Partial Least Squares Modeling in Research on Educational Achievement

Norbert Sellin

Otto Versand, Hamburg

Abstract

This paper contains a discussion of partial least squares (PLS) path modeling with latent constructs as a general method for research on educational achievement. To the extent that such research requires the analysis of comparatively large and complex models under mild supplementary assumptions, PLS is an extremely flexible and powerful tool for statistical model building. The formal specification, estimation, and evaluation of PLS models is described with special emphasis on the features that distinguish PLS from other methods for path analysis. This specifically concerns distribution-free least squares estimation and distribution-free model evaluation using jackknife techniques.

1 Introduction

Educational researchers frequently work in a situation with massive amounts of data, but relative scarcity of theoretical knowledge. In such a problem area, partial least squares (PLS) path analysis with latent constructs is a useful and flexible tool for statistical model building. The use of PLS may be considered especially when the research situation at hand demands the investigation of complex models in an exploratory rather than a confirmatory fashion.

Such situations would appear to be not uncommon especially, in educational research focusing on the complex ways in which different school, class, teacher, and student characteristics influence educational achievement. The theoretical framework of such research typically involves latent constructs such as home background and attitude toward school. A set of hypotheses concerning possible ways in which these constructs might be related to each other as well as hypotheses about possible ways in which various constructs would be expected to influence educational outcomes are included. The researcher then seeks to explore the validity of the set of initial assertions on the basis of the data being collected. This concerns the structure of the constructs as well as the relationships between constructs. That is, observed or manifest variables would

be grouped together to form specific constructs, and the relationships between constructs would be specified in terms of paths representing direct effects. The ensuing model would then be examined in terms of the extent to which the data support the initial specification and, depending on the empirical results, model modifications such as a re-definition of specific constructs and the deletion or inclusion of particular paths would be introduced to obtain a better fit. The analysis, then, can be characterized as a series of model analyses guided by theoretical considerations and empirical evidence.

The flexibility and scope of PLS facilitates the analysis and investigation of large and complex path models, specifically in the more exploratory fashion sketched above. The following presentation provides a description of partial least squares in terms of its statistical background and its use in educational research.

2 Model Specification

PLS was developed by Wold as a general method for the estimation of path models involving latent constructs indirectly measured by multiple indicators (Wold 1975, 1979, 1982). For convenience's sake, the following presentation will be restricted to what Wold (1982) calls the basic PLS design. PLS models are formally defined by two sets of linear equations called the *inner model* and the *outer model*. The inner model specifies the relationships between unobserved or latent variables (LVs), and the outer model specifies the relationships between LVs and their associated observed or manifest variables (MVs). Without loss of generality, it can be assumed that LVs and MVs are scaled to zero means so that location parameters can be discarded in the following equations. The inner model connecting latent variables can be written as:

$$\eta = \eta B + \xi \Gamma + \zeta \qquad (1)$$

where η symbolizes a (g x n) matrix of endogenous LVs and ξ a (h x n) matrix of exogenous LVs, with n denoting the number of cases. B and Γ are (g x g) and (g x h) coefficient matrices respectively, and ζ denotes the (g x n) matrix of inner model residuals. The basic PLS design assumes recursive inner structures. The LVs can then be arranged such that the matrix B is lower triangular with zero diagonal elements. The inner model (1) is subject to predictor specification:

$$E(\eta * \eta; \xi) = \eta B + \xi \Gamma \quad (2)$$

This implies $E(\xi\zeta') = \mathbf{0}$ and $E(\eta\zeta') = \zeta\zeta'$, with $\zeta\zeta'$ being a (g x g) diagonal matrix. That is, the inner model is assumed to constitute a causal chain system with uncorrelated residuals. It is also assumed that the residual belonging to a given endogenous LV is uncorrelated with the corresponding predictor LVs. The outer model equation for the endogenous LVs is given by:

$$\mathbf{y} = \prod_y \eta + \varepsilon_y \quad (3)$$

where **y** symbolizes a (m x n) matrix of manifest variables related to the LVs by the coefficients given (m x g) matrix $\prod_y$, and where ε_y denotes the associated matrix of outer model residuals. A similar equation defines the outer model relationships for the exogenous LVs involved in the model.

