Non-OLS Linear SEL Prediction: Partial Least Squares Regression

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Derived inputs and PLS

Here we continue to assume that **the columns of X have been standardized and Y has been centered**. We discuss a second prediction methodology involving some number $M < p$ of predictors derived from the original inputs x_j, so-called Partial Least Squares (PLS) Regression. The shrinking methods treated thus far take no account of **Y** in determining directions or amounts of shrinkage. PLS employs **Y**.

First PLS component

Let

$$\mathbf{z}_1 = \sum_{j=1}^{p} \langle \mathbf{Y}, \mathbf{x}_j \rangle \, \mathbf{x}_j = \mathbf{X}\mathbf{X}'\mathbf{Y}$$

For $\mathbf{w}_1 = \mathbf{X}'\mathbf{Y} / \left\|\mathbf{X}'\mathbf{Y}\right\|$, $\mathbf{X}\mathbf{w}_1 = \mathbf{z}_1 / \left\|\mathbf{X}'\mathbf{Y}\right\|$ is a linear combination of the columns of $\mathbf{X}$ maximizing, subject to the constraint that $\|\mathbf{w}\| = 1$, the quantity

$$|\langle \mathbf{Y}, \mathbf{X}\mathbf{w} \rangle|$$

which is essentially the absolute sample covariance between the variables y and $\mathbf{x}'\mathbf{w}$. (The first principal component of $\mathbf{X}$ maximizes $|\langle \mathbf{X}\mathbf{w}, \mathbf{X}\mathbf{w} \rangle|$ subject to the same constraint.)

Second PLS component

Then define $\mathbf{X}^1$ by orthogonalizing the columns of $\mathbf{X}$ with respect to $\mathbf{z}_1$. That is, define the jth column of $\mathbf{X}^1$ by

$$\mathbf{x}_j^1 = \mathbf{x}_j - \frac{\langle \mathbf{x}_j, \mathbf{z}_1 \rangle}{\langle \mathbf{z}_1, \mathbf{z}_1 \rangle} \mathbf{z}_1$$

and take

$$\mathbf{z}_2 = \sum_{j=1}^{p} \langle \mathbf{Y}, \mathbf{x}_j^1 \rangle \, \mathbf{x}_j^1 = \mathbf{X}^1 \mathbf{X}^{1\prime} \mathbf{Y}$$

(For $\mathbf{w}_2 = \mathbf{X}^{1\prime}\mathbf{Y} / \left\| \mathbf{X}^{1\prime}\mathbf{Y} \right\|$, $\mathbf{X}^1 \mathbf{w}_2 = \mathbf{z}_2 / \left\| \mathbf{X}^{1\prime}\mathbf{Y} \right\|$ is the linear combination of the columns of $\mathbf{X}^1$ maximizing

$$\left| \langle \mathbf{Y}, \mathbf{X}^1 \mathbf{w} \rangle \right|$$

subject to the constraint that $\|\mathbf{w}\| = 1$.)

PLS regression

Then for $l > 1$, define $\mathbf{X}^l$ by orthogonalizing the columns of $\mathbf{X}^{l-1}$ with respect to $\mathbf{z}_l$. That is, define the jth column of $\mathbf{X}^l$ by

$$\mathbf{x}_j^l = \mathbf{x}_j^{l-1} - \frac{\left\langle \mathbf{x}_j^{l-1}, \mathbf{z}_l \right\rangle}{\left\langle \mathbf{z}_l, \mathbf{z}_l \right\rangle} \mathbf{z}_l$$

and let

$$\mathbf{z}_{l+1} = \sum_{j=1}^{p} \left\langle \mathbf{Y}, \mathbf{x}_j^l \right\rangle \mathbf{x}_j^l = \mathbf{X}^l \mathbf{X}^{l\prime} \mathbf{Y}$$

Partial least squares regression uses the first M of these variables $\mathbf{z}_j$ as input variables.

PLS coefficient vector and predictor

The PLS predictors $\mathbf{z}_j$ are orthogonal by construction. Using the first M of these as regressors, one has the vector of fitted output values

$$\widehat{\mathbf{Y}}^{\mathrm{PLS}} = \sum_{j=1}^{M} \frac{\langle \mathbf{Y}, \mathbf{z}_j \rangle}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle} \mathbf{z}_j$$

Since the PLS predictors are (albeit recursively-computed data-dependent) linear combinations of columns of $\mathbf{X}$, it is possible to find a p-vector $\widehat{\boldsymbol{\beta}}_M^{\mathrm{PLS}}$ (namely $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{Y}}^{\mathrm{PLS}}$) such that

$$\widehat{\mathbf{Y}}^{\mathrm{PLS}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_M^{\mathrm{PLS}}$$

and thus produce the corresponding linear prediction rule

$$\widehat{f}\left(\mathbf{x}\right) = \mathbf{x}'\widehat{\boldsymbol{\beta}}_M^{\mathrm{PLS}} \qquad (1)$$

