Linear Statistical Modeling Methods with SAS

Linear Model Regularization - Part II

Xuemao Zhang
East Stroudsburg University

April 10, 2024

Outline

- Principal Components Regression
  - Principal Components Analysis
  - PCR
- Partial Least Squares
- Example

Principal Components Analysis

- The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set.
- This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

- Suppose that $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ is a vector of p random variables, and that the variances of the p random variables and the structure of the covariances/correlations between the p variables are of interest.

Principal Components Analysis

- Instead of looking at the p variances and $\frac{1}{2}p(p-1)$ covariances, we look for a few ($< p$) derived variables that preserve most of the information given by these variances and correlations/covariances.
  - PCA concentrates on variances.
- The first step is to look for a linear combination of the elements of $\mathbf{x}$ having maximum variance

$$\boldsymbol{\alpha}_1^{'}\mathbf{x} = \alpha_{11}x_1 + \cdots + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j.$$

- Next, look for a linear combination $\boldsymbol{\alpha}_2^{'}\mathbf{x}$, uncorrelated with $\boldsymbol{\alpha}_1^{'}\mathbf{x}$ having maximum variance, and so on. ... Up to p PCs could be found.
  - It is hoped, in general, that most of the variation in $\mathbf{x}$ will be accounted for by m PCs, where $m << p$.

Principal Components Analysis

- Derivation of the form of the PCs: note that

$$var(\boldsymbol{\alpha}_1^{'}\mathbf{x}) = \boldsymbol{\alpha}_1^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1,$$

where $\boldsymbol{\Sigma} = var(\mathbf{x})$. So the maximization must be subject to a normalization constraint

$$\boldsymbol{\alpha}_1^{'}\boldsymbol{\alpha}_1 = \sum_{j=1}^{p} \alpha_{1j}^2 = 1.$$

Using the technique of Lagrange multipliers (Calculus III, λ_1 is called the Lagrange Multiplier), We maximize the function

$$\boldsymbol{\alpha}_1^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda_1(\boldsymbol{\alpha}_1^{'}\boldsymbol{\alpha}_1 - 1)$$

w.r.t. $\boldsymbol{\alpha}_1$ by differentiating w.r.t. to $\boldsymbol{\alpha}_1$.

Principal Components Analysis

This results in

$$\frac{\partial}{\partial \boldsymbol{\alpha}_1}[\boldsymbol{\alpha}_1^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda_1(\boldsymbol{\alpha}_1^{'}\boldsymbol{\alpha}_1 - 1)] = 0$$

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - \lambda_1\boldsymbol{\alpha}_1 = 0$$

and thus

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1.$$

- This should be recognizable as an eigenvector equation where $\boldsymbol{\alpha}_1$ is an eigenvector of $\boldsymbol{\Sigma}$ and λ_1 is the associated eigenvalue.

Principal Components Analysis

- Which eigenvector should we choose?

$$\boldsymbol{\alpha}_1^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1^{'}\lambda_1\boldsymbol{\alpha}_1 = \lambda_1$$

- Then we should choose λ_1 to be as big as possible. So λ_1 is the largest eigenvector of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_1$ is the corresponding eigenvector.
- Then the solution to

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1$$

is the 1st PC(principal component) of $\mathbf{x}$.

Principal Components Analysis

- The second PC, $\boldsymbol{\alpha}_2^{'}\mathbf{x}$ maximizes $\boldsymbol{\alpha}_2^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2$ subject to

$$\boldsymbol{\alpha}_2^{'}\boldsymbol{\alpha}_2 = 1$$

and being uncorrelated with $\boldsymbol{\alpha}_1^{'}\mathbf{x}$:

$$cov(\boldsymbol{\alpha}_1^{'}\mathbf{x}, \boldsymbol{\alpha}_2^{'}\mathbf{x}) = \boldsymbol{\alpha}_1^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_2^{'}\boldsymbol{\alpha}_1 = \lambda_1\boldsymbol{\alpha}_1^{'}\boldsymbol{\alpha}_2 = 0.$$

- Using the technique of Lagrange multipliers with these two constraints, we maximize the function w.r.t $\boldsymbol{\alpha}_2$

$$\boldsymbol{\alpha}_2^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2(\boldsymbol{\alpha}_2^{'}\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2^{'}\boldsymbol{\alpha}_1$$

