Practical Stats Newsletter for April 2012

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 Events
2. AIC: better than r2 for comparing regression models
3. New Address for Practical Stats

1. Upcoming Events

a. Face to face courses

Applied Environmental Statistics
Statistics, down to earth.
Hypothesis tests, regression done right, and trend analysis.
June 25-29, 2012
Doubletree Tampa Bay-North Redington Beach
$1395 through May 31st, $1495 after

b. Webinars

Together, these four 2012 webinars contain much of the content of our 2-day *Nondetects And Data Analysis* course. Get the content without the travel! Multiple people can be at one site for one registration fee.

June 18: Invasive Data: Why Not Substitute One-Half the Detection Limit? $50
July 16: Computing Descriptive Statistics for Data With Nondetects. $250
Aug 20: Hypothesis Tests for Data With Nondetects. $250
Sept 17: Regression and Trend Analysis for Data With Nondetects. $250

c. Conference presentations

- **Making Sense of Nondetects**. Toronto, Canada. April 30th, 2012 at the *2012 RPIC Federal Contaminated Sites National Workshop*.
- **It Ain't Necessarily So: Urban Legends in Environmental Statistics**. Portland, OR. May 3, 2012 at the *8th National Water Quality Monitoring Conference*.

You can register for courses and webinars on our "Upcoming Classes" page at
http://www.practicalstats.com/new_classes/classes.html

2. AIC: better than r2 for comparing regression models

How should I compare and find the 'best' regression model, or logistic regression model, or other statistical models? Which variables provide the best predictive ability for my data? For years, people have used r2 (r-squared) values, picking the equation with r2 closest to 1. However there are several reasons why r2 is not a good choice for this task. First, r2 often gets larger when outliers occur. Without looking at plots, you can get a

high r2 from a very poor model. Second, r2 is a function of the units of Y used. Explaining 45% of the variation in log(Y) (r2=0.45) may be a better model than explaining 50% of the variation when using Y as the response variable. Third, r2 will get larger as the slope increases for the same error around the line. This seems somewhat unfair, as the investigator has no control over whatever the slope may be. Fourth, r2 always increases as a new explanatory variable is added to a multiple regression model, even if that variable has no relationship to Y.

Newer measures of quality of a statistical model have been developed over recent decades. One of the most widely used is the Akaike Information Criteria, or AIC. Developed by Hirotsugu Akaike in 1974, AIC is in essence a cost/benefit analysis of the model. The cost for adding new explanatory variables is 2p, where p is the number of coefficients to be estimated (the number of explanatory variables plus 1 for the intercept). So the cost of a three variable model is 8, while for one explanatory variable the cost is only 4. The benefit is the reduction in error for the model as measured by 2 times the log likelihood. The log likelihood ln(L) measures the maximum match between observed data and predicted values from the model. So

$AIC = 2p - 2\ln(L)$

where L is the maximum likelihood for the model. Computed as cost minus benefit, the AIC measures the lack of fit for the model, and the minimum AIC model is considered the best. When an explanatory variable is added that explains a lot of the behavior of the Y variable, L increases more than the unit increase in p, AIC decreases, and the more complex model with the new variable is better than the model without it. A great benefit of AIC is that models need not be nested in order to compare them. So a model with explanatory variables A, B and D can be compared to one using variables C, D, and F using their AIC values.

Variations on AIC have arisen since 1974. One is the corrected AIC, or AICc, which modifies AIC for small samples sizes. AICc approaches AIC as sample sizes become large, so always using AICc has essentially no downside. Other measures of quality not directly related to AIC such as Mallow's Cp have also been developed. While people argue over which is best for a variety of situations, it is clear that any of the new measures improve over r2. AIC and AICc, because of their simplicity and ease of application to a variety of statistical models including linear regression, is probably considered the current standard procedure.

In our June Applied Environmental Statistics course you will use the software package R to compute AICc and other new measures of quality. If you haven't taken a class in statistics for a while, and are still depending on r2 for evaluating regression models, come to North Reddington Beach FL in June. Bring your family and see if they have a better time in the Gulf waters or whether you do in our class. Consider it our challenge!

3. New Address for Practical Stats

Our surface mail address has changed to

Practical Stats

2838 Mashie Cir
Castle Rock, Colorado 80109

Email addresses and phone remain the same. See the Contact Us page on our website.

'Til next time,
Practical Stats (Dennis Helsel)
-- Make sense of your data