Linear Methods for Regression and Classification

Feng Li
feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Revision: November 7, 2013

Today we are going to talk about...

1. **Linear methods for regression**
2. **Model selection in linear models**
3. **Linear methods for classification**

Least squared estimation and Gaussian-Markov theorem

- $\hat{\beta}$ has the smallest variance among all linear unbiased estimators.
- The restriction to unbiased estimators not necessary a wise one.
- The mean squared error (MSE) of an estimator $\tilde{\theta}$ in estimating θ

$$MSE(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2 = Var(\tilde{\theta}) + (E\tilde{\theta} - \theta)^2$$

- The Gaussian-Markov theorem: the least squared estimators has the smallest MSE of all unbiased estimators.
- But there may exist a biased estimator with small MSE, i.e sacrificing a little biasness will reduce the variance.
- MSE is directly linked with prediction accuracy

$$E(Y_0 - \tilde{f}(x_0))^2 = \sigma^2 + E(x'\tilde{\beta} - f(x_0))^2 = \sigma^2 + MSE(\tilde{f}(x_0))$$

- So biased estimator may also improve prediction accuracy.

Two purposes in linear modeling

- The prediction accuracy
- The interpretation

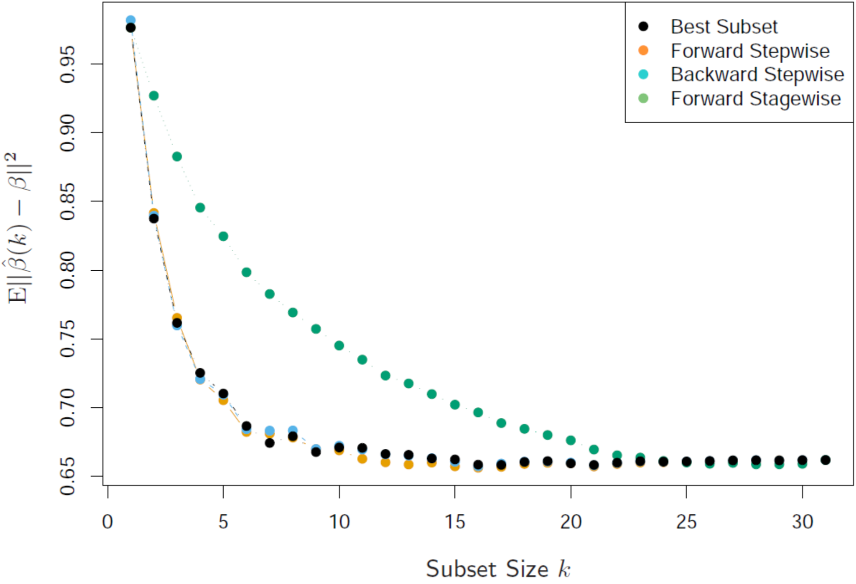
Variable selection in linear models

- **Best subset selection**
  Find the subset of size $k = 0, 1, ..., p$ with smallest residual sum of squared.
  It is eventually the procedure for searching for all possible subsets.
- **Forward/backward stepwise selection**
  Start with intercept/full covariates and add/remove a covariate that improves the fit.
  The methods depend on how you initialize the first step (sequentially).
- **Forward stagewise selection**
  Same as the forward stepwise selection.
  Each step add a covariate that is most correlated with the residual.
- **Bayesian variable selection**
  Assign a binary indicator $\mathcal{I}$ to each covariate.
  Estimate the $\mathcal{I}$: the probability of each covariate is included in the model.

Variable selection in linear models

- **Least Angle Regression**: "Democratic" version of forward stepwise regression.
  - Start without any regressors, i.e. $\hat{\beta}_1 =, ..., = \hat{\beta}_p = 0$
  - Find a predictor x_j that is most correlated with $r = y - \bar{y}$
  - Tuning the coefficient $0 < \beta_j < \grave{\beta}_j$ (where $\grave{\beta}_j$ is the OLS coefficient of the residual r with x_j) until you find another x_k has as much correlation with current residuals r with x_j
  - Now move β_j and β_k in the same way until another x_l comes.
  - Continue until all predictors come in.

