Practical Stats Newsletter for February 2021

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

In this newsletter:

A. Two Months Left to Register for Our Courses

On our online training site: https://practicalstats.teachable.com/

Our two online courses will be accepting registrations through March 31, 2021. Starting April 1, new course registration will be closed. All who have registered will continue to have complete access and support from me for one year from their signup date.

Our Nondetects And Data Analysis (NADA) course is a complete coverage of data analysis with nondetects and 'remarked data': summary statistics, regression, group testing, trend analysis and even some multivariate methods, all without substituting fabricated numbers like ½ the detection limit. One year's access to the materials costs $795. The R scripts included provide 37 new functions to make data analysis easier, and are a step forward from the NADA package in R.

Our Applied Environmental Statistics courses cover methods from simple statistics through trend analysis. They are also an introduction to using R software, the most widely used statistics software in the world. They are available in two parts, each $650 USD for a 1-year access for one person. Or get both courses together in a bundle for $1200 USD. See our online training site at the link above.

B. Three Steps to Build a Good Multiple Regression Model

Both our AES and NADA courses cover a method to build good multiple regression models. It's the closest thing to a flowchart that I teach in these courses. I've only discussed the method in our courses, and never before put the three steps to build a good regression model in a newsletter. I'm getting soft as I get closer to retirement, I suppose. Here I'll illustrate the three steps with data and graphs from our NADA course.

Step 1. Decide which scale should be used for the Y variable

Should you transform the Y variable? If so, logs, cube-roots, something else? A regression is run and the residuals tested to see if they can be fit well by a normal distribution. The scale showing the best fit is the scale to use for the Y variable in the regression. In figure 1 the regression residuals after taking logarithms of atrazine concentrations are fit better by a normal distribution (atrazine is fit well by a lognormal distribution) than with the original scale or with any other common transformation. This is shown both by the relatively straight pattern of the circles (the detected observations) and the W statistic. The left side of the plot is less populated by circles because lower values include nondetects, which are not shown on the plot but are used to compute the positions (the normal quantiles) on the X axis for the detected observations.

