

7 Calculus A, Lab 7

7.1 Introduction.

This lab is about curve fitting and linear regression. It represents a basic application of calculus to statistics and data modeling.

A common problem in applied math is to find a line or other curve passing through a set of data points. Typically the data points come from from experimental process; so they arrive containing approximations and errors. They don't exactly lie on the line or smooth curve one is seeking. The purpose of this lab is to see how we can use what we know about derivatives and optimization to find the straight line or other curve that best fits a cloud of data.

7.2 What to minimize?

Figure 1 shows a collection of data points and a straight line that might approximate them. The first question we have to ask is how we would measure the goodness of fit of this line to the data.

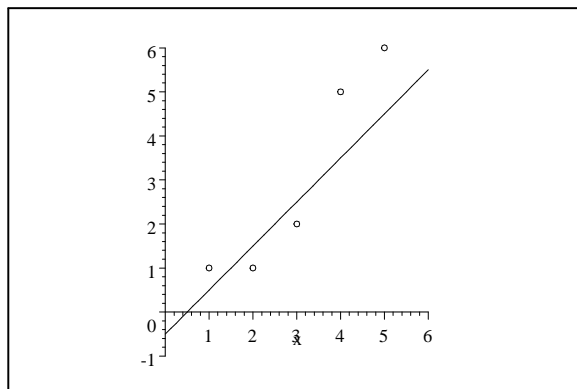


Figure 1: Approximating data points with a line.

The first answer to this question that would come to my mind would be to draw perpendicular line segments from each of the data points to the line. The sum of the lengths of these segments is a measure of the goodness of fit of the line. We would then try to minimize the sum of these lengths, which are shown in Figure 2.

There are 2 difficulties with this choice of a quantity to minimize. The first is shown in Problem 1

1. Compute the perpendicular distance from the point $(4, 5)$ to the line $y = mx + b$. I would do this by remembering that any line perpendicular to the line $y = mx + b$ has slope $-\frac{1}{m}$. This would let me write the equation of the line through $(4, 5)$ perpendicular to $y = mx + b$. I could then find where those two

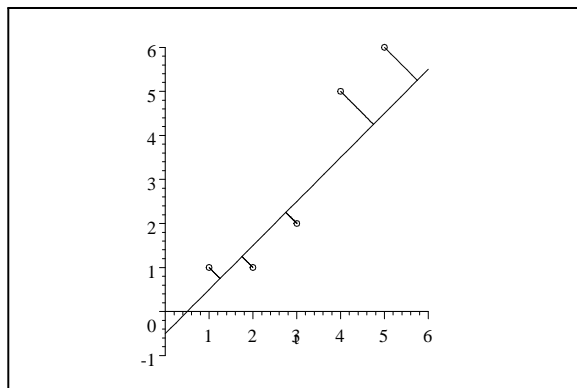


Figure 2: Perpendicular distances to the line.

lines met, and compute the distance between the point of intersection and the point $(4, 5)$.

Would you feel like computing this distance for a bunch of points, summing the results, and then minimizing the sum?

The other problem with minimizing this expression is that often the quantity on the x -axis can be measured exactly or quite accurately, and the error in the experiment is in the determination of y . If this is so, then it would seem to make more sense to use as a measure of the error in the line not the sum of the perpendicular distances to the line, but the sum of the lengths of vertical segments dropped from the points to the line, as shown in Figure 3.

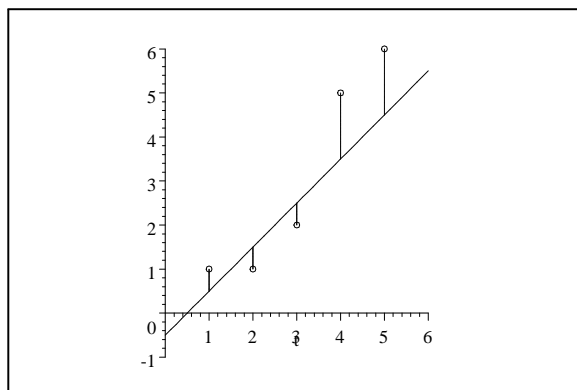


Figure 3: Vertical distances to the line.

2. Write the sum of the lengths of the vertical segments joining the points $(1, 1)$, $(2, 1)$, $(3, 2)$, $(4, 5)$, and $(5, 6)$ to the line $y = mx + b$. Your answer will

obviously depend on m and b . It should contain a bunch of absolute values.

The expression you got in Problem 2 should look a whole lot more tractable than the one from Problem 1, but there's still a difficulty. The trouble with using calculus to minimize the expression from Problem 2 is that the absolute value function doesn't always have a derivative. There isn't a simple algebraic expression for the derivative of $|x|$. One has to say that the derivative is 1 if $x > 0$, that it is -1 if $x < 0$, and that it is undefined at $x = 0$. Unless you really loved working with absolute values in school, it's therefore a good idea to avoid them here if we can.

So let's try a third approach, less intuitive than the other two. Let's use as our measure for the goodness of fit of the line not the sum of the absolute values of the vertical distances between the points and the line, but the sum of the squares of the vertical distances between the points and the line. The squares, like the absolute values, are guaranteed to be non-negative; but unlike $|x|$, x^2 has a simple derivative. We can therefore use calculus to minimize the sum of the squares of the distances, and the line we get should be a pretty good fit to the data. (There are some fancier justifications for this choice of a function to minimize, but I still think that at bottom the reason for the choice is that it results in a reasonable fit and we can do the math.) This process of finding a line that minimizes the sum of the squares of the vertical distances between the points and the line is called least squares linear regression.