Variable selection in linear models



Shrinkage methods: Ridge regression

- Set the arbitrary constrain(penalty) $\sum_{i=1}^{p} \beta_i^2 \leqslant t$
- $RSS = (y - X\beta)'(y - X\beta) + \lambda\beta'\beta$
- The target:

$$\hat{\beta} = \text{argmin}_\beta \{RSS\}$$

- $\hat{\beta} = (X'X + \lambda I)^{-1} X' y$
- The no. of effective parameters

$$tr(H) = tr(X(X'X + \lambda I)^{-1} X') = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

where $d_j > 0$ are the entries from the matrix D in **singular value decomposition** of X

$$X = UDV'$$

Shrinkage methods: LASSO

- **LASSO**: least absolute shrinkage and selection operator
- Same as ridge regression but the constrain is now

$$\sum_{i=1}^{p} |\beta_i| \leqslant t$$

- There is not closed form of $\hat{\beta}$.
- LASSO is a continuous subset selection routine
    - LASSO shrunk the least squares coefficients to exactly zero (when t is sufficient small) in order remove the effect the coefficients. Equivalent of variable selection.
    - In Bayesian framework, LASSO is same to have a linear regression while the prior of the coefficients are set as **Laplace distribution**.
- **Grouped LASSO**
    - When the predictors X belong to different categories. It is desirable to shrink the members together.
    - The constrain is now

$$\sum_{l=1}^{L} \sqrt{p_l}||\beta_l||_2 \leqslant t$$

where $\|\boldsymbol{\beta}_l\| := \sqrt{\beta_{l1}^2 + \cdots + \beta_{lk}^2}$ is the Euclidean norm.

Shrinkage or variable selection?

- Matters of personal taste.
- If interpretation is important, use variable selection
- If a lot of non-informative covariates used, shrinkage can be used.
- One may combine both shrinkage and variable selection methods.

Classification with linear discriminant analysis

- Classification: find the **decision boundaries**.
- We want to know $\Pr(G = k | X = x)$. It reads the probability of G belongs to group k conditional that X is x.
- We also know that sum of the probabilities is one a priori, i.e. $\sum_{k=1}^{K} = 1$
- Let $f_k(x)$ is the conditional density of X
- Bayes theorem shows that

$$\Pr(G = k | X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

- Knowing $f_k(x)$ implies $\Pr(G = k | X = x)$
- Linear discriminant analysis is to model each $f_k(x)$ with multivariate Gaussian while assuming the covariance matrix Σ_k are all of the same among all K classes.

Logistic regression
↪ The model

- It is essentially modeling the probability of K classes through linear function in x with log odds rations

$$\log \frac{\Pr(G = k|X = x)}{\Pr(G = K|X = x)} = \beta_{k0} + \beta_k' x$$

where $k = 1, 2, ..., K - 1$.

- Some notes:
  - There are $K - 1$ log-odds, i.e. $K - 1$ models in total.
  - The probabilities should sum to one.
  - When $K = 2$, only one model needed and the responses are binary.
- This is equivalent of

$$\Pr(G = k|X = x) = \frac{\exp\left(\beta_{k0} + \beta_k' x\right)}{1 + \sum_{l=1}^{K} (\beta_{l0} + \beta_l' x)}$$

Logistic regression
$\looparrowright$ Fitting the model with maximum likelihood

- **The idea**: using conditional likelihood with multinomial distribution
- In the $K = 2$ case, it is binomial distribution with the likelihood as

$$
\begin{aligned}
l(\beta) &= \sum_{n=1}^{N} \{y_i \log p(x_i, \beta) + (1 - y_i) \log(1 - p(x_i, \beta))\} \\
&= \sum_{n=1}^{N} \{y_i \beta' x_i - \log(1 + \exp\{1 + \beta' x_i\})\}
\end{aligned}
$$

- To obtain $\hat{\beta}$, use Newton-Raphson algorithm

$$
\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta'} \right)^{-1} \frac{\partial l(\beta)}{\partial \beta}
$$

- LASSO can be used for variable selection

$$
max_\beta \left\{ l(\beta) - \lambda \sum_{j=1}^{p} |\beta_j| \right\}
$$

- The likelihood estimation for multinomial case can be done in a similar fashion.