Lecture 12: Variable Selection (I)

Hao Helen Zhang
Outline

• Motivation for Variable Selection
• Classical Methods
  • best subset selection
  • forward selection
  • backward elimination
  • stepwise selection
• Modern Penalization Methods
  • $L_q$ penalty, ridge
  • LASSO, adaptive LASSO, LARS
  • non-negative garotte, SCAD
Problems of Least Squares Methods

- **Prediction Accuracy**
  \[ \text{MSE} = \text{Bias}^2 + \text{Var} \]
  - 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 (VS)

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

VS 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
- more
### Example: Prostate Cancer Data (Stamey et al. 1989)

<table>
<thead>
<tr>
<th>id</th>
<th>cv</th>
<th>wt</th>
<th>age</th>
<th>bph</th>
<th>svi</th>
<th>cp</th>
<th>gs</th>
<th>g45</th>
<th>psa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.56</td>
<td>16.0</td>
<td>50</td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
<td>6</td>
<td>0</td>
<td>0.65</td>
</tr>
<tr>
<td>2</td>
<td>0.37</td>
<td>27.7</td>
<td>58</td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
<td>6</td>
<td>0</td>
<td>0.85</td>
</tr>
<tr>
<td>3</td>
<td>0.60</td>
<td>14.8</td>
<td>74</td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
<td>7</td>
<td>20</td>
<td>0.85</td>
</tr>
<tr>
<td>4</td>
<td>0.30</td>
<td>26.7</td>
<td>58</td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
<td>6</td>
<td>0</td>
<td>0.85</td>
</tr>
<tr>
<td>5</td>
<td>2.12</td>
<td>31.0</td>
<td>62</td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
<td>6</td>
<td>0</td>
<td>1.45</td>
</tr>
</tbody>
</table>

... ...

**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.
Acute Leukemia Data (Golub et al. 1999)

Three types: AML (acute myeloid), ALL-B (acute lymphoblastic), ALL-T
### Stanford Heart Transplant Data

<table>
<thead>
<tr>
<th>start</th>
<th>stop</th>
<th>event</th>
<th>age</th>
<th>year</th>
<th>surg</th>
<th>plant</th>
<th>id</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>50</td>
<td>1</td>
<td>31</td>
<td>0.12</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>1</td>
<td>52</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>54</td>
<td>0.26</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>1</td>
<td>54</td>
<td>0.27</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>0</td>
<td>36</td>
<td>0</td>
<td>40</td>
<td>0.49</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>36</td>
<td>39</td>
<td>1</td>
<td>40</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>0</td>
<td>18</td>
<td>1</td>
<td>21</td>
<td>0.61</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$T_i$: failure time; $C_i$: censoring time. Data $(\tilde{T}_i, \delta_i, X_i)$, where $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. 
Notations

- data \((X_i, Y_i), i = 1, \cdots, n\)
- \(n\): sample size
- \(d\): the number of predictors, \(X_i \in \mathbb{R}^d\).
- the full index set \(S = \{1, 2, \cdots, d\}\).
- the selected index set given by a procedure is \(\hat{A}\), its size is \(|\hat{A}|\).
- the linear coefficient vector \(\beta = (\beta_1, \cdots, \beta_d)^T\).
- the true linear coefficients \(\beta_0 = (\beta_{10}, \cdots, \beta_{d0})^T\).
- the true model \(A_0 = \{j : j = 1, \cdots, d, |\beta_{j0}| \neq 0\}\).
Variable Selection in Orthogonal Design

Assume that

- $y, x_1, \ldots, x_d$ are centered
- $\langle x_j, x_k \rangle = 0$ for $j \neq k$.

Then

$$\hat{\beta}_j = \frac{\langle x_j, y \rangle}{\langle x_j, x_j \rangle}, \quad j = 1, \ldots, d.$$

Define $t_j = \hat{\beta}_j \|x_j\|^{1/2} / \hat{\sigma} = \hat{\beta}_j / [\|x_j\|^{-1/2} \hat{\sigma}]$ for $j = 1, \ldots, d$, then

$$SSR = \langle X\hat{\beta}, X\hat{\beta} \rangle = \sum_{j=1}^{d} \hat{\beta}_j^2 \|x_j\|^2$$

$$= \sum_{j=1}^{d} \hat{\sigma}^2 t_j^2 = \sum_{j=1}^{d} R_j^2.$$
The coefficient of determination

\[ R^2 = \frac{SSR}{S_{yy}} = \frac{1}{S_{yy}} \sum_{j=1}^{d} R_j^2 \]

- Each \( x_j \) contributes to \( R^2 \) regardless of other variables.
- One can use \( R_j^2 \), or \( t_j^2 \), or \( |t_j| \) to rank the importance of variables.
**Variable Selection in Non-orthogonal Design**

More practical and difficult cases: variables are correlated.

