

Standard Error Calculation for Partial Least Squares (PLS) Regression using SAS

April Grant¹, D. Keith Williams¹, Zoran Bursac¹, Geoffrey Curran²

College of Public Health, Department of Biostatistics, University of Arkansas for Medical Sciences¹
College of Public Health, Department of Epidemiology, University of Arkansas for Medical Sciences²

Abstract

Partial least squares (PLS) regression, also referred to as soft modeling or projection to latent structures, is a useful method that constructs predictive model(s) when data perhaps exceed criteria set for multiple linear regression. The basis behind PLS is that a model of Y (response) variables from X (predictor) variables is constructed. It is considered more vigorous than other soft science applications. A problem encountered when utilizing PLS regression in SAS is that standard error (SE) calculations are not easily obtained for parameter estimates. We provide a method that ‘borrows’ empirical standard error estimation algorithms in SAS PROC NLIN to calculate approximate standard error estimates for PLS model predictions. The ability to estimate SE in quick and accurate fashion will allow calculation of confidence intervals and significance testing for the responses.

Keywords: Partial least squares, standard error calculation

1. Introduction

Partial least square regression is a useful ‘soft modeling’ approach that allows prediction model construction when multiple regression criteria is not met or interrelationships among variables are not well understood.

The following excerpts by Tobias (1995), taken from An Introduction to Partial Least Squares Regression, provide a thorough overview of PLS modeling:

“PLS has been labelled as a soft modeling approach that has been used in industrial and economic settings. Generally, settings for PLS application have a set of ingredients (predictor variables) and a set of performance measurements (responses). Correlation is often present both within the predictor variables and response variables as well as among the predictor and response variables.”

“Research in science and engineering often involves using controllable and/or easy-to-measure variables (factors) to explain, regulate, or predict the behaviour of other variables (responses)...In such so-called soft

science applications, the researcher is faced with many variables and ill-understood relationships, and the object is merely to construct a good predictive model. For example, spectrographs are often used to estimate the amount of different compounds in a chemical sample. In this case, the factors are the measurements that comprise the spectrum; they can number in the hundreds but are likely to be highly collinear. The responses are component amounts that the researcher wants to predict in future samples.”

“PLS is a method for constructing predictive models when the factors are many and highly collinear. Note that the emphasis is on predicting the responses and not necessarily on trying to understand the underlying relationship between the variables. For example, PLS is not usually appropriate for screening out factors that have a negligible effect on the response. However, when prediction is the goal and there is no practical need to limit the number of measured factors, PLS can be a useful tool.”

Readily available standard error calculations would provide more potential for use in other applications. Estimation of prediction standard errors provides opportunity to calculate confidence intervals for point estimates.

2. Macro and Simulation

A macro was written to extract parameter estimates from SAS PROC PLS and input into SAS PROC NLIN in order to calculate standard errors. The macro utilizes features already available in SAS and is presented in three basic steps. Parameter estimates are generated from the PROC PLS function and saved into a separate data set. Symput statements extract the estimates and place them within PROC NLIN syntax to estimate standard error. Only one iteration is performed and standard error estimates are obtained. Iterative steps are not needed because we provide the parameter estimates obtained from PLS. With minimal revision of the macro to suit the data set in question, standard error estimates can be readily obtained.

3. Results

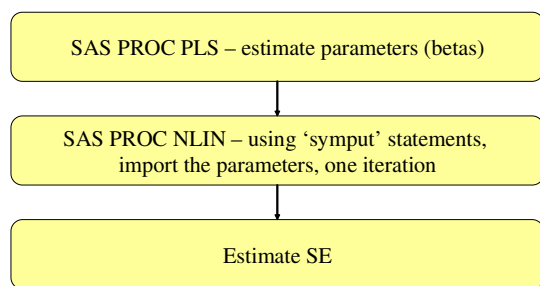


Figure 1. Macro flow chart.

Simulations were performed to investigate coverage performance. One thousand simulations were performed for each combination of latent variables, correlation estimate, and sample size. Sample sizes consisted of six levels: 20, 30, 40, 50, 100, and 120. Correlations were set at three levels: low ($r=0.2$), medium ($r=0.5$), and high ($r=0.7$). Latent variables retained from the model were set at 1, 3, and 5. This resulted in 54 possible scenarios and total of 54,000 simulations. The expected value was determined via bootstrap and coverage was determined by the proportion of true mean placement at a 95% level.

After successful simulations, the macro was applied on a public health data set, taken from the Study on the Efficacy of Nosocomial Infection Control (SENIC) project available in public domain. The premise behind the study was to assess whether infection surveillance and control programs reduced nosocomial infection rates².

