

Penalized Model

전 예 슬

목차

- 1 Limitations of Linear Regression
- 2 Cost Function
- 3 Penalized Regression Model
- 4 Extended Penalized Regression
- 5 Exercise with Python3

***1** Limitations of Linear Regression*

1-1. Linear Regression의 한계

■ 한계 1) 변수의 개수 (p) > 데이터 셋 (n)

→ ‘최소불편추정’ 원칙에 해당되는 추정값이 존재하지 않음 (\because the variance is infinite)

“The standard linear model (or the ordinary least squares method) performs poorly in a situation, where you have a large multivariate data set containing a number of variables superior to the number of samples.” (1)

■ 한계 2) 변수의 개수가 많은 경우 → ‘과적합(Overfitting) 문제 발생’

“If we have too many features, the learned hypothesis may fit the training set very well, but fail to generalize to new examples.” (2)

■ 한계 3) 변수간의 서로 연관되어 있는 경우 → ‘다중 공선성의 문제 발생’

“Maximum likelihood estimation has many wonderful properties, but is often unsatisfactory in regression problems for two reasons: Large variability: when p is large with respect to n , or when columns of X are highly correlated, the variance of $\hat{\beta}$ is large Lack of interpretability.” (1)

(1) Patrick Breheny, STA 603: Introduction to Linear Models

(2) Andrew Ng, Machine Learning Coursera

1-1. Linear Regression의 한계

Linear Regression의

‘변수가 많을 경우 발생할 수 있는 Overfitting / 다중 공선성’ 등의

한계를 해결하기 위한 방안이 필요해,

Stepwise 등 변수를 없애는

Subset Selection을 사용하면 되지 않을까?

1-2. Subset Selection의 한계

■ 한계 1) 변수의 개수 (p)가 많을 경우 → 계산해야 할 경우의 수가 많음

“One problem with this approach is that the number of [all possible subsets] grows exponentially with p and is not computationally feasible for p much larger than about 40 or 50.” (1)

■ 한계 2) 변수의 값에 영향을 많이 받음 → ‘작은 변화로도 다른 모델로 산출’

“Another problem is the fact that subset selection is discontinuous, in the sense that an infinitesimally small change in the data can result in completely different estimates.” (1)

■ 한계 3) 데이터 특성 (Dummy variable)에 영향을 받음 → ‘유의한 변수 발굴 한계’

“Some variable (especially dummy variables) may be removed from the model, when they are deemed important to be included. These can be manually added back in.” (2)

■ 한계 4) 변수를 단순히 없애는 경우, 높은 분산(Overfitting) 혹은 모델의 예측력을 개선시킬 수 없음

“However, because it is a discrete process-variables are either retained or discarded – it often exhibits high variance, and so doesn’t reduce the prediction error of the full model.” (3)

(1) Patrick Breheny, BST 764: Applied Statistical Modeling

(2) Stephanie, Sep.4,2015, Stepwise Regression Modeling

(3) The Elements of Statistical Learning

1-2. Subset Selection의 한계

■ 한계 5) 다중공선성 존재 시 → 유의한 변수 대신 무의미한 변수가 선택될 수 있음 (다중 공선성 해결 X)

“When multicollinearity is present, important variables can appear to be non-significant and standard errors can be large.”

(1)

VIF) 진단을 통해, 변수를 제거하게 되면?

다중 공선성의 문제를 해결 할 수 있지 않을까?

+ Variance Inflation Factor 설명

기존 모델 : $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$

$VIF_i = \frac{1}{(1-R_i^2)}$ (R_i^2 : X_i 가 y 일 때의 모델의 결정계수)

e.g $\rightarrow X_1 = \delta_0 + \delta_2 X_2 + \delta_3 X_3 + \delta_4 X_4 + \epsilon$

1-3. Variance inflation Factor (VIF)의 한계

▣ 한계 1) Dummy 변수(순서, 랭킹, 명목형)가 존재할 경우, 다중공선성이 아니 여도 VIF값이 크게 나올 수 있음

“The VIF may not be applicable to models where you have dummy regressors constructed from a polytomous categorical variable or polynomial regressors.” (1)

“ the indicator variables will necessarily have high VIF, even if the categorical variable is not associated with other variables in the regression models. ” (2)

▣ 한계 2) 수정 결정계수가 1에 가까울 수록, VIF가 무한으로 커질 수 있음

“As R_j^2 approaches 1, VIF approaches infinity and standard errors will blow up” (1)

(1) Fox, 2016, Applied Regression Analysis and Generalized Regression Models

(2) Paul Allison, When can you Safely Ignore Multicollinearity?