Principal Components Analysis

- Differentiation of this quantity w.r.t. $\boldsymbol{\alpha}_2$ (and setting the result equal to zero) yields

$$\frac{\partial}{\partial \boldsymbol{\alpha}_2}[\boldsymbol{\alpha}_2^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2(\boldsymbol{\alpha}_2^{'}\boldsymbol{\alpha}_2 - 1) - \phi\boldsymbol{\alpha}_2^{'}\boldsymbol{\alpha}_1] = 0$$

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1 = 0$$

- If we left multiply $\boldsymbol{\alpha}_1$ into this expression

$$\boldsymbol{\alpha}_1^{'}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_1^{'}\boldsymbol{\alpha}_2 - \phi\boldsymbol{\alpha}_1^{'}\boldsymbol{\alpha}_1 = 0$$

$$0 - 0 - \phi = 0$$

then we can see that ϕ must be zero.

Principal Components Analysis

- So we have

$$\mathbf{\Sigma}\boldsymbol{\alpha}_2 = \lambda_2\boldsymbol{\alpha}_2.$$

Furthermore,

$$var(\boldsymbol{\alpha}_2^{'}\mathbf{x}) = \boldsymbol{\alpha}_2^{'}\lambda_2\boldsymbol{\alpha}_2 = \lambda_2\boldsymbol{\alpha}_2^{'}\boldsymbol{\alpha}_2 = \lambda_2.$$

Principal Components Analysis

- Thus, the second PC $\boldsymbol{\alpha}_2'\mathbf{x}$ is the solution to

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \lambda_2\boldsymbol{\alpha}_2,$$

  where λ_2 is the second largest eigenvalue of $\boldsymbol{\Sigma}$.
- This process can be repeated for $k = 1, 2, \ldots, p$ yielding up to p different eigenvectors of $\boldsymbol{\Sigma}$ along with the corresponding eigenvalues $\lambda_1, \ldots, \lambda_p$.

Principal Components Regression

- Given data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ with each $\mathbf{x}_i, i = 1, \ldots, p$ a column vector, we can estimate $\mathbf{\Sigma}$ by the sample covariance matrix

$$\mathbf{S} = \mathbf{X}^{'}\mathbf{X}/(n-1).$$

- PCA produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- Apart from producing derived variables for use in supervised learning problems, PCA also serves as a tool for data visualization.

Principal Components Regression

- Our data consist of n observations in a p-dimensional space.
- However, not all of those p dimensions are equally useful, especially when $p >> n$.
- Many are either completely redundant (correlated features) or uninformative (noise features).
- Can we find a low-dimensional representation of the variables that captures most of the variability in the data?
- We now explore a class of approaches that transform the predictors and then fit a least squares model using the transformed variables.
- This is a dimension reduction approach.

Principal Components Regression

- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ linear combinations of our original p predictors. That is,

