STAT 480 Statistical Computing Applications

Unit 6. Model Selection

Lecture 1. Classic Model Selection

Department of Statistics, Iowa State University

Spring 2019

Problems of Least Squares Methods

- Prediction Accuracy
  - Least square estimates with full models tend to have low bias and high variance.
  - It is possible to trade a little bias with the large reduction in variance, thus achieving higher prediction accuracy.
- Interpretation
  - We would like to determine a small subset of variables with strong effects, without degrading the model fit.

Variable Selection

Variable Selection is a process of selecting a subset of predictors, fitting the selected model, and making inferences.

- include variables which are most predictive to the response;
- exclude noisy/uninformative variables from the model.

Advantages

- to build more parsimonious and interpretable models;
- to enhance the model prediction power;
- to improve the precision of the estimates.

Applications

Variable Selection is crucial to decision-making in many application and scientific areas:

- **Business:** important factors to decide credit limit, insurance premium, mortgage terms.
- **Medical and pharmaceutical industries:**
  - select useful chemical compounds for drug-making;
  - identify signature genes for cancer classification and diagnosis;
  - find risk factors related to disease cause or survival time.
- **Information retrieval:**
  - Google search, classification of text documents;
  - Email/spam filter;
  - Speech recognition, image analysis.

Example: Prostate Cancer Data (Stamey et al. 1989)

```
id cv wt age bph svi cp gs g45 psa
1 0.56 16.0 50 0.25 0 0.25 6 0 0.65
2 0.37 27.7 58 0.25 0 0.25 6 0 0.85
3 0.60 14.8 74 0.25 0 0.25 7 20 0.85
4 0.30 26.7 58 0.25 0 0.25 6 0 0.85
5 2.12 31.0 62 0.25 0 0.25 6 0 1.45
.... ...
```

- Response Y: prostate specific antigen (psa)
- Predictors X: cancer volume, prostate weight, age, benign prostatic hyperplasia amount, seminal vesicle invasion, capsular penetration, Gleason score, percent G-score 4 or 5.

Model Selection

- This is an “unsolved” problem in statistics: there are no magic procedures to get you the “best model”.
- To “implement” this, we need:
  - a criterion or benchmark to compare two models.
  - a search strategy.
- With a limited number of predictors, it is possible to search all possible models.

Best Subset Regression

- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- Possible criteria
  - R^2: not a good criterion. Always increase with model size $\Rightarrow$ "optimum" is to take the biggest model.
  - Adjusted R^2: better. It "penalized" bigger models.
  - Mallow's Cp.
  - Akaike's Information Criterion (AIC), Schwarz's BIC.
  - Cross validation (we have already seen this!)

Mallow's Cp

- For a model with p regression coefficients, (i.e., $p-1$ covariates plus the intercept β_0), define

$$C_p = \frac{RSS}{\sigma^2} - (n-2p),$$

  where

  - RSS = residual sum of squares
  - $\sigma^2 \approx$ mean square error $= \frac{RSS}{n-p}$
  - n = number of observations

- If the model is true, then $E(C_p) \approx p$. Thus one should choose models whose C_p values are low and close to p.

Best Subset Regression (Cont.)

Advantages

- Based on exhaustive search
- Check and compare all (2^p) models

Computation Limitations:

- The computation is infeasible for $p > 40$. There are over a billion models!
- Leaps and bounds procedure is efficient for $p \leq 40$

Stepwise Regression

- Basic Idea: seeking a good path through all the possible subsets
- There are three possible ways
  1. Backward elimination: starting with the full model and removing.
  2. Forward selection: starting with the intercept and adding.
  3. Stepwise selection: alternate backward elimination and forward selection.

Stepwise Regression

- This method involves adding or dropping one variable at a time from a given model based on a partial F-statistic.
- Let the smaller and bigger models be Model I and Model II, respectively. The partial F-statistic is defined as

$$\frac{RSS(Model\ I) - RSS(Model\ II)}{RSS(Model\ II)/\nu},$$

where ν is the degrees of freedom of the RSS (residual sum of squares) for Model II.

Forward Selection

- Begin with the null model – a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p–value above some threshold.