그렇다면, 어떻게 변수를 직접적으로 없애지 않고, 어떻게 문제를 해결 할 수 있을까?

변수에 Penalty를 줘서 계수의 크기를 조절하는 방향은 어떨까?

Penalized Regression / Regulation Regression

2 Cost Function

2-1. Cost Function : Linear Regression

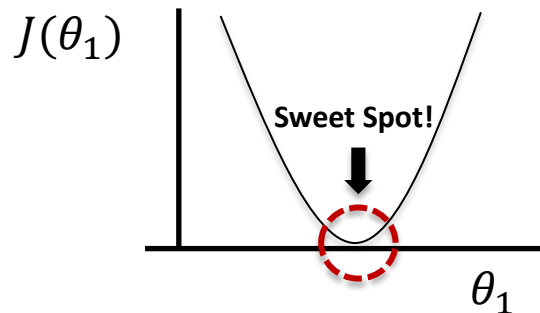
Cost function

$$\text{예측 값} - \text{실제 결과 값} \quad (\hat{y} - y)^2$$

Linear Regression Cost function

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

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



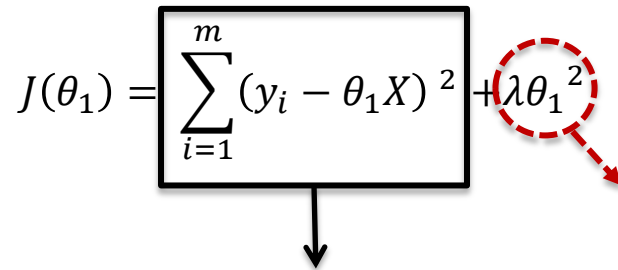
- **Gradient descent algorithm** $\theta_1 \leftarrow \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$
 - 반복적으로 수행 (learning rate 설정)
 - 변수간 데이터 크기 차이가 크면, 등고선 간격이 좁아져 최적점을 찾는데 시간이 걸림 때문에, 변수간 scaling 작업이 필요하기도 함
 - 기울기가 양의 방향이면 θ_1 값을 감소 vice versa θ_1
- **Normal Equation**
 - 변수가 많을 경우, 느려짐
 - 수학적 계산 필요
 - Scaling이 필요 없음

2-2. Cost Function : Penalized Regression

Cost function

$$\text{예측 값} - \text{실제 결과 값} \quad (\hat{y} - y)^2$$

Penalized Regression Cost function

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


The Sum of the squared residuals

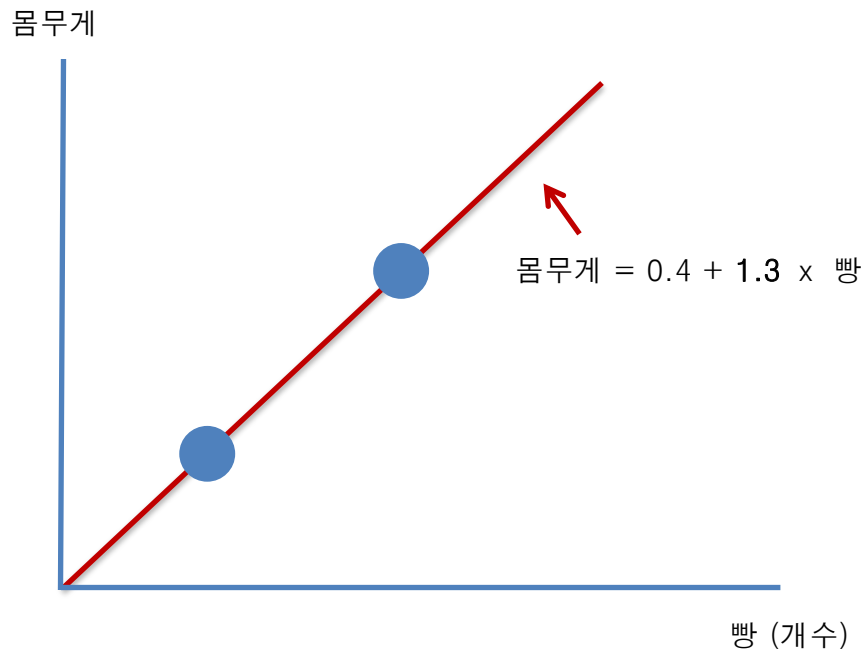
Penalty to the traditional Least Squares method

