Practical Stats Newsletter for March 2010

Subscribe and Unsubscribe: http://practicalstats.com/news
All of our past newsletters:
http://www.practicalstats.com/news/news/bydate.html

In this newsletter:
1. Upcoming Courses
2. Regression and Trends with Censored Data
3. See you at the conferences?

I'll leave the date of this newsletter as March – its "belated". You can get all kinds of belated birthday and other cards at Hallmark, so I think its OK here. I waited to release this until the date and location of our upcoming AES course was firm.

1. Upcoming Courses
Applied Environmental Statistics
Sept. 13-17, 2010. Las Vegas, NV.
(registration online in about a week)

Untangling Multivariate Relationships
and
Nondetects And Data Analysis
will be offered back to back, most likely in October 2010. Stay tuned.

You can always find our complete course listing at
http://www.practicalstats.com/new_classes/classes.html

2. Regression and Trends with Censored Data
By now you are probably used to the term 'censored data' when referring to nondetects. Censored observations are those where an individual number is not known, but it is known that the value is below or above a threshold. For nondetects, the threshold is the reporting limit of the laboratory, so that values are reported as "<10" for a reporting limit of 10. The concentration is located somewhere between 0 and 10.

Regression of data where censored values are present can be a challenge. The simplistic approach sometimes used is to fabricate values for nondetects, putting something like one-half the reporting limit in the place of each nondetect value. Two problems result from this fabrication. First, the same value is used for each observation with the same reporting limit. This produces a false consistency in the data, reducing the variation from what would have occurred in nature. Estimates of standard deviation and mean-square error used in regression computations can easily become too low as a result, producing p-values that are more significant than

they should. Second, values tied to a reporting limit such as one-half RL vary as the reporting limits vary. Limits have often gone down over time for many chemicals as methods have become more precise. The declining limits result in a decline in fabricated values. This induces an artificial trend in the data that is not necessarily present in the actual data. This 'invasive pattern' results not from what was in the sample, but from changes in lab methods, changes in the amount of sample submitted to the laboratory, or changes in other factors that influence the detection limit but are unrelated to the concentration of the chemical of interest.

To avoid these problems resulting from fabricating values that are not really there, methods for censored data can be employed. The parametric approach equivalent to standard regression is 'regression with life data' or 'tobit analysis', regression where the slope and intercept are estimated using maximum likelihood estimation (MLE) rather than the usual least-squares approach. My 2006 paper "Fabricating Data" in Chemosphere
http://dx.doi.org/10.1016/j.chemosphere.2006.04.051
discusses the problems with fabricating data and the benefits of using MLE regression instead. MLE regression produced slope estimates for the example data in that report that were as good or better than that produced for all fabricated values between 0 and the reporting limit.

There are times where parametric assumptions of regression, notably that the residuals are distributed around the regression line as a normal distribution, do not hold. In that case a nonparametric method may be preferred. The Mann-Kendall test for trend is one popular example, where the trend slope is computed by the median of all possible pairwise slopes of data. This slope (sometimes called the Sen slope) along with an estimate of intercept allow a straight line fit to the data without assuming normality. Though the name "trend" implies that the x variable is time, there is no actual restriction to that application, and the line can be used in any situation where regression might be performed. As with regression, the Mann-Kendall line was not developed for the situation of censored data. The presence of nondetects means that some of the pairwise slopes cannot be exactly determined, and so the Sen slope may not be able to be uniquely found.

In the early 1990s, Michael Akritas at Penn State developed a censored analogue to the Mann-Kendall procedure that allows a slope and intercept to be computed for censored data in a fully nonparametric fashion. The procedure is described in the textbook Nondetects And Data Analysis (Helsel, 2005), where it is given the abbreviation ATS (Akritas-Theil-Sen). In short, it uses the property of the Sen slope that when subtracted from the data, Kendall's tau on the residuals is exactly zero. Subtracting the Sen slope produces a zero correlation in what's left. The ATS line searches for a slope that produces a zero Kendall's tau correlation when subtracted from the (censored) data. No substitution is required.

