Statistical Methods

Model Selection in Regression

Dec 12, 2023

Overview

1. Variable selection
2. Toy Experiment
3. Stepwise regression
4. Principal component analysis
5. Shrinkage Method
6. Partial least squares (PLS) regression

Purpose

In reality, the true model is unknown. How to choose a good (or best) model? What does a good or best model mean?

- Precise prediction?
- Precise estimates of parameters (coefficients)?
- Related variables to the response?
- Causality of variables?

Some useful criteria in regression models are

- Prediction error
- Variance of parameter estimation
- AIC, p-value, ...
- ?

Related keywords

Related keywords:

- Variable selection
- Feature selection
- Best Model
- Model selection

Toy Experiment

Let the **true model** be

$$y_i = 10 + 0.5x_{1i} - 5x_{2i} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 0.7^2)$ and $i = 1, \ldots, 20$. Let the predictors be simulated from

$$x_{1i} \sim U(-2, 2),$$
$$x_{2i} \sim U(-1, 4).$$

We do have other variables:

$$x_{3i} = 1 + 0.8x_{1i} + e_i,$$
$$x_{4i} = 2 + 0.2x_{1i} + e_i$$
$$x_{5i} = -0.5x_{1i} + e_i,$$
$$x_{6i} = 2 + e_i$$

where $e_i \sim N(0, 0.5^2)$.

Toy Experiment

Let the observations be simulated from the true model, and then analyze it. We obtain:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.2975     0.2241  45.942  < 2e-16 ***
x1            0.5258     0.1143   4.599 0.000256 ***
x2           -5.1157     0.1004 -50.963  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5121 on 17 degrees of freedom
Multiple R-squared:  0.9935,	Adjusted R-squared:  0.9928 
F-statistic:  1304 on 2 and 17 DF,  p-value: < 2.2e-16
```

Toy Experiment

What if we put all of the variables into models?
Fit the model as

$$y = \beta_0 + \sum_{k=1}^{6} \beta_k x_k + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.33111    0.81342  12.701 1.05e-08 ***
x1           0.09593    0.24773   0.387   0.7049    
x2          -5.18012    0.11738 -44.133 1.51e-15 ***
x3           0.24966    0.20707   1.206   0.2494    
x4          -0.17613    0.27416  -0.642   0.5318    
x5          -0.46360    0.22724  -2.040   0.0622 .  
x6           0.08146    0.28703   0.284   0.7810    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4914 on 13 degrees of freedom
Multiple R-squared:  0.9954,	Adjusted R-squared:  0.9933 
F-statistic: 472.9 on 6 and 13 DF,  p-value: 1.922e-14
```

Toy Experiment

What if we fit

$$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8492     0.2519  39.102  < 2e-16 ***
x2           -5.1210     0.1128 -45.398  < 2e-16 ***
x3            0.4301     0.1169   3.678  0.00187 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5725 on 17 degrees of freedom
Multiple R-squared:  0.9919,    Adjusted R-squared:  0.991
F-statistic:  1042 on 2 and 17 DF,  p-value: < 2.2e-16
```

Toy Experiment

Why? Correlation of variables.

```
> cor(X)
           [,1]       [,2]       [,3]       [,4]        [,5]        [,6]
[1,]  1.0000000  0.1535453  0.8166749  0.6180677 -0.76215110 -0.15417098
[2,]  0.1535453  1.0000000  0.1834339  0.3592314 -0.36670917 -0.32736918
[3,]  0.8166749  0.1834339  1.0000000  0.7511628 -0.58195266 -0.18068023
[4,]  0.6180677  0.3592314  0.7511628  1.0000000 -0.50831131 -0.24291017
[5,] -0.7621511 -0.3667092 -0.5819527 -0.5083113  1.00000000  0.08623599
[6,] -0.1541710 -0.3273692 -0.1806802 -0.2429102  0.08623599  1.00000000
```

Detecting multicollinearity

Use variance inflation factors (VIFs) to examine the possible multicollinearity.

$$\mathrm{VIF}(\hat{\beta}_k) = \frac{1}{1 - R_k^2},$$

where R_k^2 is the R^2 from the regression model

$$x_k = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1} + \alpha_{k+1} x_{k+1} + \cdots + \epsilon.$$

General rule of thumb:

- $\mathrm{VIF} > 4$: further investigation.
- $\mathrm{VIF} > 10$: serious multicollinearity requiring correction.

Detecting multicollinearity

```
> vif(fit2)
      x1       x2       x3       x4       x5       x6 
5.220868 1.520642 4.403824 2.613262 2.998327 1.186896 
```

