All models are wrong, but some are useful. –An aphorism popularized by George E. P. Box

Linear least-squares data fitting

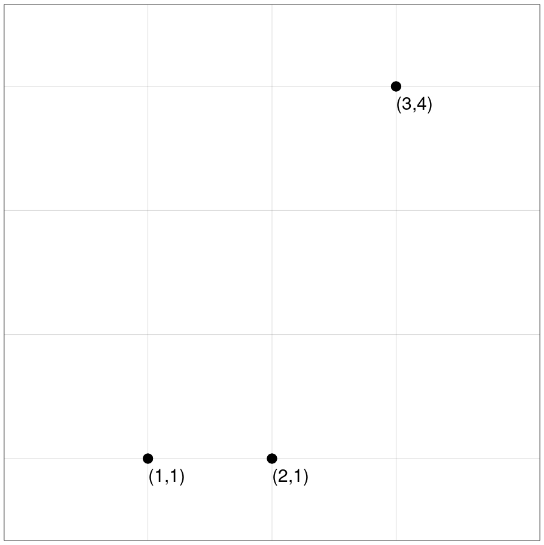
This note supplements ideas introduced in Abramson's College Algebra textbook section 4.3 *Fitting Linear Models to Data* [1], and similarly in section 2.5 of Stitz and Zeager's book [5]. The textbooks introduce the topic but then peter out, missing an opportunity to make the connection to quadratic optimization (which appears in the next chapter of Abramson's book!).

Linear least-squares data fitting is one of the most important and widely-used data fitting methods–indeed, one of the most important computational methods of any kind. Fundamentally an optimization problem, the method is typically introduced in calculus courses, or in statistics or computer-science courses using ideas from calculus. But because it's a *quadratic* optimization problem, least-squares data fitting can be discovered and worked-out simply using ideas from college algebra courses.

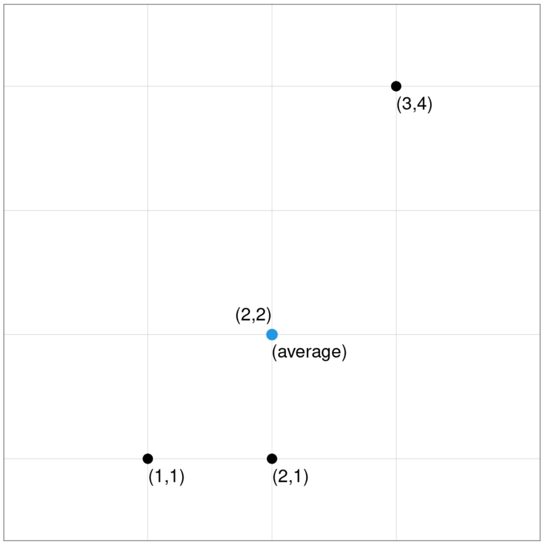
What's that mean, "data fitting?"

"Data fitting" means defining a simple mathematical function, like a point or a line for instance, that somehow approximates data reasonably well. "Data" means numeric measurements (observations) of things like time, temperature, stock prices, grades, etc. This note assumes that the data are a set of (x, y)-pairs defining points that can be plotted.[1]

For example, let's say you observe three data points: $(1, 1), (2, 1)$, and $(3, 4)$, plotted below:



Three data points.



Three data points and their average.

Suppose that three points is just too complicated and instead you want to reasonably approximate these data with only a single point. Can you think of a good, simple way to do that?

There is no one right answer to that question. But a reasonable approach might be to approximate the points by their average x and y values.[2] The average of the x-values is $(1 + 2 + 3)/3 = 2$ and the average of the y-values is $(1 + 1 + 4)/3 = 2$, so the overall average data point $(2, 2)$ looks like the blue point in the plot on the right above.

[1] Data fitting can also apply in higher-dimensions.

[2] Can you think of other reasonable ways to summarize the data?

