"It is part of the human condition that we are statistically rewarded for punishing others and punished for rewarding them" – Daniel Kahneman, Nobel Laureate in Economics, 2002, despite having never taken an economics course. (His 2011 book *Thinking Fast and Slow* cannot be recommended highly enough).

Our society is consumed with the idea collecting data, but, as you have learned in this course, data cannot be taken at face value – it needs to be properly interpreted to have any meaning, and if improperly interpreted can be extremely misleading.

One area in which data collection has become an obsession is primary and secondary education. Students take a battery of standardized tests and some of the money for a school hinges on whether its students perform better on a test in one year than they did in the last year. Many school districts are now using these tests to evaluate teachers as well. Woe betide you if you are a teacher whose students do worse than the students the year before. The LA Times has been especially vocal in pushing these tests and publishing teacher scores (see: http://tinyurl.com/p756ft5 and commentary http://tinyurl.com/27u3943 and a more general commentary on value added tests http://tinyurl.com/3ddtrs9. The references on this last web page include arguments both for and against value added testing). But what exactly can we conclude if a poorly performing school improves the next year or if a high-performing school declines?

In this lab you will explore one particular and universal feature of the relationship between two quantitative variables. This phenomenon is most clear and interpretable when the two quantitative variables are similar (e.g. heights of a parent and an offspring [collection of two heights] or the average class scores on a test for each teacher in two successive years [collection of two average test scores]).

Note: please don't read ahead to steps 9-13 until you do steps 1-8 in lab.

Activity

1. Your TA will read aloud 15 words. Try to memorize them. After 15 seconds write down all of the words from the list that you remember below.

2. Count how many you got right. Your TA will now read another set of words and again you will try to memorize them. How many do you expect to get right in the second round?

3. Write the second round words that you remember here:

4. How many did you get right? Were you correct in your prediction from (2)? If not, did you over or underestimate?

5. Open up the regression applet which you can find from this page: http://www.math.hope.edu/isi/applets.html.
Clear the data and type:
first second
on the top line. Be sure to leave a space between "first" and "second"
Now go around the classroom and each person should say what he or she got on the first quiz and on the second quiz. Enter these numbers on each line. Leave a space between them. Be sure to press "Enter" after the last one.

6. Before determining what the regression line is, what would it mean in terms of how people's score change from the first to the second quiz if:

(To answer these questions, describe what would happen on the second quiz in the case that someone did better than average on the first quiz and in the case that someone did worse than average on the first quiz.)

(a) the slope of the regression line equals 1?

(b) the slope of the regression line is greater than 1?

(c) the slope of the regression line is less than 1?

7. Click on the checkbox to show the regression line. What is the slope?

8. Write down some possible explanations for why these improvements and declines occurred.

9. Email yourself your section data so that you have it. Then go to the Datasets section of the class website and go to *Galton's Height Data*. Or just go directly to this URL:
http://labs.biology.ucsd.edu/rifkin/courses/bieb100/w14/GaltonHeightData.txt
This is one of the most famous datasets in statistics and is one that Francis Galton (Darwin's cousin) used when he discovered the concept of statistical correlation and invented regression. This specific dataset also led him to discover the phenomenon you are discovering here.

Make sure the *Show Regression Line* checkbox is unchecked. Then copy the data from the webpage into the applet. Each row contains the height of a father (first column) and the height of his son (second column).

Check the *Show **Movable** Line* checkbox and drag the blue line to where you think the regression line will be. Based on your answers in (6), what would the slope of your line imply about the relationship between father and son heights?

10. Now check the *Show Regression Line* checkbox. What is the slope of the line?

11. This data has the same form as the memory quizzes you did before. Do the explanations you wrote in (8) make sense in the context of height?

12. It turns out that **whenever** the correlation between two quantitative variables is less than perfect (1 or -1), then this phenomenon of "regression to the mean" will occur. A value far from the mean for one variable will be paired, more often than not, with a value closer to the mean for the other variable. This comes from the relationship of the regression slope and the correlation coefficient:

$$b = r\frac{SD_Y}{SD_X}$$

If the correlation is less than perfect, then extreme values in X will most likely be paired with less extreme values in Y and vice versa. As a sidenote, this idea led to a major debate between Galton's protégé Karl Pearson and his fellow *Biometricians* and those who rediscovered Mendel's laws in the early 1900s. R.A. Fisher's monumental and impenetrable paper of 1918, in which he invented the analysis of variance and the term *variance*, finally reconciled these two opposing camps and was the beginning of modern evolutionary biology and quantitative genetics.

13. One final thing to consider. Kaplan (with an office in the student center!) is a standardized test prep company. Here is their "Higher Score Guarantee": "You will score higher with Kaplan—guaranteed or your money back." Other test prep companies have similar guarantees. Based on what you have learned in this lab, is this a good business model?