Methods of Mathematics

Kenneth A. Ribet

UC Berkeley

Math 10B
April 26, 2016

There’s a “pop-up” lunch each Friday (through May 13) at the Faculty Club at 12:30PM. All are welcome.

Course evaluations for this class are open. “Although students received an invitation email and reminders along the way, previous research demonstrates that a personal reminder from the instructor and an explanation of how evaluations are used to inform your teaching can make a positive impact on response rate and quality.”

Two remaining lectures:

- Today: Least squares, linear regression
- Thursday: Dynamic programming—or maybe Euler's Method, or maybe Markov Chains. Your vote??

This class will meet for questions and review on May 3, May 5.

Our exam: Monday, May 9, 11:30AM–2:30PM.

Today's Simons Institute talk at 4PM is in Sutardja Dai Hall; refreshments will be served starting at 3:30PM.

The speaker is Martin Vingron from Berlin. The title of the talk is Recovering Biological Networks, and the Ubiquitous Inverse of the Variance-Covariance Matrix.

Cellular processes are governed by interactions among genes or their protein-products. Molecular biology has unraveled many of these interactions. With recent technological advances, extensive measurements can be made for large numbers of genes and other cellular entities. This raises the question whether from these measurements one can deduce relationships among the genes. This is called the network reverse-engineering problem. A simple mathematical object, the inverse of the variance-covariance matrix, is at the heart of a whole class of methods for attacking this problem. The talk will present biological questions and explain how the inverse variance-covariance matrix come into play. Novel applications include epigenetic factors and their interplay.

Office hours for me this week are as usual: Monday and Thursday in Evans, and Tuesday at the SLC.

Next week, however, I will be away Monday (all day) and Tuesday (morning). Therefore, my regular Monday and Tuesday office hours are cancelled during RRR week. However, I will hold office hours as follows:

- Thursday 10:30–noon,
- Friday 10:45–12:15 (lunch to follow).

Wednesday, May 4:

Field trip to the Big C.

Meet 2PM in the Foothill parking lot (south end).

Although today's class concerns statistics, we begin with a question that does not seem to involve statistics: Suppose that we have n points in the plane, say (x_1, y_1), (x_2, y_2),. . . (x_n, y_n). What line that fits these points the best?

If all of the x_i are equal, say to a, then we take the vertical line $x = a$ and we're done. For the sequel, we can and will assume that the x_i are not all the same. In particular, we will assume $n > 1$. Then we will look for a (non-vertical) line with equation $y = mx + b$; the coefficients m and b are to be selected.

Once we fix m and b, the quantities $mx_i + b$ become approximations to the y_i. The idea is to choose m and b so as to minimize the sum of the squares of the "errors" $e_i = y_i - (mx_i + b)$. (This method goes back to Gauss and is called the "least squares" method.)

To minimize $S = \sum_i (y_i - (mx_i + b))^2$ as a function of m and b, we take the derivatives of S with respect to m and b and set them both equal to 0. We get two linear equations for m and b and can solve them fairly easily.

To state the result, we introduce the averages $\bar{x}$ and $\bar{y}$ of the x_i and the y_i. It turns out that $\bar{y} = m\bar{x} + b$, so that b is determined once we know m. Here's the formula for m:

$$m = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

If we think of the values $x_1, \ldots, x_n$ as values of a random variable x, then $\bar{x}$ is the mean of x and the sum of squares in the denominator is the sample variance of x, σ_x^2. Thus we can start thinking in terms of statistics. . . .

Before we do that, we should pause for an explanation of how the formulas for b and m are derived. We want to minimize $S = \sum e_i^2$ and do this by setting to 0 the derivatives of S with respect to both m and b. The derivative with respect to b is $-2\sum e_i$, so we must have

$$0 = (\sum_i y_i - mx_i - b) = \sum y_i - m\sum x_i - nb.$$

Dividing by n, we get

$$\bar{y} - m\bar{x} = b,$$

as was stated.

To simplify, let's suppose that $\bar{y}$ and $\bar{x}$ are both 0. (We can actually reduce to this case by translating the axes over so that the origin is moved to the center of mass of the cluster of points (x_i, y_i).) Then $b = 0$; the line now passes through the origin and is given by the formula $y = mx$. We need to choose m so as to minimize the sum S.

The derivative of S with respect to m is $-2\sum e_i x_i$. Since e_i is now $y_i - mx_i$, we get

$$m \sum x_i^2 = \sum y_i x_i,$$

so $m = \dfrac{\sum x_i y_i}{\sum x_i^2}$. Note that $\sum x_i^2$ is non-zero because we are assuming that the x_i are not all equal. In particular, they can't all be 0. When we no longer assume that $\bar{x}$ and $\bar{y}$ are 0, we get the longer formula that appeared on a previous slide.