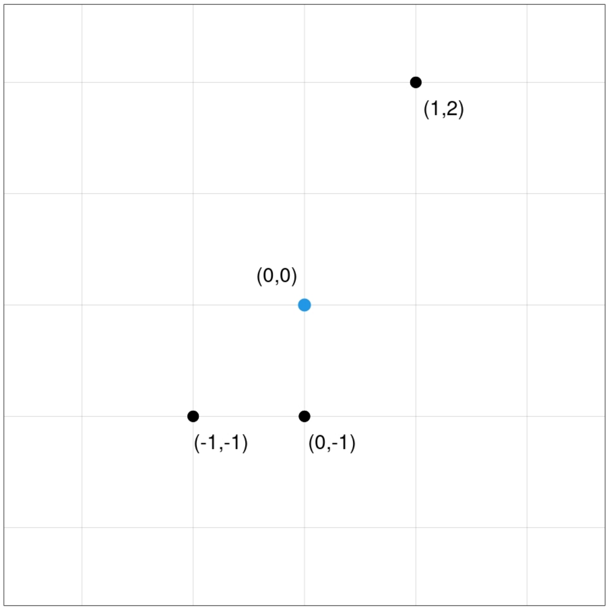
The average data point is a simple, reasonable model for the data! On average, the data look like that.

Centering data

If you subtract the average of the x-values from each x-value and also subtract the average of the y-values from each y-value, then the plot looks basically the same, as shown in the figure to the right.

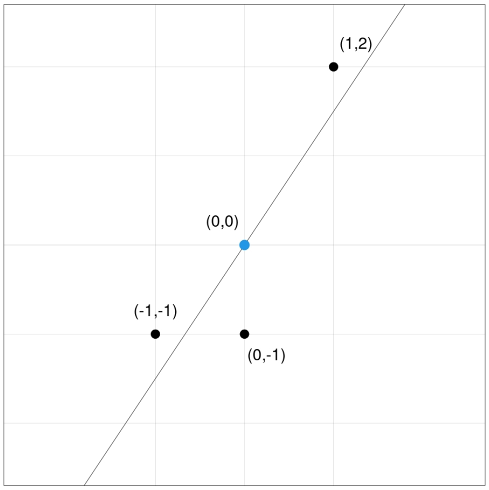
The relative positions of all the data points are the same as before, but now the average is shifted to the origin $(0, 0)$ instead of $(2, 2)$. This *translation* of the data is sometimes called "centering" for obvious reasons–the average of the data is centered at the origin. It turns out that centering the data will simplify things later. Centering data is also easy to undo–if you want the original data values back, simply add the averages back in.

Another way to think about this is that the data points are exactly the same as before, we simply re-labeled the coordinate system to put their average at the origin.



A linear model

The single-point average model of the data shown above is a fine model. It describes the overall average of the data. We can fit other models to the data too. For example, we can try to run a line through the data like the one illustrated in the next figure.

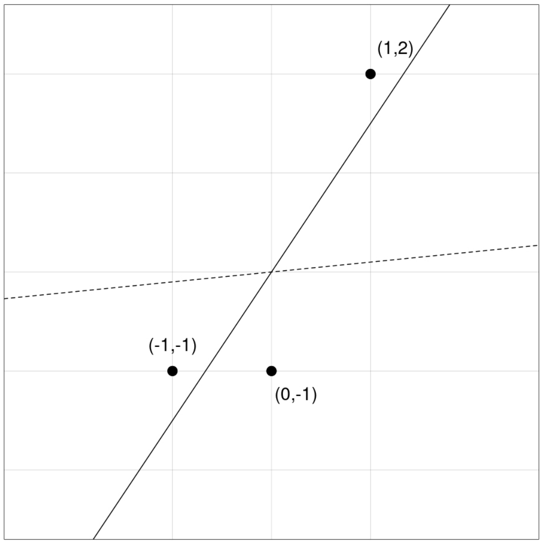


The line goes through the average of the data (the blue circle at the origin) with a slope that tries to keep the line reasonably close to the data points themselves (the black discs); the line is a linear model of the data. The rest of this note is about one way to define what keeping the line "reasonably close" to the data means.

