Partial Least Squares Regression (PLS)

Overview

PLS is sometimes called "Projection to Latent Structures" because of its general strategy. The X variables (the predictors) are reduced to principal components, as are the Y variables (the dependents). The components of X are used to predict the scores on the Y components, and the predicted Y component scores are used to predict the actual values of the Y variables. In constructing the principal components of X, the PLS algorithm iteratively maximizes the strength of the relation of successive pairs of X and Y component scores by maximizing the covariance of each X-score with the Y variables. This strategy means that while the original X variables may be multicollinear, the X components used to predict Y will be orthogonal. Also, the X variables may have missing values, but there will be a computed score for every case on every X component. Finally, since only a few components (often two or three) will be used in predictions, it does not matter that there may have been more original X variables than observations. In contrast, any of these three conditions (multicollinearity, missing values, and too few cases in relation to variables) may well render traditional OLS regression estimates unreliable (and estimates by other procedures in the general and generalized linear model families).

Partial least squares (PLS) regression/path analysis is thus an alternative to OLS regression, canonical correlation, or structural equation modeling (SEM) for analysis of systems of independent and response variables. PLS is a predictive technique which can handle many independent variables, even when there are more predictors than cases and even when predictors display multicollinearity. Like canonical correlation or multivariate GLM, it can also relate the set of independent variables to a set of multiple dependent (response) variables. However, PLS is less than satisfactory as an explanatory technique because it is low in power to filter out variables of minor causal importance (Tobias, 1997: 1).

The advantages of PLS include ability to model multiple dependents as well as multiple independents; ability to handle multicollinearity among the independents; robustness in the face of data noise and missing data; and creating independent latents directly on the basis of crossproducts involving the response variable(s), making for stronger predictions. Disadvantages of PLS include greater difficulty of interpreting the loadings of the independent latent variables (which are based on crossproduct relations with the response variables, not based as in conventional factor analysis on correlations among the manifest independents) and because the distributional properties of estimates are not known, the researcher cannot assess significance except through bootstrap induction. Overall, the mix of advantages and disadvantages means PLS is favored as a predictive technique and not as an interpretive technique, except for exploratory analysis as a prelude to an interpretive technique such as multiple linear regression or structural equation modeling.

Though developed by Herman Wold (Wold, 1981, 1985) for econometrics, PLS first gained popularity in chemometric research and later industrial applications. It has since spread to research in education, marketing, and the social sciences.

PLS may be implemented as a regression model, predicting one or more dependents from a set of one or more independents; or it can be implemented as a path model, akin to structural equation modeling. PLS is implemented as a regression model by SPSS as of SPSS Version 16 and by SAS's PROC PLS as of ver. 6.11.

Key Concepts and Terms

- **Case identifier variable** This is the case ID variable used for casewise output and saved datasets.

- **Dependent variable(s).** There may be one or multiple dependents, which may be nominal (including string), ordinal, or scalar (interval/ratio).

- **Input factors and covariates** are the categorical and scalar variables entered as the dependent(s) and independent(s). Typically there are multiple independents and multiple dependents, used to construct PLS factors and PLS responses respectively.

- **PLS factors.** These are the latent variables extracted as linear combinations of the manifest independent variables. Ordinarily the first 3 - 7 will account for 99% of the variation accounted for. Note: more factors may be extracted than manifest variables. In the SPSS PLS dialog, the user may specify how many factors to extract (five is the default).

- **PLS responses.** These are the latent variables extracted as linear combinations of the manifest response variables. Ordinarily the first 3 - 7 will account for 99% of the variation accounted for. Note: more factors may be extracted than manifest variables.

