Regression & Penalized model

김민준, 최병준, 최석영

목차
CONTENTS

Linear regression

Regression

- $y = B_0 + B_1x_1 + B_2x_2 + \cdots + \epsilon \longrightarrow \quad y = XB + \epsilon$

- 주어진 데이터로, y를 다른 x변수들의 함수 형태로 나타냄

- 가정 : 독립성, 정규성, 등분산성

- $B_0, B_1, B_2, \cdots$ 를 구하자

Linear regression

OLS

- $\hat{\beta}$ that minimizes squared loss function, which is sum of squared errors(잔차)
- e = $y - \hat{y}$ = $y - XB$

OLS Estimate $\mathbf{b} = \hat{\beta}$ of β is the estimate so that

$$\begin{aligned} Q(\beta) &= \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1}X_{i,p-1})^2 \\ &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) \end{aligned}$$

is minimized with respect to β, that is,

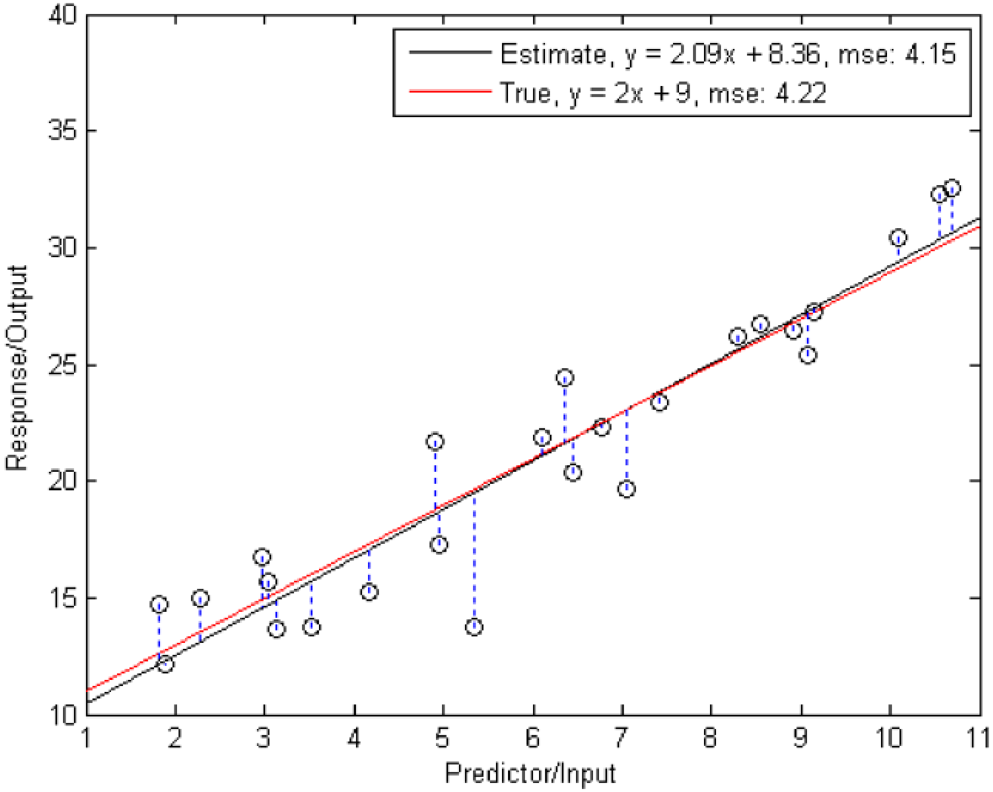
$$\begin{aligned} \mathbf{b} &= \hat{\beta} = \arg\min_{\beta}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \end{aligned}$$

BLUE
(Best Linear Unbiased Estimator)
: Among the unbiased estimators, BLUE has the smallest variance

Linear regression

OLS

- $\hat{\beta}$ that minimizes squared loss function, which is sum of squared errors(잔차)



Linear regression

MLE

- MLE, Maximum Likelihood Estimation

If

$$X_i \sim F(\Theta), i = 1 \dots n$$

then the likelihood function is

$$\mathcal{L}(\{X_i\}_{i=1}^{n}, \Theta) = \prod_{i=1}^{n} F(X_i; \Theta)$$

The likelihood function specifies **how likely the observed data is** for various possible values **for possible parameters**.

Linear regression

MLE

Under normal error assumption, minimizing squared error is equivalent to maximizing the likelihood function.

Thus, MLE and OLS estimator(Least Square Estimator) results in the same estimate

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, where\ \epsilon \sim N(0, \sigma^2)$ $\quad f\left(x \middle| \mu, \sigma^2\right) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Likelihood function

$$L(Y_1, \cdots Y_n; \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left(\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n} (Y_i - \beta_1 X_i - \beta_0)^2\right)\right)$$

Squared loss function

$$\sum_{i=1}^{n} (Y_i - \beta_1 X_i - \beta_0)^2$$

Limitation of linear regression

1. Prediction accuracy

- Prediction accuracy; the OLS estimates often have low bias but large variance.

$$E[(y - \hat{f}(x))^2] = \mathrm{Bias}[\widehat{f}(x)]^2 + Var[\widehat{f}(x)]$$

$$\mathrm{Bias}[\widehat{f}(x)] = E[\hat{f}(x) - f(x)]$$

$$Var[\widehat{f}(x)] = E[(\hat{f}(x) - E[\hat{f}(x)])^2]$$

Limitation of linear regression

2. High dimensionality problem

If $p \gg n$,
Then the problem of dimensionality arises

For linear regression, intuitively it is easy to understand by the problem of solving equation with unknown p variables(need p values at least)