For each i, let

$$a_i = x_i - \bar{x}, \quad b_i = y_i - \bar{y}$$

and let

$$a = (a_1, \ldots, a_n), \quad b = (b_1, \ldots, b_n).$$

Then

$$m = \frac{a \cdot b}{|a|^2} = \frac{a \cdot b}{|a|\,|b|} \times \frac{|b|}{|a|} = \left(\frac{a}{|a|}\right) \cdot \left(\frac{b}{|b|}\right) \left(\frac{\sigma_y}{\sigma_x}\right).$$

When we write

$$m = \left(\frac{a}{|a|}\right) \cdot \left(\frac{b}{|b|}\right) \left(\frac{\sigma_y}{\sigma_x}\right),$$

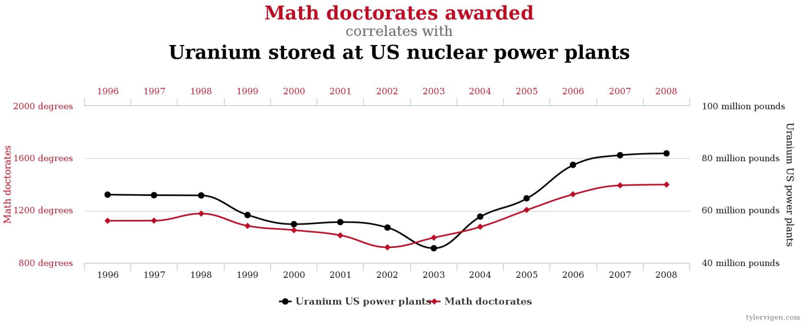
we are assuming that the y_i are not all equal—if they are equal, $m = 0$ and the line is horizontal. Recall that we assumed already that the x_i are not all equal.

The two vectors $\dfrac{a}{|a|}$ and $\dfrac{b}{|b|}$ are *unit vectors*, meaning that they have length 1. By the Cauchy–Schwarz inequality, their dot product is between -1 and 1. The slope m is then the product of this dot product and the scaling factor σ_y/σ_x. This factor compares the changes in y with the changes in x in the sense that the standard deviation measures the deviation of a random variable from its mean.

The dot product is an extremely interesting quantity. If, for example, $\frac{a}{|a|} = \frac{b}{|b|}$, the dot product is 1. If the two vectors are negatives of each other, the dot product is -1. If $a_i = 0$ for all i such that b_i is non-zero, the dot product is 0. We can think of the dot product as a measure of how well aligned the two unit vectors are with each other.

The dot product $\frac{a}{|a} \cdot \frac{b}{|b|}$ is the correlation between x and y.

Surgeon General’s warning: correlation does not imply causation. Quantities can move together because they are pushed in the same direction by a third quantity; for example, having yellow stains on one’s fingers is correlated with lung disease. (Both can come from smoking.) Things can move together for no discernible reason.



Another warning: the equation for our line was $y = mx + b$. In statistics, one apparently tends to write

$$y = \beta_0 + \beta_1 x$$

instead.

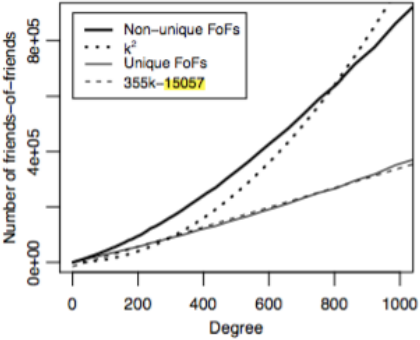
For each i, the difference $y_i - (\beta_0 + \beta_1 x_i)$ is called the ith *residual*. The least squares method minimizes the sum of the squares of the residuals.

Now what is least squares good for? The answer is that we can use collected data to posit an approximate linear relationship between two random variables and then estimate uncollected values from the linear relationship.

For example, we can take n Facebook users; for the ith user, let x_i be the number of that person's friends on Facebook and let y_i be the number of that person's "friends of friends." Then

$$y_i \approx \beta_0 + \beta_1 x_i,$$

with $\beta_1 \approx 355$
(http://arxiv.org/pdf/1111.4503v1.pdf).



Are you familiar with Packd?

Packd estimates the number of people in a room by counting the number of cell phones in the room. Presumably, the estimate is obtained by a linear expression $y = mx$, where m is calculated by observing points (x_i, y_i), where x_i is the number of phones in the room (as calculated by Packd's sensor) at the ith time and y_i is the number of people in the room (as calculated by a human).

> *We install a central hub at your location in fewer than five minutes. The hub is a small device that passively counts how many mobile phones are in a location based on WiFi signal strength. We'll toggle the radius of capture for you in order to accurately count your location. The hub correlates these counts to [the] number of people to provide realtime room occupancy estimates.*

Note the word "correlates"!

Now we can get weird. For each student in this class who took the two midterms, consider the ordered pair (first MT score, second MT score). In this way we get pairs (x_i, y_i) with $1 \le i \le 291$. The best linear fit

$$y_i \approx \beta_0 + \beta_1 x_i$$

comes with $\beta_0 = 16.21$, $\beta_1 = 0.46$. What this seems to mean is that the variation in one's first MT score accounts for only half the variation in one's second MT score. In fact, the correlation coefficient in this case is 0.46.

You can ask whether the two 0.46s are the same. Actually, $\beta_1 = 0.467499667$ (and should have been reported by me as 0.47) while the correlation coefficient is a bit smaller (0.460775). Recall that

$$\beta_1 = (\text{coefficient of correlation})\frac{\sigma_y}{\sigma_x}.$$

In this case,

$$\sigma_y \approx 5.109754049256329, \quad \sigma_x \approx 5.036253213183568.$$

The numbers match.