- **Number of latents** The same number of factors will be extracted for PLS responses as for PLS factors. The researcher must specify how many latents to extract (in SPSS the default is 5). There is no one criterion for deciding how many latents to employ. Common alternatives are:
  1. Cross-validating the model with increasing numbers of factors, then choosing the number with minimum prediction error on the validation set. This is the most common method, where cross-validation is "leave-one-out cross-validation" discussed below, prediction error is measured by the PRESS statistic discussed below, and models are computed for 1, 2, 3,..., c factors and the model with the lowest PRESS statistic is chosen.
  2. Alternatively one may use the method above to identify the model with the lowest prediction error, then choose the model with the least number of factors whose residuals are not significantly greater than that lowest prediction error model. See van der Voet (1994).
  3. In a third alternative, models may be computed for 1, 2, 3,..., c factors, and the process stopped when residual variation reaches a given level.

- **Measurement level.** Dependent and independent variables may be any level (nominal, ordinal, or scalar [interval/ratio]), but when first running PLS in SPSS, the program asks the user to specify measurement levels. This is essential in SPSS: PLS regression uses a multiple regression algorithm for scalar dependent variables and a classification algorithm for categorical dependent variables, or a mixed model if both types of predictors are used. Note
that the user can temporarily change the measurement level setting for a variable to investigate what difference it makes, say, to treat an ordinal survey item as if it were interval (however, covariates must be coded as numeric).

Categorical variable coding. Both nominal and ordinal variables are treated the same, as categorical variables, by SPSS algorithms. Dummy variable coding is used. For a categorical variable with c categories, the first is coded (1, 0, 0, ..., 0), where the last 0 is for the cth category. The last category is coded (0, 0, 0, ..., 1). In the PLS dialog, the researcher specifies which dummy variable representing desired reference category is to be omitted in the model.

When prompted at the start of the PLS run, click the "Define Variable Properties" button to obtain first a dialog letting the user enter the variables to be used, then proceed to the "Define Variable Properties" dialog, shown above. SPSS scans the first 200 (default) cases and makes estimates of the measurement level, classifying variables into nominal, ordinal, or scalar (interval or ratio). Symbols in front of variable names in the "Scanned variable list" on the left show the assigned measurement levels, though these initial assignments may be changed in the main dialog, using the dropdown menu for "Measurement Level". It is a good idea to check proper assignment of missing value codes and other settings in this dialog also. Clicking the "Help" button explains the many options available in the "Define Variable Properties" dialog.

Models. Partial least squares as originally developed in the 1960s by the Wold was a general method which supported modeling paths of causal relation between any number of "blocks" of variables (latent variables), somewhat akin to structural equation modeling. Note that SPSS and SAS implementations are not general PLS models, but rather fit only predictive partial least squares models where one block of predictors (represented in a latent variable) is used to predict a block of responses represented in another latent variable. That is, SPSS and SAS implement PLS regression rather than path models.

Regression vs. path models. PLS embraces two rather distinct modeling approaches, regression models and path models (see Vinzi et al, 2008). The former is an alternative to OLS regression or canonical correlation, while the latter is an alternative to structural equation modeling. That is, PLS path modeling is a method of modeling the relationship among latent variables. Authors may combine both approaches. For instance, Tenenhaus et al. (2004) in a marketing study, used PLS regression to obtain a graphical display of products and their characteristics, with a mapping of consumer preferences; then they used PLS path modelling to obtain a detailed analysis of each consumer group by building a causal model involving consumer preference, physico-chemical, and sensory blocks of variables.

PLS-DA models are PLS discriminant analysis models. These are simply PLS models where the dependent/response variable is binary variable or a dummy variable rather than a continuous variable.

In a simulation study of PLS regression, Temme, Kreis, & Lutz (2006: 20) found "In our study, results for simulated data, however, are very similar to those resulting from OLS regression. This issue should be the subject of a comprehensive Monte Carlo study."

SPSS. In the PLS dialog in SPSS, the "Model" tab allows the user to specify the effects to be modeled. The default is a main effects model, with all factor and covariate effects and no interactions. Clicking the "Custom" radio button allows the user to specify the exact main and interaction effects the researcher desired in the model. Interactions of input factors (categorical variables) and covariates is supported.

