Practical Stats Newsletter for July 2013

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

In this newsletter:
1. Upcoming Events
2. Time Series Methods Then and Now
3. Final Discount on Course Registration!

## 1. Upcoming Events

**In-person courses (see our Training page for registration):**
Untangling Multivariate Relationships
August 13-14, 2013
Allmendinger Center, Washington State Univ.
Puyallup, WA 98371

Time Series Methods for Frequently Collected Data
August 15-16, 2013
Allmendinger Center, Washington State Univ.
Puyallup, WA 98371
Note that discounted hotel rates at the nearby Hampton Inn expire on July 22, and course registration fees increase by $100 on July 29th. Don't wait to register for either.

Applied Environmental Statistics
*"Statistics, Down to Earth"*
Nov. 18-22, 2013
Homewood Suites
Littleton, CO 80127

**Fall Webinars (see our Webinars page for registration):**
*R: Free Software for Environmental Statistics*
September 25th (Wed) $50
R is open-source, free software for statistical analysis. It is sometimes seen as too difficult for occasional users such as environmental scientists. This webinar introduces scientists to R software with pull-down menu systems that make it no harder to use than any other statistics software. After taking this webinar, scientists should be able to download, install, and begin to use R software with ease. This could save your organization a lot in software costs -- a webinar for both managers and scientists.

*Permutation tests: Never worry about a normal distribution again!*
October 7th (Mon) $250

Permutation tests are increasingly used to provide p-values for testing means without assuming a normal distribution. Find out how they work, why they are such an advance over parametric methods like t-tests and analysis of variance, and which software performs them.

To register and for more information on all of our courses and webinars, see our Training page at http://www.practicalstats.com/training/

**2. Time Series Methods Then and Now**

If your software and your training on Time Series methods consist only of the Durbin-Watson statistic, there's much more to know these days! Why worry about serial correlation (SC) at all? If your data contain SC, and 'real time' data collected at short time intervals most always do, then hypothesis tests and regression models will be incorrect unless corrected for its effects. You may be including explanatory variables that are not actually significant, or reporting significant p-values are that actually just noise.

The classic method for detecting SC with regression models is the Durbin-Watson (DW) statistic, named after the classic statisticians James Durbin and Geoffrey Watson. The DW statistic lies between 0 and 4, indicating no SC when its value is near 2. If it falls substantially below 2, positive serial correlation (the type usually found in environmental data) is indicated. The statistic only looks for a lag-1 correlation, or AR(1) model. This is a fairly simplistic model requiring normally distributed residuals, and if residuals do not follow a normal distribution or the actual correlation is more complicated, you'll need more than the DW statistic to diagnose and remedy the situation. If DW indicates SC is present, older software might follow with the Cochran-Orcutt or Hildreth-Lu procedures, which also assume a simple AR(1) model. Much better methods are now available, and form the core of our Time Series Methods course.

Box-Jenkins time series models fit more complicated and realistic SC patterns to data. These are also called ARMA models, and are not restricted to a simple lag-1 correlation of residuals. Without going into the full development of what these models are, here is an example of differences in regression p-values when not considering SC, versus using a simple lag one AR(1) model, versus a more complete ARMA model. Total organic carbon is modeled by time (a trend test) and turbidity variables. Of interest is whether turbidity is a significant predictor, and whether that relationship is shifting over time (the trend).

Not considering SC. A linear regression of TOC vs ln(turbidity), time, and seasonal indicators sine and cosine of time produces these regression results:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -73.012023   7.670080  -9.519   <2e-16
time          0.036857   0.003825   9.637   <2e-16
lnturb        0.175817   0.008513  20.653   <2e-16
sint          0.233757   0.007155  32.671   <2e-16
cost          0.232362   0.007094  32.754   <2e-16
```

If SC is ignored it appears that all four variables are highly significant. This might mean that remediation is mandated to counteract the trend, or that some seasons are seen as worse off than others. But is this really the case?

AR(1) model: The simple AR1 model indicated by a significant Durbin-Watson test results in these corrected regression results:

| | Value | Std.Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | -75.96835 | 30.622126 | -2.480832 | 0.0132 |
| dectime | 0.03840 | 0.015266 | 2.515608 | 0.0120 |
| lnturb | 0.11492 | 0.007153 | 16.066413 | 0.0000 |
| sint | 0.24033 | 0.027858 | 8.626866 | 0.0000 |
| cost | 0.21806 | 0.028011 | 7.784756 | 0.0000 |

All of the influences still appear significant, though at much less significant p-values. If the data indicated an AR(1) model was the best one appropriate, we could stop here and write our paper! However, modern diagnostics such as AIC indicate that for these data an ARMA(1,2) model better fits the correlation structure of the regression residuals. The data contain strong memory from one observation to the next because these data were collected only minutes apart in time. After using a freely-available generalized least squares (GLS) software package, regression can be performed while correcting for the memory observed in the data. The results

| | Value | Std.Error | t-value | p-value |
|---|---|---|---|---|
| (Intercept) | -73.48256 | 128.98595 | -0.569694 | 0.5690 |
| dectime | 0.03717 | 0.06430 | 0.578056 | 0.5633 |
| lnturb | 0.11486 | 0.00721 | 15.930192 | 0.0000 |
| sint | 0.19485 | 0.11080 | 1.758628 | 0.0788 |
| cost | 0.16044 | 0.11310 | 1.418604 | 0.1562 |

show that in fact there is no trend nor any seasonal variation in these data! These were artifacts of the serial correlation. The only significant relationship is between TOC and the natural log of turbidity. Initiating remediation methods based on an incorrect trend occurrence would have been a waste of time and money. The results without GLS were wrong because the data were largely redundant over the course of many (short) time steps. They were mostly replicate values masquerading as new observations in the regression.

George Box once famously said that "All models are wrong; some models are useful." If you are collecting data at hourly or more frequent intervals and still using standard regression models, those models may be much less useful than you think. Sign up for our Time Series Methods course this August in Washington State. We guarantee that for frequently-collected data, the course will be quite useful.

### 3. Final Discount on Course Registration!

We publish a newsletter of tips on environmental statistics -- this pdf is the May newsletter, for example. Subscribing saves us a lot of time NOT answering email questions about "when is the next course/webinar being held?" So to say thank you in a Practical way, when you subscribe to our newsletter at

http://practicalstats.com/news/
you'll receive a discount code for 10% off registration costs to either our Untangling Multivariate Relationships or Time Series Methods for Frequently-Collected Data course this coming August.

This offer is also available to friends and co-workers of yours who sign up. Point them to our Newsletter page to sign up, and the discount is theirs also!

'Til next time,

Practical Stats (Dennis Helsel)
-- Make sense of your data