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

- Use least squares to fit the model

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

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

Principal Components 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

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

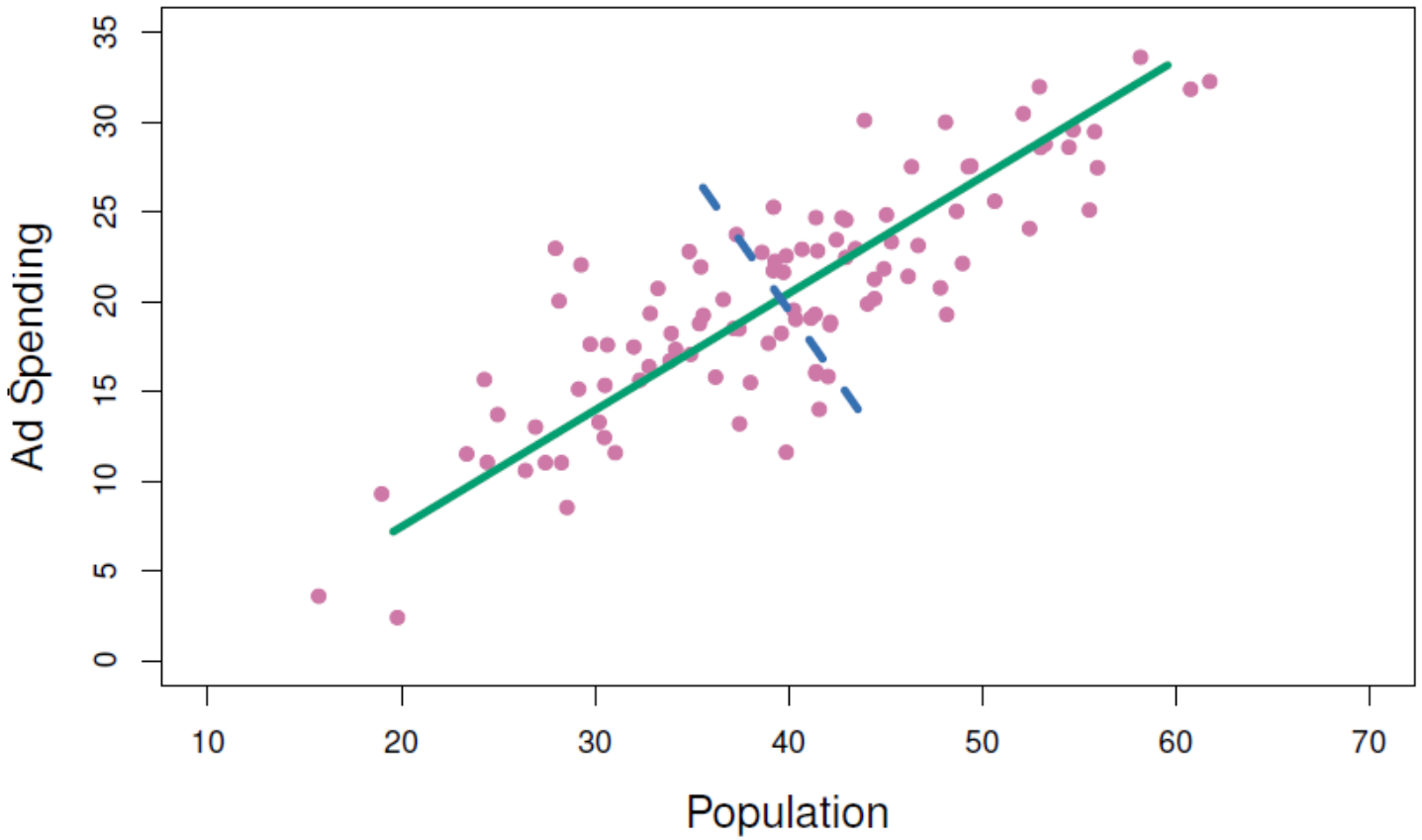
- Hence model (2) can be thought of as a special case of the original linear regression model.
- Dimension reduction serves to constrain the estimated β_j coefficients, since now they must take the form (3).
- Can win in the bias-variance tradeoff.

Principal Components Regression

- Here we apply principal components 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.

