^N y_n \end{pmatrix}. \tag{3.27}$$

Expanding gives

$$\begin{aligned} a &= \frac{N\sum_{n=1}^N x_n y_n - N\overline{x}\sum_{n=1}^N y_n}{N^2\sigma_X^2} \\ b &= \frac{-N\overline{x}\sum_{n=1}^N x_n y_n + \sum_{n=1}^N x_n^2 \sum_{n=1}^N y_n}{N^2\sigma_X^2} \\ \overline{x} &= \frac{1}{N}\sum_{n=1}^N x_i \\ \sigma_x^2 &= \frac{1}{N}\sum_{n=1}^N (x_i - \overline{x})^2. \end{aligned} \tag{3.28}$$

As the formulas for a and b are so important, it is worth giving another expression for

them. We also have

$$a = \frac{\sum_{n=1}^{N} 1 \sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} x_n \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} 1 \sum_{n=1}^{N} x_n^2 - \sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n}$$

$$b = \frac{\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} 1}. \quad (3.29)$$

Remark 3.1. *The formulas above for a and b are reasonable, as can be seen by a unit analysis. For example, imagine x is in meters and y is in seconds. Then if $y = ax + b$ we would need b and y to have the same units (namely seconds), and a to have units seconds per meter. If we substitute in the units for the various quantities on the right hand side of* (3.28)*, we do see a and b have the correct units. While this is not a proof that we have not made a mistake, it is a great reassurance. No matter what you are studying, you should always try unit calculations such as this.*

There are other, equivalent formulas for a and b; these give the same answer, but arrange the algebra in a slightly different sequence of steps. Essentially what we are doing is the following: image we are given

$$\begin{aligned} 4 &= 3a + 2b \\ 5 &= 2a + 5b. \end{aligned}$$

If we want to solve, we can proceed in two ways. We can use the first equation to solve for b in terms of a and substitute in, or we can multiply the first equation by 5 and the second equation by 2 and subtract; the b terms cancel and we obtain the value of a. Explicitly,

$$\begin{aligned} 20 &= 15a + 10b \\ 10 &= 4a + 10b, \end{aligned}$$

which yields

$$10 = 11a,$$

or

$$a = 10/11.$$

Remark 3.2. *The data plotted in Figure 1 was obtained by letting $x_n = 5 + .2n$ and then letting $y_n = 5x_n$ plus an error randomly drawn from a normal distribution with mean zero and standard deviation 4 ($n \in \{1, \ldots, 100\}$). Using these values, we find a best fit line of*

$$y = 4.99x + .48; \quad (3.30)$$

thus $a = 4.99$ and $b = .48$. As the expected relation is $y = 5x$, we expected a best fit value of a of 5 and b of 0.

While our value for a is very close to the true value, our value of b is significantly off. We deliberately chose data of this nature to indicate the dangers in using the Method of Least Squares. Just because we know 4.99 is the best value for the slope and $.48$ is the best value for the y-intercept does not *mean that these are good estimates of the true values. The theory needs to be supplemented with techniques which provide error estimates. Thus we want to know something like, given this data, there is a 99% chance that the true value of a is in $(4.96, 5.02)$ and the true value of b is in $(-.22, 1.18)$; this is far more useful than just knowing the best fit values.*

If instead we used

$$E_{\rm abs}(a,b) \ = \ \sum_{n=1}^{N} |y_n - (ax_n + b)| \, , \tag{3.31}$$

then numerical techniques yield that the best fit value of a is 5.03 and the best fit value of b is less than 10^{-10} in absolute value. The difference between these values and those from the Method of Least Squares is in the best fit value of b (the least important of the two parameters), and is due to the different ways of weighting the errors.

Exercise 3.3. *Consider the observed data $(0,0), (1,1), (2,2)$. It should be clear that the best fit line is $y = x$; this leads to zero error in all three systems of measuring error, namely* (2.10), (2.11) *and* (2.12)*; however, show that if we use* (2.10) *to measure the error then line $y = 1$ also yields zero error, and clearly this should not be the best fit line!*

Exercise 3.4. *Generalize the method of least squares to find the best fit quadratic to $y = ax^2 + bx + c$ (or more generally the best fit degree m polynomial to $y = a_m x^m + a_{m-1} x^{m-1} + \cdots + a_0$).*

While for any real world problem, direct computation determines whether or not the resulting matrix is invertible, it is nice to be able to prove the determinant is always non-zero for the best fit line (if all the x's are not equal).

