Practical Stats Newsletter for August 2017

Subscribe and unsubscribe: http://practicalstats.com/news
Archive of past newsletters http://practicalstats.com/news/archive.html

In this newsletter:
A. Practical Stats Courses
B. "Normalizing" by Ratios
C. What is Normalizing?

A. Practical Stats Courses
Our Applied Environmental Statistics courses are on our online training site:
http://practicalstats.teachable.com/
The two courses separately are each $650 for a 1-year access for one person. Or get both courses together in a bundle for $1200.
More courses, and at least one new free webinar, are also coming soon.

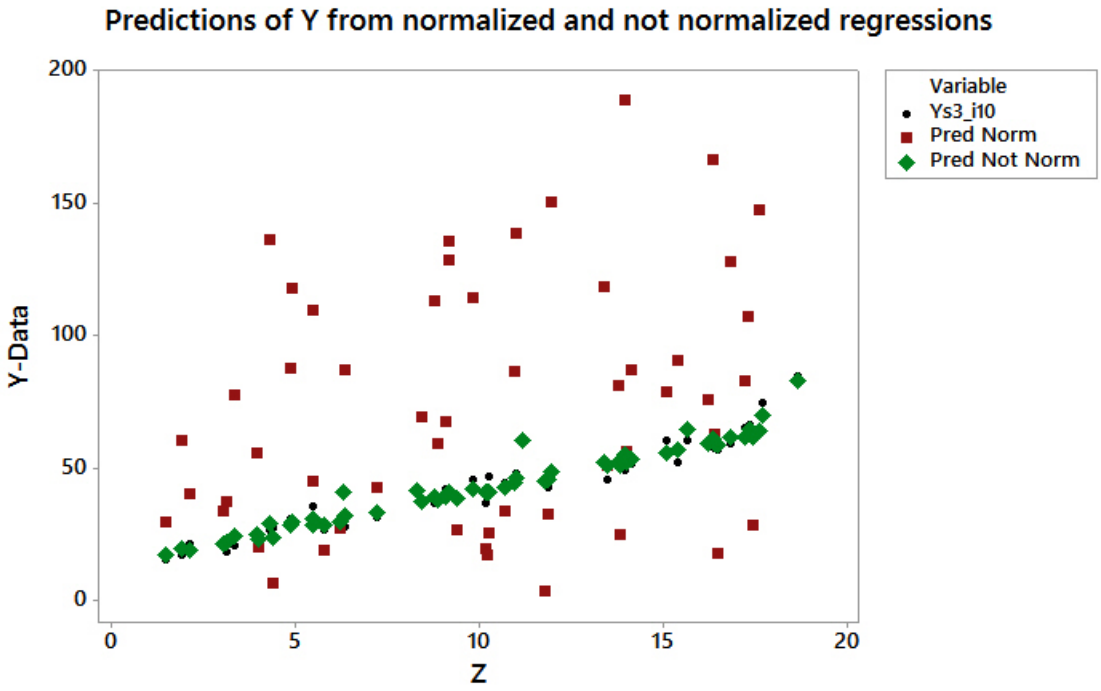
B. "Normalizing" by Ratios
Scientists sometimes divide one variable by a second, declaring that the ratio of the two has greater meaning than either individually. Examples over the years in a variety of disciplines include dividing streamflows by drainage areas; sediment loads by drainage areas; metals concentrations by the concentration of either aluminum or titanium (as a surrogate for clay percentage of a solid sample); invertebrate densities by mean depth of the lake; and contaminants in biota by lipid concentrations. There are many more examples. The intent is to equalize results so they can be compared across disparate locations and conditions. The consequences of doing so are rarely understood. A ratio is a simplistic regression equation, assuming a constant slope across all conditions and that the intercept equals zero. In natural systems neither of those two assumptions is likely to be true. The ratio function is a multiplicative scale, with unequal, nonlinear scales on either side of a ratio of 1. For example, the ratios Y/X equaling 100 and 0.01 are the same distance from the equality value 1 on a multiplicative scale, but not on the usual additive scale. Does the scientist intend that ratios of 100 and 0.01 are equivalent in intensity? Ratios above 1 are unbounded while ratio below 1 are bounded by 0 (assuming both Y and X are limited to positive values), producing an asymmetry that is likely to be unintended. This asymmetry produces biased estimates in the mean ratio and other statistics, as compared to computations in original units.

Predictions of Y computed from ratios have larger variance than would a standard regression approach. In essentially every case a better analysis would result by using the numerator as the Y variable in a regression equation, and the denominator as one of the X variables. Let the intercept be determined by the data instead of forcing it to zero. Plot the data to see if an assumption of a constant linear slope is reasonable – in natural systems the linearity will often be improved by transforming one or more of the variables in the regression.

Two examples:

Ulrich et al. (2003) found that computing enantiomeric ratios, the ratios of two organic chemical species, resulted in poor estimates of the center when using the mean ratio; added extreme outliers that were not present in original units; inhibit use of simple t-interval estimates around the mean ratio because of the induced asymmetry; and produced patterns that obscured differences between locations.

To see why this might be so, the second example involves data generated so that X and Y are linearly related to each other. The generating formula is Y = 10 + 3*Z + 0.5*X + e, where e is a set of normally distributed residuals. Note that there is a non-zero intercept and a second variable complicating the equation, both of which are realistic conditions for environmental applications. No deterministic relationships between Z and X were used ("nothing up my sleeve"). The "normalized" ratio Y/X is used as the response variable in a regression with Z as the explanatory variable. The resulting equation is Y/X = 8.4 + 1.60 Z with an r-squared of 3%. Predicting the "normalized" Y/X from Z using the regression and then multiplying by X to estimate Y values produces the estimates shown as scarlet squares in the figure below.



The regression without normalization using both X and Z as explanatory variables is Y = 11.096 + 2.8639 Z + 0.6368 X, with an r-squared of 97%. Predictions from this equation are shown as the green diamonds in the plot. Also shown are the original Y data as black dots. Note that the non-normalized predictions go right through the center of the original Y data. Also note the huge increase in variance shown by the normalized ratio predictions. In terms of prediction capability the difference is dramatic.

This topic has been a concern of mine throughout my career, primarily because of the widespread use of "normalized" ratios by geologists, hydrologists and biologists I've

worked with. I have perhaps two journal articles 'left in me' before I completely retire, and this is likely the topic of one of them. If you have a dataset that is publicly available where normalization was used, I'd like to see it to determine if it would be a good field example to complement the above simulation. If so, please send me an email at ask@practicalstats.com.

References

Ulrich, E.M., Helsel, D.R. and W.T. Foreman, 2003. *Complications with using ratios for environmental data: comparing enantiomeric ratios (ERs) and enantiomer fractions (EFs)*, Chemosphere v.53, pp. 531–538.

C. What is Normalizing?

Statisticians avoid using the term "normalizing" because it has perhaps ten different meanings. Therefore it has no meaning. Here we've discussed its use when one variable is divided by a second. You will often find it used in place of the more correct term "standardizing" when subtracting a variable's mean and dividing by its standard deviation, so that the result has a mean of zero and standard deviation of one. In none of the procedures where the term is used does the result look more like a normal distribution. The potential for confusion is obvious. I recommend that you learn what the more precise and correct term is for the process you call "normalizing", and switch to the better term.

'Til next time,

Practical Stats
-- Make sense of your data