Practical Stats Newsletter for November 2013

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

In this newsletter:

1. Upcoming Training

In-person courses (see our Training page for registration):
Applied Environmental Statistics
"Statistics, Down to Earth"
Nov. 18-22, 2013 $1395 until Nov 3, $1495 after
Homewood Suites, Littleton, CO 80127
Applied Environmental Statistics covers all of the statistical methods required for routine analysis of data. It includes how to build good regression models, a myriad of hypothesis tests including the newer permutation tests, and trend analysis. It enables you to make sense of your data.

To register (note the cost increase after 11/3) and for more information on all of our courses and webinars, see our Training page at http://www.practicalstats.com/training/

2. Bootstrapping Regression Models

What value is bootstrapping for regression models – when might you use it?
Bootstrapping is the repeated estimation of a statistic so that a distribution is not required when computing intervals or performing a hypothesis test. It treats the available data as the population and re-samples from it, selecting different sets of observations each time. Those differences in data provide an estimate of the variability of the resulting statistic, without assuming that the statistic follows a normal or other distribution. In the context of regression, we could bootstrap the data and get a series of slope estimates. If a 95% bootstrap interval of slopes does not include 0 we can state that the explanatory variable that corresponds to that slope is significant and should be included in the regression model. The bootstrap interval does not require the residuals from the line to follow a normal distribution, as does the standard interval produced by regression packages. That might be a big help in letting us avoid transforming the response (Y) variable to get a reliable p-value. Here is an example.

In our Applied Environmental Statistics course we model total dissolved solids (TDS) in the Cuyahoga River by relating it to river discharge (q). The relationship is at first curved, showing a dilution pattern (Figure 1), so fitting a linear relationship makes no sense.

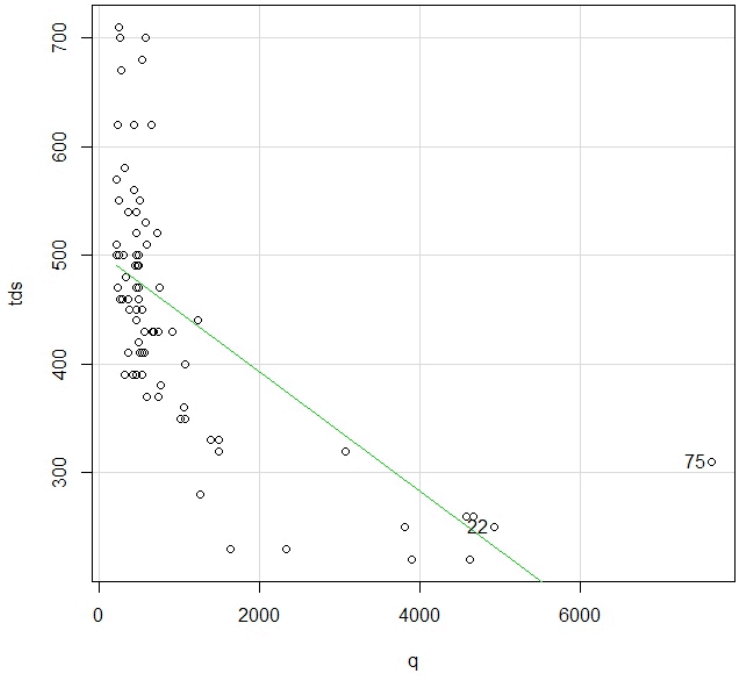

Figure 1. TDS versus river discharge (q)

To reduce the curvature, we take the log of discharge and relate TDS to lnq (Figure 2). The relationship is much straighter, sufficient to model it with regression. The equation is

TDS = 1125.46 -104.9*lnq.