Exercise 3.5. *If the x's are not all equal, must the determinant be non-zero for the best fit quadratic or the best fit cubic?*

Looking at our proof of the Method of Least Squares, we note that it was not essential that we have $y = ax + b$; we could have had $y = af(x) + bg(x)$, and the arguments would have proceeded similarly. The difference would be that we would now obtain

$$\begin{pmatrix} \sum_{n=1}^{N} f(x_n)^2 & \sum_{n=1}^{N} f(x_n)g(x_n) \\ \sum_{n=1}^{N} f(x_n)g(x_n) & \sum_{n=1}^{N} g(x_n)^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \ = \ \begin{pmatrix} \sum_{n=1}^{N} f(x_n)y_n \\ \sum_{n=1}^{N} g(x_n)y_n \end{pmatrix}. \tag{3.32}$$

Finally, we comment briefly on a very important change of variable that allows us to use the Method of Least Squares in many more situations than one might expect. Consider the

case of a researcher trying to prove Newton's Law of Universal Gravity, which says the force felt by two masses m_1 and m_2 has magnitude Gm_1m_2/r^2, where r is the distance between the objects. If we fix the masses, then we expect the magnitude of the force to be inversely proportional to the distance. We may write this as $F = k/r^n$, where we believe $n = 2$ (the value for k depends on G and the product of the masses). Clearly it is n that is the more important parameter here. Unfortunately, as written, we cannot use the Method of Least Squares, as one of the unknown parameters arises non-linearly (as the exponent of the separation).

We can surmount this problem by taking a logarithmic transform of the data. Setting $\mathcal{K} = \log k$, $\mathcal{F} = \log F$ and $\mathcal{R} = \log r$, the relation $F = k/r^n$ becomes $\mathcal{F} = n\mathcal{R} + \mathcal{K}$. We are now in a situation where we can apply the Method of Least Squares. The only difference from the original problem is how we collect and process the data; now our data is not the separation between the two masses, but rather the logarithm of the separation. Arguing along these lines, many power relations can be converted to instances where we can use the Method of Least Squares. We thus (finally) fulfill a promise made by many high school math teachers years ago: logarithms can be useful!

Exercise 3.6. *Consider the generalization of the Method of Least Squares given in* (3.32)*. Under what conditions is the matrix invertible?*

Exercise 3.7. *The method of proof generalizes further to the case when one expects y is a* linear *combination of K fixed functions. The functions need not be linear; all that is required is that we have a linear combination, say $a_1 f_1(x) + \cdots + a_K f_K(x)$. One then determines the $a_1, \ldots, a_K$ that minimize the variance (the sum of squares of the errors) by calculus and linear algebra. Find the matrix equation that the best fit coefficients $(a_1, \ldots, a_K)$ must satisfy.*

Exercise 3.8. *Consider the best fit line from the Method of Least Squares, so the best fit values are given by* (3.22)*. Is the point $(\overline{x}, \overline{y})$, where $\overline{x} = \frac{1}{n}\sum_{n=1}^{N} x_n$ and $\overline{y} = \sum_{n=1}^{N} y_n$, on the best fit line? In other words, does the best fit line go through the "average" point?*