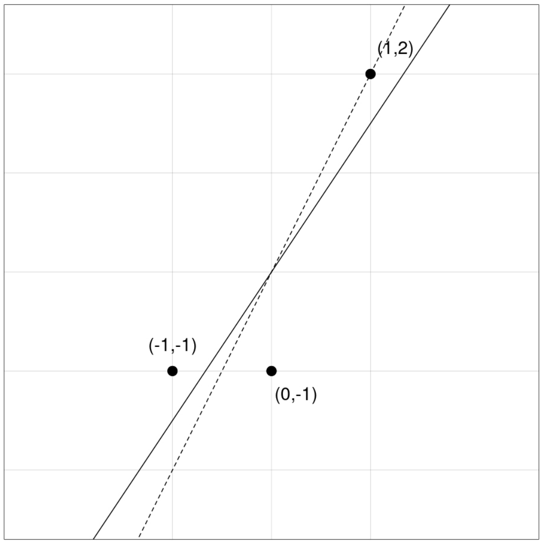
Like the simple single-point estimate of the data, this note assumes that linear models pass through the average data value (statisticians have a lot to say about this assumption). Since we've translated our data to be centered at zero, the linear models considered in this note are particularly simple and all have y-intercepts of zero.

Model error

Some lines intuitively seem to "fit" the data better than others. For instance, in the illustration below on the left, the dashed line seems to have the wrong slope compared to the solid line which looks like it "fits" the data points better. However, in the figure on the right it's not immediately obvious whether the solid line or the dashed line fits the data better.



The solid line looks like it fits better...

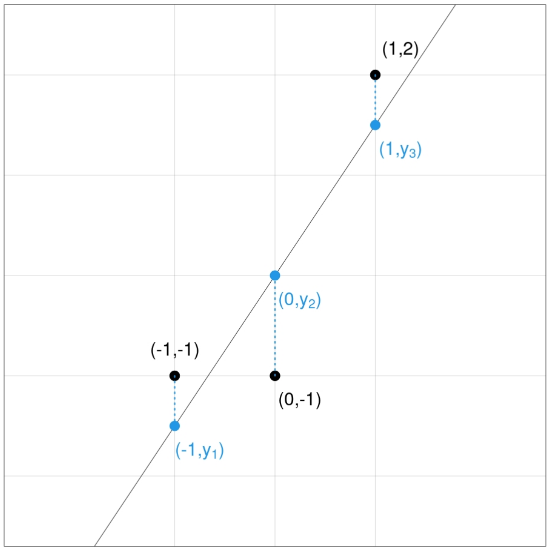


Hmmm...which line fits better, solid or dashed?

What we need is a systematic way to measure error of a model fit in order to decide whether one model fits the data better or worse than another.

One way to define the error between a linear model fit and data points is to add up the vertical distances between the line and each data point, illustrated by the dashed blue lines for an example linear model in the figure on the right (writing y_1, y_2, and y_3 as the y-coordinates of the points on the line corresponding to the data x-coordinates).

Can you think of other ways to define error between the line and the data points? What about the distances between the data points and the line that are perpendicular to the line? That's a thing called "total least squares" data fitting and in some cases has certain advantages. There are many different ways to measure data fit error!



We *could* define the error between the line and the data by adding up the lenghts of the dashed blue lines like this:

$$d_1 = |-1-y_1| + |-1-y_2| + |2-y_3|.$$

Then a "best-fit" line would be one with the smallest error value d_1. But, let's face it, nobody really enjoys working with the absolute value function. And the absolute-value expression might be tricky to optimize for a best fit line.

What is nice about absolute value distances is that the errors between the line and each point are never negative and purely additive. It would be great to find similar additive error measurements that are nicer to work with.

