

Re: Regression of correlated variables

Source: <http://sci.tech-archive.net/Archive/sci.stat.math/2006-12/msg00007.html>

- *From:* "Greg Heath" <heath@xxxxxxxxxxxxxxxxxxxx>
 - *Date:* 30 Nov 2006 09:14:56 -0800
-

Greg Heath wrote:

Tim vor der Brück wrote:

I would like to do a regression of highly correlated variables. Could somebody give me a hint about the best method. Principle

Terminology: Principal

Component Analysis?

I will assume you mean Multiple Linear Regression ($y = X*b$).

There is no general "best method". It depends what you want.

1. MATLAB provides the so-called SLASH solution $b = X \backslash y$ that has the following properties:

- a. If X is well conditioned, b , obtained using the QR decomposition of X , is a more accurate version of the traditional OLS solution obtained via $\text{inv}(X' * X) * (X' * y)$.
- b. If X is ill-conditioned or singular, b is a BASIC solution using as few nonzero regression coefficients as possible. I haven't checked the source code so I don't know exactly how the variables are chosen. I doubt if they are the same as would be chosen in a stepwise regression program.

2. MATLAB provides the truncated pseudo-inverse solution $b = \text{pinv}(X) * y$ that minimizes $(y - X*b)' * (y - X*b)$ subject to a truncation threshold for small singular values. This can be easily modified to be subject to a truncation threshold for the trace-normalized sum of singular values.

Re: Regression of correlated variables

3. Partial–Least–Squares provides a sequence of approximate solutions based on sequences of orthogonal linear combinations of the predictors constrained to maximize the covariance measure $\|X^*y\|^2$. Free MATLAB PLS Toolboxes are available from non–MATLAB sources. (Search on PLS Paige Miller).

Analysis of the above results can, in some cases, lead to a practical ad hoc method of eliminating redundant and/or irrelevant variables.

More formal methods of eliminating variables are

4. Comparing all possible predictor combinations
5. Using stepwise regression constrained by
 - a. a required subset of starting variables (including the null set)
 - b. a required subset of chosen variables (including the null set).

MATLAB's STEPWISEFIT can be used to implement method 5.

In summary, if variable subset selection is not desired, investigate methods 1 and 2. Otherwise investigate 3–5.

Please don't take this as a plug for buying MATLAB. I'm a retired engineer who found MATLAB useful. All of the the above methods can be implemented using other statistical and/or mathematical software.

Given the software package of your choice, I'm sure that others in the group can give you more specific advice.

Hope this helps.

Greg

Hope this helps.

Greg

.