Backward Selection

- Start with all variables in the model.
- Remove the variable with the largest p–value – that is, the variable that is the least statistically significant.
- The new $(p-1)$-variable model is fit, and the variable with the largest p–value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p–value defined by some significance threshold.

Stepwise Regression

- In each step, consider both forward and backward moves and make the "best" move.
- A thresholding parameter is used to decide "add" or "drop" move.
- It allows previously added/removed variables to be removed/added later.

R code

You need to install the package "`leaps`" first. The function `regsubsets()` can be used to conduct model selection by exhaustive search, forward or backward, stepwise.

```
library(leaps)
help(regsubsets)
## Default S3 method:
regsubsets(x=, y=, weights=rep(1, length(y)), nbest=1,
nvmax=8, force.in=NULL, force.out=NULL, intercept=TRUE,
method=c("exhaustive", "backward", "forward", "seqrep"),
really.big=FALSE)
```

R code: Details

- `x`: design matrix
- `y`: response vector
- `weights`: weight vector
- `nbest`: number of subsets of each size to record
- `nvmax`: maximum size of subsets to examine
- `force.in`: index to columns of design matrix that should be in all models
- `force.out`: index to columns of design matrix that should be in no models
- `intercept`: Add an intercept?
- `method`: Use exhaustive search, forward selection, backward selection or sequential replacement to search.

Fit Sequential Selection Methods in R

```
library(leaps)
n = 50 # sample size
p = 4 # data dimension
set.seed(2015)
x <- matrix(rnorm(n*p),ncol=p) # generate design matrix
y <- x[,1]+x[,2]+rnorm(n)*0.5 # true regression model
## forward selection
for1 <- regsubsets(x,y,method="forward")
## backward elimination
back1 <- regsubsets(x,y,method="backward")
summary(back1)
coef(back1, id=1:4)
## exhaustive search
ex1 <- regsubsets(x,y,method="exhaustive")
summary(ex1)
coef(ex1,id=1:4)
```

Two Information Criteria: AIC and BIC

These are based on the maximum likelihood estimates of the model parameters. Assume that

- the training data are $(\mathbf{x}_i, y_i)$, $i = 1, \cdots, n$.
- a fitted linear regression model is $\hat{\beta}^T \mathbf{x}$.

Define

- The degree of freedom (df) of $\hat{\beta}$ as the number of of nonzero elements, including the intercept.
- The residual sum of squares as $RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}^T \mathbf{x}_i)^2$.

Then

$$
\begin{aligned}
AIC &= n + n\log(2\pi) + n\log(RSS/n) + 2 \cdot df \\
BIC &= n + n\log(2\pi) + n\log(RSS/n) + \log(n) \cdot df
\end{aligned}
$$

We choose the model which gives the smallest AIC or BIC.

Compute AIC and BIC for Forward Selection

```
# four candidate models
m1 <- lm(y~x[,1])
m2 <- lm(y~x[,1]+x[,2])
m3 <- lm(y~x[,1]+x[,2]+x[,4])
m4 <- lm(y~x)
# compute RSS for the four models
rss <- rep(0,4)
rss[1] <- sum((y-predict(m1))^2)
rss[2] <- sum((y-predict(m2))^2)
rss[3] <- sum((y-predict(m3))^2)
rss[4] <- sum((y-predict(m4))^2)
```

Compute AIC and BIC for Forward Selection

```
# compute AIC and BIC
bic <- rep(0,4)
aic <- rep(0,4)
for (i in 1:4){
bic[i] = n+n*log(2*pi)+n*log(rss[i]/n)+log(n)*(1+i)
aic[i] = n+n*log(2*pi)+n*log(rss[i]/n)+2*(1+i)
}
# find the optimal model
which.min(bic)
which.min(aic)
```

Discusssions

- The models selected by forward selection, backwards elimination, and stepwise regression might not be the same, even using the same model selection criterion.
- In a forward selection or a backwards elimination procedure, BIC may result in fewer parameters in the model than AIC.
- The forward selection, backward elimination, and stepwisse regression procedures are not guaranteed to find the best model according to the AIC or BIC criterion.
- P-values in resultant models should be treated with caution, because they do not reflect the model selection process.
- Generally, there may be several models that are highly similar in the quality of the fit.