What about adding up squared distances between the data point y-values and corresponding model y-values? Something like this:

$$\begin{aligned} d &= |-1-y_1|^2 + |-1-y_2|^2 + |2-y_3|^2 \\ &= (-1-y_1)^2 + (-1-y_2)^2 + (2-y_3)^2. \end{aligned}$$

This version of error has the advantage of getting rid of those absolute value terms, replacing them with squared terms. Like the absolute values, the squared differences used by d are always non-negative. In fact, each squared difference term is simply the square of the corresponding absolute-value distance term. That means minimizing the squared differences is equivalent to minimizing the absolute value distances. And most importantly, the expression above turns out to be pretty simple to optimize. This technique was formally introduced in the very early 1800's by that wild man Legendre [3] and slightly later, Gauss [2, 4]. (These notes generally follow Legendre's more basic approach, except without using calculus.)

Adrien Legendre

Carl Gauss

Best-fit linear models using least-squares

The linear models considered here pass through the average of the data, which we previously translated to the origin $(0, 0)$. That means that the y-intercept of the line is zero and the equation of the line has a very simple form:

$$y = mx$$

for some unknown slope m.

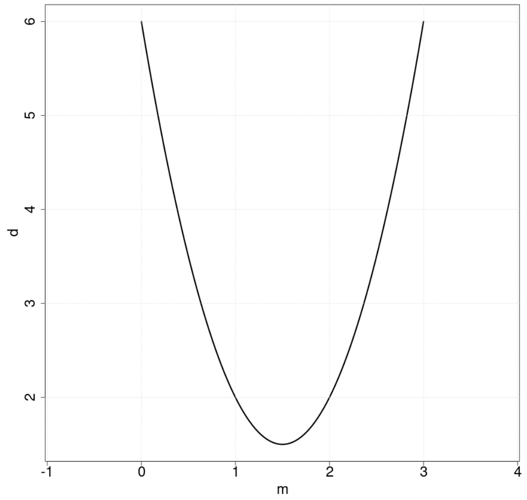
In particular, for each data-point x-value $-1, 0, 1$, the corresponding line y-values (written as y_1, y_2, y_3 above) are:

$$\begin{array}{lllllll} y_1 & = & mx_1 & = & m \cdot -1 & = & -m, \\ y_2 & = & mx_2 & = & m \cdot 0 & = & 0, \\ y_3 & = & mx_3 & = & m \cdot 1 & = & m. \end{array}$$

Plugging those values for y_1, y_2, and y_3 into the definition of the error d above yields:

$$\begin{array}{lll} d & = & (-1 - y_1)^2 + (-1 - y_2)^2 + (2 - y_3)^2 \\ & = & (-1 - -m)^2 + (-1 - 0)^2 + (2 - m)^2 \\ & = & (-1 + m)^2 + 1 + (2 - m)^2 \\ & = & m^2 - 2m + 1 + 1 + m^2 - 4m + 4 \\ & = & 2m^2 - 6m + 6. \end{array}$$

This is a quadratic function of the slope m. The graphs of quadratics are parabolic–they have either a minimum or maximum y-value depending on the sign of the leading coefficient. This one has a positive leading coefficient (2) so it opens upwards and has a minimum value. It looks like this:



Its minimum value is attained at the *vertex* of the parabola,[3] when $m = \frac{-(-6)}{2\cdot 2} = 6/4 = 3/2$ (which you can see in the picture above). This value, $m = 3/2$, is the slope of the line representing the best linear model fit of any line that passes through the average center of the data when error is measured by the squared error formula d above.

The minimum value of the sum of the squared errors is simply the value of d when $m = 3/2$, or $2 \cdot (3/2)^2 - 6 \cdot 3/2 + 6 = 9/2 - 9 + 6 = 3/2$. The process of minimizing the sum of the squared errors to determine a best linear model fit is called ordinary linear least-squares data fitting.

General solution

The last section found a best-fit line to data for a specific example. How does the linear least-squares data fitting process work in general, for generic data?