The MVs are generally assumed to be grouped into disjoint blocks, with each block representing one LV. That is, each MV is assumed to belong to just one LV and, hence, each row of $\prod_y$ contains just one non-zero element, while all other row entries are assumed to be zero. Following factor analytic terminology, these non-zero elements are called *loadings*. Since the loadings and the LVs are unknown, some standardization is necessary to avoid scale ambiguity. As a general rule, all LVs are assumed to be scaled to unit variance; i.e. $VAR(\eta_g) = 1$. Similar to the inner model, predictor specification is adopted for the outer model. As applied to equation (3), this yields:

$$E(\mathbf{y} * \eta) = \prod_y \eta \quad (4)$$

This equation involves the assumption that the outer model residuals are uncorrelated with all LVs and with the inner model residuals.

In addition to predictor specification applied to the inner and outer model, a fundamental principle of PLS modeling is the assumption that all information between observables is conveyed by latent variables. This has two implications. First, PLS models do not involve any *direct* relationships between MVs. Second, the outer residuals of one block are assumed to be uncorrelated with the outer residuals of other blocks. The latter assumption means formally that the covariances of outer model residuals can be represented by a block diagonal matrix with non-zero entries corresponding to the grouping of MVs into g + h disjoint blocks. It should be noted, however, that no restrictions are imposed on the covariances of the outer residuals within a given block of MVs.

Given the above equations, it is possible to use the inner model equation (1) to substitute the endogenous LVs involved in the outer model equation (3). The result is:

$$\mathbf{y} = \prod_y(\eta B + \xi\Gamma) + \nu \qquad (5)$$

Wold (1982) calls this substitutive elimination of latent variables, or abbreviated SELV. As can be seen from equation (5), the SELV relation connects endogenous MVs with LVs that are *indirectly* related (via the inner model), with the respective sets of manifest variables. The residuals in (5) are equal to $\nu = \prod_y \zeta + \varepsilon_y$ and are, by virtue of equations (2) and (4), uncorrelated with the corresponding predictor LVs.

3 Model Estimation

The above equations and the accompanying set of assumptions constitute the theoretical or structural form of PLS models. The LVs, the inner model coefficients, and the loadings are of course unknown and must be estimated. The PLS estimation procedure proceeds in two basic steps. The first step involves the iterative estimation of LVs as linear composites of their associated MVs. The second step involves the non-iterative estimation of inner model and outer model coefficients. For example, the estimated endogenous LVs are given by:

$$\text{est.}(\eta) = \mathbf{Y} = \mathbf{W_y y} \qquad (6)$$

where $\mathbf{W_y}$ denotes a (g x k) weight matrix. We adopt the convention of denoting the matrices of estimated LVs with capital Roman letters; the matrices of MVs are, as before, symbolized by lower case Roman letters.

Equation (6) defines the estimated LVs as linear composites of their associated MVs. Each column of $\mathbf{W_y}$ contains just one non-zero entry, and the weights are chosen so as to give the estimated LVs unit variance.

The estimated LVs defined above are in the second step of PLS estimation that is used to compute the loadings and inner model coefficients by means of standard least squares methods. The loadings are simply defined as zero-order correlations between MVs and their associated LVs. The inner model coefficients are estimated using standard path analytical procedures. That is, for recursive inner models, the respective path coefficients connecting LVs are

obtained by ordinary least squares (OLS) regression applied to each inner model relation separately.

The core of the PLS procedure is obviously the determination of the weights defining LV estimates. These weights are obtained iteratively by means of series of OLS regressions applied to each block of MVs. A distinction is made between two modes of weight estimation called *inward mode* and *outward mode*. Since it is not possible here to give a detailed description of the PLS iteration procedure, the reader is referred to Wold (1982) for an exposition of PLS weight estimation. Let it suffice to say that the estimation of outward blocks is based on an iterative sequence of simple OLS regressions where the MVs are considered dependent variables. Inward blocks are estimated by means of series of multiple OLS regressions where the MVs are considered independent variables.