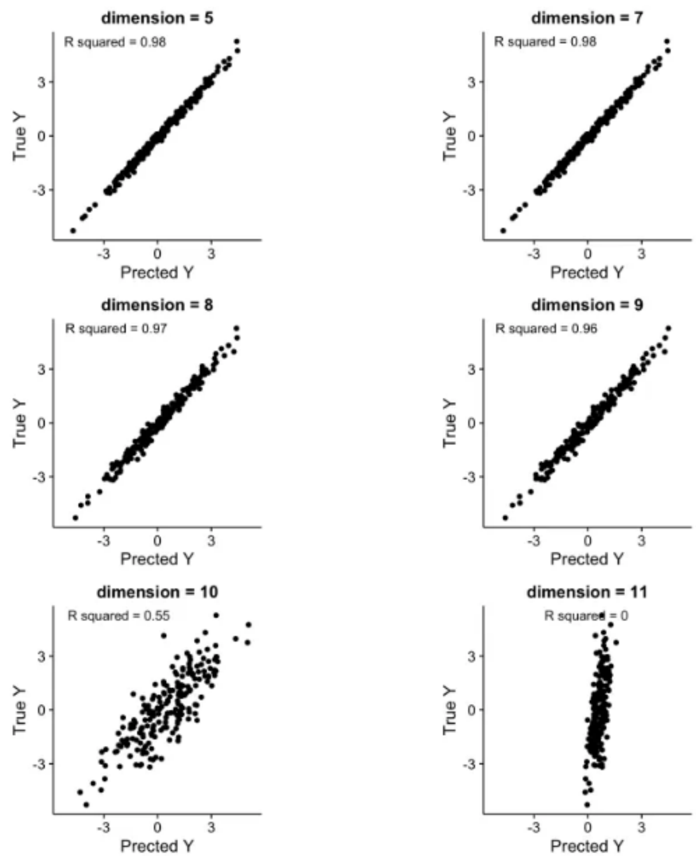
If $p \gg n$, there are more than one solution, so the parameters are can be more than one

ordinary least-squares regression (OLS), which minimizes the residual sum of squares $RSS = (\tilde{y} - XB)'(\tilde{y} - XB)$ where $\tilde{y} = y - \bar{y}1_n$, will yield an estimator that is not unique since X is not of full rank

Limitation of linear regression

2. High dimensionality problem

High dimensionality problem happens in the similar way.
→ variance on beta estimates increases!



Limitation of linear regression

3. Multicollinearity

The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model

Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model. You might not be able to trust the p-values to identify independent variables that are statistically significant

Limitation of linear regression

Solution : subset selection

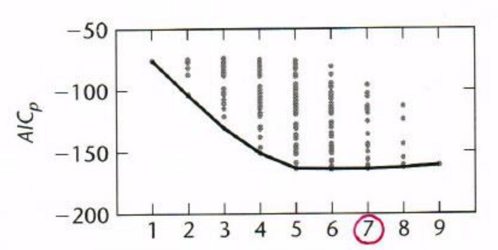
Criteria : CP, AIC, BIC, adj R^2,⋯ (MSE)

Forward selection(start from null model, +)
Backward elimination(start from full model, -)
Bidirectional elimination(combination of forward & backward)

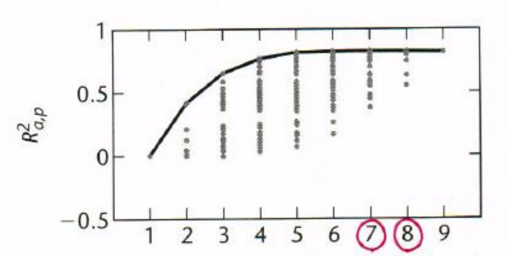
Method of choosing a subset of important explanatory variables

$$AIC = nlog(\frac{SSR}{n}) + 2p$$

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$



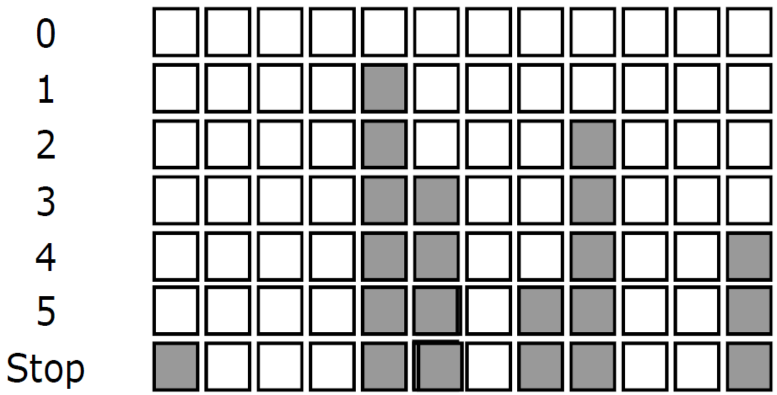
(d)