Say you've got N data points $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. (In the last section, $N = 3$.) A generic linear least-squares procedure follows these steps:

1. Compute the average of the data point x-values, $\bar{x} = (x_1 + x_2 + \cdots + x_N)/N$.
2. Compute the average of the data point y-values, $\bar{y} = (y_1 + y_2 + \cdots + y_N)/N$.
3. Subtract the averages from each of the data point x- and y-values, respectively, to create new centered data points (written below with bars over them):

$$\begin{aligned} \bar{x}_1 &= x_1 - \bar{x}, & \bar{y}_1 &= y_1 - \bar{y} \\ \bar{x}_2 &= x_2 - \bar{x}, & \bar{y}_2 &= y_2 - \bar{y} \\ &\vdots & &\vdots \\ \bar{x}_N &= x_N - \bar{x}, & \bar{y}_N &= y_N - \bar{y}. \end{aligned}$$

   The average of the shifted data points $(\bar{x}_1, \bar{y}_1), (\bar{x}_2, \bar{y}_2), \ldots, (\bar{x}_N, \bar{y}_N)$ is now centered at the origin $(0, 0)$.
4. Compute the value of the slope m that minimizes sum of the squared errors.
5. Translate the best-fit line back to the original data (undo the centering).

[3] Remember the vertex formula $-b/2a$?

The main problem is step 4: find the slope m of a line $y = mx$ going through $(0, 0)$ that best-fits the data. Just like the example in the last section, the squared difference between each centered data point y-value and corresponding model line y-value is:

$$\begin{aligned} d = & \quad (\bar{y}_1 - m\bar{x}_1)^2 \\ + & \quad (\bar{y}_2 - m\bar{x}_2)^2 \\ + & \quad \cdots \\ + & \quad (\bar{y}_N - m\bar{x}_N)^2. \end{aligned}$$

Multiply out each term...

$$\begin{array}{rcccccc} d = & & \bar{x}_1^2 m^2 & - & 2\bar{x}_1\bar{y}_1 m & + & \bar{y}_1^2 \\ & + & \bar{x}_2^2 m^2 & - & 2\bar{x}_2\bar{y}_2 m & + & \bar{y}_2^2 \\ & + & \cdots & & & & \\ & + & \bar{x}_N^2 m^2 & - & 2\bar{x}_N\bar{y}_N m & + & \bar{y}_N^2 \end{array}$$

...then collect like terms of m to get:

$$d = (\bar{x}_1^2 + \bar{x}_2^2 + \cdots + \bar{x}_N^2)m^2 - 2(\bar{x}_1\bar{y}_1 + \bar{x}_2\bar{y}_2 + \cdots + \bar{x}_N\bar{y}_N)m + (\bar{y}_1^2 + \bar{y}_2^2 + \cdots + \bar{y}_N^2).$$

This is just a quadratic equation involving m, admittedly a pretty ugly one. It has a leading coefficient a:

$$a = \bar{x}_1^2 + \bar{x}_2^2 + \cdots + \bar{x}_N^2,$$

which is definitely not negative because each term is squared. That means that either all the $\bar{x}_j$-values are zero (and all the original data x-values are identical), or the more interesting case that the quadratic opens upwards and has a minimum value. We know that minimum value will occur at the vertex of the parabola, given by the formula $m = -b/2a$ where b is this complicated-looking term:

$$b = -2(\bar{x}_1\bar{y}_1 + \bar{x}_2\bar{y}_2 + \cdots + \bar{x}_N\bar{y}_N).$$

Plugging those expressions for a and b into the vertex formula gives a value for m where the parabola achieves its minimum:

$$\begin{aligned} m &= \frac{-b}{2a} \\ &= \frac{2(\bar{x}_1\bar{y}_1 + \bar{x}_2\bar{y}_2 + \cdots + \bar{x}_N\bar{y}_N)}{2(\bar{x}_1^2 + \bar{x}_2^2 + \cdots + \bar{x}_N^2)} \\ &= \frac{\bar{x}_1\bar{y}_1 + \bar{x}_2\bar{y}_2 + \cdots + \bar{x}_N\bar{y}_N}{\bar{x}_1^2 + \bar{x}_2^2 + \cdots + \bar{x}_N^2}. \end{aligned}$$

In other words, to compute the slope of the best-fit line to the data, multiply each centered x-coordinate by each centered y-coordinate, add them up, then divide that sum by the sum of the squared centered x-coordinates.

