Statistics for the Sciences

Partial Least Squares (PLS)

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Principal Component Regression
- Partial Least Squares
- Example
- Lab

Principal Component Regression

- Our data consist of n observations with p predictors.
- However, not all of those p dimensions are equally useful, especially when $p >> n$.
- Here we apply principal component analysis (PCA) to define the linear combinations of the predictors, for use in our regression. By the theory of Principal Components,
  - The first principal component is that (normalized) linear combination of the variables with the largest variance.
  - The second principal component has largest variance, subject to being uncorrelated with the first.
  - And so on.
- Hence with many correlated original variables, we replace them with a small set of principal components that capture their joint variation.

Principal Component Regression

- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ **principal components**. That is,

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j \tag{1}$$

  - These M PCs are the linear combinations of the variables that contain as much as possible of the variability in the features.
- Then we use least squares to fit the model (regress Y on the M PCs)

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m Z_{im} + \varepsilon_i, i = 1, \ldots, n \tag{2}$$

  - In other words, we perform least squares using M new predictors $Z_1, Z_2, \ldots, Z_M$ which are the **principal components** of the predictors.

Principal Component Regression

- Notice that from definition (1),

$$\sum_{m=1}^{M} \theta_m Z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{mj} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

  where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj}. \tag{3}$$

- Hence model (2) can be thought of as a special case of the original linear regression model.
  - All predictors are in the model. No feature selection!
- Dimension reduction serves to constrain the estimated β_j coefficients, since now they must take the form (3).

Principal Component Regression

- In summary, we apply principal components analysis (PCA) to define the linear combinations of the predictors, for use in our regression.
  - Hence with many correlated original variables, we replace them with **a small set of principal components** that capture their joint variation.
- **PCR doesn't yield feature selection** - all of the original predictors are involved in the final model.
- But when M is small, then PCR can **avoid overfitting** and can give good results.
  - With $M = p$, we just get least squares regression: no dimension reduction occurs!
- PCR directions are identified in an unsupervised way, since the response Y is not used to help determine the principal component directions.
  - Consequently, PCR suffers from a potentially serious drawback: there is **no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response**.

Partial Least Squares

- Like PCR, PLS is a dimension reduction method, which first identifies a new set of features $Z_1, \ldots, Z_M$ that are linear combinations of the original features (predictors), and then fits a linear model via OLS using these M new features (predictors).
- But unlike PCR, PLS identifies these new features in a **supervised way** - that is, it makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are **related to the response**.
- Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.

Partial Least Squares

- After standardizing the p predictors, PLS computes the first direction Z_1 by setting each ϕ_{1j} in $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$ equal to the coefficient from the **simple linear regression of Y onto X_j**.
  - Z_1 is not the first PC in PCA any more.
  - One can show that this coefficient is proportional to the **correlation** between Y and X_j.
  - Hence, in computing $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$, PLS places the highest weight on the variables that are most strongly related to the response.

- Subsequent directions $Z_m = \sum_{j=1}^{p} \phi_{mj} X_j, m = 2, \ldots, M$ are found by taking residuals after regression of the original data on Z_{m-1}, and Z_m is calculated in the same way as Z_{m-1} for the residuals data (orthogonalized data), then repeating the above prescription.

Example

- Consider the data `loyn.csv` analyzed before using MLR models: Loyn (1987) selected 56 forest patches in southeastern Victoria, Australia, and related the abundance of forest birds in each patch to six predictor variables: patch area (ha), distance to nearest patch (km), distance to nearest larger patch (km), grazing stock (1 to 5 indicating light to heavy), altitude (m) and years since isolation (years).
  - Add log10 transformation of `area` and `dist`

```
##    abund area yearisol dist distl graze alt  logarea  logdist
## 1    5.3  0.1     1968   39    39     2 160 -1.00000 1.591065
## 2    2.0  0.5     1920  234   234     5  60 -0.30103 2.369216
## 3    1.5  0.5     1900  104   311     5 140 -0.30103 2.017033
## 4   17.1  1.0     1966   66    66     3 160  0.00000 1.819544
## 5   13.8  1.0     1918  246   246     5 140  0.00000 2.390935
## 6   14.1  1.0     1965  234   285     3 130  0.00000 2.369216
## 7    3.8  1.0     1955  467   467     5  90  0.00000 2.669317
## 8    2.2  1.0     1920  284  1829     5  60  0.00000 2.453318
## 9    3.3  1.0     1965  156   156     4 130  0.00000 2.193125
## 10   3.0  1.0     1900  311   571     5 130  0.00000 2.492760
```

