

## Practical Stats Newsletter for November 2005

In this newsletter:

1. Upcoming Courses
2. Walking the Line - Alternatives to Regression
3. Future topics - seeking your input

1. Upcoming Courses

Both of our current courses will be taught in Spring 2006.

### **Applied Environmental Statistics,**

our week-long survey of "how to make sense of your data", will be taught in Portland OR at the Mark Spencer Hotel on March 27-31. Registration is \$1495, which covers all course materials, digital textbook, and data. Links for registration can be found at <http://www.practicalstats.com/Pages/aes.html>

### **Nondetects And Data Analysis,**

the 2-day course on handling nondetect data using modern methods of survival analysis, will be taught May 4-5 at the Hotel de Anza in San Jose, CA. This is the Thursday and Friday just prior to the National Monitoring Conference, also in San Jose (see [www.nwqmc.org](http://www.nwqmc.org) for more detail). A great way to save some travel costs and come to both the course and conference in one trip! Registration for the course is found at: <http://www.practicalstats.com/Pages/lto.html>

Course content for each course is listed on the Practical Stats web site. For all other information, email us at [ask\[at\]sign\[practicalstats.com\]](mailto:ask[at]sign[practicalstats.com]).

### 2. Walking the Line - Alternatives to Regression

Most of our recent newsletters have focused on issues of handling nondetects. We thought we'd go back to basics this time, and look at possible problems when a regression equation is used to fit a straight line to data. The goal of regression is to fit a line that minimizes the squared differences between the points and the line, where differences are measured in the Y direction. No errors in the X direction are directly accounted for. The original development of regression was with designed experiments where the X values were set - specific temperatures, specific amounts of fertilizer applied, etc. There was no error in the X variable to consider.

When applied to data with errors in both the X and Y directions, two characteristics of the standard regression line may cause the data analyst to consider alternative approaches. First, the regression line does not necessarily go through the 'middle' of the data. This may be a problem if you are using the line to represent some underlying science about the relationship between X and Y - that is not the objective of regression. Second, a set of predicted values picked off the line will have less variability than does the original data. This may be a problem when using regression to fill-in missing data. There are alternative ways to fit a straight line to data that avoid these two problems.

### Issues with Ordinary Regression

a. Regression does not go through the 'middle' of the data. Regression is built to minimize the errors in the Y direction. Its objective is to provide the 'best' (lowest-variance) prediction of Y for a given X. To do this, it produces a slope that is generally less steep than the one you would draw if you put a central line through data on a plot. Or in other words, if you flipped the X and Y axes, regression would give you a different line. Neither regression line necessarily goes through the center of the data. See Figure 1, attached as a pdf file.

If you were expecting to compare the regression slope to a theoretical slope based on science, or from a deterministic model, don't. It won't necessarily compare well. It isn't designed to produce a slope that describes how Y and X co-vary. It describes how to get the best predictions for Y. It is Y-centric. So if your assignment of X and Y is arbitrary, regression will give you different lines depending on which you choose as Y. Another method, the Line of Organic Correlation (LOC), has been used instead of regression in fisheries science and geomorphology to relate two variables, when the point is to understand the relation between the variables rather than predicting one or the other.

b. Regression predicts values having lower variance than the original data. Regression is sometimes used to fill in missing data. A water-quality variable (Y) is related to streamflow (X), for example, through a regression equation. For times without Y data, Y is predicted from the equation using the measured flow. However, the variability of the estimated record will be lower than it would have been had the original data been measured. The difference can be seen in the attached LOC.pdf plot as the lower slope for regression compared to the LOC. The lower slope produces predictions with lower variability. If after filling-in data with regression, the variability of the combined Y data is of interest, as when asking the question "how often does the concentration exceed 100?", then the combined record will be too consistent, and the probability of exceeding the limit will be underpredicted. The answer will be wrong, using regression to fill in data. However, predictions from the LOC line would work quite well.

### Alternative methods for fitting lines

The Line of Organic Correlation (LOC) is one alternative to regression. It predicts Y data that collectively have the same variance as the original data. It gives the same equation for the line when computed as Y versus X, or X versus Y. It goes by several names (MOVE; geometric mean functional relation; and reduced major axis). Its primary usefulness has been for filling in missing data having the same variance as the measured data. It has also been used for calibration, when a single line is needed for predicting Y from X as well as X from Y. LOC minimizes the area of triangles drawn between the data and the line in both the X and y directions. So errors in both the X and Y variables are simultaneously considered. The slope of the LOC line is just the ratio of the standard deviations for the X and Y variables, multiplied by the algebraic sign of the correlation coefficient:

$$\text{LOC slope} = \text{sign}(r) * \text{sdY}/\text{sdX}$$

The intercept is found by fitting the line with the LOC slope through the point: mean of X and mean of Y.