Software.
Software for performing MLE regression is available in commercial statistics packages and in the NADA for R package that accompanies the NADA textbook mentioned above. In NADA for R, the procedure is available in the cenreg function. Links to NADA for R are on our website, at:
http://www.practicalstats.com/nada/nada/nadar.html

In contrast, the ATS procedure is not currently available in commercial software other than by the Minitab macro %ats that is part of the NADA for Minitab package available at no cost on our website.
http://www.practicalstats.com/nada/nada/downloads.html

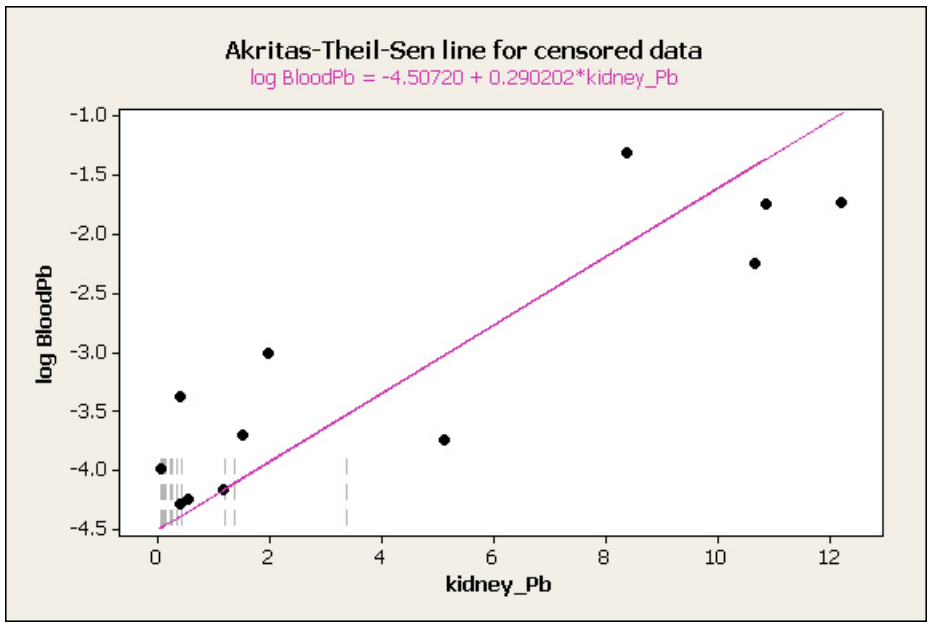
ATS is also available in NADA for R through its cenken function. Using the NADA packages for either R or Minitab allow ATS to be computed from data with multiple reporting limits. The ATS line estimates the median Y for changing values of X. With ATS, both the Y and X variables can be censored. MLE, though more widely available, assumes a specific distribution for the residuals, and allows only the Y variable to be censored.

An example use of MLE and ATS is below. The dashed gray lines on the plot below are the censored values, which are incorporated into the regression analysis. See NADA (2005) for more detail. Or attend our next Nondetects And Data Analysis training course, coming this fall.

Y variable: logarithm of lead levels in the blood of herons
X variable: lead levels in the kidneys of herons

```
Estimation Method: Maximum Likelihood
ln(Blood lead) = -4.457 + 0.2436* kidney lead
       p = < 0.0001

Estimation Method:  ATS
logBloodPb = -4.5072 + 0.290202*kidney_Pb
       p = 0.0004      Kendall's tau = 0.42
```



3. See you at the conferences?
Practical Stats will be at two upcoming conferences:
a) National Water Quality Monitoring Conference. April 26-29, Denver CO.
b) NALMS 19th Annual Southeastern Lakes Management Conference, May 4, Winston-Salem NC.

Stop by the talk "Man vs. Stats" by Dennis Helsel at either conference, and improve your survival skills in data analysis.

'Til next time,

Practical Stats (Dennis Helsel)
-- Make sense of your data