Independent variables included age, culture test-patient ratio, routine chest x-ray ratio, number of beds, medical school affiliation, geographic region, daily census, nurse count, and services.

Dependent variables include length of stay and infection risk probability. The same macro with minor modifications made to accommodate the data set from the SENIC study was utilized. Modifications include inputting data, assigning the number of parameter estimates based on the number of independent variables, and modifying the number of PROC NLIN series based on the number of dependent variables. Dummy variables were coded for nominal variable region. A total of 11 independent variables and two dependent variables were inputted into the PLS procedure.

Simulation results show that coverage area is within acceptable ranges in larger sample sizes and slightly more conservative as sample size decreases. As the number of extracted latent variables increases, estimates approach GLM obtained values. The number of retained latent variables should not be large, but optimally based on results of simulation and applied problem.

Table 1. 95% CI coverage for LV=3.

Sample Size	(r=.2) Y1	(r=.2) Y2	(r=.5) Y1	(r=.5) Y2	(r=.7) Y1	(r=.7) Y2
20	0.994	0.999	1.00	0.996	0.999	1.00
30	0.983	0.982	0.984	0.989	0.986	0.99
40	0.981	0.978	0.978	0.978	0.976	0.982
50	0.966	0.974	0.973	0.974	0.969	0.973
100	0.965	0.953	0.974	0.97	0.968	0.966
120	0.969	0.956	0.969	0.966	0.96	0.964

8

Table 2. 95% CI coverage for LV=5.

Sample Size	(r=.2) Y1	(r=.2) Y2	(r=.5) Y1	(r=.5) Y2	(r=.7) Y1	(r=.7) Y2
20	0.979	0.974	0.984	0.981	0.972	0.973
30	0.964	0.95	0.958	0.967	0.96	0.956
40	0.967	0.948	0.941	0.957	0.956	0.955
50	0.952	0.959	0.959	0.966	0.959	0.969
100	0.951	0.944	0.959	0.962	0.961	0.947
120	0.934	0.948	0.962	0.947	0.946	0.952

9

The last observation was modeled for predictions of length of stay and infection probability risk. GLM estimates and SE were obtained for comparison with PLS estimates. For the SENIC data set, comparisons of GLS and PLS parameter estimates for Y1 (length of stay) are presented in Table 3. Estimates show that as the number of PLS latent variables increase, estimates are more similar to GLM parameter estimates and SE. By looking at the variation, the optimal number of factors for extraction is five. GLM and PLS estimates for this number of variables are similar and within an acceptable range.

Group International Conference, Cary, NC: SAS Institute, pp. 1250-1257.

Table 3. Comparison of GLM versus PLS parameter estimates and SE for the SENIC data set.

GLM	PLS	PLS LV extractions
Predicted Y1 (SE)	Predicted Y1 (SE)	LV
9.56 (0.36)	8.57 (0.47)	1
9.56 (0.36)	9.15 (0.45)	2
9.56 (0.36)	9.31 (0.42)	3
9.56 (0.36)	9.93 (0.39)	4
9.56 (0.36)	9.69 (0.39)	5
9.56 (0.36)	9.51 (0.37)	9
9.56 (0.36)	9.52 (0.36)	10
9.56 (0.36)	9.56 (0.36)	11

2. Neter J, Kutner MH, Nachtsheim C, Wasserman W. *Applied Linear Statistical Models*. 4th ed. Boston: McGraw-Hill; 1996.

4. Conclusions

PLS regression is a useful tool because it is able to go beyond multiple linear regression and analyze multi-collinear data that has many correlated X-variables and at the same time create response variables. PLS regression in essence creates a model of Y (response) variables from a group of X (predictor) variables.

PLS is seen as a preliminary estimation tool. Partial least squares, when utilized properly will yield acceptable results that are similar to more computer intensive models. Estimates provide a relatively quick insight into the data. It should be stressed that PLS is a predictive modeling tool that does not aim to explain the relationship between independent and dependent variables. If a researcher is interested in a quick predictive model, PLS may be a viable option. With the availability of SE calculations from PLS parameter estimates in SAS, researchers can utilize PLS more frequently and explore potential applications in their studies.

Our proposed approach using PROC NLIN in SAS provides a quick and accurate estimate of standard error for PLS. Readily available standard error calculations make PLS an invaluable tool for purposes of CI estimation and significance testing.

A copy of the macro can be found at <http://www.uams.edu/biostat/grant/PLS%20Macro.doc>

References

1. Tobias, R. (1995), "An Introduction to Partial Least Squares Regression," Nineteenth Annual SAS Users