Regression Analysis

Chapter 10: Working With Collinear Data

Chapter 11: Variable Selection Procedures

Kyusang Yu and Sunghwan Kim

Department of Applied Statistics

Konkuk University

Spring 2019

Introduction

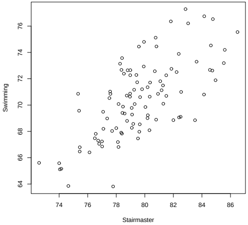
- When collinearity is present in a set of predictors, the OLS estimates tend to unstable and can lead to erroneous inferences.
- This chapter presents ways for dealing with collinearity.
  - Imposing or searching for constraints on the regression parameter
  - Principal components regression
  - Ridge regression
- Principal components regression will be covered by Multivariate data analysis.

Principal Component Analysis

- ***Principal component analysis (PCA)*** allows us to reorient the data so that the first few dimensions account for as much of the available information as possible.
- The researcher must decide how many principal components to retain for subsequent analysis, trading off simplicity against completeness.
- Each principal component is uncorrelated with all others, which has the advantage of eliminating multicollinearity when using the results in an analysis of dependence.

PCA: Motivating Example

- The scores of 100 students who were tested in stair master (X_1) and swimming (X_2):



- Correlation matrix for X_1 and X_2

$$\mathbf{R} = Corr(\mathbf{X}) = \begin{pmatrix} 1.000 & 0.693 \\ 0.693 & 1.000 \end{pmatrix}$$

PCA: Motivating Example

- Does this seem to support the idea that a single dimension (e.g., ***fitness*** component) can capture and convey most of the information contained the variables X_1 and X_2?
- In other words, the first principal component is the linear combination of X_1 and X_2 that exhibits maximum variance.
- Recall that a linear combination is simply the projection of all points in the two-dimensional space on to a single axis.

An Outline of PCA

- PCA transforms a set of correlated variable (X's) into a set of uncorrelated components (Z's).
- The principal components are linear combinations of the X's which we write as:

$$\begin{aligned} Z_1 &= u_{11}X_1 + u_{21}X_2 \\ Z_2 &= u_{12}X_1 + u_{22}X_2, \end{aligned}$$

where $u_{11}^2 + u_{21}^2 = u_{12}^2 + u_{22}^2 = 1$, and $u_{11}u_{12} + u_{21}u_{22} = 0$.

An Outline of PCA

- An important consequence of the orthogonality condition is that the total variance of the Z's is equal to the total variance of the X's:

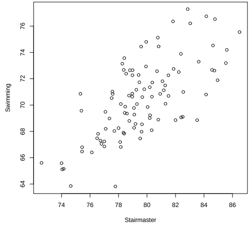
$$\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} Var(Z_i)$$

An Outline of PCA

- Finding a method of determining u_{ij}'s so that the components have the required properties is equivalent to finding eigenvalues and eigenvectors.
- There are standard algorithms which determine the weights u_{ij} and the variances of the principal components (i.e., eigenvalues).

PCA: Motivating Example

- The scores of 100 students who were tested in stair master (X_1) and swimming (X_2):



- Correlation and covariance matrix for X_1 and X_2

$$\mathbf{R} = Corr(\mathbf{X}) = \begin{pmatrix} 1.000 & 0.693 \\ 0.693 & 1.000 \end{pmatrix}, \quad \mathbf{\Sigma} = Var(\mathbf{X}) = \begin{pmatrix} 7.965 & 5.715 \\ 5.715 & 8.534 \end{pmatrix}$$

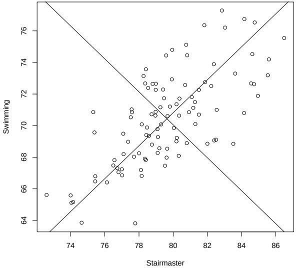
PCA: Motivating Example

- Let $\mathbf{u}_1 = (u_{11}, u_{12})$ denote a vector of unit length oriented along the longest axis of the ellipsoid so that $\mathbf{z}_1 = \mathbf{X}\mathbf{u}_1$, where $\mathbf{z}_1$ is an n-dimensional vector consisting of the values of Z_1.
- We can think of the variance of Z_1 as the variance accounted for by the first principal component.

PCA: Motivating Example

- We can now choose a second linear combination of the two variables to account for the remaining of the variance not accounted for by Z_1.
- Let $\mathbf{u}_2 = (u_{21}, u_{22})$ denote a vector of unit length oriented orthogonal to $\mathbf{u}_1$ so that $\mathbf{z}_2 = \mathbf{X}\mathbf{u}_2$.

PCA: Motivating Example



- The transformation $\mathbf{Z} = \mathbf{XU}$ has served simply to rotate the axes of the original scatter plot while preserving their orthogonality (due to the orthogonality of $\mathbf{u}_1$ and $\mathbf{u}_2$).

