PLS Regression and Multiple Imputation

Philippe Bastien* — Michel Tenenhaus**

** L'Oréal Research, 1 avenue Eugène Schueller – BP 22 – 93601 Aulnay sous Bois Cedex*
pbastien@recherche.loreal.com
*** HEC School of Management, 1 rue de la libération – 78351 Jouy-en-Josas*
tenenhaus@hec.fr

There are various ways to deal with missing data. Among the conventional methods for handling missing data "listewise deletion" could be considered as a safer or less biaised approach. However if the amount of data that must be discarded under listwise deletion is dramatic other alternatives must be considered. A first strategy is the simple imputation which will substitute a value for each missing data. For example, each missing value can be replaced by the average of the variable, the nearest neighbour (Nguyen, 2002), or by the predicted value of a regression model. These ad hoc methods are not without disadvantages, by biaising in the first two cases variance-covariances towards zero and increasing the observed correlations in the last case. Other strategies consist in using the EM algorithm (Dempster, Laird, and Rubin. 1977) for the estimation of maximum likelihood with incomplete data or the direct Maximum likelihood estimation used in program for structural equation modelling.

PLS Regression in its standard form with the use of the NIPALS algorithm can deal with missing values. Treatment of missing values with PLS-NIPALS can be implicitly associated as a simple imputation method. PLS loadings and components are iteratively calculated as slopes of least squares lines passing through the origin on the available data. Missing data are in fact estimated with these simple regressions.

The PLS Kernel algorithm proposed by Ränner, Geladi, Lindgren, and Wold is based on a simplified version of the EM algorithm for the calculation of covariances matrices when missing data are present. An improved version of this algorithm using the complete EM algorithm is now very easy to implement with the new SAS procedures MI and PLS.

Simple imputation which treats the missing values in a deterministic way does not reflect uncertainty associated with their prediction. If the data are missing at random, and if the parameters of the model do not depend on the process generating the missing values, Rubin in the 1970's proposed to replace each missing value with a set of plausible values that represent the uncertainty about the right value to impute. This process leads to valid statistical inferences that properly reflect the uncertainty due to missing values.

Multiple imputation inference involves three distinct phases :

- The missing data are filled in m times to generate m complete data sets.
- The m complete data sets are analysed by using standard procedures.
- The results from the m complete data sets are combined for the inference.

We compare NIPALS, EM , and MI in the treatment of missing data in PLS regression. Based on simulation studies, the three methods give comparable results when no more than 30% of the data are missing. With more missing data EM or MI seem to give better results compare to NIPALS. Furthermore using EM or MI it is possible to include information from variables outside the model to estimate covariance matrices.

References

Allison, P.D. (2002), *Missing Data*, Series : Quantitative Applications in the Social Sciences, Sage Publication.

Dempster,A.P., Laird, N.M. and Rubin, D.B. (1977*) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion*). Journal of the Royal Statistical Society Series B, 39,1-38.

Efron B., Tibshirani R.J. (1993) – *An introduction to the Bootstrap*. Chapman and Hall, New York.

Little, R.J.A. and Rubin, D.B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.

Nguyen, D.V., Wang,N., and Carroll, R.J.(2002) *Evaluation of missing value estimation for array data* (Manuscript)

Ränner S., Geladi P., Lindgren F. And Wold S. (1995). *A PLS Kernel Algorithm for data sets with many variables and few objects. Part II : cross-validation, missing data and examples*. Journal of Chemometrics, Vol 9, 459-470 (1995)

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.

Tenenhaus, M. (1998), *La régression PLS* , éditions Technip, Paris

Wold S., Martens & Wold H. (1983) : *The multivariate calibration problem in chemistry solved by the PLS method.* In *Proc. Conf. Matrix Pencils*, Ruhe A. & Kåstrøm B. (Eds), March 1982, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, p. 286-293.

Yang C. Yuan (2002) Multiple Imputation for Missing Data:Concepts and New Development, SAS Institute Inc., Rockeville, MD