λ determines how severe that penalty is.

λ can be any value from 0 to positive infinity.

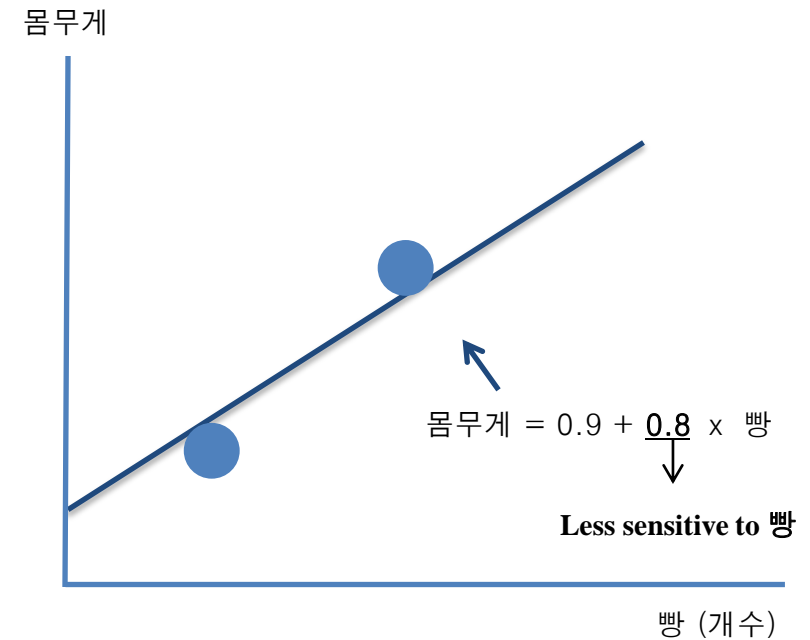
2-3. Cost Function - Comparison

Linear Regression



$$\begin{aligned} &\text{the sum of squared residuals} + \lambda \times \beta_1^2 \\ &= 0 + 1.69 = 1.69 \quad (\text{set } \lambda = 1) \end{aligned}$$