Cross-validation. The PLS model is developed for all the cases save one, then tested on that hold-out case. This is repeated n times, with each case used as the validation case in turn. In addition to this leave-one-out form of cross-validation (also called "full cross-validation"), software (e.g., SAS) may support cross-validation by splitting the data into blocks or by reserved test set validation. Leave-one-out crossvalidation is recommended for small datasets; split sample and reserved test set methods require larger samples.

Goodness of fit. Compared to SEM, PLS lacks the variety of goodness of fit measures to assess overall model fit. However, cross-validation indices (cv-redundancy and cv-communality) are supported by some existing PLS software programs.

PRESS. The predicted residual sum of squares (PRESS) for cross-validation is computed for the 0-factor model, the 1-factor model, the 2-factor model, etc. When PRESS declines only insignificantly when an additional factor is extracted, factor extraction is stopped.

Statistical output in SPSS PLS

Proportion of variance explained by latent factor. This section of the output lists latent factors by rows (here, there are the default 5). The "cumulative X variance" is the percent of variance in the X variables accounted for by the latent factors. The "cumulative Y variance" is the percent of variance in the Y variables accounted for by the latent factors. Both are interpreted as cumulative R-square in regression. Note that a model may explain variance more in the X variables than the Y variables, more in the Y variables than the X variables, or equally. A well-fitting model would explain most of the variation in both the X and Y variables (not the case in this example). "Adjusted R-square" penalizes for model complexity, accounting for the fact that in the example below, adjusted R-square actually declines with increasing numbers of factors because the added variance explained is less than the complexity penalty factor.

The more a factor explains of the variation in the Y variables, the more powerful it is apt to be in explaining the variation in a new sample of dependent values. The more a factor explains in the variation of the X variables, the more it will reflect the observed values of the set of independent variables.
PRESS (predictive error sum of squares). Not shown above, but when cross-validation is undertaken, the variance explained table will also contain a column for the PRESS statistic. The row with the lowest PRESS statistic corresponds to the number of useful factors. However, when PRESS values differ only by very small amounts, on parsimony grounds the researcher may opt to re-run the model requesting the number of factors for the row corresponding to the fewest factors for the set of rows all comparatively small in PRESS levels.

Latent factor weights and loadings: Loadings or weights indicate how much each independent variable contributes to the axis representing the loadings below are the X-loadings, representing the correlation of the X-variables with the Y-scores. The loadings below are the X-loadings, representing the directions of the lines for each independent in X-space. Typically, X-weights and X-loadings are quite similar and serve similar interpretative uses. Loadings, as in principal components factor analysis, are used to impute meanings for the factors - something which can be difficult when, as below, there is not a simple factor structure devoid of important cross-loadings of variables on more than one factor. By one rule of thumb in confirmatory PLS factor analysis, loadings should be .7 or higher to confirm that independent variables identified a priori are represented by a particular factor (Hulland, 1999: 198). However, the .7 standard is a high one and real-life data may well not meet this criterion, which is why some researchers, particularly for exploratory purposes, will use a lower level such as .4 for the central factor and .25 for other factors (Raubenheimer, 2004). In any event, factor loadings must be interpreted in the light of theory, not by arbitrary cutoff levels. If factors can be convincingly labeled, then one may use R-square contributions from the proportion of variance explained table above to assign relative importance to, say, occupational prestige (factor 1) vs. race (factor 5). To the extent that variables are crossloaded, such comparisons may be unwarranted or misleading. Weights are used to plot the position of independent and dependent variables in factor space, as illustrated below in the section on chart and plot output.