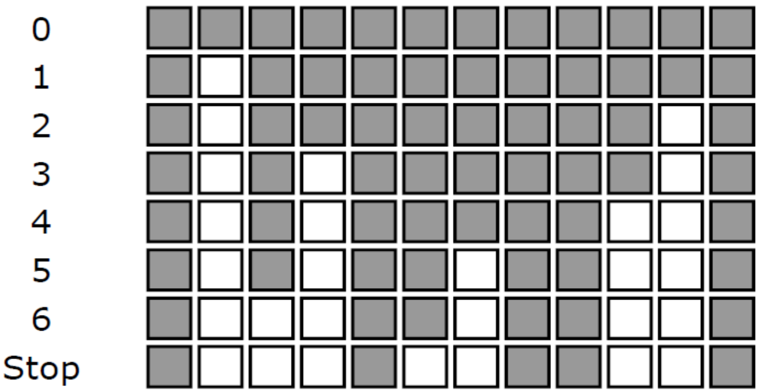
(b)

Limitation of linear regression

Solution : subset selection



Forward selection



Backward elimination

Limitation of linear regression

Limitation of subset selection

1. Computational infeasibility
The number of possible subsets grows exponentially with p
Today's computers can only search all possible subsets for p around 40~50.
-> Using forward/backward/stepwise selection don't have this problem, but they don't search for all possible subsets.

2. Instability
Subset selection is discontinuous, so even the small change in data can result in completely different estimates.
Especially in high dimensions, subset selection is thus often unstable and highly variable

3. Multicollinearity problem
When multicollinearity is present, important variables can appear to be non-significant and standard errors can be large

4. Overfitting problem
Because it is a discrete process, where variables are either retained/discarded, it often exhibits high variance, and so doesn't reduce the prediction error of the full model.

Limitation of linear regression

Limitation of subset selection in statistical viewpoint

It is a common practice to report inferential results from ordinary least squares models following subset selection as if the model had been prespecified from the beginning.

This is unfortunate, as the resulting inferential procedures violate every principle of statistical estimation and hypothesis testing:

Test statistics no longer follow t/F distributions
Standard errors are biased low, and confidence intervals falsely narrow
p-values are falsely small
Regression coefficients are biased away from zero

Limitation of linear regression

Solution : VIF

VIF is a diagnostic tool, so it is usually used after the model decision

Variance Inflation Factor(VIF)

Original model: $y = B_0 + B_1 x_1 + B_2 x_2 + \cdots + \epsilon$

$$VIF_i = \frac{1}{(1 - R_i^2)}$$

Where R_i^2 is coefficient of determination of X_i when it is plugged in as y
Ex) R_1^2 : $coefficient\ of\ determination\ of\ \mathrm{X}_1 = \beta_0 + \beta_2 x_2 \cdots + \epsilon$

Limitation of linear regression

Limitation of VIF

1. Cutoff needs to be determined

A cutoff value of 4 or 10 is sometimes given for regarding a VIF as high.
But it is important to evaluate the consequences of the VIF in the context of the other elements of the standard error, which may offset it (such as sample size...)

2. Difficult to use in mixed data set

The indicator variables will necessarily have high VIF, even if the categorical variable is not associated with other variable in the regression models.