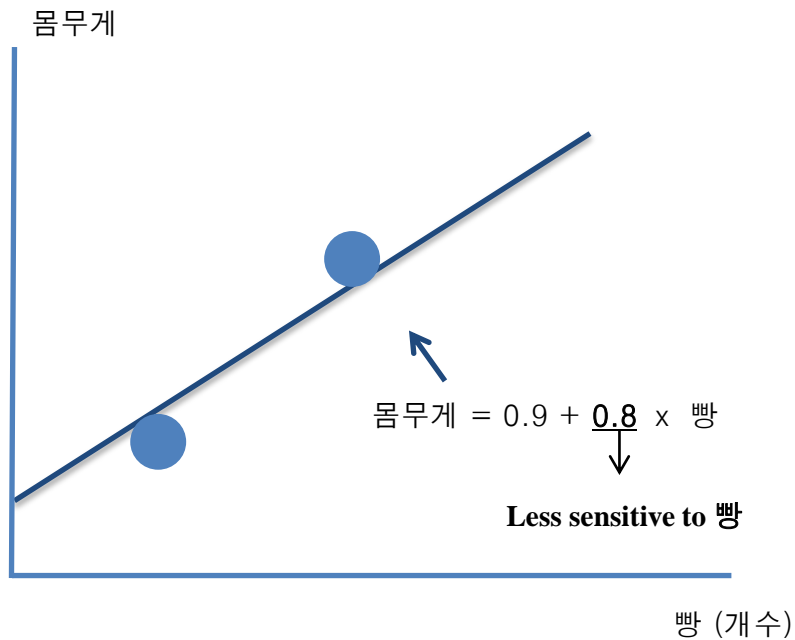
Penalized Regression



$$\begin{aligned} &\text{the sum of squared residuals} + \lambda \times \beta_1^2 \\ &= 0.09 + 0.01 + 0.64 = 0.74 \quad (\text{set } \lambda = 1) \end{aligned}$$

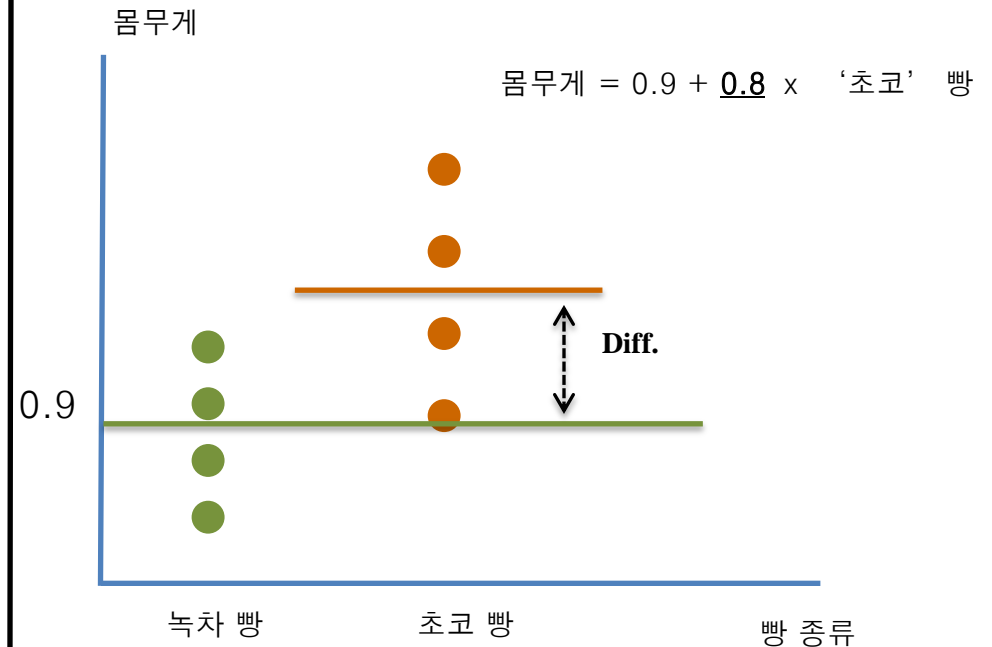
2-3. Cost Function – Discrete variable

Penalized Regression – continuous variable



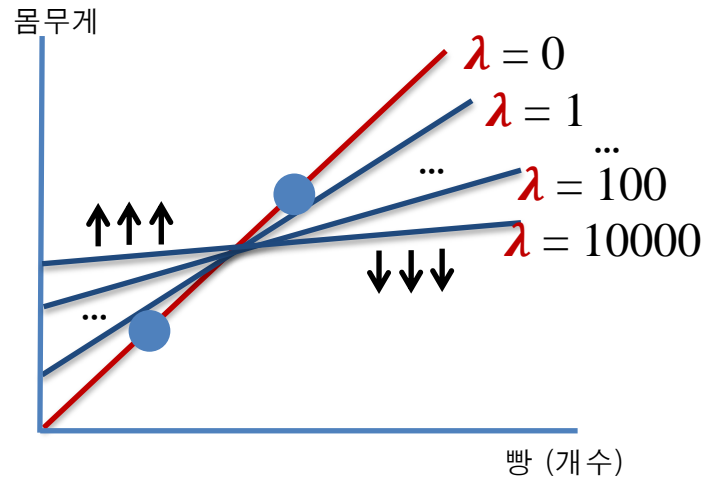
the sum of squared residuals + $\lambda \times \beta_1^2$