More information on LOC is given in the textbook by Helsel and Hirsch (2002), which is freely available from the US Geological Survey at

<http://pubs.usgs.gov/twri/twri4a3/>

where you can download either just Chapter 10, or the entire book. LOC was applied to environmental chemistry in the article: Hirsh, R. and E.J. Gilroy, 1984. Methods of fitting a straight line to data: examples in water resources; Water Resources Bulletin 20, 705-711.

A second possible line through the center of data uses principal component analysis (PCA). In two dimensions, the first principal component is the line that minimizes the distances between the points and the line in a direction perpendicular to the line. Unlike regression, the line is not optimized in the Y direction only. It is sometimes called the 'major axis'. PCA is often close to the line you would draw by eye. The human eye-brain system sees 'error' on a plot perpendicular to the line's slope. This line and the LOC line are very similar, but not exactly the same, to one another.

More information on this topic is available in chapter 10 of the Helsel and Hirsch text (link above).

3. Future topics - seeking your input

Topics we're looking at for our 'feature' (section #2) discussions in 2006:

Free software for statistics

Statistical methods for regulatory decisions

The quantile test - testing percentages

What to do when all your data are nondetects

Statistics add-ons for Excel

Important sites on the web for practical stats

Let us know what you think of these, and if there are others you are interested in. We want your involvement. Email us at: [ask\[at\]practicalstats.com](mailto:ask[at]practicalstats.com)

Replace the stuff inside and inclusive of the brackets with the @ sign.

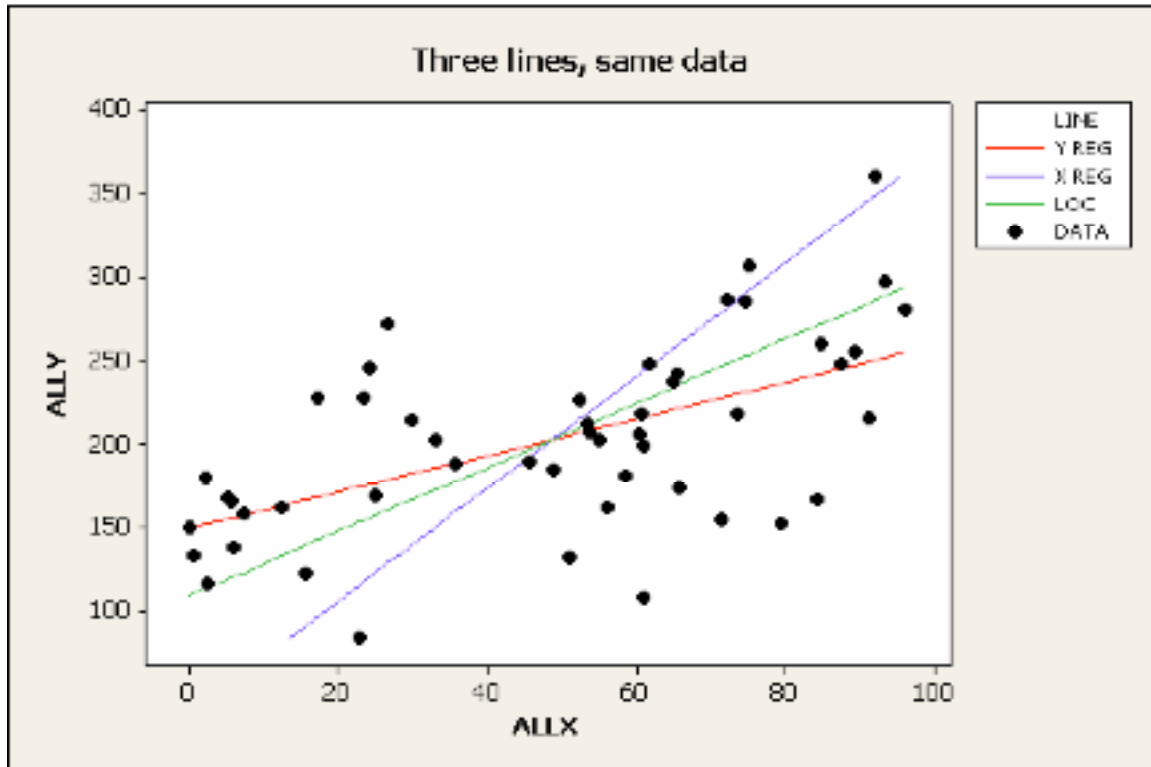
Speaking of involvement, if you've found this newsletter helpful, consider sharing it with others in your office or with contacts interested in environmental statistics. We would like a larger signup base to make the time we put into it more worthwhile. We never give your email address to anyone else. We send this only to people who request it - see the signup link at the top of the page. Our goal is to make each issue informative, worthwhile reading.

'Til next time,

Practical Stats

<http://www.practicalstats.com>

-- Make sense of your data



The red line is the usual regression line using Y as the y variable.

The blue line is the inverse regression line, with X as the y variable.

The green line is the LOC (Line of Organic Correlation).