Practical Stats Newsletter for June 2008

Subscribe and Unsubscribe: http://practicalstats.com/news
Archive of past newsletters: http://practicalstats.com/news

In this newsletter:

1. Practical Stats Consulting Services

Practical Stats now offers consulting services to agencies and companies, providing expert statistical analysis for environmental issues through subcontracting with us, rather than hiring an employee. Services include statistical investigations (data analysis, risk analysis, vulnerability studies), report review and preparation, and individualized advice and training. Services are provided primarily by Dennis Helsel, Ph.D. More information, including costs and qualifications, can be obtained by email http://www.practicalstats.com/contactus ,
directly to Dennis Helsel: dhelsel[at sign]practicalstats.com ,
or by phone (303) 870-4921.

2. Logistic Regression – predicting probabilities/risk

Logistic regression is the only statistical procedure that is taught in each of our 3 courses on applied statistics. In our *Applied Environmental Statistics* course it is used to predict the probabilities of some event happening, such as an exceedance of a standard. In *Nondetects And Data Analysis* it is used to perform regression on data classified as either 'below the detection limit' or 'above the detection limit'. In *Untangling Multivariate Relationships* it is used as an alternative to discriminant analysis, classifying data measured with multiple variables into one or another group. You can see that logistic regression is quite a useful and broadly applicable method.

Logistic regression predicts the probability of an event or group assignment; thus the y-variable is predicted as a value between 0 and 1. The explanatory variables are any variables that would be used in a standard regression equation. Unlike the more familiar least-squares regression, logistic regression requires no normality assumptions about regression residuals. It employs likelihood ratio tests to determine if one or more explanatory variables significantly predicts the pattern of occurrence/assignment of the y variable. Examples might include predicting the probability of a landslide occurring based on rainfall intensities and slope; the probability of exceeding a water quality standard or detection limit as a function of streamflow and upstream use of a chemical contaminant; or the probability of being classified into one of several lithologic facies groups as a function of presence or absence of a series of fossil types and several geochemical measures.

One of the complicating issues in multiple regression carries over into logistic regression – multicollinearity. This is correlation among the explanatory variables, inflating the variance of the slope coefficients and so obscuring the significance of individual variables. With multicollinearity present, some of the explanatory variables may be dropped from the equation when they should not be. This is particularly a problem with stepwise approaches to both ordinary and logistic regression. To detect the problem, compute the variance inflation factor (VIF) or related statistics for each explanatory variable by using your multiple regression package.

Slope coefficients in logistic regression have a different meaning than with ordinary regression. Slopes are multipliers of the natural logs of the odds ratio, the ratio of the probability of an event divided by the probability of a non-event. This isn't very intuitive. Exponentiating, a column usually called "Odds ratio" in logistic regression output is the multiplier of the odds ratio itself. A value of 1 for this multiplier indicates there is no effect for that explanatory variable. A value different from 1 indicates the effect this variable has on the odds of occurrence. A value of 2, for example, indicates that a unit change in that explanatory variable results in a doubling of the odds ratio. If the y-variable were a detect/nondetect determination, a doubling of the odds ratio indicates that the probability of a detect divided by the probability of a nondetect is doubled. The rate of change for the probability of a detect itself changes as a function of the probability. As a result, a graph of the probability of a detected value versus values for the explanatory variable looks like an S-shaped curve rather than a straight line.

There are several resources for learning more about logistic regression. For an example of its use, the probability of nitrate concentrations in groundwater across the United States exceeding a value of 3 milligrams per liter were predicted using logistic regression (Nolan et al., Environ. Sci. Technol. 1997, 31, 2229-2236). For more on the procedure itself, chapter 15 of *Statistical Methods in Water Resources*, the free textbook for our Applied Environmental Statistics course, has an overview of the method. The textbook is available for download at
http://www.practicalstats.com/aes/aesbook/AESbook.html

There are also entire textbooks on the method, though not usually with examples from environmental science, including Hosmer and Lemeshow's 2000 book *Applied Logistic Regression*. The regression chapter in *Nondetects And Data Analysis*
NADA textbook on Amazon
also has a discussion on its use as applied to modeling detection frequencies.

3. Upcoming Courses and Talks

For courses, we're teaching the *Nondetects And Data Analysis* course in Seattle to Region 10 USEPA and guests on July 29-30. It is open to government agency personnel in Idaho, Oregon, and Washington State. It may be open to others working in those states. To find out, contact Diane Ruthruff (~~diane.ruthruff[at sign]epa.gov~~).

All our open-enrollment courses are listed at

http://www.practicalstats.com/new_classes/classes.html.
Next up is a one-day *Introduction to Practical Statistics* on Sept. 24, 2008, just before the California Groundwater Resources Association annual meeting in Costa Mesa CA. It will introduce some of the concepts in our *Applied Environmental Statistics* course, looking at them from the perspective of groundwater quality. For more information, see http://www.grac.org/stats.asp . If you know a ground water scientist who has been alienated by statistics, this workshop will begin the reconciliation.

Nondetects And Data Analysis, the course that illustrates methods for correctly handling data with nondetects, will be held at the Mainsail Suites and Conference Center in Tampa, FL on Nov. 11-12, 2008. Registration is $795 online at http://www.practicalstats.com/new_classes/classes.html. The Conference Center is near the Tampa International Airport, so no car rental (and its associated gas prices) are necessary. Special rates on hotel suites are also available - see the above registration link.

Immediately following on Nov. 13-14 at the same location is *Untangling Multivariate Relationships*. This course covers the multivariate methods of primary interest to environmental science, focusing on what each method is designed to do, when to use them, and when not to. More detail on course content is on our website. Registration is $895 through the same link, above. The following week in Tampa is the SETAC (Society for Environmental Toxicology and Chemistry) annual meeting.

As for talks, you can hear Dennis Helsel at the following locations:

American Statistical Association annual meeting
August 3-7 Denver CO
"Nanostatistics"

UseR conference
August 12-14, Dortmund, Germany
"NADA for R software"

Linköping University, Sweden
Week of October 6, 2008
"Survival Analysis Methods for Left-Censored (Nondetect) Data"

SETAC (Society for Environmental Toxicology and Chemistry)
Nov. 16-20, Tampa FL
"Summing Nondetects"

'Til next time,

Practical Stats

-- Make sense of your data