Penalized Regression – Discrete variable



the sum of squared residuals + $\lambda \times \text{Diff}^2$

2-4. Cost Function - λ

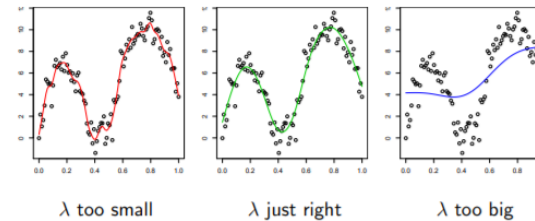


0에 가까워 짐

the sum of squared residuals + $\lambda \times \beta_1^2$

Larger λ gets, our predictions for 몸무게

Become less and less sensitive to 빵

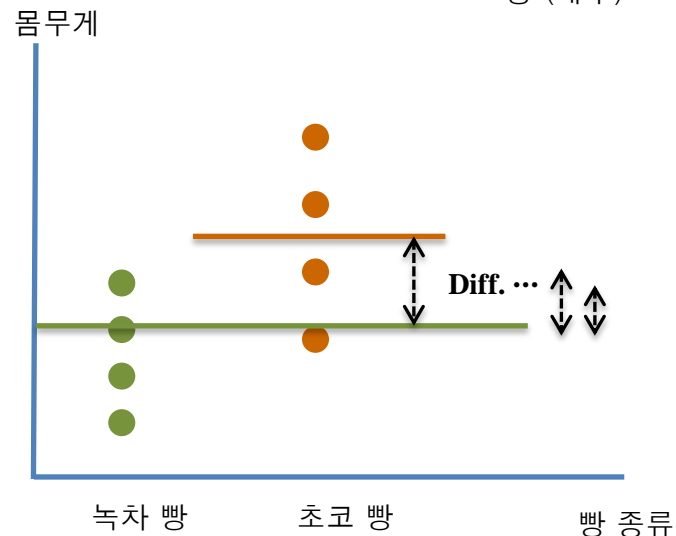


the sum of squared residuals + $\lambda \times Diff^2$

Larger λ gets, our predictions for 초코 빵 몸무게

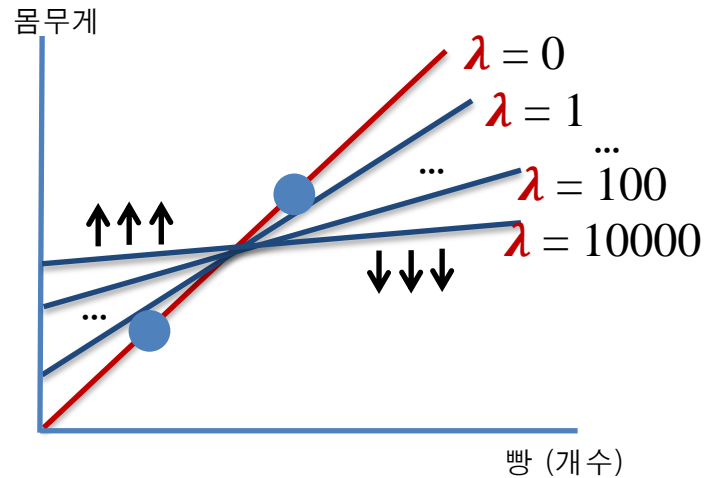
become less sensitive to the difference between

the 녹차빵 and 초코빵



왜 λ 값이 커질 수록 계수는 작아지는 것일까?

