Practical Stats Newsletter for November 2012

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

In this newsletter:
1. Upcoming Events
2. Forcing Regression Through the Origin (no-intercept model)
3. 2013 webinars

We're releasing this month's newsletter a few days early because the time limit is fast approaching for the discount on our upcoming AES course. Register now and avoid paying more! The special discount on room reservations at the Embassy Suites where the course will be held also expires Nov 12th.

1. Upcoming Events

Applied Environmental Statistics *Statistics, down to earth*
(our 1-week survey of stats for environment/natural resources)
December 3-7, 2012. Phoenix AZ.
$1395 until Nov 12th, $1495 after.

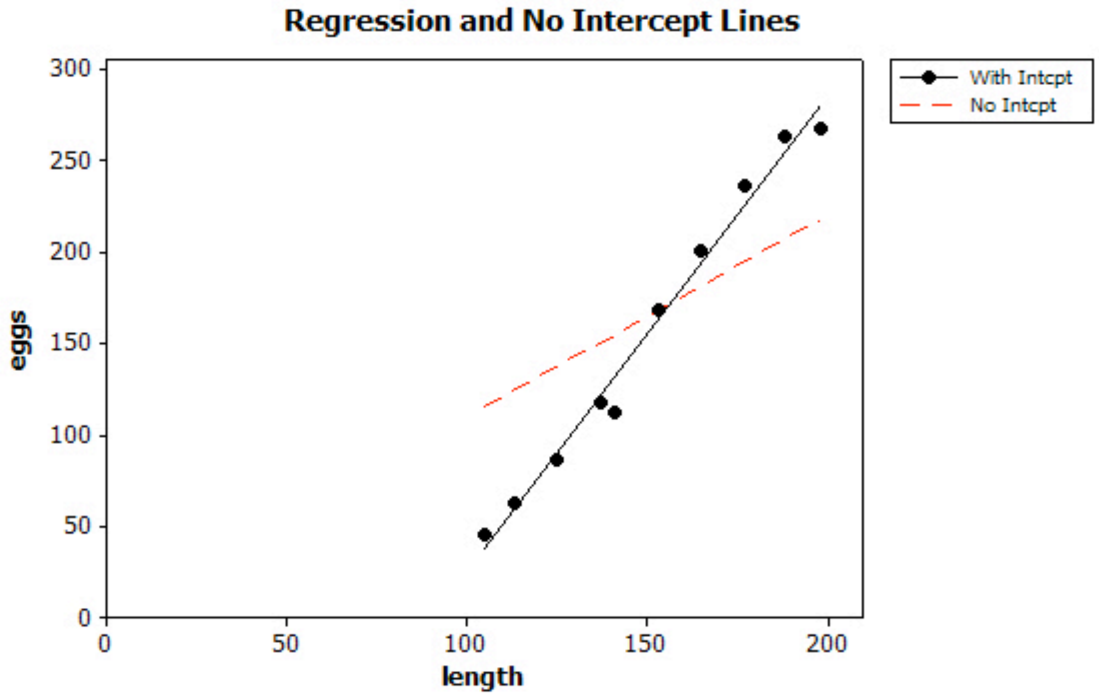
You can register for courses and webinars through our "Upcoming Classes" page at http://www.practicalstats.com/new_classes/classes.html

2. Forcing Regression Through the Origin (the no-intercept model)

One question we are regularly asked in our Applied Environmental Statistics course is "When should I force the regression line to go through zero?" or the equivalent "Should I delete the intercept when it is not significantly different from zero?" The short answers are "Never" and "No", but here's three reasons why.

a. A straight line through the origin often doesn't fit the data well

Many regression relationships of environmental data involve variables that when X is zero, Y must also be zero. Concentrations in surface water (Y) cannot be anything but zero when streamflow (X) goes to zero. Fertility of a fish species (Y) is zero when the body length of a fish (X) is zero. Yet examples like these do not guarantee that the direction of a relationship between Y and X form a linear pattern heading straight for the origin. Often the relationship heads towards a non-zero intercept, especially when data are a long way from the origin, even though theoretical considerations would say otherwise. Forcing a regression to go through the origin can pull the line away from the pattern of the data itself. Below is a plot of fish fertility (Y) versus body length of the fish (X), a classic data set available on the web. A no-intercept model fits poorly.