### Weights

<table>
<thead>
<tr>
<th>Variables</th>
<th>Latent Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>sex=Male</td>
<td>-0.04</td>
</tr>
<tr>
<td>race=White</td>
<td>0.297</td>
</tr>
<tr>
<td>race=Black</td>
<td>-0.301</td>
</tr>
<tr>
<td>age</td>
<td>0.493</td>
</tr>
<tr>
<td>preglift</td>
<td>0.779</td>
</tr>
<tr>
<td>happy=Very Happy</td>
<td>-0.113</td>
</tr>
<tr>
<td>happy=Pretty Happy</td>
<td>-0.069</td>
</tr>
</tbody>
</table>

### Loadings

<table>
<thead>
<tr>
<th>Variables</th>
<th>Latent Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>sex=Male</td>
<td>0.006</td>
</tr>
<tr>
<td>race=White</td>
<td>0.534</td>
</tr>
<tr>
<td>race=Black</td>
<td>-0.373</td>
</tr>
<tr>
<td>age</td>
<td>-0.379</td>
</tr>
<tr>
<td>preglift</td>
<td>-0.845</td>
</tr>
<tr>
<td>happy=Very Happy</td>
<td>2.062</td>
</tr>
<tr>
<td>happy=Pretty Happy</td>
<td>-0.207</td>
</tr>
</tbody>
</table>

Variable importance in projection (VIP): VIP coefficients reflect the relative importance of each X variable for each factor in the prediction model. VIP coefficients thus represent the importance of each X variable in fitting both the X- and Y-scores, since the Y-scores are predicted from the X-scores. As a rule of thumb advanced by Wold (1994), the researcher may wish to drop from the model any independent variable which (a) has a small VIP (< .8) and (b) also has a regression coefficient (see "parameter estimates" below) small in absolute size. Here, both sex and race are candidates for deletion.

### Variable Importance in the Projection

<table>
<thead>
<tr>
<th>Variables</th>
<th>Latent Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>sex=Male</td>
<td>-0.049</td>
</tr>
<tr>
<td>race=White</td>
<td>0.663</td>
</tr>
<tr>
<td>race=Black</td>
<td>0.374</td>
</tr>
<tr>
<td>age</td>
<td>0.134</td>
</tr>
<tr>
<td>preglift</td>
<td>1.739</td>
</tr>
</tbody>
</table>

Regression parameter estimates by dependent variable: In the example below, there dependent was a categorical variable with three levels (1=very happy, 2=pretty happy, 3=not very happy). The third level became the reference category, which is why only the first two dummy dependents are shown in the example below. The parameter estimates are the regression coefficients used in conjunction with the independent variables, which are both categorical and covariate variables, to predict the dependent variables. The sign of the coefficient indicates the direction of the effect. For instance, below, higher occupational prestige is associated positively with "very happy" and negatively with "pretty happy" as compared to persons in the "not very happy" category.

### Parameters

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Dependent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>happy=Very Happy</td>
</tr>
<tr>
<td>(Constant)</td>
<td>0.049</td>
</tr>
<tr>
<td>sex=Male</td>
<td>0.015</td>
</tr>
<tr>
<td>race=White</td>
<td>0.034</td>
</tr>
<tr>
<td>race=Black</td>
<td>-0.020</td>
</tr>
<tr>
<td>age</td>
<td>0.001</td>
</tr>
<tr>
<td>preglift</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Charts/plots. In SPSS, the following plots are available.

- Plots of variable importance in projection (VIP) by latent factor: This plot simply reproduces in graphic form the cumulative VIP scores from the "Variable importance in projection" table above.
Plots of latent factor scores: In these plots, the X score for a model with a given number of factors is plotted on the X axis and the Y score for a model with a given number of factors is plotted on the Y axis. In the example below, each of the three dependent variables (which were dummy variables) are graphed in each plot, represented by different colors. These plots show the directions toward which each PLS X-factor is projecting with respect to the Y factors. For instance, in the example below, the X-scores for factor 1 are associated with increasing values on Y-scores for factor 2, but little effect on Y-scores for factor 1. This reflects a weak model, with a low percent of variance explained. These plots may also display grouping patterns among the cases, as below between the X-scores for factors 4 and 5 with the Y-scores for factors 2-5. Grouping, particularly on the factors explaining the bulk of the variance, may indicate clusters of cases not well represented by variables in the model and which need to be analyzed separately.