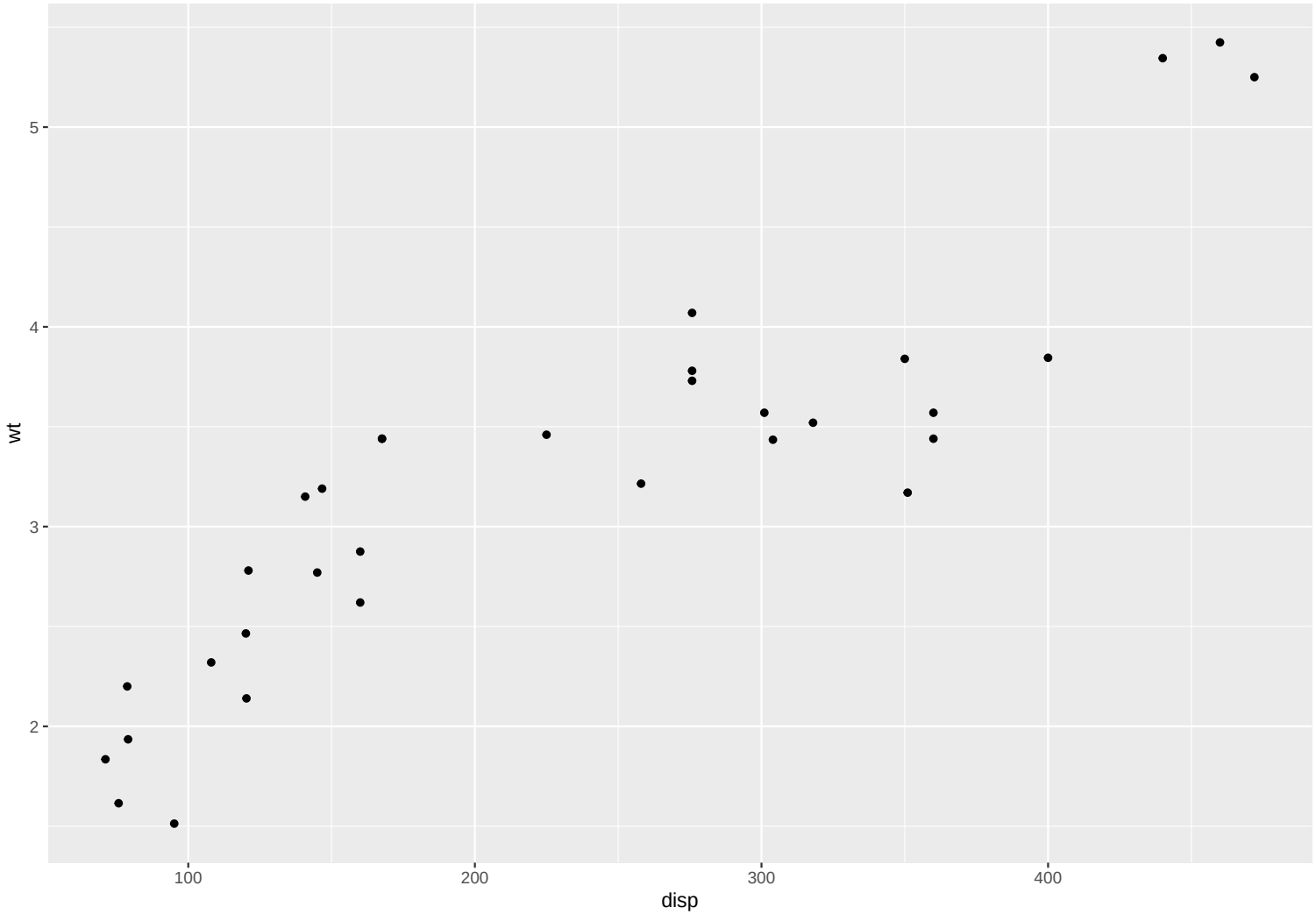
Principal Components Regression

- The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.

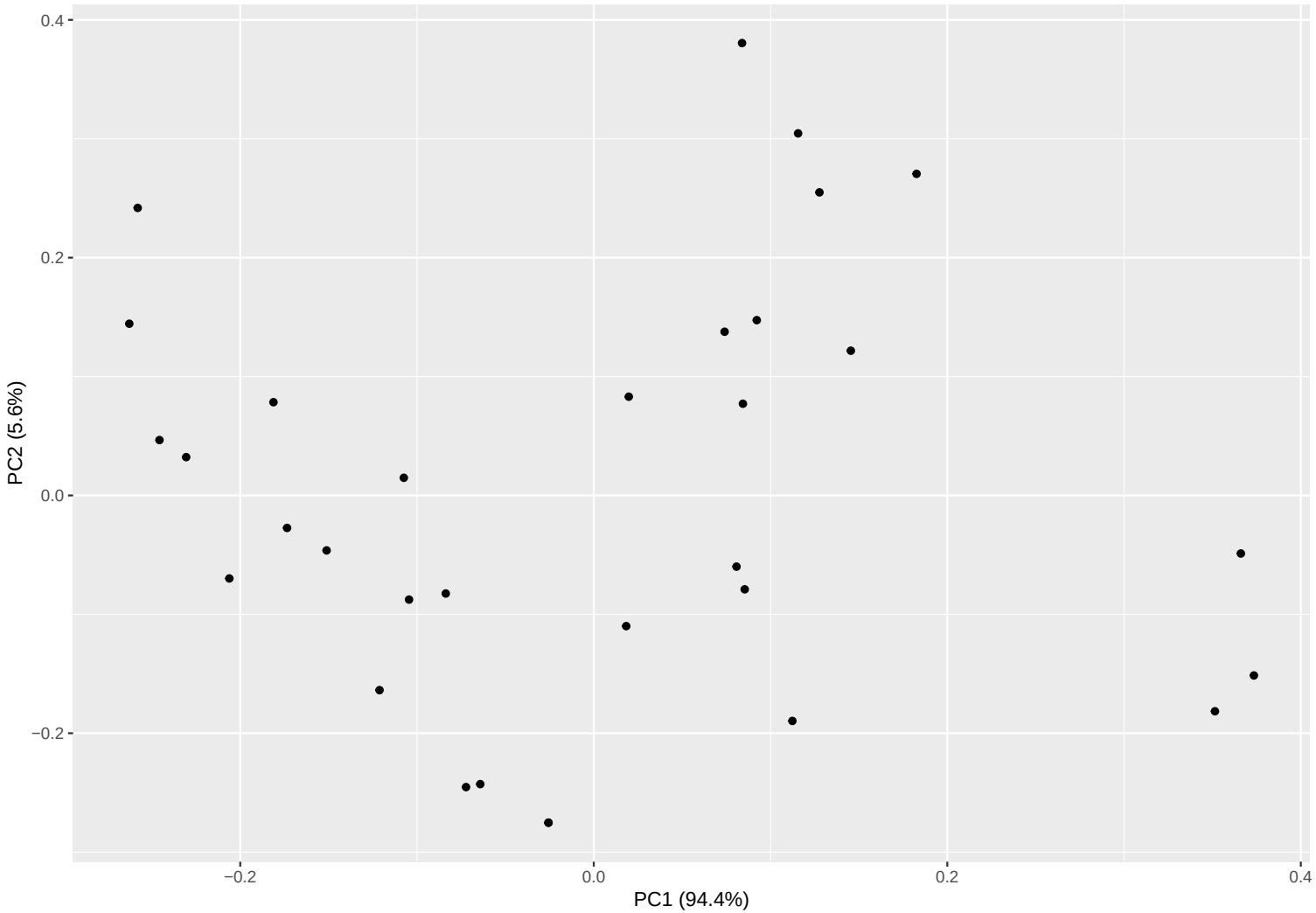


Principal Components Regression

- Consider the `mtcars` data.



Principal Components Regression



Principal Components Regression

- PCs are the linear combinations of the variables that contain as much as possible of the variability in the features.
- **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.
- Choose M by cross-validation or validation set approach.
  - Also consider the percentage of variance in X or Y by the predictors
- With $M = p$, we just get least squares regression: no dimension reduction occurs!

Principal Components Regression

- PCR directions are identified in an unsupervised way, since the response Y is not used to help determine the principal component directions.
- That is, the response does not supervise the identification of the principal components.
- 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**.

Principal Components Regression

- In SAS, the The **PLS** Procedure is used to fit Principal Components Regression.
- Read The PLS Procedure https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_pls_syntax01.htm

A simulation study

```
proc iml;
call randseed(1); /* Set seed for reproducibility */

/* Generate 100 by 100 predictor matrix xtr with random normal values */
xtr = j(100, 100);
call randgen(xtr, "Normal");

/* True coefficients for simulation study */
beta = j(100, 1, 0); /* 100 by 1 matrix of zeros */
beta[1:10, ] = 1;

/* Generate response variable ytr as a linear combination of xtr with added noise */
noise = randfun(100, "Normal");
ytr = xtr * beta + noise;

/* Combine ytr with xtr into a single matrix */
DataMatrix = ytr || xtr;

/* Display combined data */
create DataGen from DataMatrix[colname={"Response"}];
append from DataMatrix;
close DataGen;

quit;
```

A simulation study

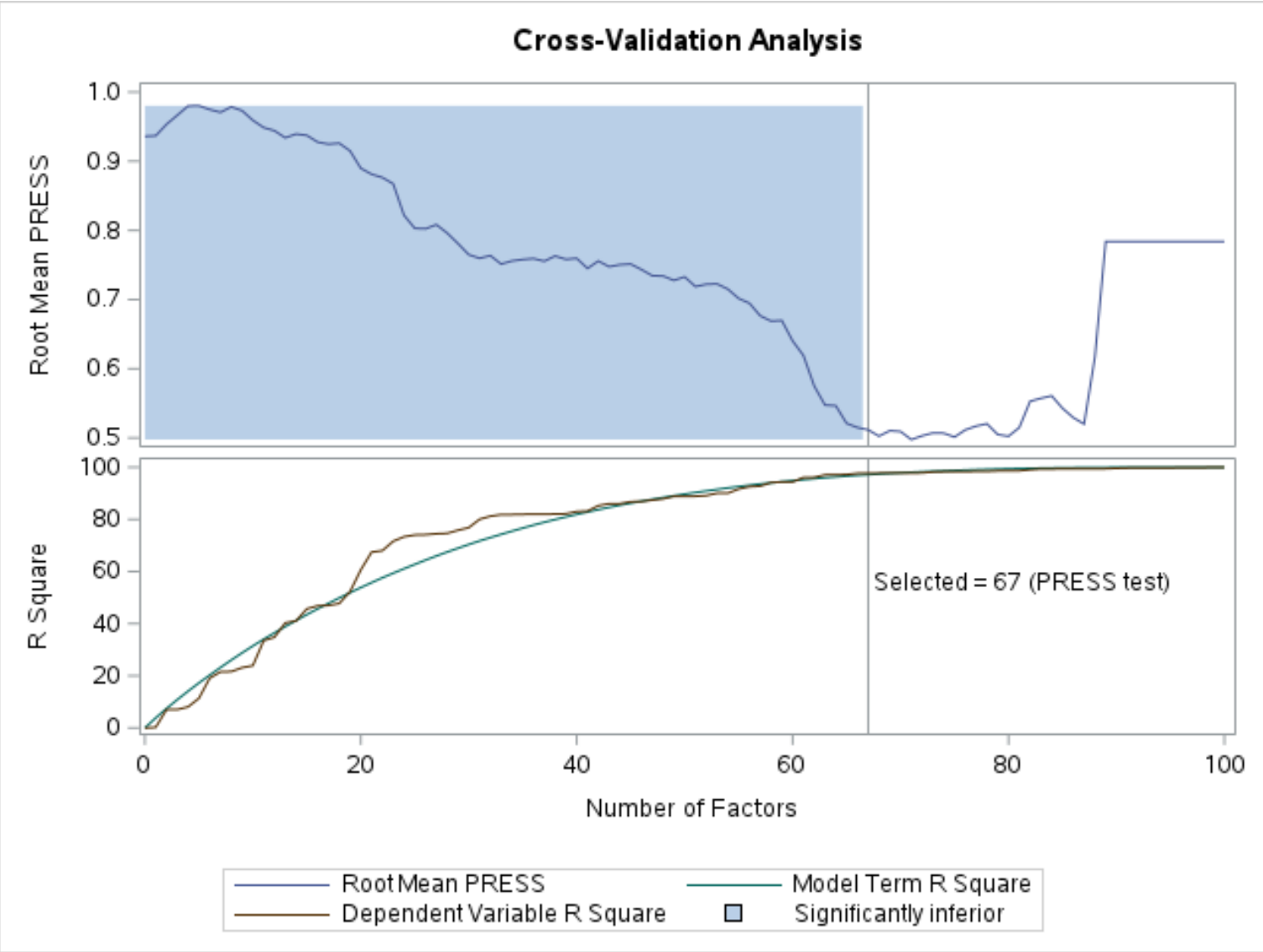
```
proc print data=DataGen;
run;
```

A simulation study

- `nfac=100`: all 100 possible factors
- `cv=random`: k-fold Cross-validation
- `cvtest(stat=press)`: specifies the test statistic for the model comparison.

```
proc pls data=DataGen method=pcr nfac=100 cv=random(seed=123)
cvtest(stat=press) plots=cvplot;
  model Response = COL2 - COL101;
run;
```

A simulation study



Partial Least Squares

- PCR identifies linear combinations, or directions, that best represent the predictors $X_1, \ldots, X_p$.
- These directions are identified in an **unsupervised way**, since the response Y is not used to help determine the principal component directions.
- That is, the response does not supervise the identification of the principal components.
- 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, and then fits a linear model via OLS using these M new features.
- 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.

Details of 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**.
- 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 Hitters data again