Limitation of linear regression

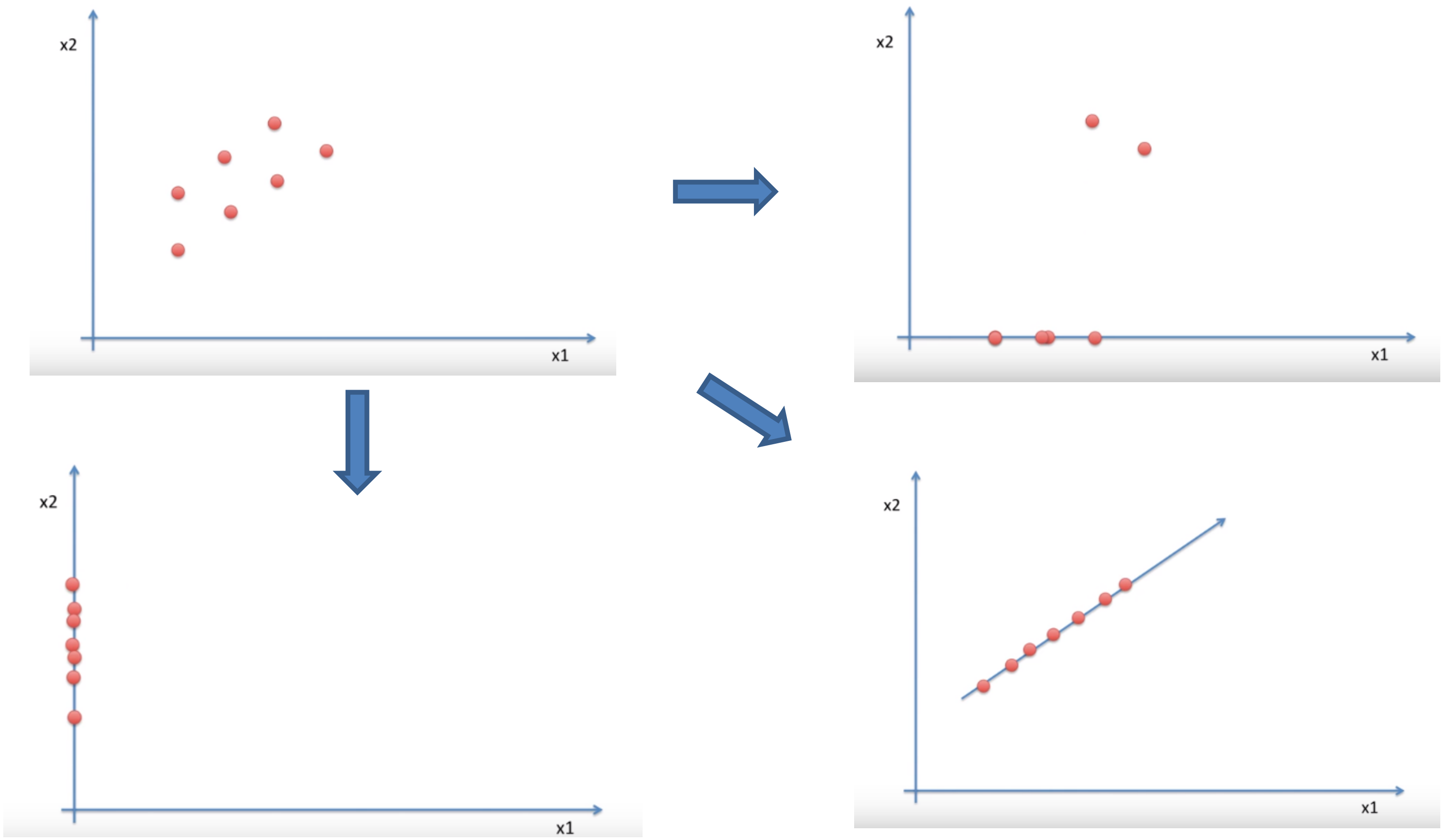
Solution : PCA

Principal component analysis(PCA)

- Variable reduction technique
- Used when variables are highly correlated
- Principal components retained account for a maximal amount of variance of observed variables
- Reduces the number of observed variables to a smaller number of principal components which account for most of the variance of the observed variables
- A large sample procedure
- Components are uninterpretable(no underlying constructs)
- Feature Scaling

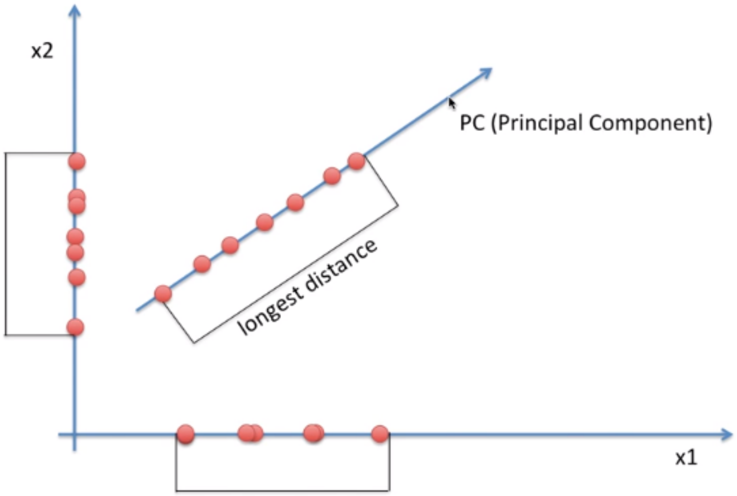
Limitation of linear regression

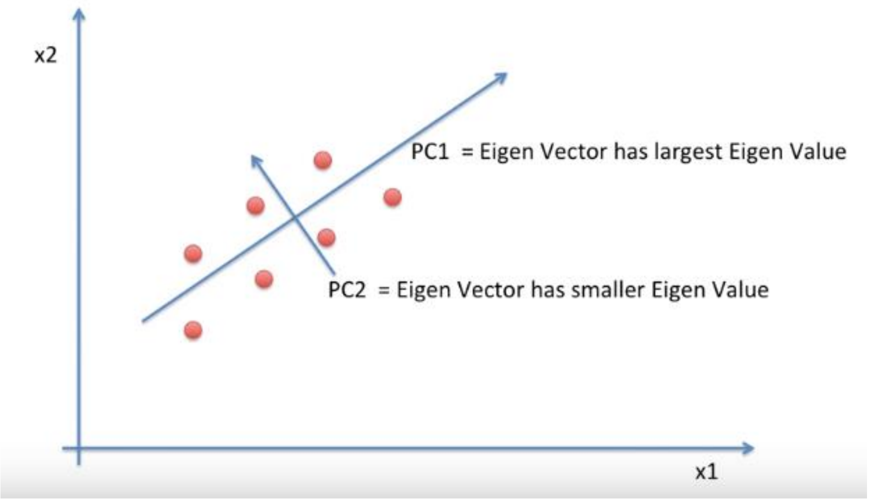
Solution : PCA



Limitation of linear regression

Solution : PCA





Limitation of linear regression

Limitation of PCA

1. Difficult to interpret
Using PCA for dimension reduction in regression ignores the relationship between X and y
Often difficult to interpret p variables and the derived principal components

2. Difficult to use in mixed data set
Cf) FAMD(PCA method for mixed data)

Penalized model

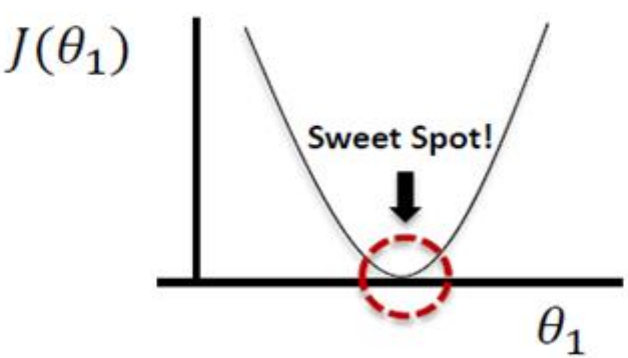
Cost function

Cost function : (예측 값 – 실제 값)의 크기를 나타내는 함수

Linear regression cost function = squared loss function

$$\hat{y}_i = \theta_0 + \theta_1 x$$

$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2$$



Penalized model

Cost function

First, consider the simplest form of the penalized regression cost function, where the penalty is given in squared form

Penalized Regression Cost function

$$J(\theta_1) = \boxed{\sum_{i=1}^{m} (y_i - \theta_1 X)^2} + \lambda \theta_1^2$$

The Sum of the squared residuals

Penalty to the traditional Least Squares method

λ determines how severe that penalty is.