3. Write the sum S of the squares of the lengths of the vertical segments joining the points $(1, 1)$, $(2, 1)$, $(3, 2)$, $(4, 5)$, and $(5, 6)$ to the line $y = mx + b$. Don't bother to simplify your answer, which will depend on m and b , and which will contain a bunch of squares.

The goal now is to find values for m and b that make the sum from Problem 3 as small as possible. There are a variety of ways to do this, but here's a naive one that works. Think of both m and b as variables, and think of the sum S of squares of vertical distances as a function of m and b . If there are values of m and b that minimize S , then at the minimum, we can regard b as a constant and m as a variable; and we have to be at a minimum of S as a function of m . Alternatively, we can regard m as a constant and b as a variable; and we have to be at a minimum of S as a function of b . So for the best-fit line, we expect that

$$\frac{dS}{dm} = \frac{dS}{db} = 0.$$

4. Regard m as a constant and b as a variable, set $\frac{dS}{db} = 0$, and solve for b .

The equation you get for b in terms of m has a simple geometric interpretation. The center of mass of the points we are working with is the point $(3, 3)$ whose x coordinate is the average of the x coordinates of the 5 points and whose y coordinate is the average of their y coordinates. You should be able to interpret the equation in Problem 4 as saying that the line $y = mx + b$ must pass through the center of mass of the points.

5. Now that you know the minimum value for b , you can plug this value into your formula for S , compute $\frac{dS}{dm}$, and set it equal to 0 as well. Solve for m and

b. Plot the line $y = mx + b$ together with the 5 data points. The line ought to look as if it fits the points at least as well as any other straight line would. If it doesn't, then go back and look for your mistake.

An alternative approach would have been to set both $\frac{dS}{db}$ and $\frac{dS}{dm}$ equal to 0 simultaneously, and to solve. This wouldn't have given us the geometric interpretation in terms of the center of mass, but it might have involved less thinking.

Now let's try a fancier problem. Suppose we don't want to run a straight line through our data points, but a parabola $y = ax^2 + bx + c$. We might do this if we expected for theoretical reasons that our data points should follow a curve of this form. Since minimizing the sums of the squares of the lengths of vertical segments joining the data points to the curve worked before, why not try the same idea again?

6. Write the sum S of the squares of the lengths of the vertical segments joining the points $(1, 1)$, $(2, 1)$, $(3, 2)$, $(4, 5)$, and $(5, 6)$ to the parabola $y = ax^2 + bx + c$. Don't bother to simplify your answer, which will depend on m and b , and which will contain a bunch of squares.

The goal now is to find values for a , b , and c that make the sum from Problem 6 as small as possible. Proceed just as before: think of a , b , and c as variables, and the sum S as a function of a , b , and c . For the best-fit parabola, we expect

$$\frac{dS}{da} = \frac{dS}{db} = \frac{dS}{dc} = 0.$$

7. Compute $\frac{dS}{da}$, $\frac{dS}{db}$, and $\frac{dS}{dc}$. Set them all equal to 0, and solve for a , b , and c . I'd use Maple to solve them together, but it is also just a high school algebra problem. Then plot the parabola $y = ax^2 + bx + c$ together with the 5 data points. The parabola ought to look as if it fits the points at least as well as any other parabola would. If it doesn't, then go back and look for your mistake.

At this point, it should seem highly likely that given a collection of points, we could fit a curve of any give degree to that set of points by using the same approach we used for lines (curves of degree 1) and parabolas (curves of degree 2).

There are really only two bits left to finishing up a completely general theory of linear regression. They are a bit more notationally demanding than what we've done so far, but if you stay calm, you should be fine. You're still just working with linear equations; it's just that the coefficients will involve slightly messy symbolic expressions.

Suppose you want to fit a straight line to the three data points (x_1, y_1) , (x_2, y_2) , and (x_3, y_3) . Suppose also that

$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}$$

is the average of x_1, x_2, x_3 , and that \bar{y} is the average of y_1, y_2, y_3 . The sum of the squares of the vertical distances from the three data points to the line

$y = mx + b$ is then

$$S = (mx_1 + b - y_1)^2 + (mx_2 + b - y_2)^2 + (mx_3 + b - y_3)^2.$$

8. (a) Regard m as a constant, and set $\frac{dS}{db} = 0$. Show that at the end of all the algebra, you have

$$b = \bar{y} - m\bar{x}. \quad (1)$$

(b) Plug this value for b into S and you get

$$S = [m(x_1 - \bar{x}) - (y_1 - \bar{y})]^2 + [m(x_2 - \bar{x}) - (y_2 - \bar{y})]^2 + [m(x_3 - \bar{x}) - (y_3 - \bar{y})]^2.$$

Now set $\frac{dS}{dm} = 0$ and solve for m . You should get

$$m = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + (x_3 - \bar{x})(y_3 - \bar{y})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2}. \quad (2)$$

Formulas (1) and (2) for b and m , which generalize immediately to more than 3 data points, show that one can get best fit lines without doing the calculus anew for each set of data. You just compute the averages and plug into the formulas. At the end of all the calculus, we therefore have simple procedures for fitting lines to data, which can easily be implemented in a calculator or simple statistics program.

9. Finally, an open-ended question to think about. Can you design a way to measure how close a set of points comes to lying on a straight line? Ideally, you would like a measure that doesn't change if you shift all the points a constant distance horizontally or vertically, and which doesn't change when you multiply every x coordinate by a fixed constant, or when you multiply every y coordinate by a fixed constant. Such a measure would let you say numerically how good your best fit line is as an approximation to the data points.