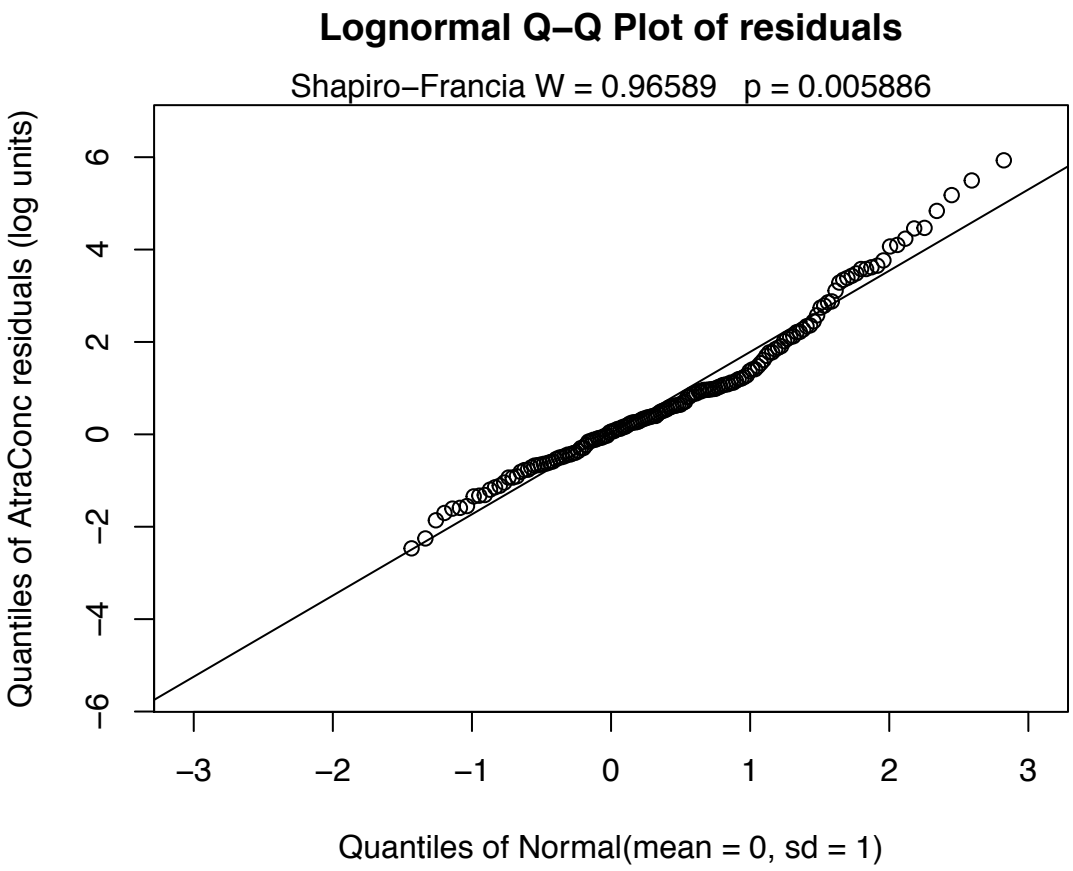

Figure 1. QQ plot of residuals from a multiple regression of atrazine concentrations (censored Y variable) and 8 explanatory variables. Note the Shapiro-Francia W is 0.965, very close to the max of 1.

Step 2. Decide whether to transform the scale of each X variable

Both numerical and graphical methods can be used to determine if an X variable should be transformed. The goal is to have a linear relationship between the Y variable and each X variable. In figure 2 are partial plots for the first six X variables in the regression. Curved relationships indicate that a transformation of the X variable is called for. One of the variables (PctCorn) is judged by a numerical method to be sufficiently curved to be transformed. While the smooth curve for Temp appears to contain waves, a straight-line relationship is judged to be as good of a fit -- the smooths added to the plots sometimes curve due to only a few points, especially at the right and left edges of the plots. The smooths incorporate both censored and detected observations.