Example

- After the log transformation of `area` and `dist`
  - Response `abund`
  - 6 predictors

```
## 'data.frame':    56 obs. of  7 variables:
##  $ abund   : num  5.3 2 1.5 17.1 13.8 14.1 3.8 2.2 3.3 3 ...
##  $ yearisol: int  1968 1920 1900 1966 1918 1965 1955 1920 1965 19
##  $ distl   : int  39 234 311 66 246 285 467 1829 156 571 ...
##  $ graze   : int  2 5 5 3 5 3 5 5 4 5 ...
##  $ alt     : int  160 60 140 160 140 130 90 60 130 130 ...
##  $ logarea : num  -1 -0.301 -0.301 0 0 ...
##  $ logdist : num  1.59 2.37 2.02 1.82 2.39 ...
```

Example - PCR method

```
## Data:    X dimension: 56 6
##  Y dimension: 56 1
## Fit method: svdpc
## Number of components considered: 6
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
## CV           10.83    6.658    6.747    6.746    6.953    7.078
## adjCV        10.83    6.604    6.733    6.717    6.924    7.012
##
## TRAINING: % variance explained
##        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X        37.50    65.08    78.14    89.95    95.63   100.00
## abund    62.72    62.94    63.88    64.07    68.57    68.77
```

Example - PCR method

- Here Cross-Validation method is used: the smallest cross-validation error occurs when $M = 1$ component is used.
- Using $M = 1$ captures 37.50% of all the variance, or information, in the predictors.
  - 62.72% of variation of the response can be explained by the first PC.
- Using MLR regression, 68.77% of variation of the response can be explained by all predictors (6 PCs).

Example - PCLS method

```
## Data:    X dimension: 56 6
##  Y dimension: 56 1
## Fit method: kernelpls
## Number of components considered: 6
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
## CV           10.83    6.744    7.117    7.034    6.992    6.989
## adjCV        10.83    6.725    7.039    6.976    6.937    6.933
##
## TRAINING: % variance explained
##        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X        37.37    47.33    66.56    83.57    91.53   100.00
## abund    64.88    68.20    68.66    68.77    68.77    68.77
```

Example - PCLS method

- The smallest cross-validation error occurs when $M = 1$ component is used.
- Using $M = 1$ captures 37.37% of all the variance, or information, in the predictors.
  - 64.88% of variation of the response can be explained by the first PC which is a little better than PCR.

Lab

- SPSS does not support PCR directly, you may manually save the scores produced by the PCA and then regress the response Y on the scores using MLR models.

Lab

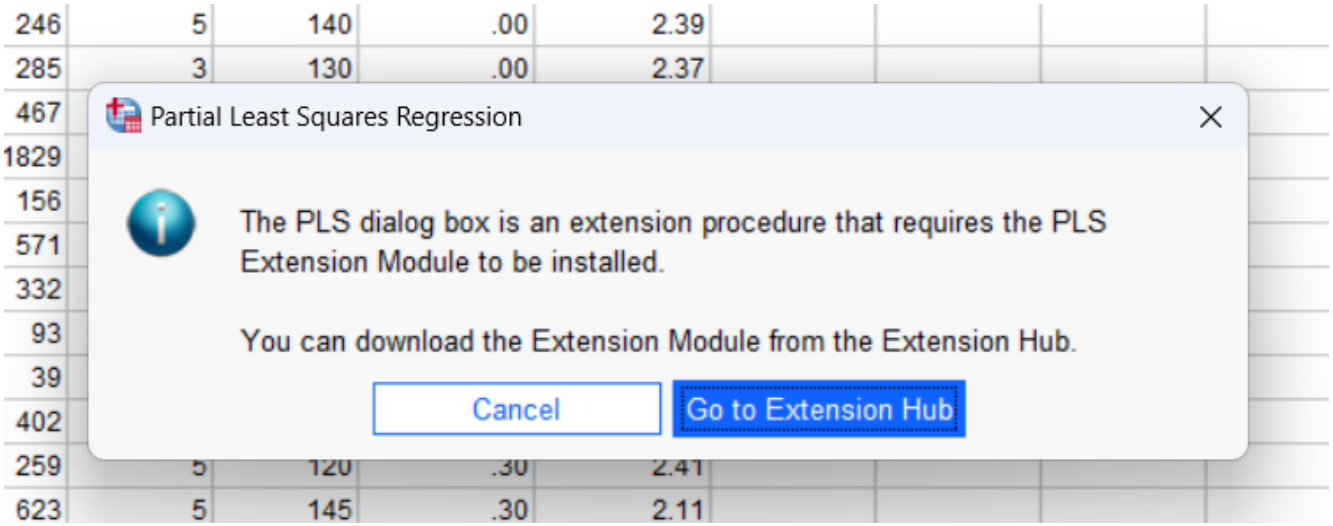
- R PLS code

```
library(dplyr)
loyn$logarea= log10(loyn$area)
loyn$logdist= log10(loyn$dist)
loyn=loyn%>%dplyr::select(-area, -dist)

library(pls)
set.seed(1)
pls.fit=plsr(abund~., data=loyn, scale=TRUE, validation="CV")
summary(pls.fit)
```

