# Variable Selection and Regularized Methods [1]

Shaobo Li

University of Kansas

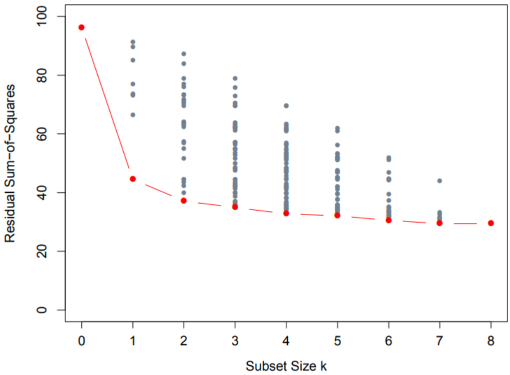
[1] Partially based on Hastie, et al. (2009) ESL, and James, et al. (2013) ISLR

# Introduction

- Variable selection – excluding unnecessary variables
  - Interpretation and simplicity
  - Prediction stability and accuracy
  - Bias-variance tradeoff
- Common approaches
  - Stepwise selection (for parametric models)
  - Shrinkage method (also called *regularization*)
  - Dimension reduction (project $p$ predictors to an $m$-dimensional subspace)
- It can be subjective that some times domain knowledge may force certain variables to be included in the model.

# Best Subset Selection

- Search through all subset predictors




- Computationally expensive even infeasible (not commonly used)
  - How many models do we need to evaluate for 10 candidate $X$'s?

# Selection Criteria – AIC, BIC

- Akaike information criterion (AIC), the smaller the better

$$AIC = -2\log(\hat{L}) + 2p$$

- Bayesian information criteria (BIC), the smaller the better

$$BIC = -2\log(\hat{L}) + \log(n)p$$

  where $\hat{L}$ is estimated likelihood function
- For linear regression, $-2\log(\hat{L})$ is equivalent to RSS
- BIC weighs more on $p$ comparing to AIC. What does this mean?

# Forward, Backward, and Stepwise Selection

- Computationally less expensive than best subset
  - practically useful when the number of $X$ is not too large
- Iteratively adding or dropping one variable at a time
- Forward/backward is **greedy** procedure. That is, they won't adjust any added/dropped variables in previous step
- Stepwise: start with forward, and then iteratively add and drop variables
- Commonly used selection criteria: AIC, BIC
- R function: `"step()"`
- An illustration: click here

# High-Dimensional Regression

- Number of predictor is very large (even larger than sample size)
- Ultra-high dimension $p \gg n$
- It is very common for gene expression and image data
- Sparsity assumption: only a few predictors are relevant
- OLS fails when $n < p$. Why?
- LASSO or similar methods provide sparse solution

# Shrinkage Methods

- Also called penalized or regularized method.
- Shrink the regression coefficients toward 0 by constraints (regularization)
- Computationally efficient (especially for high dimensional data)
- Estimates are usually *biased*
- A game of bias-variance tradeoff
- Popular shrinkage methods:
  - $L_1$ penalty: LASSO, adaptive Lasso, SCAD, MCP
  - $L_2$-norm penalty: Group Lasso (when $X$ has group structure)
  - $L_2$ penalty: Ridge regression (not for variable selection)
  - $L_1 + L_2$ penalty: Elastic Net

# LASSO

- Least absolute shrinkage and selection operator (LASSO)
- Introduced by Tibshirani (1996)
- One of the most popular variable selection methods
- It estimates the coefficients and selects variables simultaneously.
- A tuning parameter $\lambda$ controls the "power" of selection.
- Need to standardize all predictors in shrinkage estimation. Why?

# LASSO

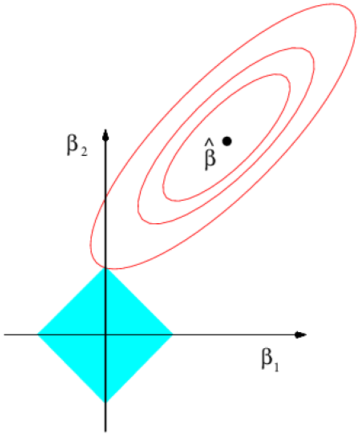
- LASSO solves the *$(L_1)$ penalized least square*

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

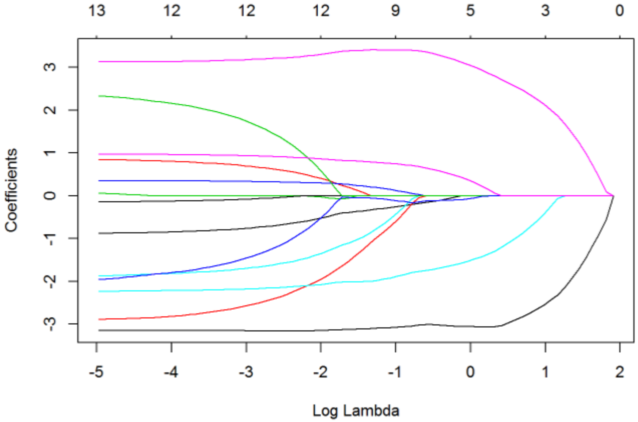
- It is a *convex optimization* problem
- It is equivalent to solve a constrained optimization problem

$$\begin{aligned} \min & \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ \text{s.t. } & \sum_{j=1}^{p} |\beta_j| = a \end{aligned}$$

# An Illustration

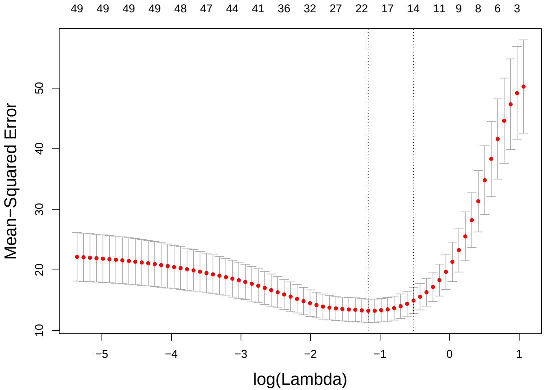


# LASSO Regression Solution Path – Boston Housing Data



# Tuning Parameter $\lambda$ Selection

- $\lambda$ controls the shrinkage level (different *lambda* associates with different estimated model)
- Cross-validation
  - In R, use the function `cv.glm()` in package `glmnet`



# Ridge Regression

- Recall least square. We solve the optimization

$$\hat{\beta}_{LS} = \arg\min_{\beta} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

- Ridge regression solves a *($L_2$) penalized least square*

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- $\lambda$ is a tuning parameter, called shrinkage parameter
- Writing in matrix form, we can get the analytical solution

$$\hat{\beta}_{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad \text{(Exercise: Show it!)}$$

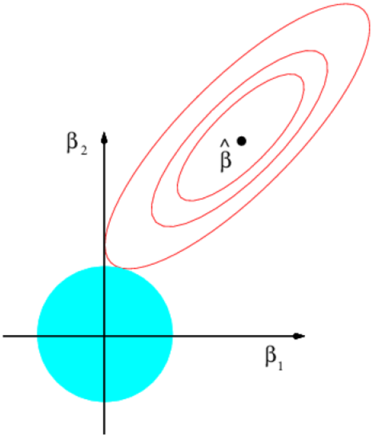
# Ridge Regression

- It is equivalent to solve a constrained optimization problem

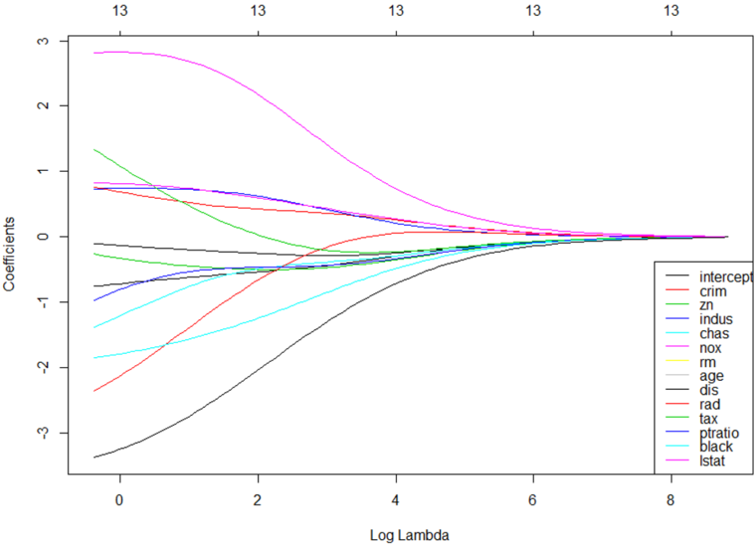
$$\min \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$
$$\text{s.t.} \ \sum_{j=1}^{p} \beta_j^2 = a$$

- $a$ corresponds to the tuning parameter $\lambda$

# An Illustration



# Ridge Regression Solution Path – Boston Housing Data



# Elastic Net Regression

- Introduced by Zou and Hastie (2005)
- Combination of Ridge and LASSO

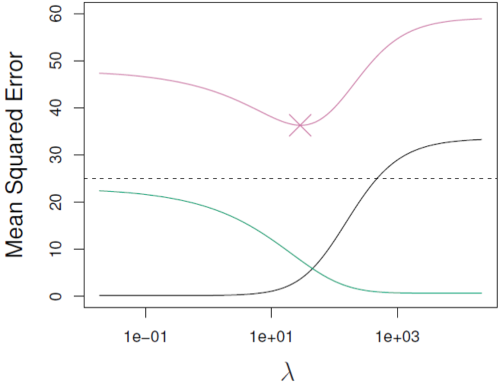
$$\hat{\beta}_{EN} = \arg\min_{\beta} \sum_{i=1}^{n} \left( y_i - \mathbf{x}_i^T \beta \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

- Convex optimization
- Ridge and LASSO are special cases of Elastic Net
- It incorporates the advantages of both Ridge and LASSO
  - Ridge regression: lower variance; multicollinearity
  - LASSO: variable selection (selects at most $n$ variables if $p > n$)

# Some Variants of LASSO

- $L_1$ penalty:
  - LASSO (Tibshirani, 1996)
  - Adaptive-LASSO (Zou, 2006)
  - SCAD (Fan and Li, 2001)
  - MCP (Zhang, 2010)
- $L_2$-norm penalty (when $X$ has group structure):
  Group Lasso (Yuan and Lin, 2006)
- $L_1 + L_2$ penalty:
  - Elastic Net (Zou and Hastie, 2005)

# Bias-Variance Tradeoff




Simulated data with n=50 observations, p=45 predictors, all having nonzero coeficients. Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set.