```
> fit3 <- lm(y~x1+x2+x4+x5+x6)
> vif(fit3)
      x1       x2       x4       x5       x6 
3.325623 1.518739 1.851395 2.933347 1.186818 
> fit4 <- lm(y~x1+x2+x4+x6)
> vif(fit4)
      x1       x2       x4       x6 
1.633968 1.247091 1.846439 1.144239 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.92127    0.86778  11.433 8.35e-09 ***
x1           0.54143    0.15181   3.566  0.00281 ** 
x2          -5.09171    0.11644 -43.728  < 2e-16 ***
x4          -0.02013    0.25244  -0.080  0.93749    
x6           0.18664    0.30872   0.605  0.55450    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5383 on 15 degrees of freedom
Multiple R-squared:  0.9937,	Adjusted R-squared:  0.992 
F-statistic: 590.1 on 4 and 15 DF,  p-value: 2.689e-16
```

Stepwise regression

Strategies:

- Forward selection
- Backward selection
- Both selection

Criterion:

- Akaike An Information Criterion (AIC):

$$-2 \log \text{likelihood} + 2df,$$

$$n \log \frac{\text{RSS}}{n} + 2df$$

  where df is the number of parameters in the model.
- F-test (or p-value)

Alternative method

- Dimension reduction on X: by the **principal component method (PCA)**
- PCA is a statistical technique that uses an **orthogonal transformation** to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables (called principal components).
- Idea and concept: the transformation is defined in such a way that **the first principal component has the largest possible variance**.
- Note that PCA is sensitive to the relative scaling of the original variables.

Some background of PCA

The random vector $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix $\boldsymbol{\Sigma}$. Consider a linear combinations

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p = \boldsymbol{a_1'X}$$

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p = \boldsymbol{a_2'X}$$

$$\vdots$$

$$Z_p = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p = \boldsymbol{a_p'X}$$

Then, we obtain

$$\mathrm{Var}(Z_i) = \boldsymbol{a_i'\Sigma a_i} \quad i = 1, \ldots, p$$

$$\mathrm{Cov}(Z_i, Z_k) = \boldsymbol{a_1'\Sigma a_k} \quad i, k = 1, \ldots, p$$

PCA: Find PCs

The principal components are those **uncorrelated linear combinations** $Z_1, \ldots, Z_p$ whose variance are **as large as possible**.

- The first principal component (PC) is the linear combination with **maximum variance**.
- First PC = linear combination $\boldsymbol{a}_1'\boldsymbol{X}$ that maximizes $\mathrm{Var}(\boldsymbol{a}_1'\boldsymbol{X})$ subject to $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$.
- Second PC = linear combination $\boldsymbol{a}_2'\boldsymbol{X}$ that maximizes $\mathrm{Var}(\boldsymbol{a}_2'\boldsymbol{X})$ subject to $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$ and

$$\mathrm{Cov}(Z_1, Z_2) = \mathrm{Cov}(\boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{a}_2'\boldsymbol{X}) = 0.$$

- The third PC to the p^{th} PC are the same as previous step.

Result of PCA

Let the pairs of eigenvalues and eigenvector of $\boldsymbol{\Sigma}$ be $(\lambda_1, \boldsymbol{e_1})$, ..., $(\lambda_p, \boldsymbol{e_p})$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Then, the i^{th} PC is

$$Z_i = \boldsymbol{e_i'}\boldsymbol{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p,$$

$$\text{Var}(Z_i) = \boldsymbol{e_i'}\boldsymbol{\Sigma}\boldsymbol{e_i} = \lambda_i \quad i = 1, \ldots, p$$

$$\text{Cov}(Z_i, Z_k) = \boldsymbol{e_1'}\boldsymbol{\Sigma}\boldsymbol{e_k} = 0 \quad i \neq k.$$

Keywords: eigenvalues and eigenvector of $\boldsymbol{\Sigma}$.

Procedure of analyzing the result

1. Find the eigenvalues and eigenvector of $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$.
2. Choose the first few large eigenvaules and the corresponding eigenvectors to be the coefficients of the linear combination.
3. The rule of thumb is to choose the PCs with $\lambda_i > 0.7$ or use a scree plot of λ_i's.
4. Interpret the PC loadings (coefficients) in each PC.
5. Finally, use the PC scores which are Z_i's to complete the statistical analysis.

PCA on X of the toy experiment