λ can be any value from 0 to positive infinity.

Penalized model

Objective function

The concept of minimizing the cost function is equivalent to maximizing the likelihood function

In Penalized model, instead of maximizing likelihood function $l(\theta|x)$, we maximize the function

$$M(\theta) = \ell(\theta|\mathbf{x}) - \lambda P(\theta)$$

P is a function that penalizes what one would consider less realistic values of the unknown parameters

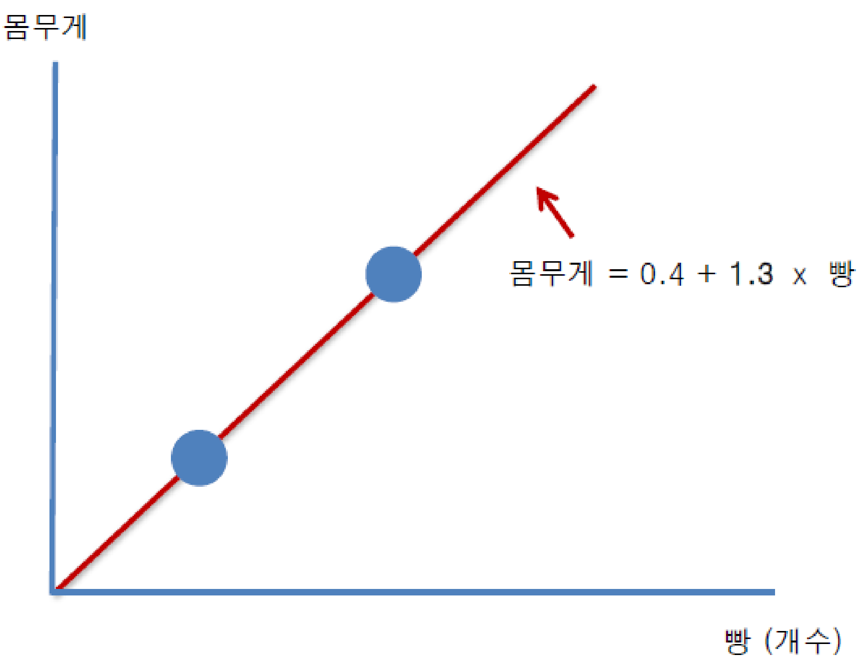
λ controls the tradeoff between the two parts

The function M is called the *objective function*

Penalized model

Cost function

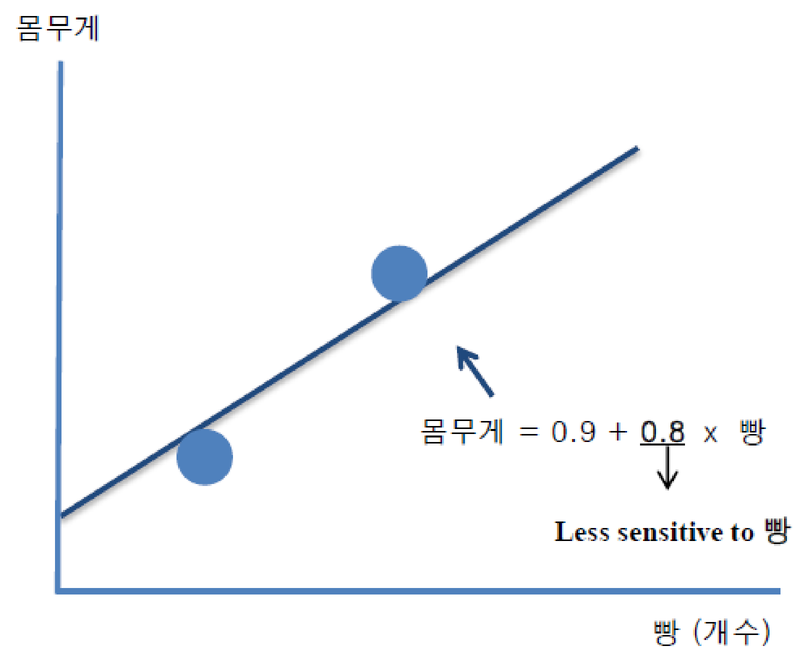
Linear Regression



the sum of squared residuals + λ x $\beta_1{}^2$

= 0 + 1.69 = 1.69 (set λ =1)

Penalized Regression



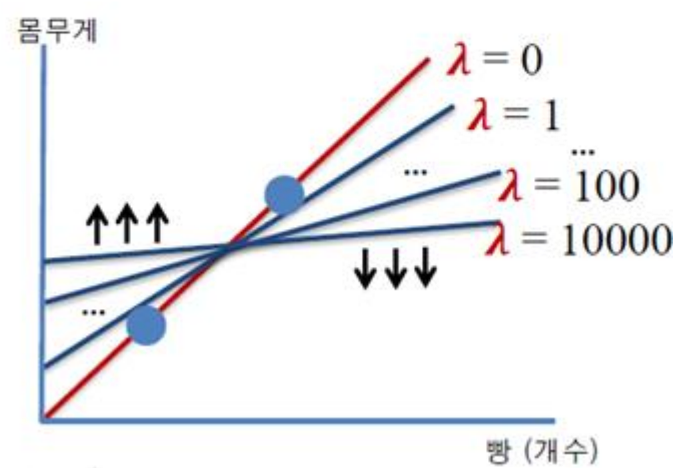
the sum of squared residuals + λ x $\beta_1{}^2$