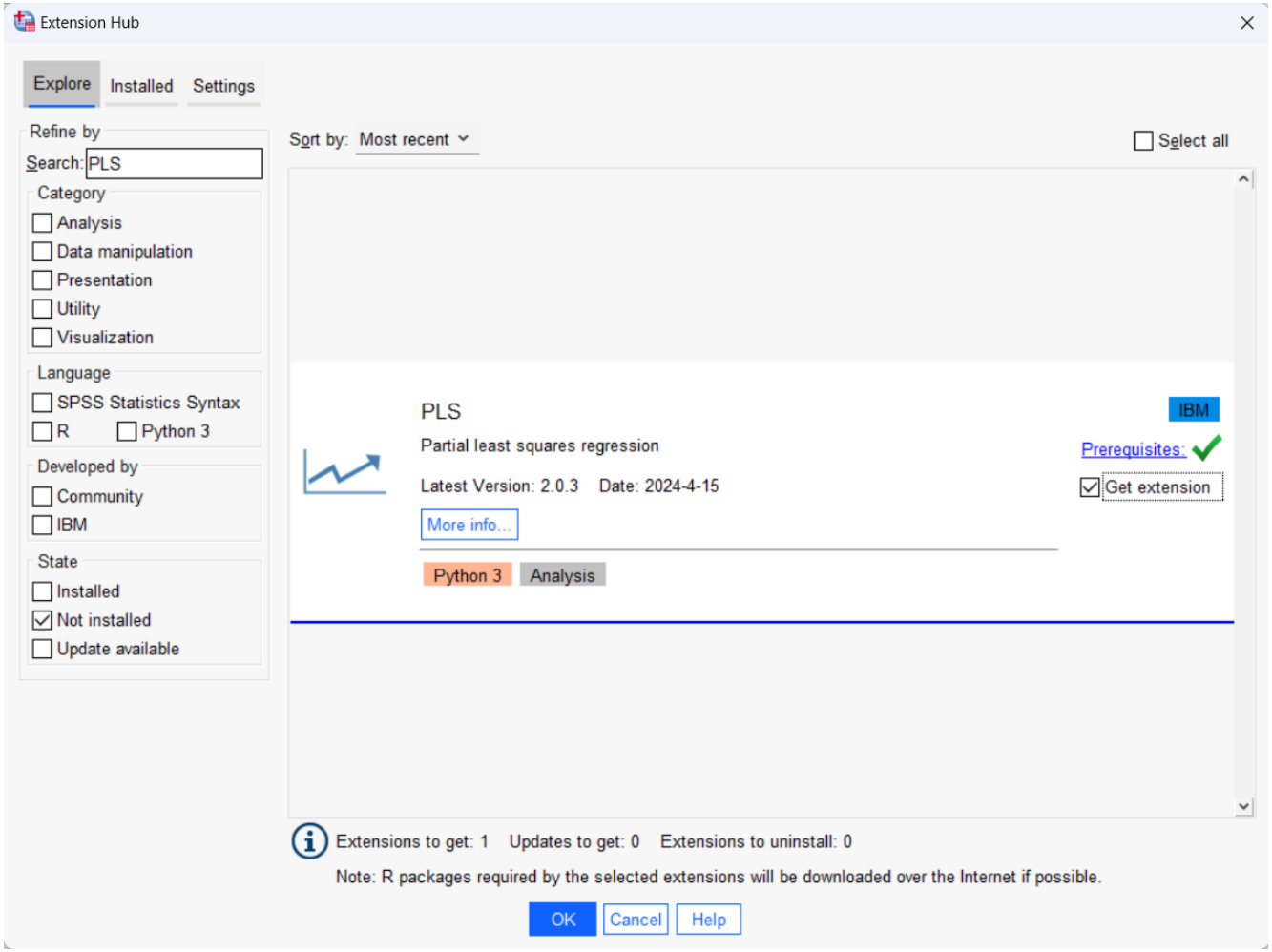
Lab

- Let's make log10 transformation of `area` and `dist` after importing the data; Also change the measure of `graze` from `Nominal` to `Scale`
  - Or you can create dummy variables for `graze`
- Click on `Analyze` → `Regression` → `Partial Least Squares...`. If it is your first time to run PLS in SPSS, you will see