- It is also possible to plot X scores from one PLS factor against the X scores for a different PLS factor. In the example below, X-scores for factor 1 are negatively related to X-scores for factor 2, and unrelated to those for factor 3.
*Plots of latent factor weights:* In SPSS, these are plotted for the first three latent factors. Variables located near the 0,0 mark contribute little to model fit. Variables further from the origin contribute more to differentiating the X factors (dimensions). Variables which contribute little on the X factors which explain the bulk of the variance in Y are variables which the researcher may wish to drop from the model; dropping them may improve the Y-variance explained.
Saving variables. Clicking the "Options" tab in the SPSS PLS main dialog allows the user to save variables to a SPSS dataset. The dataset name should be unique or the file will be overwritten. One may save predicted values, residuals, distance to latent factor model, latent factor scores, latent factor loadings, latent factor weights, regression parameter estimates, and variable importance to projection (VIP) coefficients.

Residual and normal quantile plots, based on saved residual values, are a means for identifying outlier observations. Residual plots and normal quantile plots also help identify non-normality and heteroscedasticity of error terms. A cloud-like residual plot with most cases in the middle third of the plot indicates heteroscedasticity and normality assumptions are not violated. In the case of normal quantile plots, points on a straight line indicate the normality assumption is not violated.

Software for PLS

- PLS Regression
Assumptions

- Multicollinearity. Since the PLS factors are orthogonal, by definition mathematical multicollinearity is not a problem in PLS. This is a major reason why PLS models may be selected over OLS regression models or structural equation modeling. Note, however, that this does not mean that multicollinearity just “goes away.” To the extent that the original X variables are multicollinear, PLS will lack a simple factor structure and the factor cross-loadings will mean PLS factors will be difficult to label, interpret, and distinguish.

- Distribution-free. PLS is a maximum likelihood approach to regression and path modeling, unlike structural equation modeling using the usual maximum likelihood estimation method, which assumes multivariate normality (see Lohmoller, 1989: 31).

- Independence of observations is not required. (See Lohmoller, 1989: 31).

- Bootstrap estimates of significance. As the distribution of PLS is unknown, conventional significance testing is impossible. However, testing may be accomplished by bootstrap methods such as the jackknife, as illustrated by Davies (2001) and implemented in Unscrambler software. Resampling methods like the jackknife do not have specific sample size requirements, but the smaller the sample, the more likely that fitted confidence limits will be fitted to noise in the data rather than to a true underlying distribution.

- Proper use of dummy variables. The dummy variable representing desired reference category must be omitted in the model. Coefficients for the remaining dummy variables in the set for a given categorical variable must be interpreted with respect to the reference category.

- Standardized variables. When interpreting output, keep in mind that all variables in the model have been centered and standardized, including dummy variables for categorical variables.

- Assumptions of linear regression. As an extension of the multiple linear regression model, PLS shares most of the other assumptions of multiple regression, with the notable exception that lack of multicollinearity among the independents is not required in PLS. In particular, the researcher must be concerned with outliers and nonlinear data relationships.

Frequently Asked Questions

- Why is PLS sometimes described as a ‘soft modeling’ technique? This is an allusion to assumptions of multiple linear regression compared to PLS. OLS regression makes “hard” assumptions, including lack of multicollinearity among the independent variables, which in turn often implies few independent variables with well-understood relationships to the dependent variable. “Soft modeling” refers to modeling when these assumptions cannot be met. PLS can handle multicollinearity, many independent variables, and because its focus is prediction, not explanation, lack of well-understood relationships to the independents to the dependent is not critical.

- You said PLS could handle large numbers of independents, but can’t OLS regression do this too? In principle, yes. In practice, the larger the number of independents, the greater the likelihood of multicollinearity, which renders OLS regression results unreliable. Moreover, when the number of variables is as great or greater than the number of cases, OLS regression will “overfit” the data and one will have a spurious perfect fit model which will fail to have predictive power for a new set of data.