2-4. Cost Function - λ 증명



OLS

$$\begin{aligned} RSS(\theta_1) &= \sum_{i=1}^m (Y_i - \theta_1 X)^2 \\ &= (Y - \theta_1 X)^T (Y - \theta_1 X) \\ &= Y^T Y - \theta_1 X^T Y - \theta_1 Y^T X + \theta_1^2 X^T X \end{aligned}$$

θ_1 에 대해 미분

$$0 = -2X^T Y + 2\theta_1 X^T X$$

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

Ridge

$$\begin{aligned} PRSS(\theta_1) &= \sum_{i=1}^m (Y_i - \theta_1 X)^2 + \lambda \theta_1^2 \\ &= (Y - \theta_1 X)^T (Y - \theta_1 X) + \lambda \theta_1^2 \\ &= Y^T Y - \theta_1 X^T Y - \theta_1 Y^T X + \theta_1^2 X^T X + \lambda \theta_1^2 \end{aligned}$$

θ_1 에 대해 미분

$$0 = -2X^T Y + 2\theta_1 X^T X + 2\lambda \theta_1$$

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

2-5. Cost Function - λ 증명 + OLS 추정치와의 관계

$$PRSS(\theta_1) = \sum_{i=1}^m (Y_i - \theta_1 X)^2 + \lambda \theta_1^2$$

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

$$= (X^T X + \lambda)^{-1} X^T Y$$

$$= (X^T X + \lambda)^{-1} X^T X [(X^T X)^{-1} X^T Y]$$

$$\text{let, } X^T X = R$$

$$= (R + \lambda)^{-1} R [(R)^{-1} X^T Y]$$

$$= [R(I_p + \lambda R^{-1})]^{-1} R [R^{-1} X^T Y]$$

$$= (I_p + \lambda R^{-1})^{-1} R^{-1} R [R^{-1} X^T Y]$$

$$= (I_p + \lambda R^{-1})^{-1} R^{-1} R [\underline{(X^T X)^{-1} X^T Y}]$$

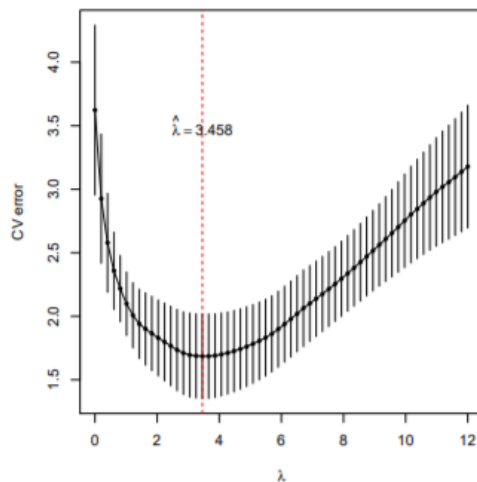
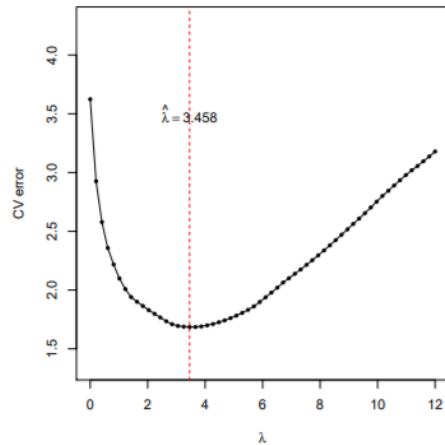
$$= (I_p + \lambda R^{-1})^{-1} \widehat{\theta_{ls}}$$

$$\widehat{\theta_{ridge}} = \frac{X^T Y}{X^T X + \lambda} = [I + \lambda (X^T X)^{-1}]^{-1} \widehat{\theta_{ls}}$$

$$\therefore \widehat{\theta_{ridge}} = \frac{\widehat{\theta_{ls}}}{(I_p + \lambda R^{-1})}$$

그렇다면, 적절한 λ 값은 어떻게 구할 수 있는 것일까?