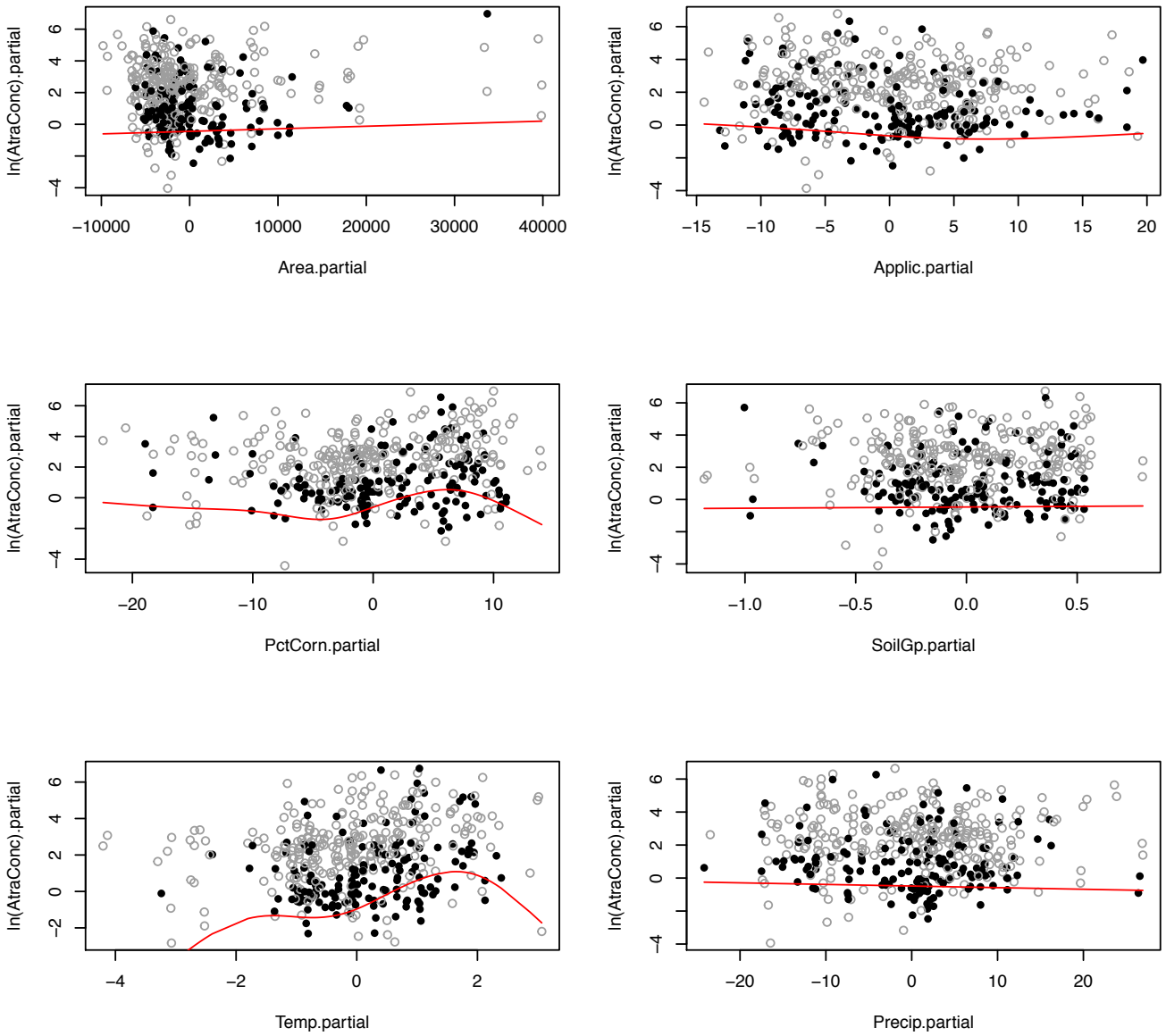

Figure 2. Partial plots of six explanatory variables. These plots help guide the decisions of whether or not to transform one or more X variables. Partial values for detected data are plotted as solid circles and for nondetects as gray open circles. Residuals for nondetects are computed at the detection limit value and the smoothing method knows that this is the maximum and computes the smooth accounting for the probabilities of lower residuals for nondetects.

Step 3. Delete unimportant X variables to find the lowest AIC model

Modern regression software usually includes a routine named something like "all possible regressions". These routines run …. regressions for all possible models built from the variables being considered. The best models are those with low AIC, a measure of error left unexplained by the model adjusted for the number of variables used. A human can then choose from among the top models. For eight variables there are 2^8 -1 = 255 possible regression models. After running all 255, the 10 models with the lowest AIC are printed.

```
n.xvars                                      model.xvars      aic
   5                 Applic Temp Dyplant Pctl logPctCorn 796.4067
   4                        Temp Dyplant Pctl logPctCorn 796.5552
   5                   Area Temp Dyplant Pctl logPctCorn 797.1077
   6             Area Applic Temp Dyplant Pctl logPctCorn 797.1513
   6          Applic SoilGp Temp Dyplant Pctl logPctCorn 798.0970
   6          Applic Temp Precip Dyplant Pctl logPctCorn 798.1015
   5                 Temp Precip Dyplant Pctl logPctCorn 798.3151
   5                 SoilGp Temp Dyplant Pctl logPctCorn 798.4167
   7     Area Applic Temp Precip Dyplant Pctl logPctCorn 798.8639
   6            Area Temp Precip Dyplant Pctl logPctCorn 798.8797
```

The lowest AIC model uses 5 X variables, Applic, Temp, Dyplant, Pctl and logPctCorn. Yet the second-best model with only 4 variables could be chosen instead if, for example, Applic was expensive to collect. The difference in AIC between the first two models, one without Applic, is negligible. The routines to compute partial plots and all possible regressions for censored data are demonstrated more fully in our NADA training course, so register soon if you'd like more guidance on these methods.

The routines shown above will also be part of the NADA2 package for R when it becomes available within a month or so. The training on how to use them is only available by registering for either AES or the NADA course by March 31st.

C. NADA2 Workshop in April at the Natn. Monitoring Conference

The upcoming NADA2 package for data analysis with nondetects using R statistical software will be demonstrated in a workshop I'll conduct at the 12th National Monitoring Conference (https://www.nalms.org/2021nmc/) during the week of April 19-23, 2021. The conference will be held entirely online. There is no extra charge beyond the conference registration fee to attend the workshop. The NADA2 package consists of methods for statistical analysis for data with nondetects used in our Nondetects And Data Analysis (NADA) course, including those above for building multiple regression models. For more information on NADA2 see our August 2020 newsletter.

'Til next time,

Dennis Helsel
ask@practicalstats.com
Practical Stats LLC
-- Make sense of your data