

# 고급바이오정보학

## 5강. Genetic Population Analysis : Penalized Logistic Analysis

---

출처 : 방송통신대학교 바이오정보·통계학과

# 목차

---

1. Introduction to Genetic Association Studies
2. Penalized Regression Analysis
3. Statistical Software

# 1) Genetic Association Studies

---

- 주요 목적
  - 어떤 질병이나 특성에 관련된 원인 유전자 또는 유전자 지역들을 식별을 위해서..
    - 암, 백혈병, 비만 등...
    - 혈압, 콜레스테롤 수치, 지질단백질, 체질량 지수( BMI)..
- Genome-wide Association Study (GWAS) <- 최근 활발히 연구되는 분야로
  - 질병 또는 특성과 관련된 일반적인 유전자 변이를 식별
  - Single nucleotide polymorphisms( SNP )
  - 일반적인 165개의 인간의 복잡한 질병과 특성은 1,200 개 이상의 유전자 변이와 연관 됨.( Lander, 2011 Nature)

## 1. Introduction to Genetic Association Studies

# 2) Type of Genetic Data

---

- 유전자/유전체 데이터  $\leq$  x변수 또는 feature
  - 단일 염기 다형성( SNP )
  - Microarray 유전자 발현 데이터
  - DNA methylation 데이터
  - 유전자 복제수 변이
  - 차세대 유전자 서열 데이터 ( 한사람의 유전자 전체 데이터 )
- 표현형 데이터  $\leq$  y변수 또는 목표변수
  - 양적 특성
  - 사례-대조군 데이터
  - Matched case-control data : 한 사람의 암세포와 정상세포의 데이터
  - 생존 시간

## 3) Statistical Challenges

---

- High-dimensional genetic/genomic data
  - 고차원의 문제( 변수 개수가 너무 많음)
  - 유전자(변수)수가 샘플수보다 많음.
  - 유전자들간의 상호 상관이 높음.
  - 소수의 유전자만이 표현형에 영향을 줌.( Sparisty of true singals )
- Statistical Analysis
  - 개별 유전자 Test 시 Multiple testing problem이 발생 => 8강, 9강에서 다룸.
  - Regularization Procedures가 필요
    - Tuning parameter selection problem

## 2. Penalized Regression Analysis

# 1) Penalized Likelihood for Regression

---

- Penalized likelihood

$$Q_{\lambda}(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + p_{\lambda}(\boldsymbol{\beta})$$

- $l(\boldsymbol{\beta})$  is a log-likelihood
- $\boldsymbol{\beta}$  is a parameter of interest (regression coefficients)
- $\lambda$  is a tuning parameter for sparsity
- $p_{\lambda}(\boldsymbol{\beta})$  is a penalty function

- Data for regression model

- $X$  : 유전자 데이터
- $Y$  : 표현형

# 2) Linear Regression Model

---

- Analysis of quantitative outcomes : blood pressure, BMI, ...
- The phenotype outcome follows a Normal distribution

$$Y_i \sim N(\beta_0 + X_i \beta, \sigma^2), \quad i = 1, 2, \dots, n$$

- $\beta_0$  is intercept parameter
  - $\beta$  is a vector of  $p$ -dimensional regression coefficients
  - $\sigma^2$  is a constant variance
  - $p_\lambda(\beta)$  is a penalty function
- Penalized likelihood using least squared loss

$$Q_\lambda(\beta_0, \beta) = \frac{1}{2} \sum_{i=1}^n (Y_i - \beta_0 - X_i \beta)^2 + p_\lambda(\beta)$$

## 2. Penalized Regression Analysis

### 3) Penalized Likelihood for Regression

---

- Analysis of binary outcomes : case-control outcomes
  - Presence or absence of a disease
  - $Y_i = 1$  for cases and  $Y_i = 0$  for controls
- Penalized likelihood using logistic regression

$$Q_\lambda(\beta_0, \beta) = -l(\beta_0, \beta) + p_\lambda(\beta)$$
$$l(\beta_0, \beta) = \sum_{i=1}^n [Y_i \log p_i(\beta_0, \beta) + (1 - Y_i) \log(1 - p_i(\beta_0, \beta))]$$

and

$$p_i(\beta_0, \beta) = \frac{e^{\beta_0 + X_i \beta}}{1 + e^{\beta_0 + X_i \beta}}$$



## 2. Penalized Regression Analysis

### 4) Conditional Logistic Regression Model

---

- Analysis of matched binary outcomes : matched case-control outcomes
    - Presence or absence of a disease for matched individuals
    - $Y_i = 1$  for cases and  $Y_i = 0$  for controls
    - $\delta_i \in 1, 2, \dots, K$  : the stratum of the  $i$ -th individual
  - Penalized likelihood using conditional logistic likelihood
- 한사람의 암세포 샘플은 1개 이고, 정상세포 샘플은 여러 개 일때

$$Q_\lambda(\beta) = -\log L(\beta | n_1, \dots, n_K) + p_\lambda(\beta)$$

$$L(\beta) = \prod_{i=1}^K \frac{\exp(\sum_{j \in \Delta_k} X_i \beta Y_j)}{\sum_{j \in \Delta_k} \exp(X_i \beta)}$$

with  $\Delta_k = \{j \leq n: \delta_j = k\}$  consists of  $m + 1$  indices  
for 1:  $m$  matched design.

# 5) Penalty Functions: Shrinkage Estimate

---

- Lasso (Least Absolute Shrinkage and Selection Operator)

- $l_1$  norm penalty induces sparsity
- Most of regression coefficients  $\beta$  are exactly to zero.

$$p_{\lambda}(\beta) = \lambda \sum_{j=1}^p |\beta_j|$$

- Ridge regression estimate

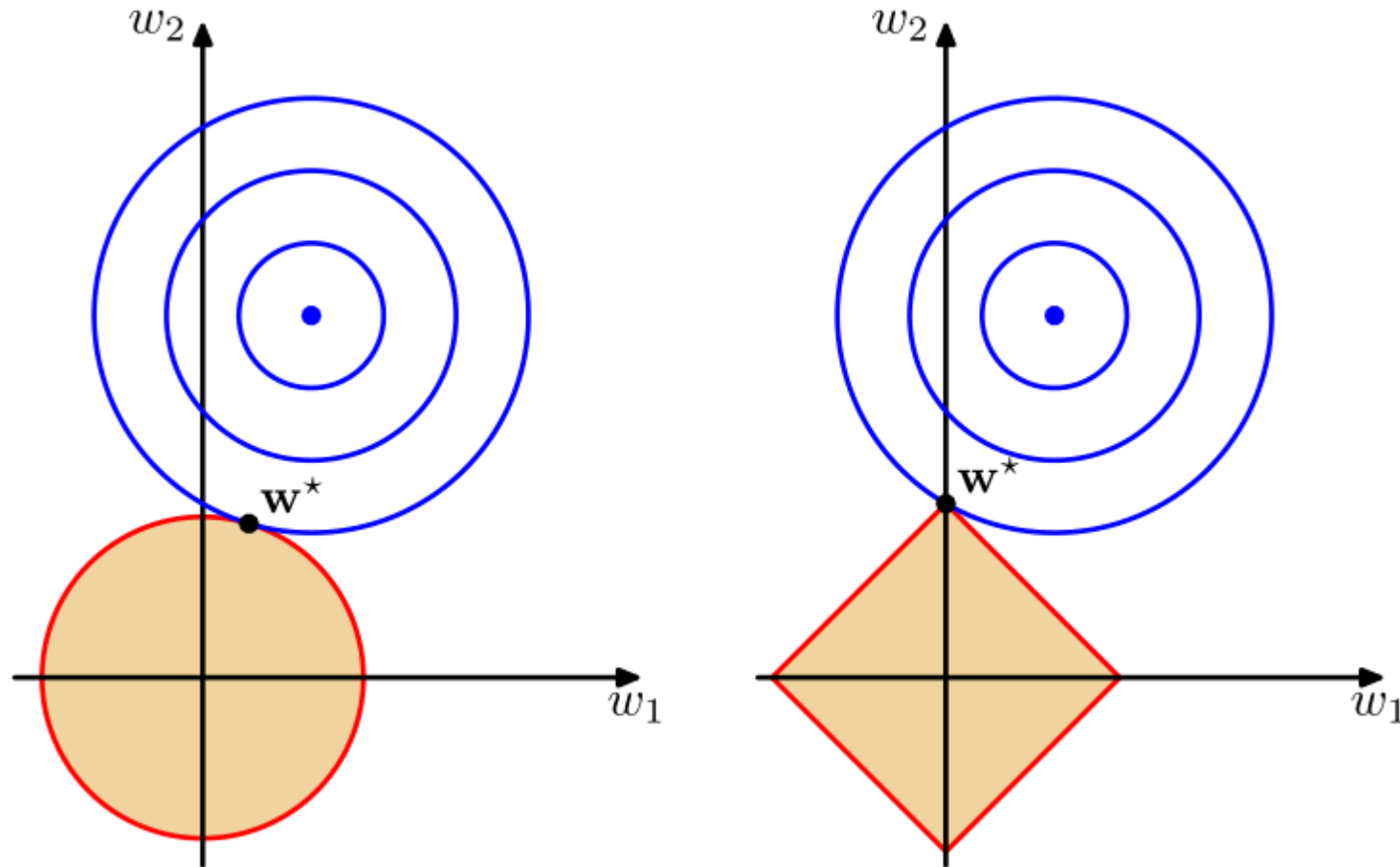
- $l_2$  norm penalty induces smoothness
- Regression coefficients for correlated variables are shrinked to each other

$$p_{\lambda}(\beta) = \lambda \sum_{j=1}^p \beta_j^2$$

## 2. Penalized Regression Analysis

### 5) Penalty Functions: Shrinkage Estimate

---



# 6) Elastic Net Regularization

---

- Elastic net regularization procedure
  - It combines  $l_1$  norm penalty and  $l_2$  norm penalty

$$p_{\lambda}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2$$

- $l_1$  part generate a sparse model
- $l_2$  part encourages grouping effects for correlated variables
- No limit on the number of selected variables
- If  $\lambda_1 = 0$ , the elastic net reduces to ridge
- If  $\lambda_2 = 0$ , the elastic net reduces to lasso

# 7) Network-based Regularization

---

- Elastic net regularization including a Laplacian matrix : 이미 알고 있는 유전자들간의 관계를 반영

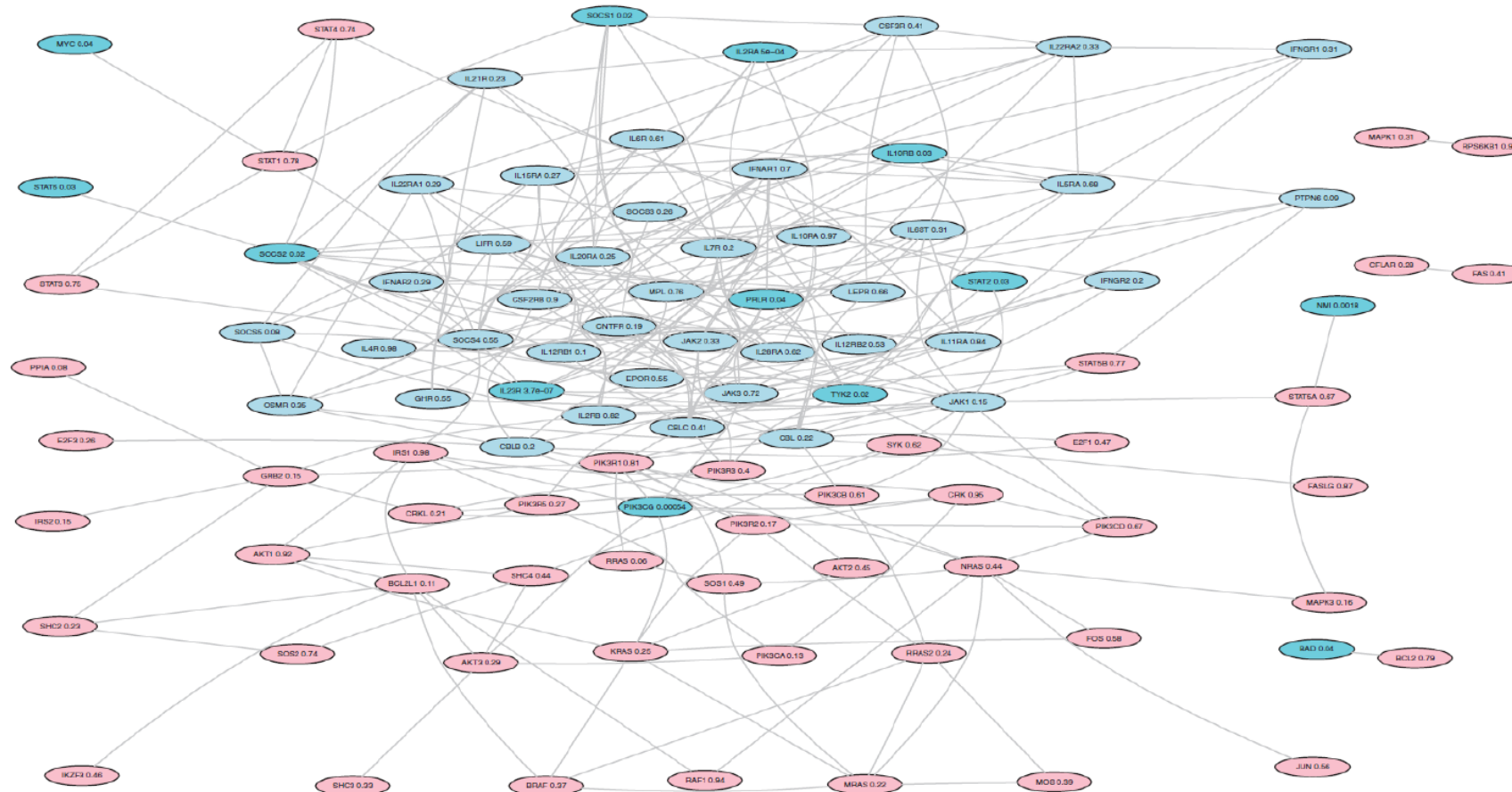
- A Laplacian matrix  $L$  represents a complex network structure of genes

$$p_{\lambda}(\beta) = \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \beta^T L \beta$$

- If  $L = I$ , it is exactly the same as elastic net
- Smoothness w.r.t linked structure of the regression coefficients is induced.
- Biological network information can be incorporated into regularization procedure.
  - Gene regulatory pathway
  - Metabolic pathway

## 8) Example of Genetic Pathway


## IL-2 receptor Beta chain in T cell activation pathway



## 2. Penalized Regression Analysis

# 9) Biological Network Database

- Biological network information is available at
  - KEGG ([www.genome.jp/kegg/](http://www.genome.jp/kegg/))
  - BioCarta ([www.biocarta.com](http://www.biocarta.com))
  - GenMAPP ([www.genmapp.org](http://www.genmapp.org))
  - HPRD ([www.hprd.org](http://www.hprd.org))



KEGG Home  
Release notes  
Current statistics  
Plea from KEGG

KEGG Database  
KEGG overview  
Searching KEGG  
KEGG mapping  
Color codes

KEGG Objects  
Pathway maps  
Brite hierarchies

KEGG Software  
KegTools  
KEGG API  
KCOM

KEGG FTP  
subscription

GenomeNet  
DSGET/LinkDB  
Feedback

Kanahisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (See Release notes for new and updated features).

New article: Data, information, knowledge and principle: back to metabolism in KEGG.

Main entry point to the KEGG web service

KEGG2 KEGG Table of Contents Update notes

Data-oriented entry points

KEGG PATHWAY KEGG pathway maps [Pathway list]

KEGG BRTE BRTE functional hierarchies [Brite list]

KEGG MODULE KEGG modules [Module list]

KEGG ORTHOLOGY Ortholog groups [KO system]

KEGG GENOME Genomes [KEGG organisms]

KEGG GENES Genes and proteins [Release history]

KEGG COMPOUND Small molecules [Compound classification]

KEGG REACTION Biochemical reactions [Reaction modules]

KEGG DISEASE Human diseases [Cancer | Infectious disease]

KEGG DRUG Drugs [ATC drug classification]

KEGG MEDICUS Health information resource integrating drug labels

Organism-specific entry points

KEGG Organisms Enter org code(s) [Go] hsa hsa eco

Analysis tools

KEGG Mapper KEGG PATHWAY/BRTE/MODULE mapping tools

KEGG Atlas Navigation tool to explore KEGG global maps

KAAS KEGG automatic annotation server

BLAST/FASTA Sequence similarity search

SIMCOMP Chemical structure similarity search

PathPred Biodegradability/biosynthesis pathway prediction

GenMAPP  
Gene Map Annotator and Pathway Profiler

Home About Us Contact Us Downloads

GenMAPP  
INTRO

Welcome to GenMAPP

GenMAPP is a free computer application designed to visualize gene expression and other genomic data on maps representing biological pathways and groupings of genes. Integrated with GenMAPP are programs to perform a global analysis of gene expression or genomic data in the context of hundreds of pathway MAPs and thousands of Gene Ontology Terms (GO terms). Import lists of genes/proteins to build new MAPs (MAPBuilder), and export archives of MAPs and expression/genomic data to the web.

The main features underlying GenMAPP are:

- Draw pathways with easy-to-use graphics tools
- Color genes on MAP files based on user-imported genomic data
- Query data against MAPs and the Gene Ontology

Learning to Use GenMAPP

To learn how to import your gene data, create pathway MAPs and/or visualize your data in the context of GenMAPP pathways check out our interactive tutorial.

GenMAPP Interactive Tutorial

GenMAPP Resources

To get information on existing functions in GenMAPP, advanced use cases, frequently access questions, see our online help resources and GenMAPP Workshop Resources page.

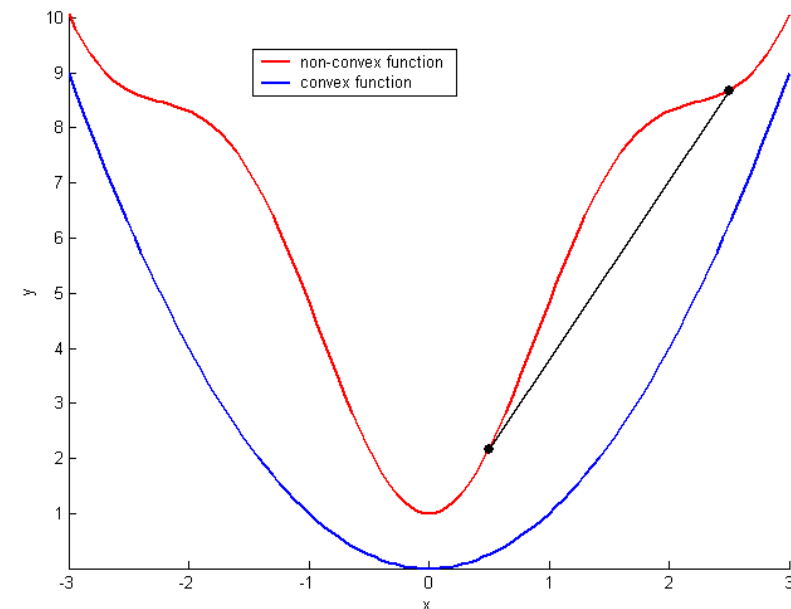
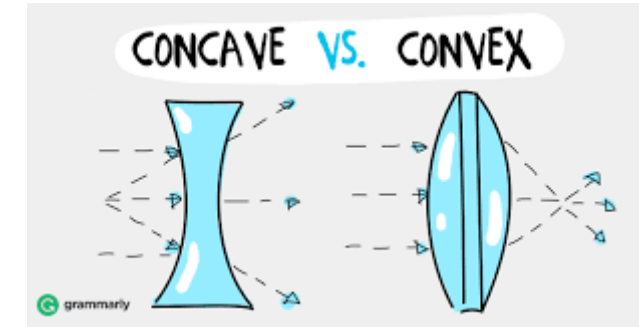
GenMAPP Help  
GenMAPP Workshop Resources

Home | About Us | Contact Us | Help | Downloads | News | Journals | GenMAPP

# 10) Convex Optimization

- Regression coefficients estimates
  - For given  $\lambda$ , we need to minimize the penalized likelihood
$$Q_{\lambda}(\beta) = -l(\beta) + p_{\lambda}(\beta)$$
  - The closed form solution to  $\beta$  does not exist.
  - The penalized likelihood is **convex**.
  - Computational algorithms for convex optimization can solve the equation.
  - Cyclic coordinate descent algorithm provide an efficient and fast solution for high-dimensional genomic data.
  - Statistical software such as an R package is publicly available.

[https://en.wikipedia.org/wiki/Coordinate\\_descent](https://en.wikipedia.org/wiki/Coordinate_descent)

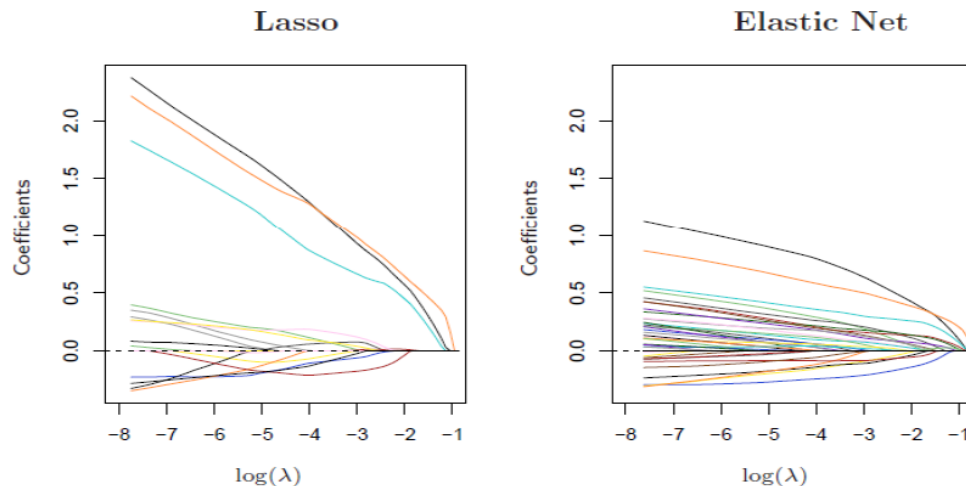




## 1) Publicly Available Statistical Software

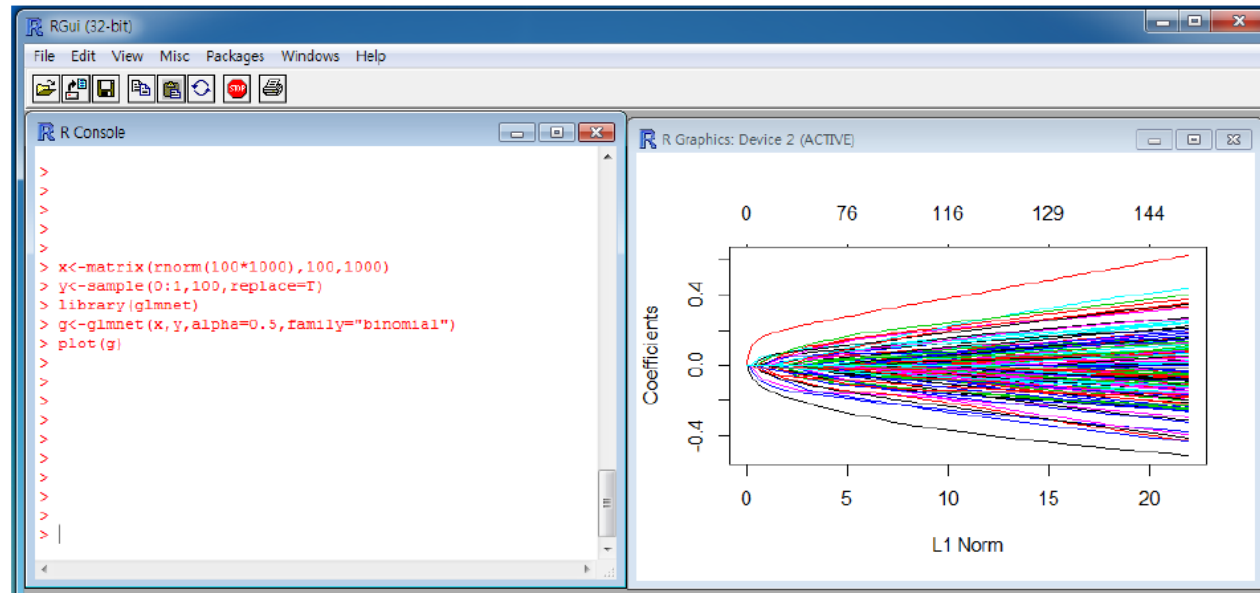
---

- R packages for regularization procedures
  - ‘glmnet’ : elastic net solution with a regularization path
  - ‘grpreg’ : regularization with concave penalty functions
  - ‘lars’ : lasso solution with a regularization path
  - ‘penalized’ : elastic net and fused lasso solution
  - ‘pclogit’ : network-based penalized logistic regression



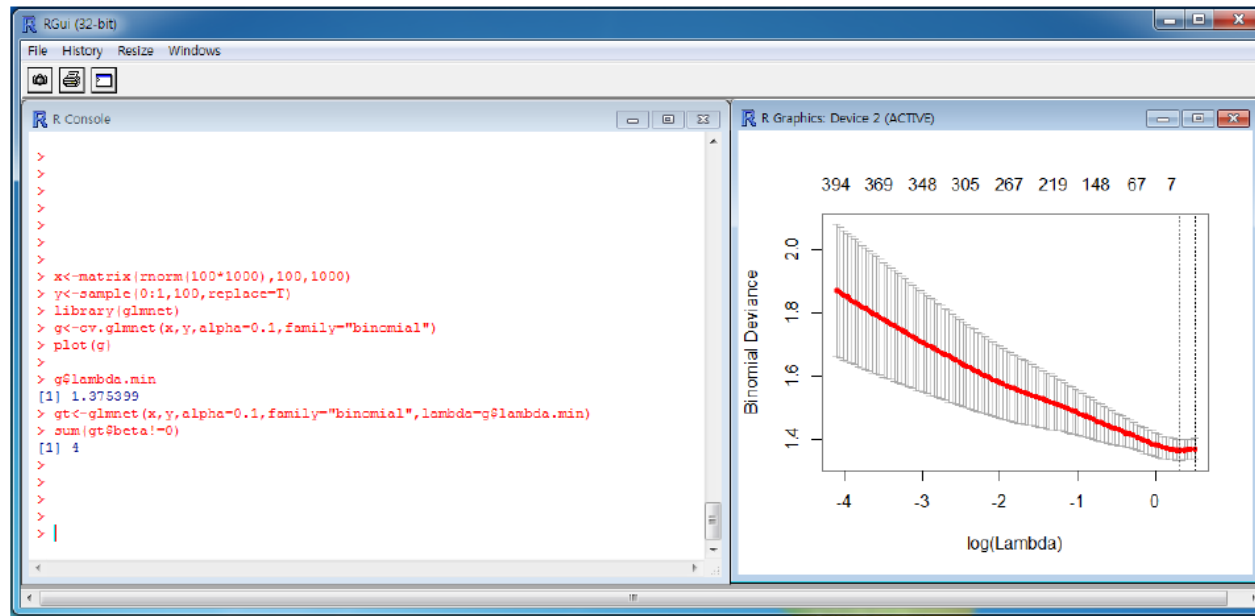
## 2) R package 'glmnet'

- High-dimensional data analysis with 'glmnet'
  - $x$  :  $n \times p$  matrix (genomic data)
  - $y$  :  $n$ -dimensional vector (phenotype outcome)
  - alpha : controls smoothness between 0 (ridge) and 1 (lasso)
  - family : “binomial” for case-control data



## 3) Tuning Parameter Selection

- Cross validation
  - Data is separated by training set and test set
  - K-fold cross validation
    - For a fine grid of  $\lambda$ , deviance is compared.
    - Pick the best  $\lambda$  that minimizes deviance or other criterion.



## 4) Code

---

```
# install.packages("glmnet")  
library(glmnet)
```

```
set.seed(101010)
```

```
x <- matrix( rnorm(100*1000), 100, 1000 )  
y <- sample(0:1, 100, replace = T)  
g <- glmnet(x, y, alpha = 0.5, family = "binomial")  
plot( g )
```

```
g <- cv.glmnet(x, y, alpha = 0.1, family = "binomial")  
plot( g )  
g$lambda.min  
gt <- glmnet(x, y, alpha = 0.1, family = "binomial", lambda = g$lambda.min )  
sum( gt$beta != 0 )
```