

Re: stepwise regression by GENSTAT

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2008-08/msg00289.html>

- *From:* Old Mac User <chendrixstats@xxxxxxxxxx>
 - *Date:* Sun, 17 Aug 2008 13:51:00 -0700 (PDT)
-

On Aug 17, 9:33 am, hubert.ruyp...@xxxxxxxxxx wrote:

On 17 aug, 01:47, RichUlrich <rich.ulr...@xxxxxxxxxxxx> wrote:

On Sat, 16 Aug 2008 07:40:57 -0700 (PDT), Old Mac User

<chendrixst...@xxxxxxxxxx> wrote:

On Aug 16, 3:20 am, hubert.ruyp...@xxxxxxxxxx wrote:

I am using GENSTAT for performing a multiple linear regression. To select the most suitable explanatory variables this program provides a "stepwise regression method". However I do not understand how this method leads to a result.

According to my handbook it drops the variable which gives the lowest mean square (Residual) in an ANOVA set-up.

What it drops is the variable that makes the least improvement in the mean square (Residual). That is the variable that has the least effect, at that point. If your handbook says otherwise, you need a new handbook.

Is "the least improvement in the mean square (Residual)" not the same

Re: stepwise regression by GENSTAT

as "the variable that results in the highest value of the mean square (Residual)?"

If your handbook makes a general recommendation, that stepwise methods are desirable, you need a new handbook — there was a spell in the 1970s, when computers were first available to them, that some practitioners generated a fad for Stepwise; but wiser statisticians (and their own experiences) shortly convinced them that there was no magic in it. You can Google groups <stepwise group:sci.stat.* author:ulrich > for comments and threads on the subject.

My handbook considers only stepwise regression as a method to select the explanatory variables (Mc Conway, Jones and Taylor, Statistical Modelling using Genstat").

And here I am losing the logic. I was convinced that a lower mean square (Residual) resulted in a higher variance ratio (which is the m.s.(Regression) divided by the m.s.(Residual) and thus in a lower F-probability.

Can some-one help me out ?

Hubert

I respectfully submit that if you do not understand multiple regression and it's related offshoots, you should get some assistance from a practicing statistician. Building models from data has many facets, it can be tricky, and the risk of getting silly results is high.
OMU

Right.

Who am I to contest this. However some knowledge of the statistical possibility's (and limitations) is never wrong (I think)

--

Rich Ulrich

OK, let me try again. Rich Ulrich is right. The idea is to find the minimum subset of variables that explains the "significant" variation in the data... leaving behind only "random variation" in the residuals (residuals = (Observed - Predicted by the Model)). To say it another way, minimize the variation in the residuals (residual sum of squares) with the fewest possible variables (parsimony... stingy to the point of starvation). If you have to use many regression coefficients... with each "just barely significant"... to explain variation in the data, then you have a very poor model. It may explain variation in the data, but it will be a terrible predictor.

But if you have several candidate variables (or even many candidate variables) wherewith to explain that variation, problems mount. In principle, the first task is to identify those candidate variables that offer promise. Then on to which subset of these to use as predictors. Then... in many instances... concerns about possible interactions, curvature, etc. This is where the fun begins. For instance, if values of your response (dependent variable), approach a "natural boundary" (an upper or lower limit... or both) then you need to do a special transformation of the dependent variable.

These are some... but not all... of the reasons I suggest "get help". Especially if you intend to use this model in a meaningful way. If, for

Re: stepwise regression by GENSTAT

instance, you need a "special transformation" of the dependent variable and you don't do that, your finished model may predict impossible outcomes for certain values of the predictors. That can cause a loss of credibility when the model fails to pass the giggle test.

Rich, just fyi, I use stepwise regression and use it a lot. A quick pass with stepwise tells me the approximate number of predictor variables the data will support. But, recognizing that there may be several or even many combinations of variables each of which will explain about the same amount of variation with about the same number of predictors... I then use "all combinations regression" concentrating on using about that pre-screened number of predictors. This reduces the number of potential models to a realistic level.

I will not, under any circumstance, attempt an analysis of multivariable data without access to the complete correlation matrix... (complete = for all the variables). I insist on getting the determinant of the variables (in regression" (determinant of the correlation coefficients among the predictors) or all bets are off. In short, I use my own software because no commercial software that I've seen will do all that I want done... and do that cleanly. I don't want to open a window over here to get the correlation matrix and another window to examine the residuals, etc. As a passing comment, those who design, produce, and sell commercial software are reluctant to explain "you can create silly models using our software". Silly = won't pass the giggle test. This is partly because they don't want to reveal that the King has no clothes, and partly for legal reasons. Software that "designs experiments" is among the worst offenders.

There are instances when the correlation matrix looks "pretty good" but there is a constraint among some or all of the predictors. This is notably true of "mixture experiments". Those need extra help.

There comes a point at which any selection of the combination of

Re: stepwise regression by GENSTAT

Re: stepwise regression by GENSTAT

variables

to be used needs to be consistent with the underlying technology that produced the data. So I often create and examine all of the likely models

... again to see which ones will pass the giggle test. In many instances

... as I'm sure you know... we end up with several models that are "just

as good, one or the other" and that is a strong argument for getting advice

from the parties that produced the data. The fact is, that advice should

be sought before even beginning an analysis of the data.

So many people seem to believe "now I've got some software... stand back and

watch me while I do this." Well, I often spend between 4 and 12 hours pondering,

analyzing, comparing models, discussing the data with the "owners", etc. and

even more time documenting the possible models for a presentation.

All of the is a great argument for using designed experiments. Those are usually

analyzed and wrapped up in about an hour... sometimes in minutes.

To the original poster: If you want additional advice, I suggest that you

tell us more about the data. How many candidate predictors are you trying to

screen? A general description of the data and the origin of the data would

be helpful.

OMU

.