Let's quickly check this result using the example from the last section with centered data points $(-1, -1), (0, -1), (1, 2)$:

$$m = \frac{(-1)(-1) + (0)(-1) + (1)(2)}{(-1)^2 + 0^2 + 1^2} = \frac{3}{2},$$

matching what we calculated before.

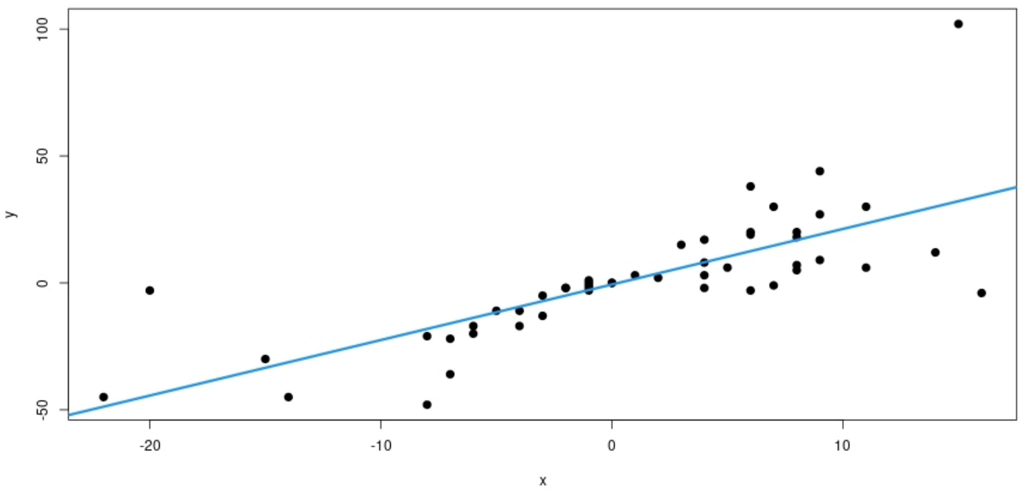
The equation of the best fit line $y = mx$ fits the *centered* data, not the original data. But it's easy to simply undo the translation (centering) of the data applied to the line using the average $\bar{x}$ and $\bar{y}$ values computed above:

$$\begin{aligned} y &= m(x - \bar{x}) + \bar{y} \\ &= mx + (\bar{y} - m\bar{x}). \end{aligned}$$

The slope of the best-fit line to the original data is still m, but instead of zero, the intercept is $\bar{y} - m\bar{x}$.

An example with more data

This illustration shows 100 somewhat random data points and a linear fit (blue line) computed using the procedure in the last section:



College algebra is powerful

It's hard to overstate the importance of linear least-squares data fitting, one of the most widely-used computational methods. Linear least-squares data fitting also exhibits many remarkable statistical properties that you can learn about in stats courses. Applications range from astronomy to zymology and everything in between including finance, health care, engineering, advertising, logistics, and many more.

The fact that such a useful method can be developed in a simple way illustrates the relevance and explanatory power of college algebra.

References

[1] Abramson, Jay. "College Algebra (OpenStax)." (2018).

[2] Gauss, Carl Friedrich. "Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss." sumtibus Frid. Perthes et IH Besser, 1809.

[3] Legendre, Adrien Marie. "Nouvelles méthodes pour la détermination des orbites des comètes; par AM Legendre." chez Firmin Didot, libraire pour lew mathematiques, la marine, l'architecture, et les editions stereotypes, rue de Thionville, 1806.

[4] Stigler, Stephen M. "Gauss and the invention of least squares." the Annals of Statistics (1981): 465-474.

[5] Zeager, Jeff, and Carl Stitz. "College Algebra." (2016).