Sometimes a log transform is taken of both X and Y due to skewness and a large range in the scale of variables, resulting in a good linear fit by a regression in log units. The resulting power function back in original units is $Y = e^{intcpt} * X^{slope}$ where intcpt is the intercept in the log-log regression, and slope is the slope of the log-log regression. This equation will be a curve in the original units that goes through the origin.

b. The r^2 statistic changes meaning, and is artificially high after deleting the intercept

The coefficient of determination, r^2, equals the sum of squares of the predictions divided by the total sum of squares. The "squares" with an intercept are the squared differences between the predicted or observed values and the mean $\bar{Y}$ for the entire data set.

$$r^2 = \frac{\sum\left(\hat{Y}-\bar{Y}\right)^2}{\sum\left(Y-\bar{Y}\right)^2}$$

The r^2 for the regression line with intercept equals 0.98 for the data above.

When the intercept is set to zero this equation uses zero instead of $\bar{Y}$, so distances are from zero to the predicted and observed values. Unless data actually have a mean of zero, which environmental data will not because they generally cannot go negative, the result is "absurdly" close to 1 (for the dashed red line above it equals 0.93). The adjective "absurdly" is a quote from several statistical texts. Though the reported r^2 is close to 1, it has no meaning. A better equation for an r^2 when setting the intercept equal to zero is

$$r^2 = 1 - \frac{SSerror}{SStotal} = 1 - \frac{\sum\left(Y - bX\right)^2}{\sum\left(Y - \bar{Y}\right)^2}$$

but this is not printed in commercial statistical software. For the above no-intercept (dashed red) line this r^2 equals 0.62. The r^2 for the regression with intercept is much higher, showing the intercept model to clearly be a better fit to the data.

An alternative measure to compare a with-intercept and no-intercept regression model is the standard error of regression (s), which is always printed out in commercial software and does not suffer from the above problem. A lower standard error is a better model. For the with-intercept line, s equals 10.7. Without the intercept it is 49.6, showing much greater error.

Do not judge whether to delete the intercept based on the reported r^2. It is a bogus value when the intercept is set to zero.

c. Confidence intervals are usually a bad fit after deleting the intercept
Because the value for Y at X=0 is "known" in the no-intercept model, the confidence interval at that point has a zero width. Confidence intervals increase in width as X increases, forming a cone-shaped pattern around the no-intercept (dashed-line) model. This is quite different than the usual regression confidence intervals when the intercept is estimated, and quite a bit less realistic. If you set the intercept to zero, be sure to understand the implications of what you are doing.

So ignore the significance test for the regression intercept and let the intercept be the intercept. For more information on methods for computing a good regression equation, and as importantly on things not to do, take our in-person "Applied Environmental Statistics" course Dec. 3-7 in Phoenix, AZ. If you know someone who needs this information, please tell them about the course! Much is new in statistics since the time many of us took our last stats course -- AES covers the new as well as the standard. For more information and to register:
http://www.practicalstats.com/new_classes/classes.html

3. 2013 webinars
Our 2012 training webinars on methods for nondetect data and on ProUCL4 were a great success, with lots of interest and attendance. We expect to hold even more webinars in 2013. Look for both live webinars (held on a specific date and time), as well as some as on-demand recordings.

Our first webinar in 2013
Urban Legends in Environmental Statistics
The most commonly-made errors in environmental data analysis
January 14, 2013
11:30 am Mountain time, 1:30 Eastern

This webinar is designed for those who want to avoid the most common statistical mistakes made by environmental scientists. No prerequisites are needed, this is an introductory level webinar. All scientists interpreting data and their supervisors are welcome. Registration will be available soon at http://www.practicalstats.com/webinars

'Til next time,

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