

Re: Regression significance conundrum

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2005-10/msg00312.html>

- *From:* "Paige Miller" <paige.miller@xxxxxxx>
 - *Date:* 20 Oct 2005 05:44:13 -0700
-

Andy Spragg wrote:

- > I'm revisiting some old bivariate data (66 values) with a fresh pair
- > of statistical spectacles (last time I didn't /have/ any statistical
- > spectacles). My eye told me then. and still tells me now, that there
- > is a good linear correlation between the two variables. Last time I
- > just fitted a straight line, got an R-squared of 0.79, and was quite
- > happy.
- >
- > Now I know better. This time, I fitted a straight line and discovered
- > that the constant and gradient are both statistically highly
- > significant (p values both 0 to 3dp). I checked the residuals. They
- > are beautifully normally distributed. Only two (when standardized) are
- > unusual in 95% confidence terms (and with 66 data points, I expect two
- > or three unusual residuals anyway). No pattern when they're plotted
- > against the order of the data, or against the fitted value. So my
- > original correlation was far more legitimate than I realised.
- >
- > Then I managed to rain on my own parade. I decided that actually,
- > there might be a slight curvature in the data, and I might do better
- > if I fitted a quadratic. So I tried it. I expected the constant and
- > the gradient to remain highly significant, and that the stats would
- > tell me whether or not the additional term was also statistically
- > significant.
- >
- > What I actually found is that in the quadratic fit, /none/ of the
- > three coefficients are significant at the 95% level (p values 0.084,
- > 0.161 and 0.404 respectively, for constant, linear and quadratic terms
- > respectively)! However, the R-squared is the same as for the linear
- > regression, and all the observations about the residuals remain valid.
- > The only difference is three unusual residuals rather than two, and
- > the observation at each end of the data set is flagged as having large
- > influence.
- >
- > So what's going on here? If I had started with the quadratic
- > regression, I would apparently have concluded with 95% confidence that
- > my data set was random noise about a mean value of 0. How come the
- > stats don't show that a linear regression is highly significant and
- > that a quadratic fit does not confer significant additional benefit?

Re: Regression significance conundrum

If you do the statistics properly, then you probably won't have any problems. You should most likely start with a plain old scatterplot of the data, that's going to tell you that you do/might have a correlation and a linear fit.

Without an a priori reason to suspect curvature, you always fit the linear model first and examine the residuals.

If, upon fitting a linear model, there is curvature in a plot of X versus the residuals, then you would try to fit a quadratic (or higher order) polynomial. Sounds like you initially thought there was no curvature, but with upon further review (I learned that phrase from NFL referees) you thought there was some slight curvature. By the way, you didn't specify which residual plot you looked at nor which plot you looked at to decide there might be some slight curvature.

So next you fit a quadratic model. You should probably perform the "extra sum of squares test" which will tell you that adding the quadratic term had no statistically significant effect, and that you should stick with the linear model. Your comment that the R-squared was not improved on the quadratic model is another way to know that adding the quadratic term didn't help (although comparing R-squared values is not the equivalent of a formal hypothesis test).

HTH

--

Paige Miller
paige.miller@xxxxxxx

.

- **References:**

- ◆ **[Regression significance conundrum](#)**

- ◇ *From:* Andy Spragg

- Prev by Date: **[Re: Who is this "Reef Fish"?](#)**
- Next by Date: **[Re: Regression significance conundrum](#)**
- Previous by thread: **[Regression significance conundrum](#)**
- Next by thread: **[Re: Regression significance conundrum](#)**
- Index(es):
 - ◆ **[Date](#)**
 - ◆ **[Thread](#)**