2-5. Cost Function - Cross Validation for λ

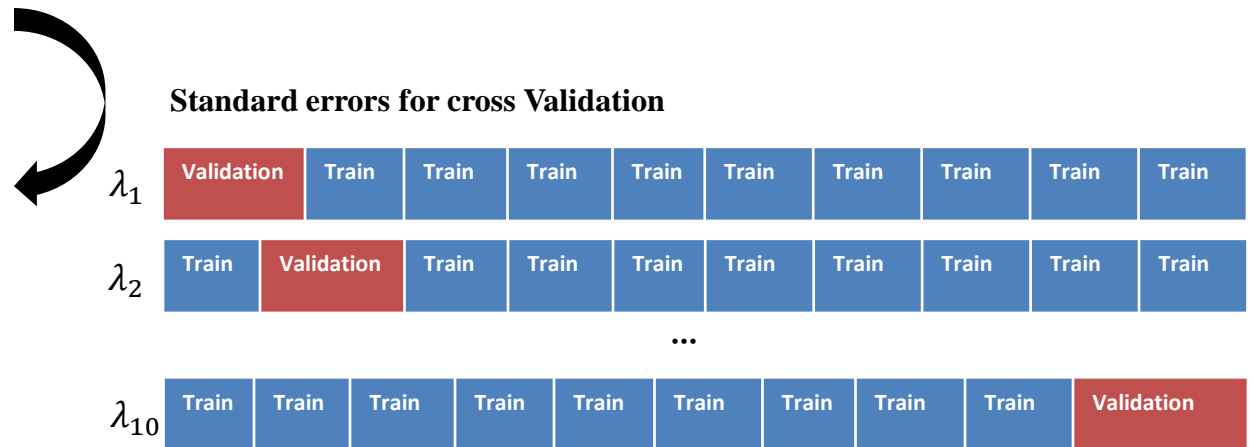


“Cross Validation is a simple, intuitive way to estimate prediction error.” (1)

“Cross Validation is choosing the lambda that provides best predictive accuracy.” (2)

“We just try a bunch of values of lambda and use Cross validation, **typically 10-fold Cross Validation**, to determine which one results in the lowest Variance.” (3)

Standard errors for cross Validation



$CV_1(\theta)$ = validation error, λ_1

$$SD(\theta) = \sqrt{\text{var}(CV_1(\theta), \dots, CV_K(\theta))}$$

(1) Josh Starmer, StatQuest: Regularization Part1 : Ridge Regression

(2) University of Washington, Machine Learning: Regression

(3) D Zurell, 2012 , Collinearity- Damaris Zurell research

***) Cost Function - Caution for Cross Validation**

▣ 데이터가 계절성 특징을 가지고 있을 경우

“The data contained a yearly seasonality and just taking June as a hold out might be naïve” (1)

▣ Cross Validation에 사용되는 Test 및 Train 데이터 셋도 결국엔, ‘Known’ 데이터 셋

“A cross Validation only validates what is in her. Hyper parameter optimization is usually done outside of validation.” (1)

▣ Cross Validation을 통해 선택하는 경우, Overfitting이 될 경우는 아닌지 살펴볼 필요가 있음

“Despite its wide applicability, traditional cross validation methods tend to select overfitting models, due to the ignorance of the uncertainty in the testing sample. .” (2)

