The method of least squares
Math 218, Mathematical Statistics
D Joyce, Spring 2016

Linear regression started out as the *method of least squares*, a method developed at the beginning of the 19th century to find the closest line to a bivariate data set. The method of least squares started out with Legendre's creation of the the method of least squares as published in 1806 in his work *Nouvelles méthodes pour la détermination des orbites des cometes.* (Gauss claimed he discovered earlier, but he was young at the time and had not communicated his discovery until 1809. Laplace also wrote about it in 1812.)

Here's the idea of the method of least squares. You have n data points

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

and you want to find the linear function $y = ax + b$ whose graph is closest to the points. Since we're treating y as a function of x, the error for a given point (x_i, y_i) is measured by how far off the actual value y_i is from the predicted value for the function $y = ax + b$ which would be $y = ax_i + b$. Thus, the error for that point is

$$|(ax_i + b) - y_i|.$$

Legendre's idea was to use the square of the error instead of the error itself, then add all the squared errors together to get the total error $\mathcal{E}(a, b)$ for the given line $y = ax + b$:

$$\mathcal{E}(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2.$$

Using standard methods from calculus, we can find the line with the smallest total error $\mathcal{E}(a, b)$ by taking the derivatives with respect to a and b and setting them to 0 to find the critical point, which, in this case there's only one (assuming $n \geq 2$ and the x values are not all the same). The minimum will occur that critical point.

Here are the details for that computation. Note that each summation is $\sum_{i=1}^{n}$.

$$\frac{\partial \mathcal{E}}{\partial a} = 2 \sum (ax_i + b - y_i) \cdot x_i$$
$$= 2 \left(a \sum x_i^2 + b \sum x_i - \sum x_i y_i \right)$$
$$\frac{\partial \mathcal{E}}{\partial b} = 2 \sum (ax_i + b - y_i)$$
$$= 2 \left(a \sum x_i + bn - \sum y_i \right)$$

When these two partial derivatives are each set to 0, we get the pair of simultaneous linear equations

$$\begin{cases} a \sum x_i^2 + b \sum x_i = \sum x_i y_i \\ a \sum x_i + bn = \sum y_i \end{cases}$$

in the two unknowns a and b.

One way to solve any pair of linear equations in two unknowns is to use Cramer's rule. It says that the pair of simultaneous linear equations

$$\begin{cases} pa + qb = r \\ sa + tb = u \end{cases}$$

is

$$a = \frac{rt - uq}{pt - sq}$$
$$b = \frac{pu - sr}{pt - sq}$$

For our pair of equations, that gives

$$a = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$
$$b = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

For statistics, it's most useful to express the answer in terms of sample means $\overline{x} = \frac{1}{n} \sum x_i$ and

$\overline{y} = \frac{1}{n} \sum y_i$. Dividing the numerators and denominators of our solution equations by n gives us

$$
\begin{aligned}
a &= \frac{\sum x_i y_i - n\,\overline{x}\,\overline{y}}{\sum x_i^2 - n\,\overline{x}^2} \\
b &= \frac{(\sum x_i^2)\overline{y} - \overline{x}(\sum x_i y_i)}{\sum x_i^2 - n\,\overline{x}^2}
\end{aligned}
$$

If we add a little notation, namely,

$$
\begin{aligned}
S_{xy} &= \sum(x_i - \overline{x})(y_i - \overline{y}) \\
&= \sum x_i y_i - n\,\overline{x}\,\overline{y} \\
S_{xx} &= \sum(x_i - \overline{x})^2 \\
&= \sum x_i^2 - n\,\overline{x}^2 \\
S_{yy} &= \sum(y_i - \overline{y})^2 \\
&= \sum y_i^2 - n\,\overline{y}^2
\end{aligned}
$$

then we can write our answer even more succinctly as

$$
\begin{aligned}
a &= \frac{S_{xy}}{S_{xx}} \\
b &= \overline{y} - a\,\overline{x}
\end{aligned}
$$

Thus, we have found the *least squares line*. It's also called the *line of regression* or the *regression line*. It is the line

$$
y = ax + b
$$

where a and b are as just found.

Math 218 Home Page at
   `http://math.clarku.edu/~djoyce/ma218/`