It should be noted that the distinction between outward and inward blocks corresponds to the differentiation between reflective and formative indicators made by Hauser (1973). Following this differentiation, outward indicators are assumed to reflect rather than to determine a latent dimension. An example would be a set of attitude items which are used as indicators of some attitudinal dimension such as a more positive or more negative attitude toward school. Such indicators are 'reflective' because changing student answers to some attitude items would not cause changes of the attitude being measured. Inward or formative indicators, on the other hand, can be assumed to form or produce the associated latent dimension. A typical example would be a specific teaching style measured by a variety of teacher behaviors. Such indicators can be considered inward because changes of the manifest teaching behaviors would cause changes in the teaching style.

As shown by Wold (1982), one-block PLS models estimated by the outward mode are numerically and analytically equivalent to the first principal component. Also, two-block PLS models where both blocks are estimated by the inward mode are equivalent to a canonical correlation analysis in that the ensuing correlation between the two LVs is equal to the first canonical correlation. Being special cases of PLS, principle component and canonical correlation analysis can be considered basic modules on which the analysis of larger models is based. As noted by Noonan and Wold (1988), PLS has in fact been developed as a generalization of these methods toward the formulation and estimation of more complex path models.

It should also be noted that two-block PLS models, where the exogenous block is estimated by the inward mode and where the endogenous block is estimated by the outward mode, can be considered another basic module of

PLS. This type of model has been considered by Hauser and Goldberger (1971) a so called multiple causes - multiple indicators model. While Hauser and Goldberger developed maximum likelihood estimates, it can be shown that PLS applied to this type of model provides a least squares solution equivalent to what Van den Wollenberg (1977) calls redundancy analysis.

In general terms, the PLS approach merges the basic modules mentioned above in order to estimate more complex path models involving more than two LVs. The estimation process is entirely based on least squares methods being applied under the restrictions imposed by the formulation of inner and outer model relationships. As the focus is on least squares prediction of LVs and MVs, PLS is generally referred to as a 'prediction oriented' approach. Furthermore, since PLS makes use of least squares methods which do not necessarily require stringent assumptions about the distribution of variables, residuals and parameters, Wold (1982) refers to PLS as a 'soft modeling' approach in the sense that PLS does not require restrictive assumptions prevalent to other methods of latent variable path analysis of which the most important is LISREL (Jöreskog 1973; Jöreskog & Sörbom 1978).

4 Model Evaluation

As mentioned above, PLS basically aims to predicit least squares of endogenous LVs and MVs, subject to constraints imposed by the specification of inner and outer model relationships. It will also be recalled from the preceding discussion that predictor specification as applied to inner and outer model relationships constitutes an integral part of the theoretical form of PLS models. Hence, apart from the examination of point estimates (i.e., weights, loadings, and inner model path coefficients), an important part of model evaluation is the examination of fit indices reflecting the predictive power of estimated inner and outer model relationships. Such fit indices can be derived readily from the various inner and outer model equations presented above. For example, R^2 values familiar from multiple regression can be obtained for the inner model relationships. Similarly, so-called communality and redundancy coefficients can be obtained for outer model relations. Communality coefficients are equal to the squared correlations between MVs and their associated LVs and are thus similarly defined as the communalities familiar from standard factor analysis procedures. Redundancy coefficients are derived from the afore mentioned substitutive elimination of latent variables and reflect the joint predictive power

of the inner and outer model relationships being investigated (see, for example, Lohmoeller 1981 for a complete discussion of fit indices).
The statistics referred to above can be used in essentially the same way as the familiar R^2 computed for multiple regression equations. They reflect the relative amount of 'explained' or 'reproduced' variance of LVs and MVs. Two options are available if the researcher wants to go beyond model evaluation in purely descriptive terms. One option is to adopt the distributional assumptions on which the computation of classical estimates of standard errors and F-tests are based and to apply standard significance tests. This approach has been used by Noonan and Wold (1983), for example. It should be reiterated, however, that the classical distributional assumptions, notably normality and independence of residuals, do not constitute prerequisites for PLS estimation. It may thus be regarded as inappropriate to adopt them *post hoc* in order to evaluate the model results. In practice, it is also often the case that the classical assumptions would appear highly unrealistic so that it would make little sense to employ standard test statistics.