= 0.09 + 0.01 + 0.64 = 0.74 (set λ =1)

Penalized model

Understanding Cost function of penalized model(=objective function)

$$M(\theta) = \ell(\theta|\mathbf{x}) - \lambda P(\theta)$$



What does hyperparameter λ do?
: decide the sensitivity of $\hat{y}$ to the variable x
- If λ is large, then $\hat{y}$ is less sensitive to the change of value x

How?

$$\therefore\ \hat{\theta} = \frac{X^T Y}{X^T X + \lambda}$$

How do we choose λ?
: By cross-validation

Penalized model

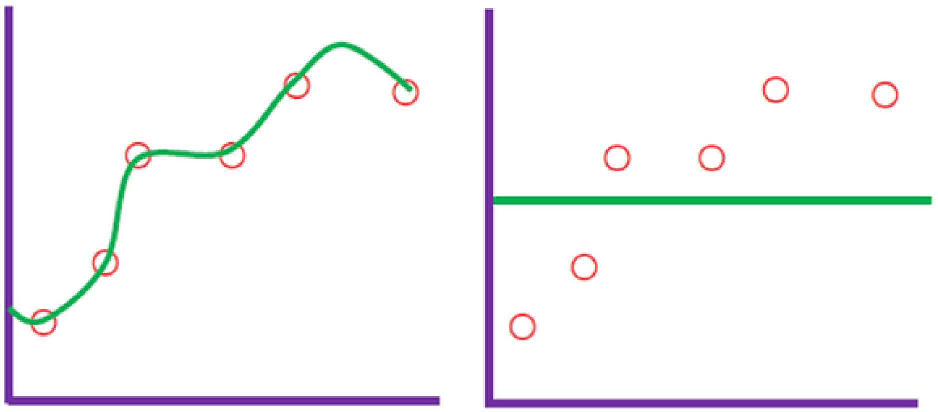
Advantages of Penalized model

Multicollinearity problem relieved

By reducing sensitivity of parameters to multicollinearity

Variance of estimators is reduced by giving up the small bias of the estimator (Bias-variance trade-off)

We call this technique **regularization**



Penalized regression

Comparison

Linear regression

minimize $\mathrm{RSS} = \sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2$

Isn't good estimator when

1. Number of parameters is large (high dimension)
2. Columns of X are highly correlated

Penalized linear regression

Minimize $\mathrm{RSS}(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}),$

$P_\lambda(\boldsymbol{\beta})$: penalty function,

λ : regularization parameter

Penalty function에 따라 다양한 종류

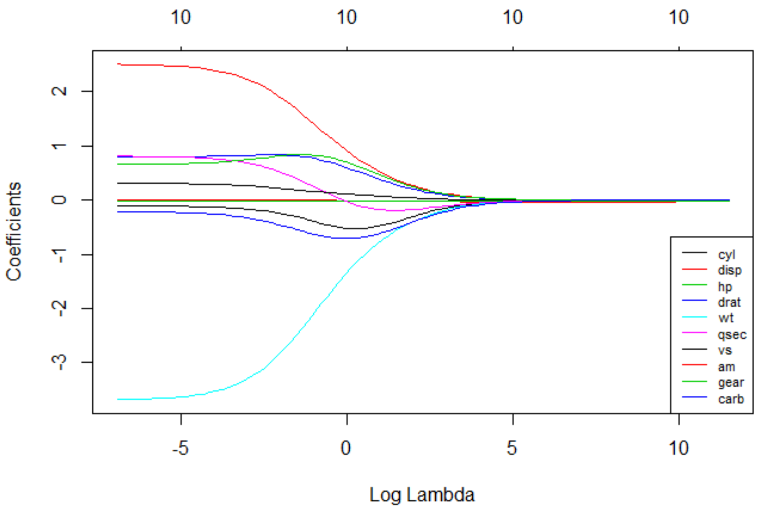
Penalized regression

Ridge

- Minimize $RSS(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2$ $\longrightarrow$ $\hat{\beta} = (X'X + \lambda I)^{-1} X'y$

(L2 Regularization, L2 penalty
전개 후 beta에 대해 미분)

As $\lambda \to 0$, $\widehat{\beta}^{\text{ridge}} \to \widehat{\beta}^{\text{OLS}}$
As $\lambda \to \infty$, $\widehat{\beta}^{\text{ridge}} \to \mathbf{0}$



