Subset selection for prostate cancer data set

Mingliang Wang

1 Best subset and Ridge regression

Consider a linear regression model

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \tag{1}$$

In this model, some β_i's in $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_n]^T$ are zeros. Our task is to identify those zero elements in $\boldsymbol{\beta}$ so that we can find those x_i in $X = [x_1, \cdots, x_n]$ that are key factor in the model.

The best subset method performs regression or classification using every possible models and choose the one with minimum prediction errors. The total number of possible combinations of models is 2^n which can be computationally costive when n is large.

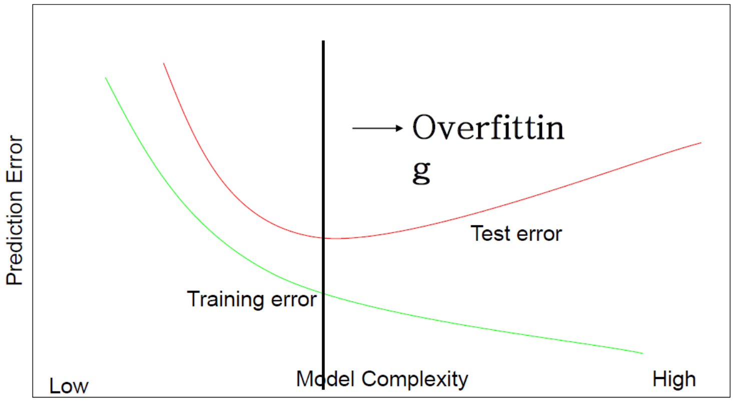
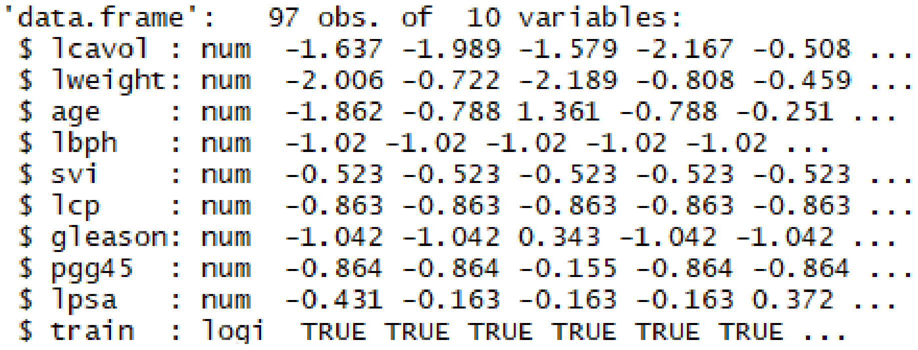


Figure 1: Model complexity vs. model performance

The ridge regression is a way to regularized the model by penalized the model complexity. The cost function is given as

$$\min \|y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \tag{2}$$

where λ is the penalty factor, $\|\cdot\|$ is the L2 norm. In this way, we can find a model that with suitable model complexity and find the key factors. The relation of model complexity and the model performance is demonstrated in Figure 1 which shows that the larger model complexity is always beneficial for the training of the model. However, when the model complexity exceeds that of the data, the testing errors will increase along with the model complexity.



```
'data.frame':   97 obs. of  10 variables:
 $ lcavol : num  -1.637 -1.989 -1.579 -2.167 -0.508 ...
 $ lweight: num  -2.006 -0.722 -2.189 -0.808 -0.459 ...
 $ age    : num  -1.862 -0.788 1.361 -0.788 -0.251 ...
 $ lbph   : num  -1.02 -1.02 -1.02 -1.02 -1.02 ...
 $ svi    : num  -0.523 -0.523 -0.523 -0.523 -0.523 ...
 $ lcp    : num  -0.863 -0.863 -0.863 -0.863 -0.863 ...
 $ gleason: num  -1.042 -1.042 0.343 -1.042 -1.042 ...
 $ pgg45  : num  -0.864 -0.864 -0.155 -0.864 -0.864 ...
 $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
 $ train  : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

Figure 2: zipdata description

2 Data description

The data is stored in a file named “zprostate”. As shown in Figure 2, the data contain 97 observations of 10 variables. The variables are indicators of the potential patients. The variables are as follows:

1. “lcavol” = log(cancer volume)

2. ‘lweight’ = log(prostate weight),

3. “age” = age of the patient,

4. “lbph” = log(benign prostatic hyperplasia amount),

5. “svi” = short seminal vesicle invasion,

6. “lcp” = log(capsular penetration),
7. “gleason” = Gleason score
8. “pgg45” =log percentage Gleason scores 4 or 5
9. “lpsa” = log(prostate specific antigen)

The indication of patient has a prostate cancer are stored in “train” , “True” indicates the patient has a cancer. We aims to model the “lpsa” of patients with prostate cancers.

3 Results

The above two methods are implemented for selecting the variables. The result of the two methods are given in the Table 1 and 2, respectively.

Table 1: Results of best subset

variables	coefficients	MSE
(Intercept)	2.4773573	0.4924823
“lcavol”	0.7397137	
“lweight”	0.3163282	

As shown in Table 1, the key variable is “lcavol” and “lweight”. The prediction error, calculated as

$$\frac{1}{N}\sum_{k=1}^{N}(y[k]-\text{pred}[k])^2 \tag{3}$$

where pred is the prediction of the model, N is the total number of observation. The above two methods are implemented for selecting the variables. The result of the two methods are given in the Table 1 and 2.

As shown in Table 2, the key variables are still “lcavol” and “lweight” (with the largest two coefficients in magnitude). Commonly used approach for selecting λ is the cross validation method.

$$\frac{1}{N}\sum_{k=1}^{N}(y[k]-\text{pred}[k])^2 \tag{4}$$

Table 2: Results of Ridge regression

variables	coefficients	MSE
(Intercept)	1.00000000	2.748148
"lcavol"	0.60774607	
"lweight"	0.28434026	
"age"	-0.11036319	
"lbph"	0.20050020	
"svi"	0.28330007	
"lcp"	-0.16236691	
"gleason"	0.01204058	
"pgg45"	0.20593909	

where pred is the prediction of the model, N is the total number of observation. In this case, the best $\lambda = 4.9$. By comparing MSE, we can see that best subset gives better performance. So when the dimension of available variables is not very large, the best subset is always a better choice.