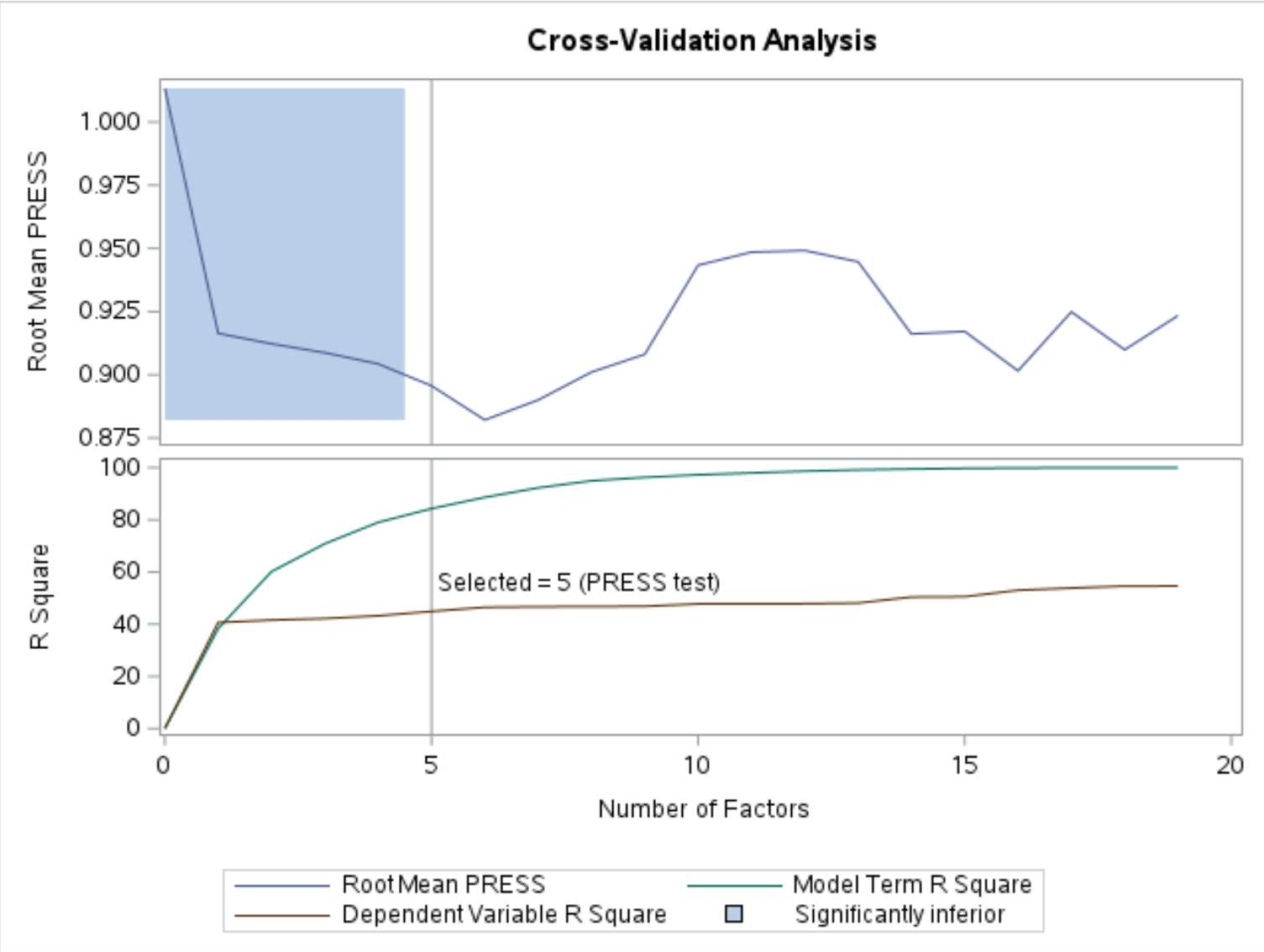
```
PROC IMPORT
DATAFILE='/home/u5235839/my_shared_file_links/u5235839/Hitters.csv'
    DBMS=CSV
    OUT=Hitters;
    GETNAMES=YES;
RUN;

data Hitters1;
set Hitters;
/* Create binary indicators */
Division_1 = (Division = 'E');
League_A= (League= 'A');
NewLeague_A= (NewLeague='A');
run;
```

Example

- PCR

```
proc pls data=Hitters1 method=pcr nfac=19 cv=random(seed=123)
cvtest(stat=press) plots=cvplot;
model Salary = AtBat    Hits    HmRun   Runs    RBI Walks
Years   CAtBat  CHits   CHmRun  CRuns   CRBI    CWalks  League_A
Division_1  PutOuts Assists Errors  NewLeague_A;
run;
```