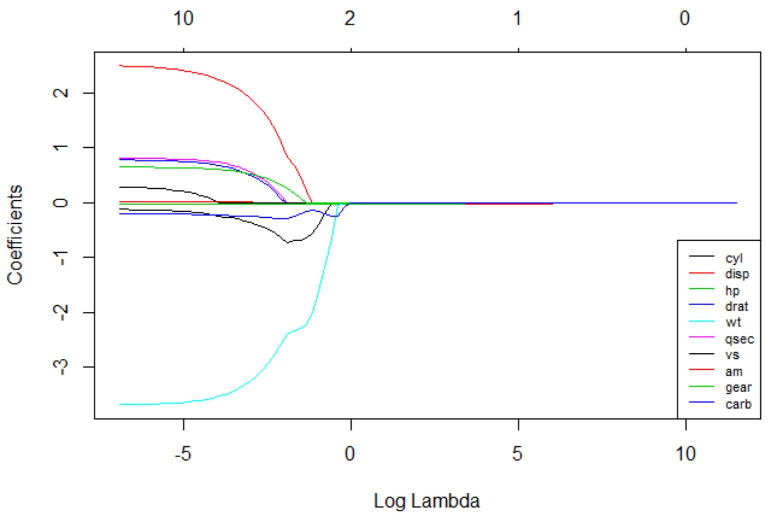
Penalized regression

Lasso

- least absolute shrinkage and selection operator
- Minimize $RSS(\beta) + \lambda \sum |\beta_j|$ $\longrightarrow$ $\textit{let}\ \boldsymbol{\beta} \geq \mathbf{0},$ $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}(\mathbf{X'Y} - \boldsymbol{\lambda})$

(L1 Regularization, L1 penalty
전개 후 beta에 대해 미분)

As λ → X'Y, $\hat{\beta}^{Lasso}$→ 0
(As λ increases, $\hat{\beta}^{Lasso}$→ 0)



Penalized regression

Ridge vs Lasso

Ridge

Beta can asymptotically close to 0

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\lambda I})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

→ cannot perform variable(feature) selection

L2 regularization

Prediction + overfitting + multicollinearity
(다양한 유의미한 변수 존재할 경우)

Lasso

Beta can be 0

$$\boldsymbol{let\ \beta \geq 0},$$
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\mathbf{Y} - \boldsymbol{\lambda})$$

→ can perform variable(feature) selection

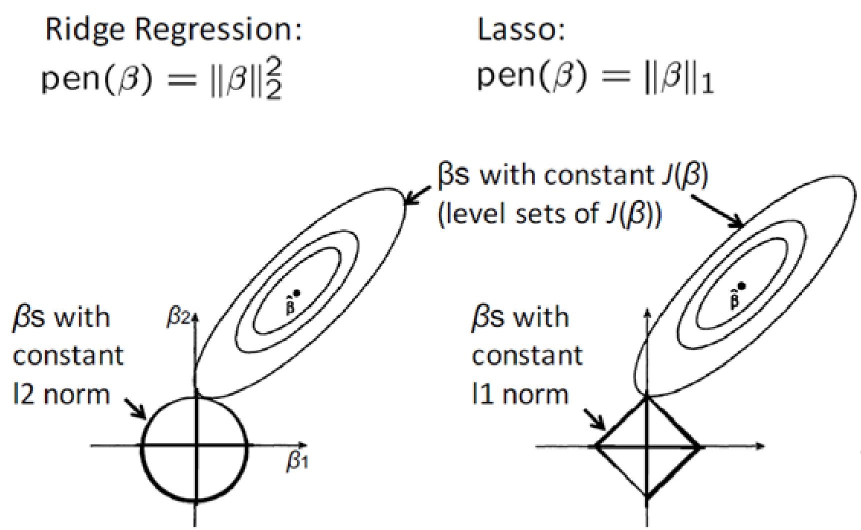
L1 regularization

Prediction + overfitting + variable selection
(많은 변수일 경우)

Penalized regression

Ridge vs Lasso

Ridge	Lasso
$RSS(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2$	$RSS(\beta) + \lambda \sum \lvert\beta_j\rvert$
$\longrightarrow$ $RSS(\beta) + \lambda(B^2{}_1 + B^2{}_2 + B^2{}_3 + \cdots)$	$\longrightarrow$ $RSS(\beta) + \lambda(\lvert B_1\rvert + \lvert B_2\rvert + \lvert B_3\rvert + \cdots)$



Penalized regression

Elastic Net

- When high dimensional & correlated data?
  Ridge : still complex model(can't eliminate variables)
  Lasso : when variables are correlated, eliminate many variables(data loss)

Ridge + Lasso!

Penalized regression

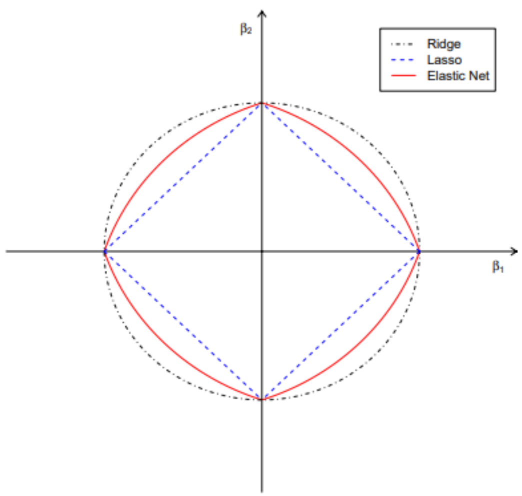
Elastic Net

- Minimize $RSS(\beta) + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum |\beta_j|$

Ridge, L2 Lasso, L1

parameter λ_1, λ_2

2-dimensional illustration $\alpha = 0.5$



Penalized regression

Selection of λ

- Penalized regression depends heavily on the regularization coefficient λ

how should we select λ ?

⟶ based on how well predictions $\hat{\beta}_\lambda$ do at predicting actual Y

같은 데이터로 모델을 만들고 성능을 평가하면 의미 X, Overfitting의 위험성

⟶ split the data set into two fractions, use one to fit(train set) and the other to evaluate(test set)

만약 데이터의 크기가 작다면?

