Practical Stats Newsletter for May 2016

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

In this newsletter:
A. Upcoming Webinars and Talks
B. Deciding whether to transform X variables in multiple regression
C. Statistics in Water Resources

A. Upcoming Webinars and Talks

Our webinar "Seven Perilous Errors in Environmental Statistics" is currently available for free at our Online Training Center (http://practicalstats.teachable.com).

Our "Permutation Test and Bootstrapping" course will be at the Online Training Center by July, on demand at any time you want to start.

We are glad to come to your site and teach any of the six courses we offer in-person. See http://practicalstats.com/training/ for details.

B. Deciding whether to transform X variables in multiple regression

Our Applied Environmental Statistics course, which will be offered online this coming fall, covers methods for building good multiple regression models, among many other topics. One helpful guide for whether to transform the potential explanatory (X) variables in your equation is called a component+residual plot, or 'partial plot'. Its goal is to picture the relationship between the Y variable and that one X variable while accounting for the effects of all other X variables in the equation. The common practice of simply plotting Y versus X cannot do this, as all the effects of the remaining X variables are clouding the picture. Instead, an adjusted $Y_{adj} = e_i + b_j x_{ij}$ is plotted on the Y axis, removing the confounding relationships. Once we get a clear picture of the effect of an individual X variable, we can judge whether the relationship is straight or not. If straight, use the X variable as is. If the relationship is curved, transforming X to straighten it out will improve your model, its predictions for Y, and increase the r-squared for all hunters of maximum r-squared.

A component+residual plot (crplot) represents the regression line "edge-on in the direction of the corresponding predictor" (Fox and Weisberg, 2011), plotting the observed data around it. In this way it uses the same observational task as a probability plot – the proposed model is shown as a straight line, and the degree to which the data around it plot as a straight line defines how well the model (using the current units of X) fits.

For example, a multiple linear regression of total dissolved solids (tds) versus time and streamflow (Q) was computed, and a crplot of the model visualized:

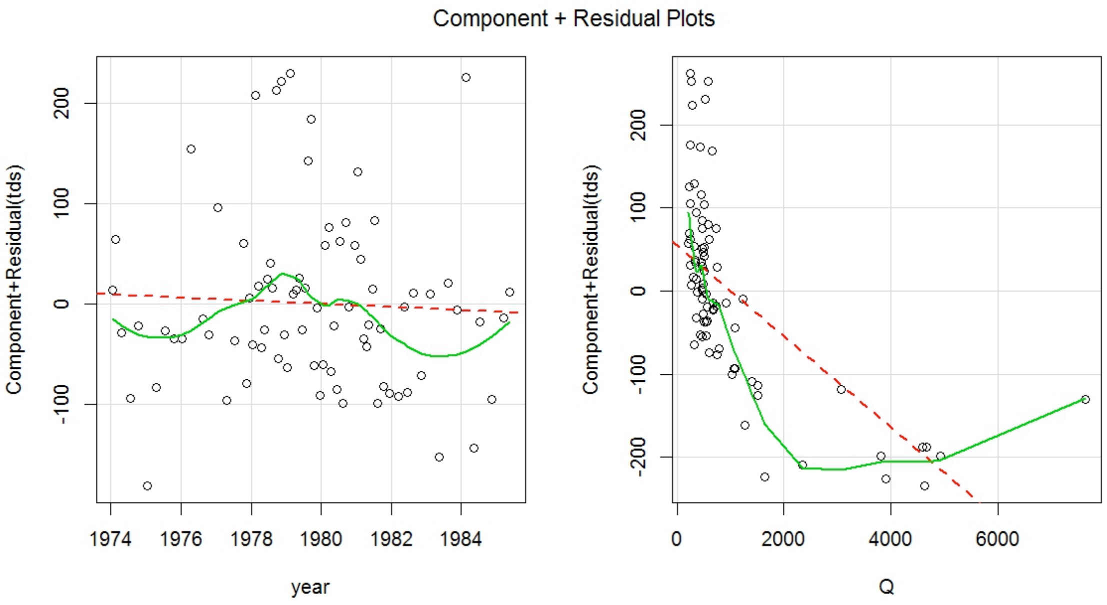

Figure 1 crplots for the regression of tds versus year and Q

Dashed red lines represent the multiple regression, showing the adjusted Y values versus values for each of the two X variables. The crplot to the right demonstrates that the relationship between Y and Q, after adjusting for the year variable, is curved. A better regression should be possible by transforming Q (which is often true for streamflows). Figure 2 shows crplots using log(Q) along with year in the regression. The data are much closer to a linear pattern around the dashed representation of the multiple regression line. The regression using log units for streamflow should be preferred.

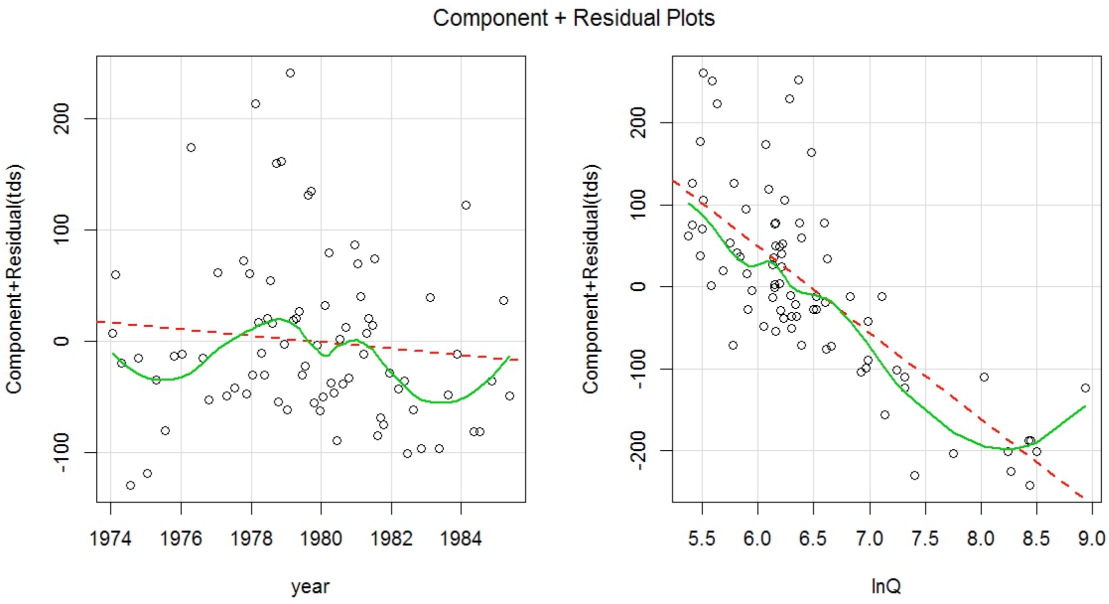

Figure 2. crplots for the regression of tds versus year and log(Q)

In the car package of R software, the crplot command will produce these plots.

Reference: J. Fox and S. Weisberg, 2011, An R Companion for Applied Regression, 2nd Edition. Sage Publications.

C. Statistics in Water Resources

The new edition of the classic textbook is on track to be finished by the end of this year. It will be available as a pdf and (we think) as a hardback textbook. To preview the in-depth information on water resources that will be contained in it, see this page: http://xkcd.com/1662/

'Til next time,

Practical Stats
-- Make sense of your data