Since PLS provides estimated case values of LVs and estimated case values of inner and outer residuals, less demanding statistical techniques such as jackknifing (Tukey 1977) can be used. Wold (1982) specifically proposed the general use of the Stone-Geisser test of predictive relevance (Stone 1974; Geisser 1974). This test basically produces jackknife estimates of residual variances while jackknife standard errors of point estimates can be obtained as a by-product. The general idea is to omit one case at a time, to re-estimate the model parameters on the basis of the remaining cases, and to reconstruct or predict omitted case values using the re-estimated parameters. The extent to which this prediction exercise is successful can be measured by the Q^2 statistic proposed by Ball (1963). As applied to $i = 1,2,\ldots,n$ cases and the familiar case of multiple regression with k regressors, Q^2 is computed as:

$$Q^2 = 1.0 - \sum_n (Y_i - \sum_k X_{ki} b_{k(i)})^2 / \sum_n (Y_i - Y_{\cdot(i)})^2 \qquad (7)$$

where $b_{k(i)}$ is the set of regression coefficients obtained when the i-th case is omitted while $Y_{\cdot(i)}$ denotes the mean of the dependent variable computed without the i-th case. It can be seen from the equation above that Q^2 is nothing else than the jackknife analogue of the familiar R^2. The higher Q^2 is, the higher the predictive relevance of the tested model equation. Q^2 values, contrary to R^2 values, may increase when predictors are deleted from the equation. This would indicate that 'noise' emanating from irrelevant of unstable predictors was removed. Q^2 values may also turn out to be negative. The corresponding model

relation is then said to be misleading because the trivial prediction in terms of the sample mean of the dependent variable is superior to the prediction derived from the model equation being tested.
The concept of measuring predictive relevance by means of the Q^2 statistic allows straightforward extension to PLS models (see, for example, Lohmoeller 1981; Wold 1982; Sellin 1991). The corresponding procedures provide indices of predictive relevance for inner and outer model relations as well as jackknife standard errors of inner and outer model coefficients. These statistics do not require any distributional assumptions and can be used to evaluate the predictive power of the model being investigated.

5 Comparative Comments on LISREL and PLS

As pointed out by Wold (1982), PLS and LISREL are complementary rather than competitive methods for the estimation of the same type of path models. The LISREL approach assumes that observations are governed by a specified multivariate distribution and offers, on this basis, a general framework for (a) maximum likelihood estimation, (b) hypothesis testing leading to either rejection or non-rejection of the tested model, and (c) assessment of standard errors for the model parameters. Least squares estimation, including PLS, is distribution-free except for predictor specification. As compared with LISREL, the complementary characteristics of PLS are (a) least squares estimation by means of the PLS algorithm, (b) tests of predictive relevance using jackknife procedures leading to either non-relevance or some positive degree of predictive relevance, and (c) jackknife estimation of standard errors (Noonan and Wold 1983).

In general, PLS is useful in research situations where exploratory model analyses without restrictive distributional assumptions would seem appropriate. LISREL, on the other hand, is a powerful and highly flexible statistical tool in situations where distributional assumptions would seem justified and where theoretical knowledge is so strong that a confirmatory analysis strategy is in order. It should be noted, however, that PLS may still be operational in situations when LISREL can not be used. These include the analysis of large and complex path models where LISREL often fails to converge within reasonable time limits as well as model analyses based on small data sets where the sample covariance matrix is not positively definite (e.g., when the number of MVs exceeds the number of cases).

