Cross Validated

How exactly are standardized residuals calculated

Asked 8 years, 10 months ago Modified 6 years, 4 months ago Viewed 40k times

8

I'm working on a model for something and at the moment I prefer working solely in Excel. I've been double checking the results of the linear model in JMP, Minitab, and Statistica, and (more or less) been getting the same answers.

One thing that's coming out odd though is my standardized residuals, I'm getting much different answers than Excel's regression routine, and I know it has to do with how I am calculating them:

The standard deviation of our population varies relative to the output, so we work in terms of the relative standard deviation. We have an assumed %RSD of 5% (based on a lot of previous work, we also have reason to assume normality). From this I standardize the residuals by saying $\frac{(x-u)}{u \cdot RSD}$ where x = the observed value and u = the predicted value, so x-u = the residual.

Note that $u \cdot RSD = s$. Simple z-score. Problem is that the values Excel is giving me for the standardized residuals are much different than mine. This isn't exactly surprising since I am using a varying standard deviation. But their values don't seem to be tied to the reality of the data. One observation could be off by as much as 50% (around 6 standard deviations away) and the standardized residuals I'm given are only like 2 or 3.

Anyways, I'm having a really hard time finding out exactly *how* the residuals are standardized in a linear regression. Any help would be appreciated

regression excel residuals

Share Cite Improve this question Follow

edited Dec 11, 2017 at 18:01
kingledion
802 2 10 21

asked Aug 10, 2015 at 18:41
emorris1000
81 1 1 2

Which version of Excel and what implementation of regression (`LINEST` ? The analysis toolpack? Something else)? Are these *standardized* or *studentized* residuals? Could you explain what it means to "work in terms of the relative standard deviation"? How exactly is that modifying the usual least squares regression model? – whuber ♦ Aug 10, 2015 at 18:50

using the analysis toolpack. Standardized residuals (z not t). My understanding was that many systems assumed a static standard deviation that was independent of the scale. So say you have a stdev = 500, it would be 500 if your u was 2000 or 20,000. Is that how it works in linear regression? Honestly I don't know. I don't have the best background in stats: a decent knowledge of it but just enough to get me in trouble. ed: is this related to homoscedasticity and heteroscedasticity? – emorris1000 Aug 10, 2015 at 18:55

It's ok not to have a stats background, but you still ought to read the replies at stats.stackexchange.com/questions/3392 concerning using Excel. Your question uncovers a particularly egregious error in the Analysis ToolPak (which has wholly inadequate documentation). – whuber ♦ Aug 10, 2015 at 19:16

1 I didn't know that Analysis ToolPak actually had documentation. – Aksakal Aug 10, 2015 at 19:29

@Aksakal To paraphrase an early '70's meme, having no documentation means never having to admit you're wrong ;-). – whuber ♦ Aug 10, 2015 at 19:52

2 Answers

Sorted by: Highest score (default)

9

The statistical tools in Excel have always been black boxes. There's nothing for it but to do some forensic reverse-engineering. By performing a simple regression in Excel 2013, involving the data $x = (1, 2, 3, 4, 5, 6, 7, 8, 9)$ and $y = (2, 1, 4, 3, 6, 5, 9, 8, 7)$, and requesting "standardized residuals" in the dialog, I obtained output that states

- The "Standard Error" is $1.3723\ldots$.
- There are 9 observations.
- The residuals r_i are $(0.5333\ldots, -1.35, \ldots, 0.35, -1.533\ldots)$.
- The corresponding "Standard Residuals" are $(0.4154\ldots, -1.0516\ldots, \ldots, 0.2726\ldots, -1.1944\ldots)$.

Since "standardized" values are typically numbers divided by some estimate of their standard error, I compared these "Standard Residuals" to the residuals and to the "Standard Error." Knowing that various formulas for variances are sums of squares of residuals r_i divided variously by n (the number of data) or $n - p$ (the number of data reduced by the number of variables, in this case two: one for the intercept and a second for the slope), I squared everything in sight. **It became immediately obvious that Excel is computing the "Standard Residual" as**

$$\frac{r_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n} r_i^2}}.$$

This formula reproduced Excel's output *exactly*--not even a trace of floating point roundoff error.

The denominator is what would be computed by Excel's STDEV function. For residuals *from a mean*, it is an unbiased estimate of their variance. For residuals *in a regression*, however, it has no standard meaning or value. It's garbage! But now you know how to compute it... .

Share Cite Improve this answer Follow

answered Aug 10, 2015 at 19:28

whuber ♦
328k 61 763 1.3k

I double checked this and you're absolutely right. But is it completely without value? Lets assume a system is homoscedastic. So I position myself so that I am looking down the line of fit, if that makes sense, so that I've reduced a dimension from the system and now I have a mean, the line of fit, and the 1-dimensional scatter. In that sense how is it different from a normal standard deviation? This assumes homoscedacity, if the standard deviation is dependent on the scale of the result then you couldn't do this. I'm not a statistician by any means so I could be way off. – emorris1000 Aug 11, 2015 at 13:52

1 It is not "completely without value," but it is misleading. Residuals do not behave like "normal" data *because they are mutually (negatively) correlated.* When $n - p$ is large, this correlation is of little import, so the Excel calculation yields almost the same results as a correct calculation would. But worse, *the residuals are not homoscedastic* even when the error model is! The standard error of a residual depends on the values of its independent variables. The further those values are from their means, the greater is the SE. Excel has no excuse for not doing valid calculations. – whuber ♦ Aug 11, 2015 at 14:01

2

In R :

```
modeGlob <- lm(rnorm(100)~ abs(rnorm(100))) #Your model.

hii <- hatvalues(modeGlob) # hat matrix.

rst <- modeGlob$residuals / (summary(modeGlob)$sigma * sqrt(1-hii)) # manually calculate standardized
residuals.

identical(rstandard(modeGlob) , rst) # check, this must be TRUE.

plot(rstandard(modeGlob) , rst) # check it graphically.
```

Share Cite Improve this answer Follow

edited Feb 22, 2018 at 13:33

answered Feb 22, 2018 at 13:22

Omran Allatif
21 3

4 Welcome to the site. Be aware that this is not an R Q&A site, but a statistics one. Not everyone will use, or even be able to read R. In light of that, can you add some text to explain this? – gung - Reinstate Monica Feb 22, 2018 at 13:49