Example

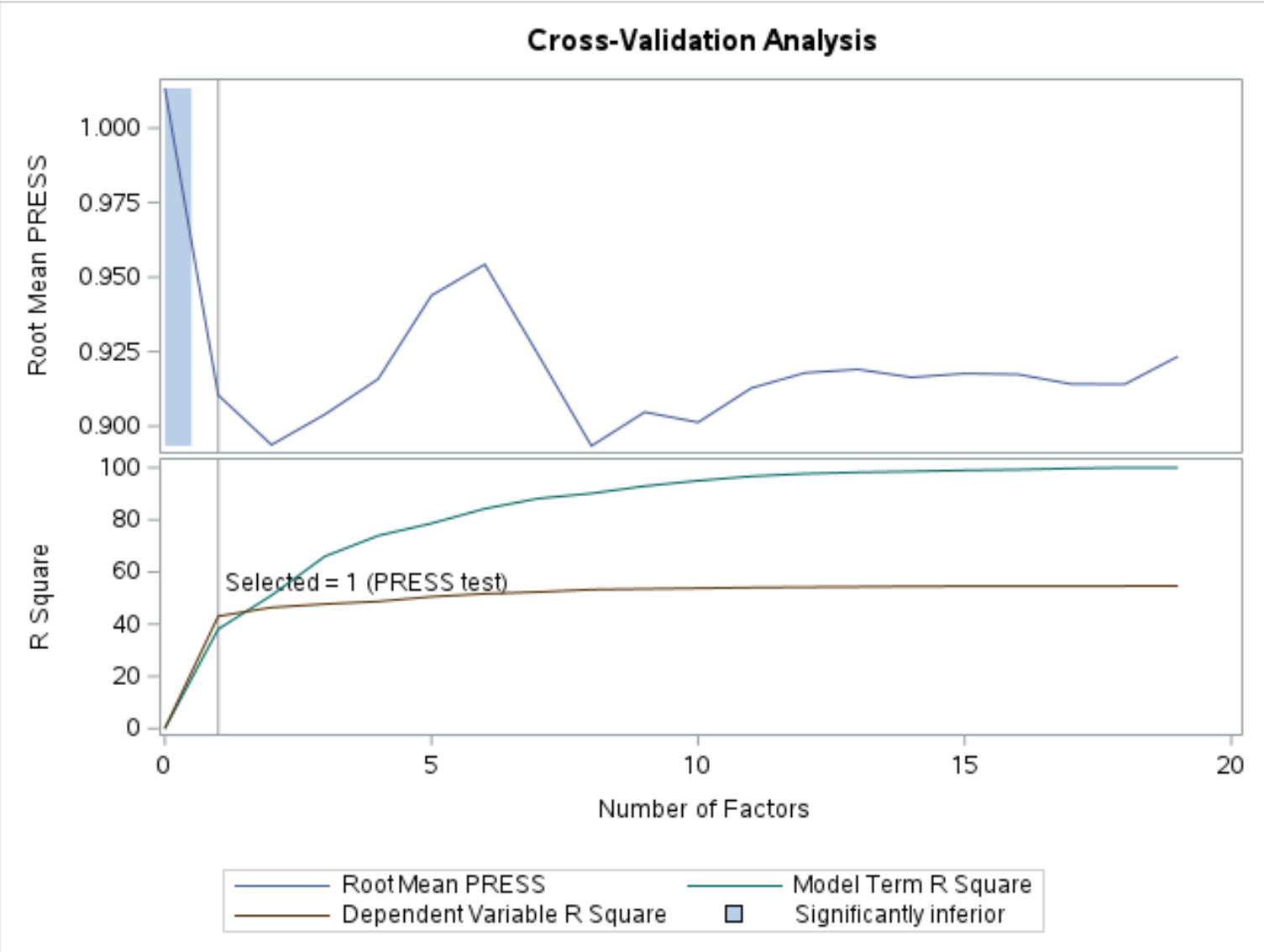

Cross-Validation Analysis
Root Mean PRESS
1.000
0.975
0.950
0.925
0.900
0.875
R Square
100
80
60
40
20
0
Selected = 5 (PRESS test)
0
5
10
15
20
Number of Factors
Root Mean PRESS
Dependent Variable R Square
Model Term R Square
Significantly inferior

Example

- PLS

```
proc pls data=Hitters1 method=pls nfac=19 cv=random(seed=123)
cvtest(stat=press) plots=cvplot;
model Salary = AtBat    Hits    HmRun   Runs    RBI Walks
Years   CAtBat  CHits   CHmRun  CRuns   CRBI    CWalks  League_A
Division_1  PutOuts Assists Errors  NewLeague_A;
run;
```

Example


Cross-Validation Analysis
Root Mean PRESS
1.000
0.975
0.950
0.925
0.900
R Square
100
80
60
40
20
0
Selected = 1 (PRESS test)
0
5
10
15
20
Number of Factors
Root Mean PRESS
Model Term R Square
Dependent Variable R Square
Significantly inferior

License