6 Use of Partial Least Squares

The PLS approach has been widely used in educational research and specifically in large international research projects conducted by the International Association for the Evaluation of Educational Achievement (IEA). Examples of the application of PLS in educational research are provided by Noonan, Wold (1983) and Keeves (1986), among others. Using the same body of data, Keeves (1986) specifically compared PLS with four other approaches to path analysis including LISREL with the result that PLS provided the most flexible and most appropriate approach.

The Classroom Environment Study (Anderson, Ryan & Shapiro 1989) can be considered another particularly interesting example of the use of PLS. This IEA study involved nine countries and attempted to examine the influences of a variety of instructional practices as well as school, teacher and student characteristics on student achievement. Fairly large amounts of data were obtained from various instruments including questionnaires, achievement tests, and extensive classroom observation. After all, data for about 200 variables were available for each country. The initial design of the study did not involve an explicit theoretical framework as to the ways in which the variables would be expected to be related to each other and the specific ways in which student outcomes would be expected to be influenced. However, with data collection under way, the need for such a conceptual framework was recognized and a general structural model, termed the *core model*, was developed (Anderson et al. 1989, pp. 22-24). This model provided a theoretical structure for the study in that the many variables were grouped into fifteen constructs and in that the relationships between constructs were specified. For a variety of reasons, it was not possible to test the core model empirically. However, the core model served as a point of departure for the formulation of various submodels which were examined in terms of their congruence with the data being collected in diverse countries.

One of these submodels focused primarily on variables and constructs related to individual students such as home background, educational aspiration, attitude toward the school and subject, entry achievement, and student perceptions of instructional behaviors of their teachers. These models also involved a nonrecursive or feedback relation between attitude toward the school subject being studied and cognitive achievement in that subject. This specification was introduced because it seemed reasonable to expect that attitude at one point in time would influence later achievement, and that achievement, in turn, would influence later attitude (Anderson et al. 1989, pp. 168-169). This type of

continuous feedback processes over time can be approximated by so-called nonrecursive path models. The basic PLS design described above was modified by the use of two-stage least squares estimation to facilitate the examination of nonrecursive path models. Depending on the availability of data in the participating countries, the ensuing models involved more than ten constructs and between sixty and eighty manifest variables. PLS was chosen as the primary data analysis approach, not only because other methods would not permit an examination of path models of that size, but also because distributional assumptions prevalent to virtually all alternative methods were judged untenable for the study.

Another, and in many ways quite different, submodel of the Classroom Environment Study focused on variables which reflected instructional practices and events as derived from classroom observations. The examination of these models involved two major data analytical problems. Firstly, the observational variables reflecting instructional practices were defined at the class level and the number of cases involved in the analysis was thus equal to the number of classes. Since not more than thirty classes were observed in most of the countries, the number of cases involved in the analysis was extremely small. Secondly, the frequency distribution of many of the observational variables was extremely skewed because specific instructional behaviors were exhibited by just a few teachers or because these behaviors occurred very infrequently. As a consequence, the possibility that statistical relationships could be strongly influenced by seemingly outlying cases had to be considered.

With regard to the analysis of teaching behaviors, the initial design of the Classroom Environment Study was closely related to what is commonly referred to as process-product research (see, for example, Dunkin & Biddle 1974). This type of research typically focuses on the examination of zero-order correlations between variables reflecting teaching practices (process) and specific measures of educational outcomes (product). Statistically significant correlations are then often clustered together to suggest that the corresponding behaviors represent effective teaching styles.

For the Classroom Environment Study, the use of PLS allowed going beyond an examination of simple zero-order correlations in that more complex and theoretically more adequate models could be tested at least for those countries involving more than 30 classes. Also, extensive use was made of jackknife techniques as a safeguard against drawing conclusions from statistical relationships which were heavily influenced by just a few outlying cases. These jackknife methods were also seen as more appropriate than standard statistical

tests because the underlying assumptions required to carry out such tests were clearly not fulfilled.
The Classroom Environment Study demonstrates the broad scope of PLS analyses, especially in situations when more common methods of data analysis can hardly be applied. This specifically concerns the flexibility in more exploratory analyses of larger models as well as path analyses which have to deal with relatively small numbers of cases and somewhat problematical characteristics of the data to be analyzed.