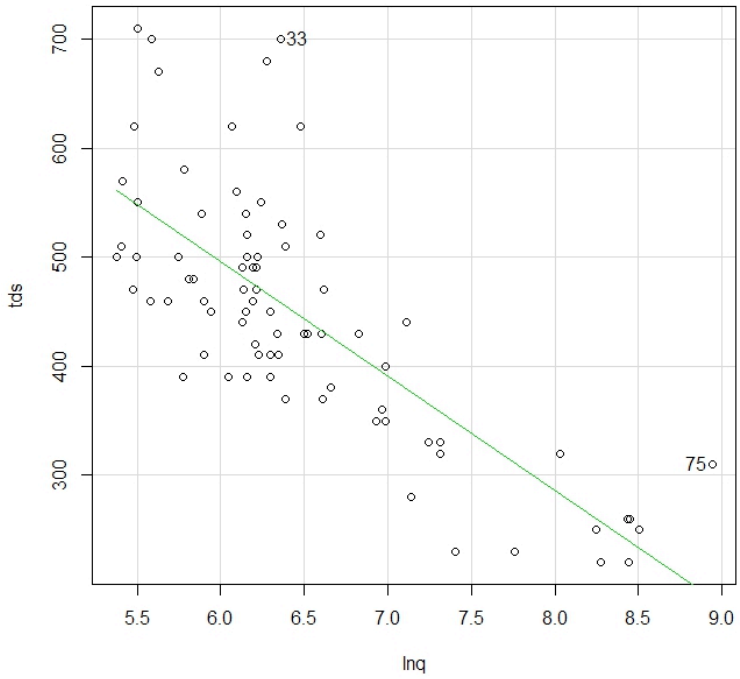

Fig 2. TDS versus natural log of river discharge (lnq)

However the residuals from this relationship are non-normal, calling the p-value of the test on the slope for lnq into question. The solution we often use in class is to also transform TDS, the y variable, and run a log-log regression. One problem with this is that we then will be predicting the geometric mean of TDS with the regression, rather than the mean TDS. This difference may be an issue if we are estimating the total mass over the year, or in some other way want to end up with the mean TDS rather than a geometric mean / median estimate. Transformations of the y variable in regression may have unwanted consequences.

Instead, we can bootstrap the relation between the variables, and examine the resulting slope estimates. If there are 2000 repetitions performed, for example, we want to know if the central 95% of the 2000 slopes includes zero or not. If zero is within the 95% two-sided interval, zero is a plausible 95% estimate of the true slope, and we cannot conclude that the true slope of the relationship differs from zero. In other words, the p-value for lnq will be above 0.05. However if the 95% interval does not include zero, zero is not a plausible 95% estimate of the slope. In that case the p-value for lnq would be smaller than 0.05 and we conclude that lnq is a significant predictor for TDS. Its slope is significantly different from zero at an alpha of 0.05.

Using R we bootstrap the regression relationship 2000 times. Slightly different sets of n=80 TDS-lnq pairs are chosen in each repetition, and 2000 regression equations computed. The 2000 slopes resulting are shown as a histogram in Figure 3.

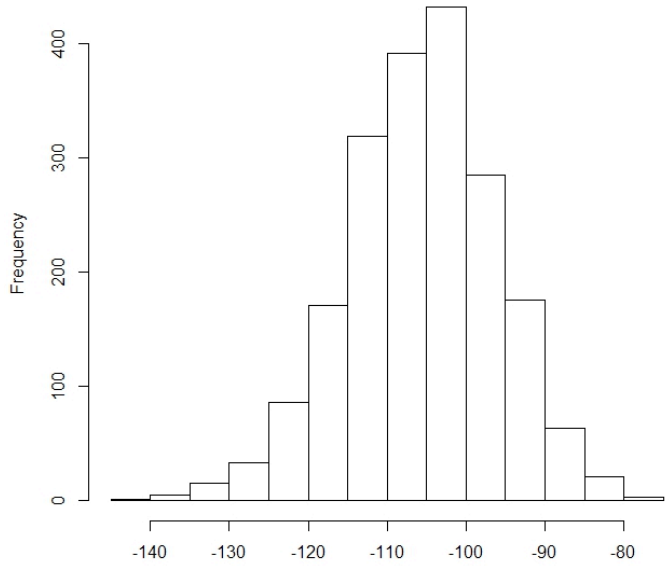

Figure 3. Histogram of 2000 bootstrapped slope estimates

The lower end of the 95% confidence interval for the slope is the 0.025 quantile of the 2000 slope estimates. The upper end is the 0.975 quantile. Their values

```
     2.5%
-125.2892
    97.5%
-87.89792
```

do not include zero between them. Therefore the slope for lnq is significantly different from zero at the 0.05 level, and this evaluation did not require an assumption that regression residuals follow a normal distribution. No transformation of TDS was required. We can predict TDS from lnq with confidence using the original regression equation.

If bootstrapping is new to you, if the idea that transforming the y variable predicts the geometric mean rather than the mean, or if the procedure for building good regression models is still a bit foggy, take our Applied Environmental Statistics course on November 18-22 in the Denver, CO area.

3. Twick or Tweets

If you'd like to receive more frequent but very short items of info on environmental statistics, follow @PracticalStats on Twitter. You'll also find out "in real time" when and where upcoming courses and webinars are scheduled. Just go to PracticalStats.com and click on the "Follow @PracticalStats" button. I'll be tweeting from Singapore over the next several weeks – their water agency is a world leader in the capture and reuse of every drop that falls on the island nation.

'Til next time,

Practical Stats
-- Make sense of your data