Penalized regression

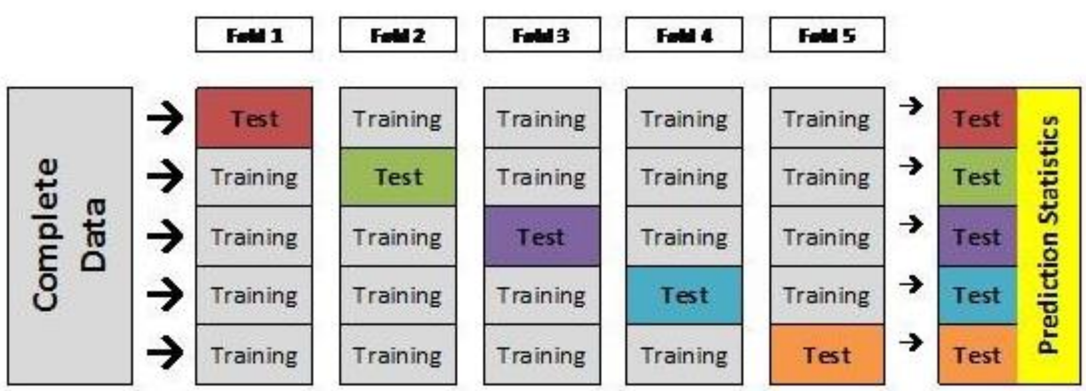
Cross validation

- 데이터의 크기가 작으면 test set에 대한 신뢰성 저하

→ 모든 데이터를 한 번씩 test set으로 사용하여 test set을 증가

cross-validation splits the data into K folds, fits the data on K-1 of the folds and evaluates risk on the fold that was left out

데이터에 계절성이 있는 경우, overfitting, 계산 많음

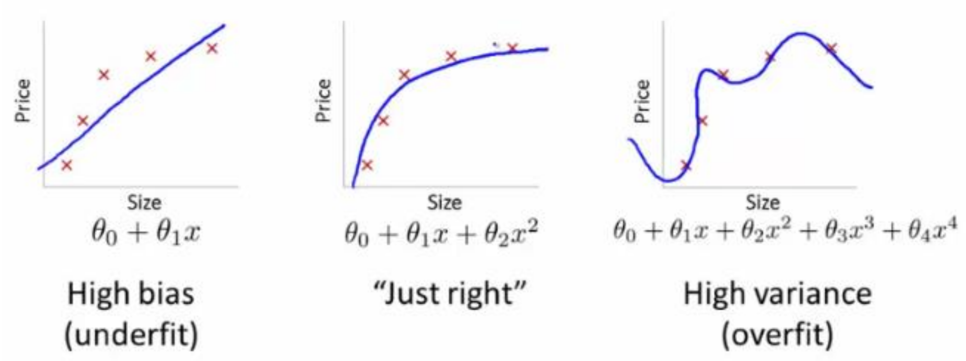


Penalized regression

Advantages

Penalty를 추가해 beta를 작게 만듦 → 더 단순한 모형으로

→ Overfitting 해결



Ridge

$$\hat{\beta} = (X'X + \lambda I)^{-1} X' y$$

Lasso

$$let\ \beta \geq 0,$$

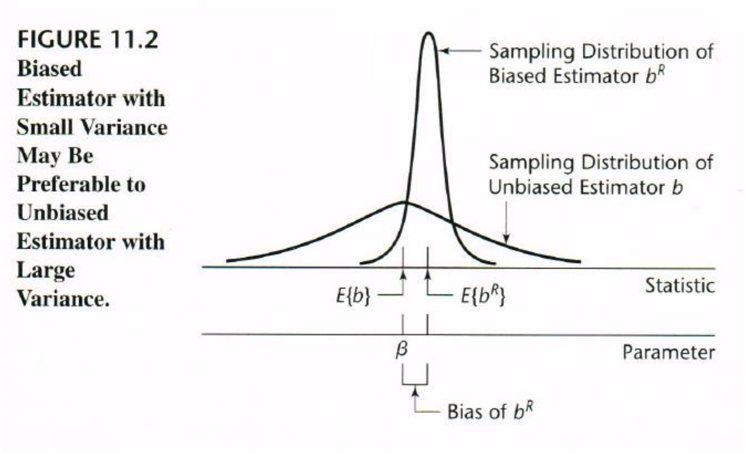
$$\hat{\beta} = (X'X)^{-1}(X'Y - \lambda)$$

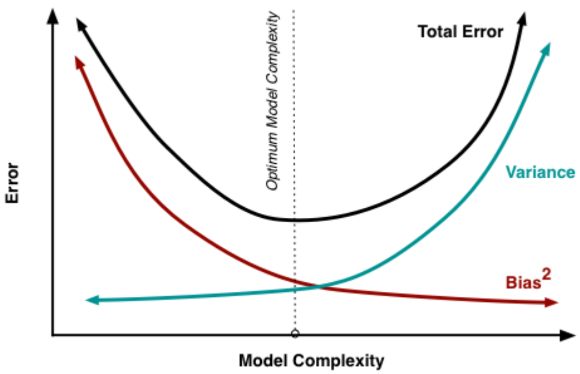
Penalized regression

Advantages

Multicollinearity : high variance on variables

→ Reduce model complexity, lower variance & higher bias





감사합니다