Exercise 3.9 (Kepler's Third Law)**.** *Kepler's third law states that if T is the orbital period of a planet traveling in an elliptical orbit about the sun (and no other objects exist), then $T^2 = CL^3$, where L is the length of the semi-major axis. I always found this the hardest of the three laws; how would one be led to the right values of the exponents from observational data? One way is through the Method of Least Squares. Set $\mathcal{T} = \log T$, $\mathcal{L} = \log L$ and $c = \log \mathcal{C}$. Then a relationship of the form $T^a = CL^b$ becomes $a\mathcal{T} = b\mathcal{L} + c$, which is amenable to the Method of Least Squares. The semi-major axis of the the 8 planets (sadly, Pluto is no longer considered a planet) are Mercury 0.387, Venus 0.723, Earth 1.000, Mars 1.524, Jupiter 5.203, Saturn 9.539, Uranus 19.182, Neptune 30.06 (the units are astronomical units, where one astronomical unit is $1.496 \cdot 10^8$ km); the orbital periods (in years) are 0.2408467, 0.61519726, 1.0000174, 1.8808476, 11.862615, 29.447498, 84.016846 and 164.79132. Using this data, apply the Method of Least Squares to find the best fit values of a and b in $T^a = CL^b$ (note, of course, you need to use the equation $a\mathcal{T} = b\mathcal{L} + \mathcal{C}$).*

Actually, as phrased above, the problem is a little indeterminate for the following reason. Imagine we have $T^2 = 5L^3$ or $T^4 = 25L^6$ or $T = \sqrt{5}L^{1.5}$ or even $T^4 = 625L^{12}$. ***All of***

these are the same equation! *In other words, we might as well make our lives easy by taking* $a = 1$*; there really is no loss in generality in doing this. This is yet another example of how changing our point of view can really help us. At first it looks like this is a problem involving* ***three*** *unknown parameters,* a*,* b *and* C*; however,* ***there is absolutely no loss in generality in taking*** $a = 1$*; thus let us make our lives easier and just look at this special case.*

For your convenience, here are the natural logarithms of the data: the lengths of the semi-major axes are

$$\{-0.949331,\ -0.324346,\ 0,\ 0.421338,\ 1.64924,\ 2.25539,\ 2.95397,\ 3.4032\}$$

and the natural logarithms of the periods (in years) are

$$\{-1.42359,\ -0.485812,\ 0.0000173998,\ 0.631723,\ 2.47339,\ 3.38261,\ 4.43102,\ 5.10468\}.$$

The problem asks you to find the best fit values of a *and* b*. In some sense this is a bit misleading, as there are infinitely many possible values for the pair* (a, b)*; however, all of these pairs will have the same* ***ratio*** b/a *(which Kepler says should be close to 3/2 or 1.50). It is this ratio that is truly important. The content of Kepler's Third Law is that the square of the period is proportional to the cube of the semi-major axis. The key numbers are the powers of the period and the length (the* a *and the* b*), not the proportionality constant. This is why I only ask you to find the best fit values of* a *and* b *and not* C *(or* $\mathcal{C}$*), as* C *(or* $\mathcal{C}$*) is not as important. If we take* $a = 1$ *then the best fit value of* $\mathcal{C}$ *is 0.000148796, and the best fit value of* b *is almost* 1.50*.*

Our notes above have many different formulas to find the best fit values a *and* b *for a relation* $y = ax + b$*. For us, we have* $\mathcal{T} = \frac{b}{a}\mathcal{L} + \frac{\mathcal{C}}{a}$*. Thus, for this problem, the role of* a *from before is being played by* $\frac{b}{a}$ *and the role of* b *from before is being played by* $\frac{\mathcal{C}}{a}$*. Therefore if we want to find the best fit value for the ratio* $\frac{b}{a}$ *for this problem, we just use the first of the two formulas from* (3.29)*.*

References

[BD] P. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco, 1977.

[CaBe] G. Casella and R. Berger, *Statistical Inference*, 2nd edition, Duxbury Advanced Series, Pacific Grove, CA, 2002.

[Du] R. Durrett, *Probability: Theory and Examples*, 2nd edition, Duxbury Press, 1996.

[Fe] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd edition, Vol. II, John Wiley & Sons, New York, 1971.

[Kel] D. Kelley, *Introduction to Probability*, Macmillan Publishing Company, London, 1994.

[LF] R. Larson and B. Farber, *Elementary Statistics: Picturing the World*, Prentice-Hall, Englewood Cliffs, NJ, 2003.

[MoMc] D. Moore and G. McCabe, *Introduction to the Practice of Statistics*, W. H. Freeman and Co., London, 2003.