- Is PLS always a linear technique? While PLS is a linear technique, nonlinear PLS (NLPLS) has been advanced by Malthouse, Rasmussen, and others. This nonlinear approach
combines PLS with feedforward neural network analysis (or in a variant with radial basis function neural network analysis). See Malthouse et al. (1997).

- **How is PLS related to principal components regression (PCR) and maximum redundancy analysis (MRA)?**

  All three are similar in sharing the strategy of indirect modeling, using factors derived from the X variables to predict factor scores of the Y variables, which are then used to construct predictions of the raw Y variables. Where the PLS algorithm maximizes the strength of the relation of the X and Y factor scores, PCR maximizes the degree to which the X factor scores explain the variance in the X variables; and MRA maximizes the degree to which the Y factor scores explain the variance in the Y variables. Whereas the PLS algorithm chooses X-scores of the latent independents to be paired as strongly as possible with Y-scores to explain the maximum proportion of the factor variation. Often this means that PCR latents are less related to dependent variables of interest to the researcher than are PLS latents. On the other hand, PCR latents do not draw on both independent and response variables in the factor extraction process, with the result that PCR latents are easier to interpret.

  PLS generally yields the most accurate predictions and therefore has been much more widely used than PCR. PLS may also be more parsimonious than PCR. In a chemistry setting, Westerholt & Vega (2003: 257) conducted simulations to compare PLS and PCR, finding "in all cases, except when artificial constraints were placed on the number of latent variables retained, no significant differences were reported in the prediction error reported by PCR and PLS. PLS almost always required fewer latent variables than PCR, but this did not appear to influence predictive ability."

  Attempts have been made to improve the predictive power of PCR. Traditional PCR methods use the first k components (first by having the highest eigenvalues) to predict the response variable, Y. Hwang & Nettleton (2003: 71 ) note, "Restricting attention to principal components with the largest eigenvalues helps to control variance inflation but can introduce high bias by discarding components with small eigenvalues that may be most associated with Y. Jolliffe (1982) provided several real-life examples where the principal components corresponding to small eigenvalues had high correlation with Y. Hadi and Ling (1998) provided an example where only the principal component associated with the smallest eigenvalue was correlated with Y."

  Recall variance inflation (measured in regression by the variance inflation factor, VIF) indicates multicollinearity: while a multicollinear model may explain a high proportion of variance in Y, but redundancy among the X variables leads to inflated standard error and inflated parameter estimates. Minimizing variance inflation may not minimize mean square error (MSE). To deal with the tradeoff between variance inflation and MSE, some researchers employ an "inferential approach", which uses only components whose regression coefficients significantly differ from zero (Mason & Gunst, 1985). More recently, Hwang & Nettleton (2003) have proposed a PCR selection strategy which selects components which minimize mean square error (MSE) demonstrating through simulations studies that their estimator performed superior to traditional PCR, inferential PCR, or even traditional PLS (which ranked second in the simulation, among many variants tested). However, it appears that Hwang-Nettleton estimators are not employed by current software.

  - **What are the SIMPLS and PCR methods in proc PLS in SAS?**

    PROC PLS in SAS supports a METHODS= statement. With METHODS=PLS one gets standard PLS calculations. As this is the default method, the statement may be omitted. With METHODS=SIMPLS, one gets an alternative algorithm developed by de Jong (1993). SIMPLS is identical to PLS when there is a single response (dependent) variable but is computationally much more efficient when there are many responses. Results under PLS and SIMPLS are generally very similar.

    PCR. With METHODS=PCR one is asking for principal components regression, which predicts response factors from variables underlying the predictor variables. Latents created with PCR may not predict Y-scores as well as latents created by PLS or SIMPLS.

  - **What are the NIPALS and SVD algorithms?**

    There is more than one way to compute PLS coefficients. NIPALS is the nonlinear iterative partial least square method and is the most common. SVD is the singular value decomposition method, sometimes used because it is computationally faster. Results are very similar by either algorithm.

Bibliography