```
Loadings:
     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
[1,]  0.542  0.265  0.297  0.430  0.150  0.584
[2,]  0.309 -0.920                0.210       
[3,]  0.594  0.270 -0.514 -0.114  0.351 -0.418
[4,]  0.304        -0.356 -0.361 -0.724  0.355
[5,] -0.403        -0.701  0.270  0.262  0.452
[6,]                0.161 -0.774  0.466  0.380
```

```
> ### PCA on X
> pca <- princomp(X)
> summary(pca)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.6901414 1.1141005 0.63101485 0.40024536 0.37103159 0.31737610
Proportion of Variance 0.5836223 0.2535914 0.08135139 0.03272943 0.02812597 0.02057947
Cumulative Proportion  0.5836223 0.8372137 0.91856512 0.95129455 0.97942053 1.00000000
```

Regression on PCs of X

```
> fit.by.pca1 <- lm(y~z1+z2)
> summary(fit.by.pca1)

Call:
lm(formula = y ~ z1 + z2)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.22320 -0.31090 -0.03002  0.37515  1.70838 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.5948     0.1727   3.444   0.0031 ** 
z1           -1.2739     0.1022 -12.465 5.61e-10 ***
z2            4.8563     0.1550  31.322  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7725 on 17 degrees of freedom
Multiple R-squared:  0.9853,	Adjusted R-squared:  0.9835 
F-statistic: 568.2 on 2 and 17 DF,  p-value: 2.703e-16
```

```
> fit.by.pca2 <- lm(y~z1+z2+z3)
> summary(fit.by.pca2)

Call:
lm(formula = y ~ z1 + z2 + z3)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.01762 -0.31748 -0.04178  0.36181  1.35603 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.59483    0.14765   4.029 0.000972 ***
z1          -1.27393    0.08736 -14.583 1.17e-10 ***
z2           4.85629    0.13253  36.643  < 2e-16 ***
z3           0.63074    0.23399   2.696 0.015915 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6603 on 16 degrees of freedom
Multiple R-squared:  0.9899,	Adjusted R-squared:  0.988 
F-statistic: 520.9 on 3 and 16 DF,  p-value: 3.701e-16
```

Regression on PCs of X

```
##
## Call:
## lm(formula = y ~ z1 + z2 + z3 + z4 + z5 + z6, data = data.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7952 -0.3342  0.1027  0.4271  1.0735
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6028     0.1949   8.224 1.65e-06 ***
## z1            2.4274     0.1090  22.263 9.78e-12 ***
## z2           -4.1962     0.1525 -27.519 6.56e-13 ***
## z3            0.4358     0.3305   1.319    0.210
## z4           -0.4675     0.3702  -1.263    0.229
## z5            0.6914     0.4942   1.399    0.185
## z6           -0.8195     0.8107  -1.011    0.331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8716 on 13 degrees of freedom
## Multiple R-squared:  0.9898, Adjusted R-squared:  0.9851
## F-statistic: 209.9 on 6 and 13 DF,  p-value: 3.602e-12
```

Regression on PCs of X

Correlation matrix of PC scores:

```
> cor(pca$scores)
              Comp.1        Comp.2        Comp.3        Comp.4        Comp.5        Comp.6
Comp.1  1.000000e+00 -5.975531e-16 -3.197900e-16 -6.474717e-16 -6.044973e-17 -1.213539e-15
Comp.2 -5.975531e-16  1.000000e+00  1.053321e-15  5.214905e-16 -2.522863e-18  3.536800e-16
Comp.3 -3.197900e-16  1.053321e-15  1.000000e+00  5.431297e-16  7.632493e-17  3.695276e-16
Comp.4 -6.474717e-16  5.214905e-16  5.431297e-16  1.000000e+00  1.451978e-16  8.981136e-16
Comp.5 -6.044973e-17 -2.522863e-18  7.632493e-17  1.451978e-16  1.000000e+00  1.346402e-15
Comp.6 -1.213539e-15  3.536800e-16  3.695276e-16  8.981136e-16  1.346402e-15  1.000000e+00
```

Short summary

- Check the correlation between variables.
- Use the correlation coefficient or VIF to examine possible multicollinearity.
- Multicollinearity may cause the insignificance of important variables.
- Solution 1: Drop the high correlated variables.
- Solution 2: Use PCA technique to summarize the similarity of variables, and the use the PC scores to be the predictors.
- It is better to construct the model matrix X with orthogonal property (ie, uncorrelated).

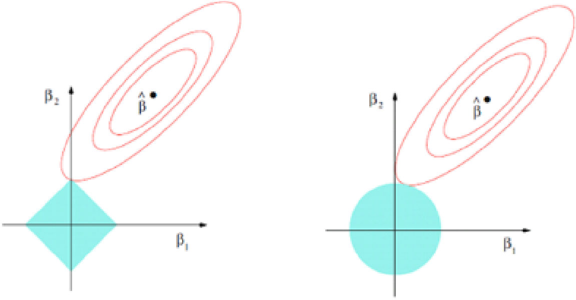
Purpose

In reality, the true model is unknown. What are the key variables depending on the response?
Methodologies:

- Best-subset model: Stepwise regression
- Shrinkage methods: Lasso regression and Ridge regression

Shrinkage Method

- It is related to the constrained optimization problem.
- It is called regularization or shrinkage.



Common methods

Let

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 .$$

- Lasso regression

$$\hat{\beta}^{Lasso} = \arg\min_{\beta} \quad \mathrm{RSS}(\beta)$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

- Ridge regression

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \quad \mathrm{RSS}(\beta)$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

Lagrange multiplier method

- Lasso regression

$$\hat{\beta}^{Lasso} = \arg\min_{\beta} \quad L_{lasso}(\beta)$$

where $L_{lasso}(\beta) =$

- Ridge regression

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \quad L_{ridge}(\beta)$$

where $L_{ridge}(\beta) =$

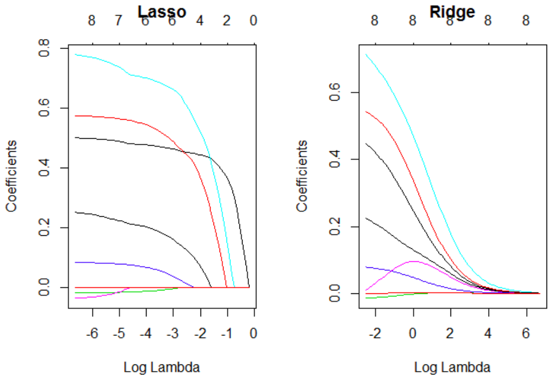
Shrinkage Method

- λ is called the penalty parameter.
- The methods lead variables to be sparsity.
- The estimates might not be exact values, but the important variables related to the response may be extracted correctly.
- What is the suitable value of λ? (λ is also called the tuning parameter.)

Shrinkage Method

A quick question:

What is the suitable value of λ?



Choose λ

- Cross-validation (CV): also known as the leave-one-out method. Split the training pairs into K parts or "folds", denoted by $F_1, \ldots, F_K$. Treat each group as the testing group at a time and fit the model by the other groups, denoted by $\hat{f}_\lambda^{-k}(x)$, $k = 1, \ldots, K$. Evaluate the prediction error of the testing group by

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \notin F_k} (y_i - \hat{f}_\lambda^{-k}(x_i))^2$$

- Bayesian information criterion (BIC)

$$\mathrm{BIC} = -2 \log \text{likelihood} + df \log n,$$

where df is the number of variables in the model.

Toy Experiment

Let the **true model** be

$$y_i = 10 + 0.5x_{1i} - 5x_{2i} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 0.49)$ and $i = 1, \ldots, 20$. Let the predictors be simulated from

$$x_{1i} \sim U(-2, 2),$$
$$x_{2i} \sim U(-1, 4).$$

We do have other variables:

$$x_{3i} = 1 + 0.8x_{1i} + e_i,$$
$$x_{4i} = 2 + 0.2x_{1i} + e_i$$
$$x_{5i} = -0.5x_{1i} + e_i,$$
$$x_{6i} = 2 + e_i$$

where $e_i \sim N(0, 0.25)$.

Idea

It is related to the principal components regression, but it is not just to find hyperplanes of maximum variance between the independent variables. It considers a linear regression model by projecting the predicted variables and the observable variables to a new space.

Strategy:
PLS is used to find the relations between two matrices (X and Y).

When to use?

- The matrix of predictors has more variables than observations.
- There is multicollinearity among X values.

The general model

$$X = TP^t + E,$$

$$Y = UQ^t + F,$$

where X is an $n \times p$ matrix of predictors, Y is an $n \times m$ matrix of responses, T and U are $n \times l$ matrices of projections of X and Y, respectively. P and Q are orthogonal loading matrices, and E and F are error terms.

Purpose:
PLS regression aims to incorporate information on both X and Y in the definition of the scores and loadings. Hence, the decompositions of X and Y are made by maximizing the covariance between T and U.

Questions?

- What are differneces between PC regression and PLS regression?
- How to implement the PLS regression in R?

Reference:
https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf