BSAN 400 Introduction to Machine Learning

Lecture 3. Linear Regression and Variable Selection[1]

Shaobo Li

University of Kansas

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

Linear Regression – a fundamental learning algorithm

- Supervised learning method
- It assumes the dependence of Y on X is linear
- Largely used in many disciplines
- Simple and interpretable
- Fundamental in data science

Linear Regression Models

- Simple linear regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Multiple linear regression

$$Y = \beta_0 + \beta_1 X + \ldots + \beta_p X_p + \epsilon$$

- Y: dependent variable (response, outcome)
- X's: independent variable (covariates, explanatory variable)
- β's: regression coefficients
- ϵ: random error (irreducible error)

Linear Regression Models

- Using matrix format

$$Y = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$ is called design matrix with first column being 1's
- The *estimated linear regression model* is

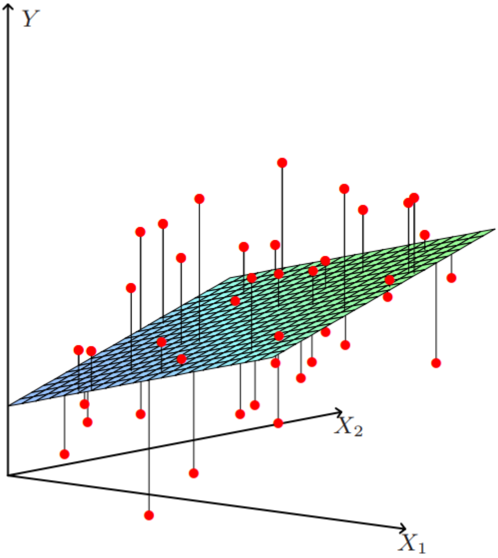
$$\hat{Y} = \mathbb{E}(Y|X) = \mathbf{X}\hat{\beta}$$

- Goal: estimate regression coefficient β

Model Assumptions

- $\mathbb{E}(Y|X)$ is a linear function of X or its basis expansion such as $X_1^2, X_2^3, \ldots$
- The error term $\{\epsilon_i, \ldots, \epsilon_n\} \overset{i.i.d.}{\sim} N(0, \sigma^2)$

Least Square Solution



Least Square Solution

- We want to minimize residual sum squares

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2$$
$$= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

- Take first-order derivative with respect to β and set to 0

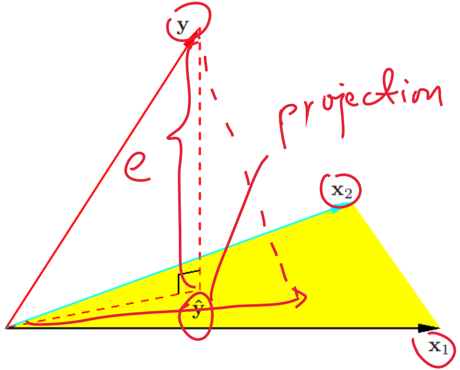
$$0 = \frac{\partial RSS(\beta)}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$
$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta$$

$$\underset{p\times 1}{\beta} = \underset{p\times p}{(X^TX)^{-1}} \underset{p\times n}{X^T} \underset{n\times 1}{y}$$

- This is called *normal equation*.

Least Square Solution

- By assuming $p < n$, the solution is $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- The predicted value is $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called hat matrix or projection matrix



Model Assessment – R Square and MSE

- It is proportion of variation in Y explained by the model

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 increases monotonically as number of explanatory variable increasing.

- Adjusted R^2

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}\frac{RSS}{TSS}$$

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n-p-1} \times RSS$$

It is an unbiased estimate of σ^2 for irreducible error ϵ.

Model Assessment – AIC, BIC, Cp, CV

- 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. → RSS

- Mellow's C_p is the same as AIC for linear regression
- Cross-validation error (CV score)

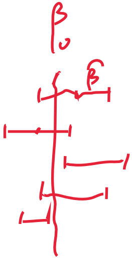
Hypothesis Testing – Test individual coefficients

Is a specific X relevant?

- Testing for individual β
  - H_0: $\beta_j = 0$; H_1: $\beta_j \neq 0$
  - Using T-test since the true variance is unknown

$$T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \sim t_{n-p-1}$$

  where v_j is the jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$
  - Reject H_0 if p-value $< \alpha$ or $|T| > T_{1-\alpha}^{(n-p-1)}$
- Confidence interval: $\hat{\beta} \pm se(\hat{\beta}) \times T_{1-\alpha}^{(n-p-1)}$



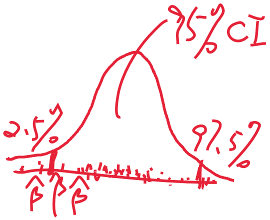
Hypothesis Testing – Test multiple coefficients

- F-test for overall significance
  - H_0: $\beta_1 = \ldots = \beta_p = 0$; H_1: at least one $\beta \neq 0$
  - F statistics

$$F^* = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$, is total sum squares

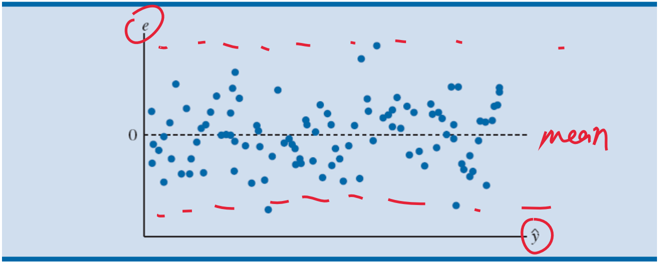
Bootstrap



- Resampling method
- A powerful tool to quantify uncertainty
  - standard error
  - confidence interval
- Random sampling with replacement
- Example:
  - train a model with 1000 bootstrap samples
  - store all the parameter estimates
  - calculate standard error and confidence interval

Model Diagnostics

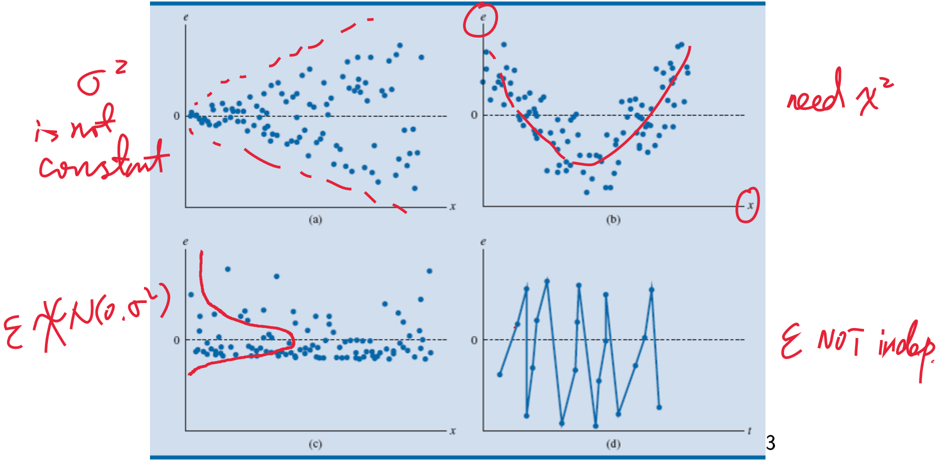
- Check the assumptions on the error term
  - Independent normal distribution?
  - $\mathbb{E}(\epsilon_i) = 0$?
  - $Var(\epsilon_i) = \sigma^2$ =constant?
- Residual plot (an ideal residual plot looks like this)[2]



[2] source: Camm, et al., *Essentials of Business Analytics*

Residual plot

Which type of assumption is violated?



[3]

[3] source: Camm, et al., *Essentials of Business Analytics*

Other Diagnostic Plots

- Normal Quantile-Quantile Plot
  - It plots the standardized residual vs. theoretical quantiles
  - An easy way to visually test the normality assumption
  - If residual follows normal distribution, you should expect all dots lie on the diagonal straight line.
- Residual-Leverage Plot
  - This plot checks if there are any influential points, which could alter your analysis by excluding them
  - The points that lie outside the dashed line, Cook's distance, are considered as influential points

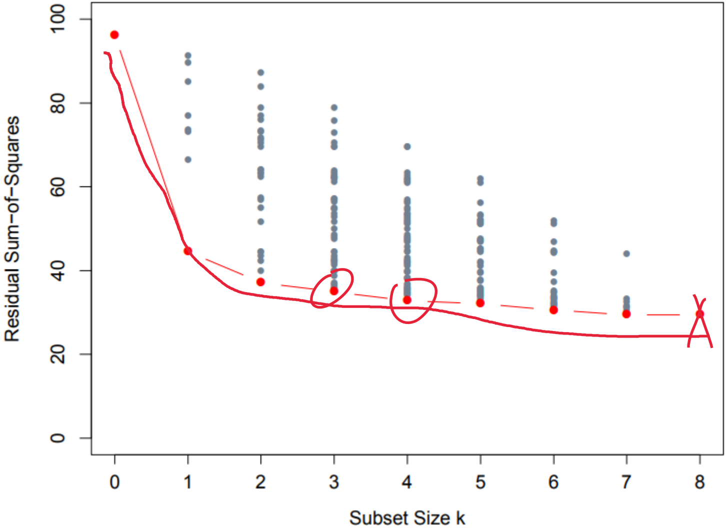
Variable Selection Methods For Linear Regression

- Why? Try to improve the model: exclude unnecessary predictors
  - Interpretation and simplicity
  - Prediction stability and accuracy
  - Less computational cost
  - Bias-variance tradeoff
- Common approaches
  - Subset selection
  - Shrinkage (also called *regularization*)
  - Dimension reduction (project p predictors into a M-dimensional subspace)
- Some times it is subjective, and needs domain knowledge so that certain variable has to be in the model.

Best Subset Selection

- Select the best subset of predictors such that the model is optimal in terms of a certain assessment metric
- Computationally expensive even infeasible
  - *leaps and bounds* (an R package "leaps") algorithm makes it feasible for p as large as 30 or 40.
- Suppose there are 10 predictors. How many models need to be fitted and evaluated?

Example of Best Subset Selection



Forward, Backward, and Stepwise Selection

- Computationally less expensive than best subset
- 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
- Selection criteria: AIC or BIC
- R package: `"step"`
- An illustration: click here

Shrinkage Methods

- Also called penalized estimation
- Shrink the regression coefficients toward 0 by constraints (regularization)
- Shrinkage methods are always preferred over best subset or stepwise methods. Why?
- A game of bias-variance tradeoff
- We discuss two popular shrinkage methods:
  - Ridge regression
  - LASSO

Ridge Regression

- Recall least square. We solve the optimization

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

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

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

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

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

Ridge Regression

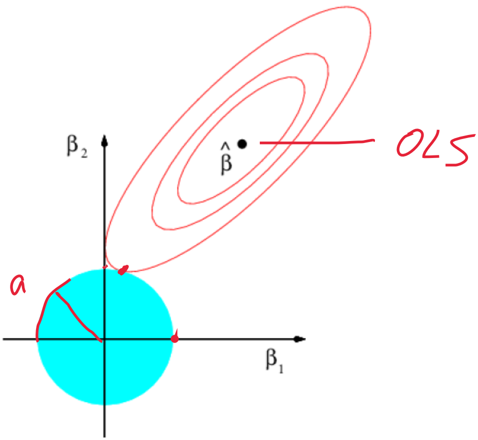
- It is equivalent to solve a constrained optimization problem

$$\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad LS$$

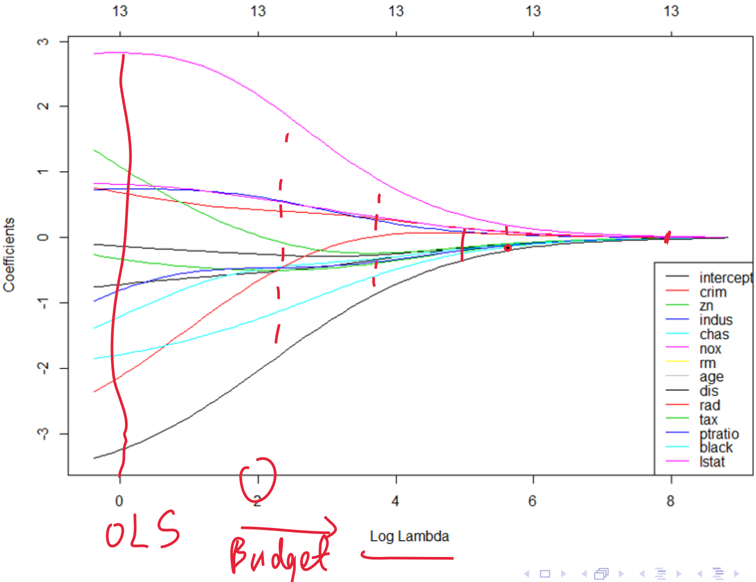
$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 = a \quad \text{constraint}$$

- a corresponds to the tuning parameter λ

An Illustration



Ridge Regression Solution Path – Boston Housing Data



LASSO

- Least absolute shrinkage and selection operator (LASSO)
- Introduced by Tibshirani (1996)
- Shrinkage estimation
- It estimates the coefficients and perform variable selection simultaneously

LASSO

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

Ridge $|\beta_j|^2$

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

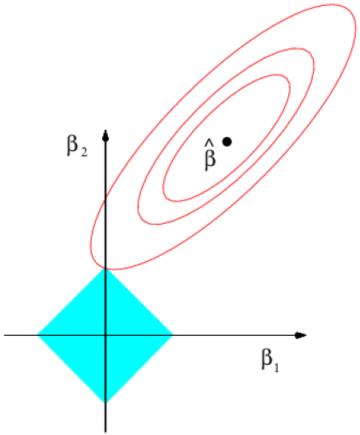
- It is a *convex optimization* problem
- It is equivalent to solve a constrained optimization problem

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

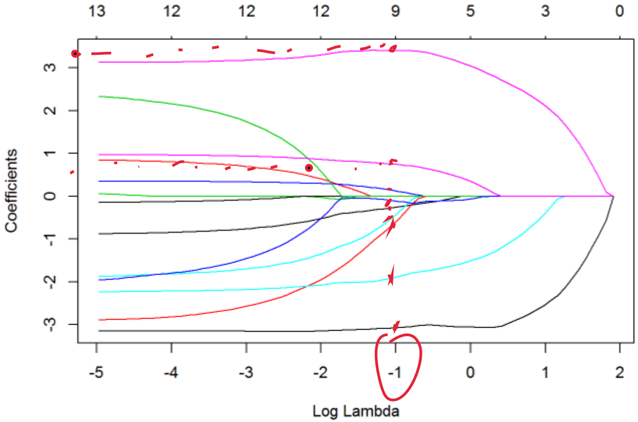
$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| = a$$

Budget Constraint

An Illustration

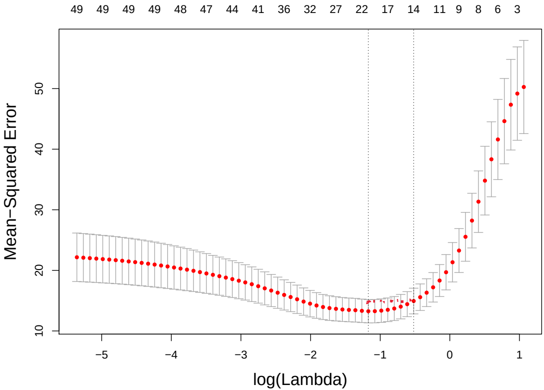


LASSO Regression Solution Path – Boston Housing Data



Tuning Parameter λ Selection

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



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

Elastic Net Regression

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

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

- 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 Comments

- It is recommended to standardize all predictors in shrinkage estimation. Why?
- Solution of Ridge regression is equivalent to the posterior of Bayesian estimates
- There are many other type of penalized estimators with different penalty functions that can perform variable selection.
  - Group Lasso (Yuan and Lin, 2006)
  - Adaptive-LASSO (Zou, 2006)
  - SCAD (Fan and Li, 2001)
  - MCP (Zhang, 2010)