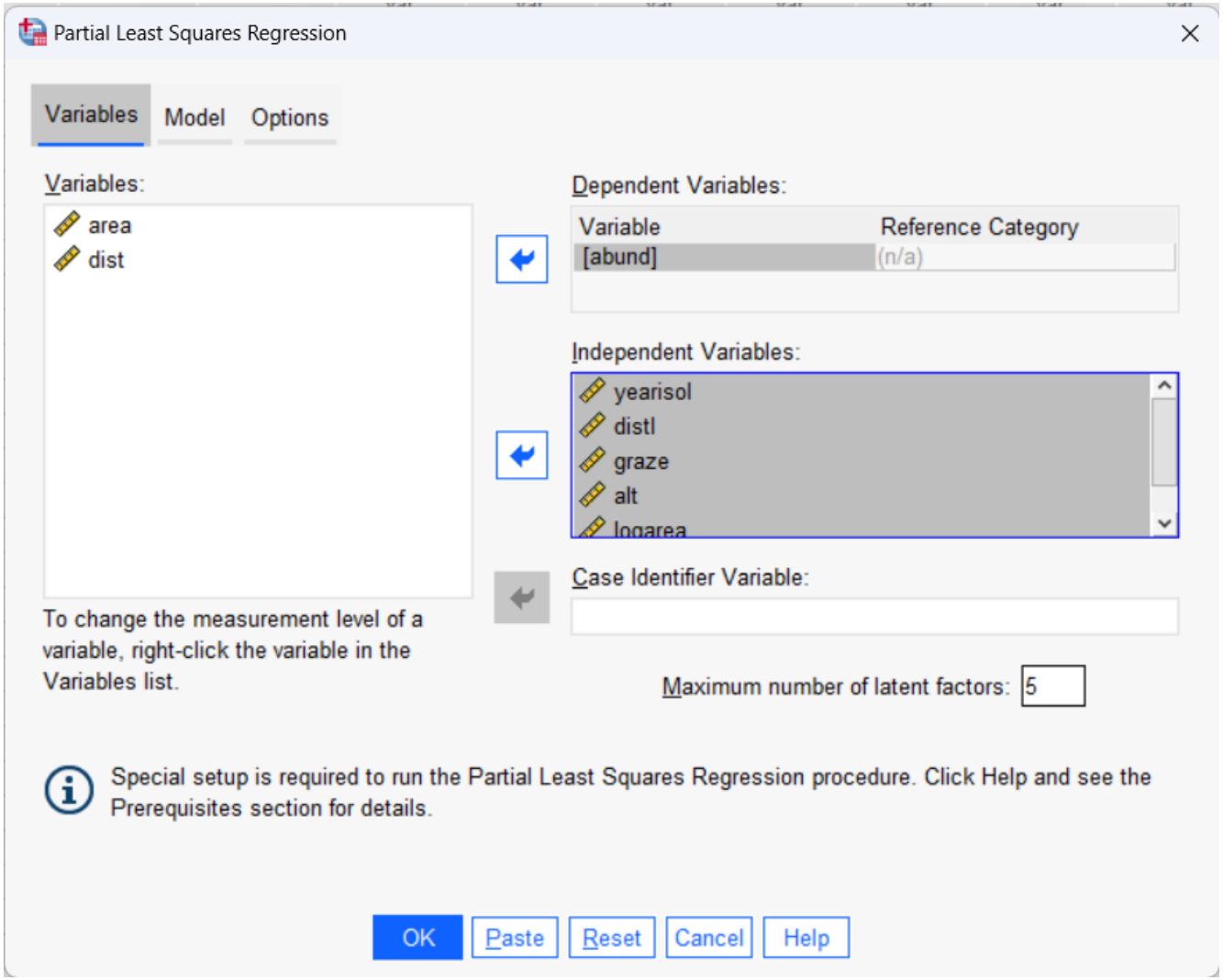
Lab

- Click `OK` to install it



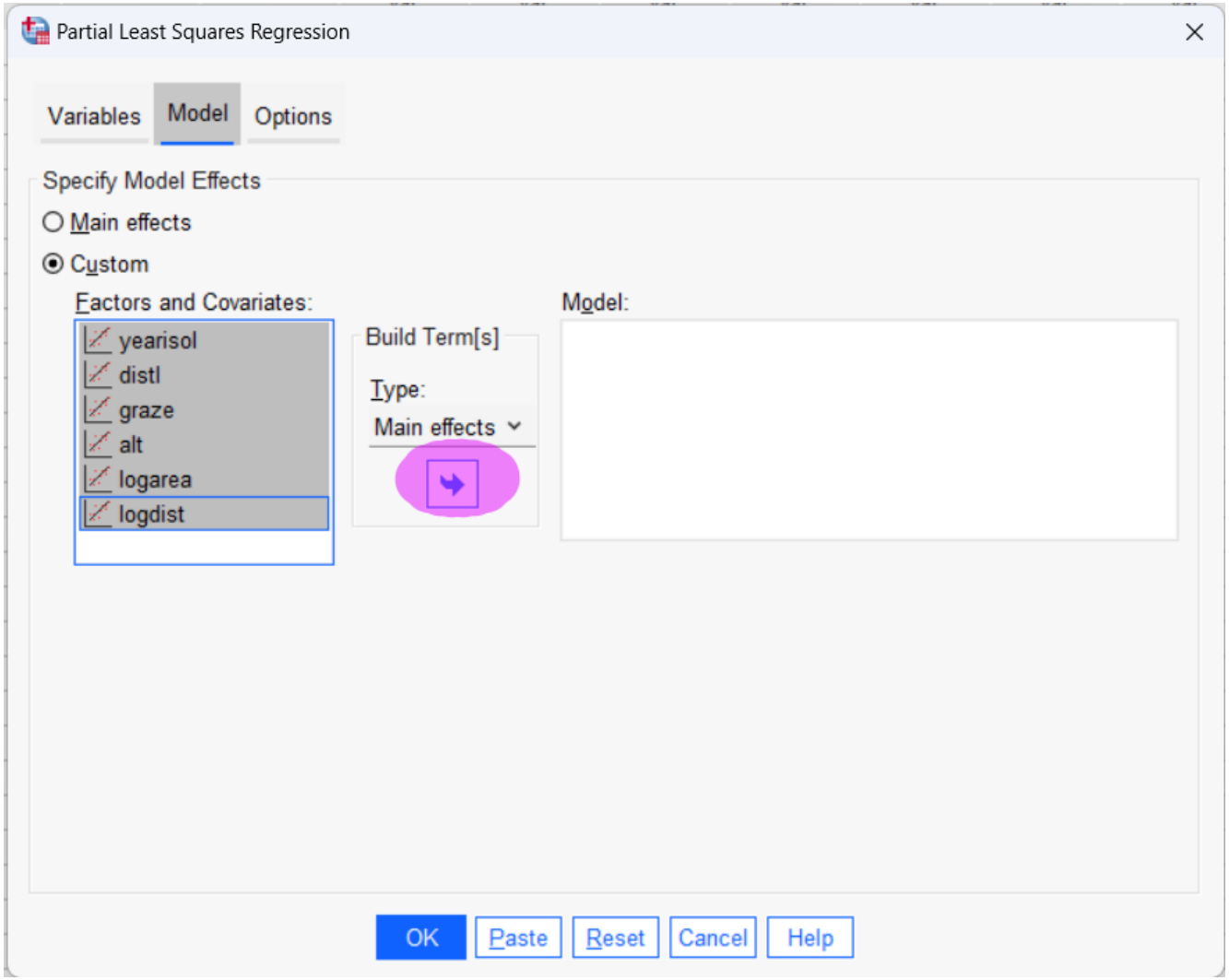
Lab

- Once the extension is installed, on `Analyze` → `Regression` → `Partial Least Squares...`; choose dependent and independent variables



Lab

- Build the model; Click `OK` run the analysis.



License