PLS - Exercise 2

In the following we'll focus on a small data set with n=5 observations so that the math can be done by calculator. It can of course be done in Unscrambler alternatively. Data is given below.

i	$\mathbf{X}_{1i}$	$\mathbf{X}_{2i}$	$\mathbf{y}_i$
1	1	0	5
2	4	2	7
3	3	6	9
4	6	4	10
5	5	10	12

PLS - Exercise 2

Plot **y** against $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively as well as $\mathbf{x}_1$ against $\mathbf{x}_2$ to get a feel of the data! Comment what you see!

Calculate means and standard deviations for $\mathbf{x}_1$, $\mathbf{x}_2$ and **y**! Center or standardize data prior to the PLS regression! Which/why/when?

Calculate RMSEC (Root Mean Squared Errors Calibration) for a one component PLS regression! Compare it to RMSEC for the MLRs: $\mathbf{y}= f(\mathbf{x}_1)$, $\mathbf{y}= f(\mathbf{x}_2)$ and $\mathbf{y}= f(\mathbf{x}_1, \mathbf{x}_2)$! Look into the earlier MLR exercises for some help.

Are these RMSECs comparable at all?

For a little help on the PLS math see the next slides.

The PLS algorithm in short

PLS components are found in successive steps for centered or standardized data:

1. Find the loading weights **w** using the remaining variation/errors in **X** and **y**. Scale **w** to unit length.
2. Project **X** on **w** – that is find the scores **t** on that component.
3. Estimate unit length loadings **p** for **X** and **y** respectively using the **X** and **y** errors and the just found scores **t**.
4. Calculate the new **X** and **y** errors – that is the variations not accounted for by the found component(s).

Finally the regression parameters can be calculated for an appropriate number of components for the original data.

PLS math (standardized or centered data)

Loading weights: $$w = \frac{1}{\sqrt{y_{res}^{'} X_{res} X_{res}^{'} y_{res}}} X_{res}^{'} y_{res}$$

Scores: $$t = X_{res} w$$

Loadings: $$p_y = y_{res}^{'} t/(t't)$$ $$p_X = X_{res}^{'} t/(t't)$$

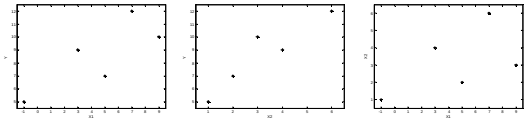
New Residuals: $$X_{res} = X_{res} - tp_X^{'}$$ $$y_{res} = y_{res} - tp_y^{'}$$

(res stands for residuals/errors/remaining variations)

Solution

Both X-variables show positive correlations to **y**. $\mathbf{x}_1$ more than $\mathbf{x}_2$. There is also some correlation between them indicating that the variations in the X-variables could be due to a common latent variable.

PLS regression with one component seems a sensible choice.



	$\mathbf{X}_1$	$\mathbf{X}_2$	**y**
mean	4.6	3.2	8.6
var	14.8	3.7	7.3

Corr.	**y**	$\mathbf{X}_1$	$\mathbf{X}_2$
y	1	0.79852	0.93321
$\mathbf{X}_1$	0.79852	1	0.55405
$\mathbf{X}_2$	0.93321	0.55405	1

Solution

If you made the MLRs on original or mean centered data and the PLS on standardized data these RMSECs are not comparable. Preferably you should report the RMSECs in original units. Therefore multiply standardized RMSECs by std(**y**).

RMSECs	$\mathbf{y}=\mu$	PLS (one component)	MLR $\mathbf{y}=f(\mathbf{x}_1)$	MLR $\mathbf{y}=f(\mathbf{x}_2)$	MLR $\mathbf{y}=f(\mathbf{x}_1, \mathbf{x}_2)$
Standardized data	0.894	0.142	0.538	0.321	0.109
Original units	2.417	0.383	1.455	0.868	0.294

In this case one component PLS regression does almost as good a job as the full MLR and surely better than the two simple univariate regressions. Including a second component the PLS RMSEC will be the same as for the full MLR.

In PCR **X** alone determines the decomposition whereas in PLS regression the **y** helps guide the decomposition. **X** still has a saying which is why centering and standardization of **X** will yield different decompositions. The **x** with the largest variation ($\mathbf{x}_2$) will be favored! In this case mean centering instead of standardization will result in a one component PLS RMSEC of 1.041 which is considerably worse.