Example: PCA

- Correlation matrix:

$$\mathbf{R} = Corr(\mathbf{X}) = \begin{pmatrix} 1.00 & 0.69 \\ 0.69 & 1.00 \end{pmatrix}$$

$$= \begin{pmatrix} 0.71 & -0.71 \\ 0.71 & 0.71 \end{pmatrix} \begin{pmatrix} 1.69 & 0 \\ 0 & 0.31 \end{pmatrix} \begin{pmatrix} 0.71 & 0.71 \\ -0.71 & 0.71 \end{pmatrix}$$

- Covariance matrix:

$$\mathbf{\Sigma} = Var(\mathbf{X}) = \begin{pmatrix} 7.97 & 5.72 \\ 5.72 & 8.53 \end{pmatrix}$$

$$= \begin{pmatrix} 0.69 & 0.72 \\ 0.72 & -0.69 \end{pmatrix} \begin{pmatrix} 13.97 & 0 \\ 0 & 2.52 \end{pmatrix} \begin{pmatrix} 0.69 & 0.72 \\ 0.72 & -0.69 \end{pmatrix}$$

Uses of Regression Equations

1. Description and model building
    - Used to describe a given process or as a model for a complex system
    - Try to choose the smallest number of predictors that accounts for the most substantial part of the variation in the response (principle of parsimony)

Uses of Regression Equations

2. Estimation and Prediction
    - Predict the value of a future observation or mean response
    - Try to select a set of predictors to minimize MSE
3. Control
    - To determine the magnitude by which the value of a predictor variable must be altered to obtain a specified value of the response
    - Try to minimize the standard errors of the estimates of the coefficients

What is the "best set" of variables?

- There is no unique "best set" of variables.
- A good variable selection procedure should point out several "adequate" subsets of variables rather than generate a single "best" set.
- The process of variable selection should be viewed as an intensive analysis of the correlation structure of the predictors; and how they individually and jointly affect the response under study.

Criteria for evaluating equations

1. Residual mean square (RMS)
2. Mallows C_p
3. Information criteria
   - AIC
   - BIC
   - AIC^C

Residual mean square

- With the p-term equation (includes a constant and $p-1$ variables), the RMS is defined as

$$\mathrm{RMS}_p = \frac{SSE_p}{n-p}.$$

- The smaller RMS is preferred.

Mallows C_p

- The standardized total mean squared error of prediction for the observed data is measured by

$$J_p = \frac{1}{\sigma^2} \sum_{i=1}^{n} MSE(\hat{y}_i).$$

- To estimate J_p, Mallows (1973) uses the statistic

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + (2p - n).$$

- The deviation of C_p from p can be used as a measure of bias.

Information criteria

- AIC for a p-term equation is given by

$$AIC_p = n\ln(SSE_p/n) + 2p.$$

- BIC is defined as

$$BIC_p = n\ln(SSE_p/n) + p(\ln n).$$

- AIC^c is given by

$$AIC_p^c = AIC_p + \frac{2(p+2)(p+3)}{n-p-3}.$$

- The model with smaller value of ICs are preferred.

Multicollinearity and variable selection

- Two situations:
  1. No multicollinearity
  2. Multicollinearity (VIFs are greater than 10)
- Different approaches to variable selection procedures depending on these situations

Evaluating all possible equations

- Very direct and equally well to both collinear and noncollinear data
- Fit all possible subset equations (2^q equations).
- Pick out the three "best" (R^2, C_p, RMS_p) equations and analyze them to arrive at the final model.
- Practically infeasible when q is large.

Variable selection methods

- Greedy approaches
  - Forward selection procedure
  - Backward elimination procedure
  - Stepwise Method

Forward selection procedure

1. Pick a cutoff t_F, and start with an equation containing no predictor variables.
2. Include the variable X_1 having the highest simple correlation with Y.
   - If the absolute value of the t-value for X_1 is larger than t_F, it is retained; otherwise, stop the procedure.

Forward selection procedure

3. Include the second variable X_2.
   - Compute the residuals when Y is regressed on X_1.
   - X_2 is one which has the highest simple correlation with the residual.
   - If the absolute value of the t-value for X_2 is larger than t_F, it is retained; otherwise, stop the procedure.
4. The procedure is terminated when the last variable entering the equation is insignificant.

Backward elimination procedure

1. Pick a cutoff t_B, and start with the full equation having q predictors.
2. Find the most insignificant variable X_1 having the smallest absolute value of the t-value among the full set or having the minimum reduction of SSE.
   - If all the t-tests are significant (checked by t_B), stop; otherwise, drop X_1.
3. Fit the model with X_1 deleted.

Backward elimination procedure

4. Find the most insignificant variable X_2 among the remaining $(q-1)$ predictors.
   - If all the t-tests are significant (checked by t_B), stop; otherwise, drop X_2.
5. The procedure is terminated when all variables are significant (checked by t_B) or all variables are deleted.

Stepwise method

- A forward selection but with deletion at each stage
- A variable entered in earlier stage may be eliminated at later stage
- Different levels of cutoffs (typically, $t_F < t_B$)

General remarks

- An effective stopping rule
  - In FS: Stop if minimum t-test is less than 1
  - In BE: Stop if minimum t-test is greater than 1
- Recommend BE over FS
  - It depends on situations.
  - Stepwise method is recommended when q is large.
- Several equations are generated by selection procedures.
  - Final model is chosen by AIC or BIC.
  - Diagnostic should be carried out.

A possible strategy

- Examine the variables one at a time (e.g., transformation to induce symmetry and reduce skewness).
- Construct pairwise scatterplots.
- Fit the full linear regression model and delete insignificant variables.
  - Check residuals
- Examine if additional variables can be dropped.
  - AIC/BIC would be good criteria for examining non-nested models.
- For the final model, check VIF.
- Attempt should be made to validate the fitted model.