PLS predictor complexity

It seems like in (1), the number of components, M, should function as a complexity parameter. But then again there is the following. When the $\mathbf{x}_j$ are orthogonal, it's fairly easy to see that $\mathbf{z}_1$ is a multiple of $\widehat{\mathbf{Y}}^{\text{OLS}}$, $\widehat{\beta}_1^{\text{PLS}} = \widehat{\beta}_2^{\text{PLS}} = \cdots = \widehat{\beta}_p^{\text{PLS}} = \widehat{\beta}^{\text{OLS}}$, and all steps of partial least squares after the first are simply providing a basis for the orthogonal complement of the 1-dimensional subspace of $C(\mathbf{X})$ generated by $\widehat{\mathbf{Y}}^{\text{OLS}}$ (without improving fitting at all). That is, here changing M doesn't change flexibility of the fit at all.

Presumably, when the $\mathbf{x}_j$ are nearly orthogonal, something similar happens, and one might thus expect PLS to be most effective as a shrinkage method where there are substantial correlations among columns of $\mathbf{X}$.

PLS complexity and effective df

This observation about PLS in cases where predictors are orthogonal has another simple implication. That is that there will be no naive form for effective degrees of freedom for PLS. Since with $\mathbf{z}_j$ the jth principal component of $\mathbf{X}$ and, say,

$$\mathbf{Z}^M = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_M)$$

we have

$$\widehat{\mathbf{Y}}^{\mathrm{PCR}} = \mathbf{Z}^M \left(\left(\mathbf{Z}^M \right)' \mathbf{Z}^M \right)^{-1} \left(\mathbf{Z}^M \right)' \mathbf{Y}$$

principal components regression on M components has effective degrees of freedom M. But the fact that the "$\mathbf{Z}^M$" matrix corresponding to PLS depends upon $\mathbf{Y}$ makes PLS *nonlinear in* $\mathbf{Y}$. And the "orthogonal $\mathbf{X}$" argument shows that a PLS predictor with $M = 1$ can have effective degrees of freedom as large as *rank* $(\mathbf{X})$.

PLS, PCR, and OLS

PLS is a kind of compromise between principal components regression and ordinary least squares. To see this, note that maximizing

$$|\langle \mathbf{Y}, \mathbf{X}\mathbf{w} \rangle|$$

subject to the constraint that $\|\mathbf{w}\| = 1$ is equivalent to maximizing the sample covariance between $\mathbf{Y}$ and $\mathbf{X}\mathbf{w}$ i.e.

$$\left(\begin{array}{c}\text{sample standard}\\ \text{deviation of } y\end{array}\right) \cdot \left(\begin{array}{c}\text{sample standard}\\ \text{deviation of } \mathbf{x}'\mathbf{w}\end{array}\right) \cdot \left(\begin{array}{c}\text{sample correlation}\\ \text{between } y \text{ and } \mathbf{x}'\mathbf{w}\end{array}\right)$$

or equivalently

$$\left(\begin{array}{c}\text{sample variance}\\ \text{of } \mathbf{x}'\mathbf{w}\end{array}\right) \cdot \left(\begin{array}{c}\text{sample correlation}\\ \text{between } y \text{ and } \mathbf{x}'\mathbf{w}\end{array}\right)^2 \qquad (2)$$

subject to the constraint.

PLS, PCR, and OLS cont.

Now if only the first term (the sample variance of $\mathbf{x}'\mathbf{w}$) were involved in (2), a first principal component direction would be an optimizing $\mathbf{w}_1$, and $\mathbf{z}_1 = \left\|\mathbf{X}'\mathbf{Y}\right\| \mathbf{X}\mathbf{w}_1$ a multiple of the first principal component of $\mathbf{X}$. On the other hand, if only the second term were involved, $\hat{\beta}^{\mathrm{OLS}} / \left\|\hat{\beta}^{\mathrm{OLS}}\right\|$ would be an optimizing $\mathbf{w}_1$, and $\mathbf{z}_1 = \hat{\mathbf{Y}}^{\mathrm{OLS}} \left\|\mathbf{X}'\mathbf{Y}\right\| / \left\|\hat{\beta}^{\mathrm{OLS}}\right\|$ a multiple of the vector of ordinary least squares fitted values. The use of the product of two terms can be expected to produce a compromise between these two.

This logic applied at later steps in the PLS algorithm then produces for $\mathbf{z}_l$ a compromise between a first principal component of $\mathbf{X}^{l-1}$ and a suitably constrained multiple of the vector of least squares fitted values for $\mathbf{Y}$ based on the matrix of inputs $\mathbf{X}^{l-1}$. $\mathbf{X}^l$ has columns that are the corresponding columns of $\mathbf{X}$ minus their projections onto the span of $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l\}$ and $\mathcal{C}(\mathbf{X}) \supset \mathcal{C}(\mathbf{X}^1) \supset \mathcal{C}(\mathbf{X}^2) \cdots$