- There are no natural orderings of importance for the input variables
- The role of a variable can only be measured relative to the other variables in the model.
  - Example: highly correlated variables
- It is essential to check all possible combinations.
For each $k \in \{0, 1, \ldots, d\}$, find the subset of size $k$ that gives smallest residual sum of squares

- Search through all $(2^d)$ possible subsets:
  - When $d = 10$, we check 1024 combinations.
  - When $d = 20$, more than one million combinations.
- The larger $k$, the smaller RSS. (see the following picture)
Figure 3.5: All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.
How to Choose the best $k$

This question involves
- the tradeoff between bias and variance,
- the more subjective desire for parsimony

In practice, we can use a number of model selection criteria
- cross validation; prediction error on the test set
- Mallow’s $C_p$, F-statistic
- Generalized Information Criteria (GIC):

$$\text{GIC}(\text{model}) = -2 \cdot \text{loglik} + \alpha \cdot \text{df},$$

$\text{df}$ is the model size (or the number of effective parameters).
This question involves

- the tradeoff between bias and variance,
- the more subjective desire for parsimony

In practice, we can use a number of model selection criteria

- cross validation; generalized cross validation (GCV)
- prediction error on the test set
- Mallow’s $C_p$, F-statistic
- Generalized Information Criteria (GIC):

$$\text{GIC}(\text{model}) = -2 \cdot \text{loglik} + \alpha \cdot \text{df},$$

df is the model size (or the number of effective parameters).
About Best Subset Selection

Advantages:

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

Computation Limitations:

- The computation is infeasible for \(d \geq 40\).
- Leaps and bounds procedure (Furnival and Wilson 1974) is efficient for \(d \leq 40\)
- There is a contributed package \textit{leaps} in R.
Searching Methods

**Basic Idea**: seeking a good path through all the possible subsets

- forward selection
- backward elimination
- stepwise selection
Forward Selection

- Starting with the intercept, sequentially add one variable that most improves the model fit
  - If there are $k$ variables in the model and the parameter estimate is $\hat{\beta}$, and we add in one variable, resulting in the estimate $\tilde{\beta}$.
  - The improvement in fit is often based on the $F$ statistic

$$F = \frac{\text{RSS}(\hat{\beta}) - \text{RSS}(\tilde{\beta})}{\text{RSS}(\tilde{\beta})/(n - k - 2)}.$$

  Typical add the variable which produces the largest value of $F$.

- Stops when no variable produces an $F$-ratio greater than the 90th or 95th percentile of the $F_{1,n-k-2}$ distribution
Forward Selection for Prostate Cancer Data

The *leaps* function in R produces the sequence:

\[
\begin{array}{ccccccccc}
\text{cv} & \text{wt} & \text{age} & \text{bph} & \text{svi} & \text{cp} & \text{gs} & \text{g45} \\
\text{step1} & x & & & & & & \\
\text{step2} & x & x & & & & & \\
\text{step3} & x & x & x & & & & \\
\text{step4} & x & x & x & x & & & \\
\text{step5} & x & x & x & x & x & & \\
\text{step6} & x & x & x & x & x & x & x \\
\text{step7} & x & x & x & x & x & x & x & x \\
\text{step8} & x & x & x & x & x & x & x & x & x
\end{array}
\]

Forward selection for prostate cancer data: \( \hat{A}_{AIC} = \{1, 2, 3, 4, 5\} \), \( \hat{A}_{BIC} = \{1, 2, 5\} \).
Backward Elimination

- Starting with the full model, sequentially drop one variable that produces the smallest $F$ value.
- Stops when each variable in the model produces an $F$-ratio greater than the 90th or 95th percentile of the $F_{1,n-k-2}$.
- Can only be used when $n > d$. 
Stepwise Selection

- 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.
Pros and Cons

Advantages:

- intuitive; simple to implement; work well in practice
- May have lower prediction error than the full model

Limitations:

- Greedy-search type algorithms are fast, but locally optimal
- Highly variable due to discreteness (Breiman, 1996; Fan and Li, 2001)
- Hard to establish asymptotic theory and make inferences.
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,

```r
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)
```
Details

Arguments:

- \( x \): 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.
library(leaps)

# sample size
n = 50
# data dimension
p = 4

# generate design matrix
set.seed(2015)
x <- matrix(rnorm(n*p),ncol=p)
# true regression model
y <- x[,1]+x[,2]+rnorm(n)*0.5

## forward selection
for1 <- regsubsets(x,y,method="forward")
summary(for1)
coef(for1, id=1:4)
```r
## backward elimination
back1 <- regsubsets(x,y,method="forward")
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 \((x_i, y_i), i = 1, \cdots, n\).
- a fitted linear regression model is \(\hat{f}(x)\).

Define

- The degree of freedom (df) of \(\hat{f}\) as the number of effective parameters of the model, or the model size.
- The residual sum of squares as \(\text{RSS} = \sum_{i=1}^{n} [y_i - \hat{f}(x_i)]^2\)

Then

\[
\begin{align*}
AIC &= n \log(\text{RSS}/n) + 2 \cdot \text{df}, \\
BIC &= n \log(\text{RSS}/n) + \log(n) \cdot \text{df},
\end{align*}
\]

We choose the model which gives the smallest AIC or BIC.
AIC and BIC for Linear Regression Models

Assume that

- the training data are \((x_i, y_i), i = 1, \cdots, n\).
- a fitted linear regression model is \(\hat{f}(x) = \hat{\beta}^T x\).
- For example, \(\hat{\beta}\) can be the regression coefficients given by the OLS, Lasso, forward selection.

Define

- The degree of freedom (df) of \(\hat{\beta}\) as the number of nonzero elements in \(\beta\) (model size), including the intercept
- The residual sum of squares as \(\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{\beta}^T x_i)^2\)

\[
\text{AIC} = n \log(\text{RSS}/n) + 2 \cdot \text{df},
\]

\[
\text{BIC} = n \log(\text{RSS}/n) + \log(n) \cdot \text{df},
\]

We choose the model which gives the smallest AIC or BIC.
By definition, the BIC for the model $M$ is formally defined as

$$BIC = -2 \log \hat{L} + \log(n) \cdot df,$$

where

- $L$ is the likelihood function of the model parameters;
- $\hat{L}$ is the maximumized value of the likelihood function of the model $M$.

Special example: Consider the regression model:

$$Y_i = X_i^T \beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

The observations $Y_i \sim N(X_i^T \beta, \sigma^2), i = 1, \cdots, n$ are independent.
Example: BIC in Regression Case

For model $M$, the design matrix $X_M = \{X_{ij} : i = 1, \cdots, n; j \in M\}$. The likelihood

$$L(\beta | y, X_M) = (2\pi \sigma^2)^{-n/2} \exp\left\{ -\frac{(y - X_M\beta)^T (y - X_M\beta)}{2\sigma^2} \right\},$$

and the log likelihood is

$$\log L(y | x_1, \cdots, x_n) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{(y - X_M\beta)^T (y - X_M\beta)}{2\sigma^2}.$$  

The MLE is given by

$$\hat{\beta}_{MLE} = (X_M^T X_M)^{-1} X_M^T y, \quad \hat{\sigma}^2_{MLE} = \frac{RSS}{n},$$

where $RSS = (y - X_M\hat{\beta})^T (y - X_M\hat{\beta})$. 
Then

$$-2 \log \hat{L} = n \log(2\pi) + n \log(\hat{\sigma}^2) + n = n \log(2\pi) + n \log\left(\frac{RSS}{n}\right) + n.$$ 

Removing the constant, we get

$$BIC = n \log\left(\frac{RSS}{n}\right) + \log(n) \cdot |M|.$$
Compute AIC and BIC for Forward Selection

```r
# 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
bic <- rep(0,4)
aic <- rep(0,4)

for (i in 1:4){
  bic[i] = n*\log(rss[i]/n)+\log(n)*(1+i)
  aic[i] = n*\log(rss[i]/n)+2*(1+i)
}

# find the optimal model
which.min(bic)
which.min(aic)