7 Conclusion

Partial least squares is a flexible and extremely powerful technique for the examination of path models with latent constructs measured by multiple indicators. It is distribution-free except for predictor specification and, thus, requires much less stringent assumptions than other approaches to latent variable path analysis. PLS also allows the use of distribution-free jackknife techniques for the evaluation of statistical relationships. These methods do not require the strict assumptions prevalent to classical significance testing. These unique features of PLS facilitate the analysis of complex models even under circumstances that would cause other methods to fail to produce reasonable results. As such circumstances would appear to be quite common in research on educational achievement, PLS may be considered useful for many researchers working in the field.

Bibliography

Anderson, L.W., Ryan, D.W. & Shapiro, B.J. (1989). *The IEA Classroom Environment Study*. Oxford: Pergamon.

Ball, R.J. (1963). The Significance of Simultaneous Methods of Parameter Estimation in Econometric Models. *Applied Statistics, 12,* 14-25.

Dunkin, M.J. & Biddle, B.J. (1974). *The Study of Teaching.* New York: Holt, Rinehart & Winston.

Geisser, S. (1974). A Predictive Approach to the Random Effects Model. *Biometrica, 61,* 101-107.

Hauser, R.M. (1973). Disaggregating a Social-Psychological Model of Educational Attainment. In A.S. Goldberger & O.D. Duncan (Eds.), *Structural Equation Models in the Social Sciences* (pp. 255-284). New York: Wiley.

Hauser, R.M. & Goldberger, A.S. (1971). The Treatment of Unobservable Variables in Path Analysis. In A.L. Costner (Ed.), *Sociological Methodology* (pp. 81-117). San Francisco: American Sociological Association, Jossey Bass.

Jöreskog, K.G. (1973). A General Method for Estimating a Linear Structural Equation System. In A.S. Goldberger & O.D. Duncan (Eds.), *Structural Equation Models in the Social Sciences* (pp. 85-112). New York: Wiley.

Jöreskog, K.G. & Sörbom, D. (1978). *LISREL IV. A General Computer Program for Estimation of Linear Structural Equation Systems by Maximum Likelihood Methods*. Department of Statistics, University of Uppsala, Sweden.

Keeves, J.P. (1986). Aspiration, Motivation and Achievement: Different Methods of Analysis and Different Results. *International Journal of Educational Research, 10*(2).

Lohmoeller, J.B. (1981). LVPLS 1.6. Program manual. Hochschule der Bundeswehr, Research Report 81.04, München.

Noonan, R. & Wold, H. (1983). *Evaluating School Systems Using Partial Least Squares Evaluation in International Education, 7*. Oxford: Pergamon.

Noonan, R. & Wold, H. (1988). Partial Least Squares Path Analysis. In J.P. Keeves (Ed.), *Educational Research, Methodology and Measurement: An International Handbook*. Oxford: Pergamon.

Sellin, N. (1991). *Statistical Model Building in Research on Teaching: The Case of a Study in Eight Countries*. Unpublished thesis, Faculty of Education, University of Hamburg.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, 36*, 111-147.

Tukey, J. (1977) *Exploratory Data Analysis*. Reading: Addison Wesley.

Van den Wollenberg, A.L. (1977). Redundancy Analysis: An Alternative to Canonical Correlation Analysis. *Psychometrica, 42* (2), 207-219.

Wold, H. (1975). Path Models with Latent Variables. The NIPALS Approach. In H.M. Blalock (Ed.), *Quantitative Sociology* (pp. 307-357). New York: Seminar Press.

Wold, H. (1979). Model Construction and Evaluation when Theoretical Knowledge is Scarce. Cahier 79.06, Department of Econometrics, University of Geneva.

Wold, H. (1982). Soft Modelling: The Basic Design and some Extensions. In K.G. Jöreskog & H. Wold (Eds.), *Systems Under Indirect Observation, Part II*. Amsterdam: North Holland Press.