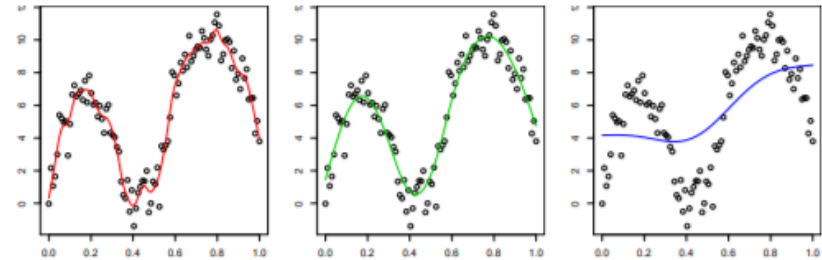
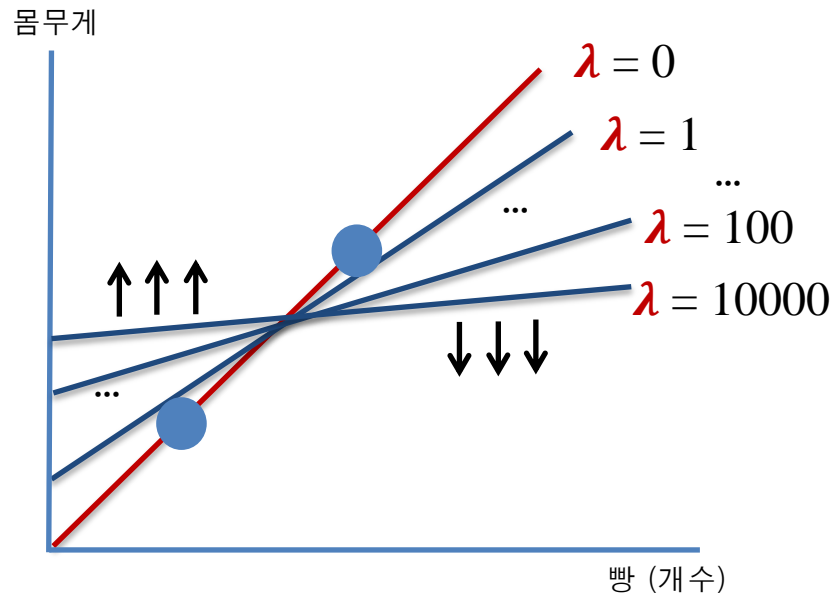
(1) Martin Schmitz, PhD “When Cross Validation Fails”

(2) Hadi Ebrahimnejad, Machine Learning at Accolade

그렇다면,
앞에서 말한 이슈들을
변수에 Penalty를 주는 방식이 어떻게 해소가 되는 것일까?

- 1) Overfitting의 이슈는?
- 2) 다중 공선성 이슈는 ?

2-6. Penalized Cost Function – How to avoid overfitting



λ too small

λ just right

λ too big

$$\begin{aligned}
 RSS(\theta_1) &= \sum_{i=1}^m (Y_i - \theta_1 X)^2 + \lambda \theta_1^2 \\
 &= (Y - \theta_1 X)^T (Y - \theta_1 X) + \lambda \theta_1^2 \\
 &= Y^T Y - \theta_1 X^T Y - \theta_1 Y^T X + \theta_1^2 X^T X + \lambda \theta_1^2
 \end{aligned}$$

θ_1 에 대해 미분

$$0 = -2X^T Y + 2\theta_1 X^T X + 2\lambda \theta_1$$

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

2-7. Penalized Cost Function – Solution for Multicollinearity

Problem with multicollinearity

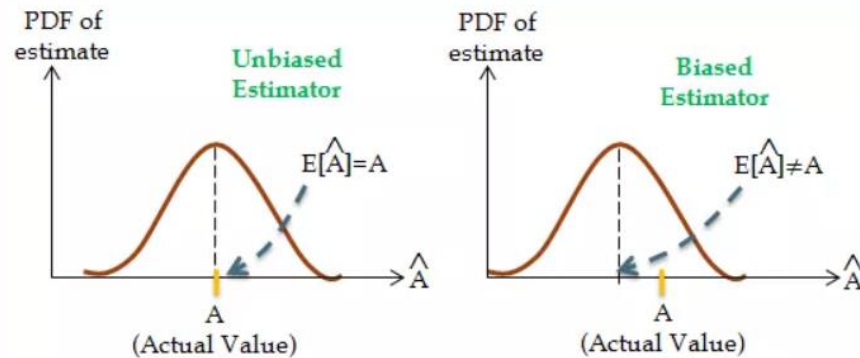
- 다중공선성은 계수의 분산을 크게 만든다. 이렇게 커진 계수의 분산은 의미 있는 계수를 찾아내기 어렵다.

Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0. In other words, by **overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant.** Without multicollinearity (and thus, with lower standard errors), those coefficients might be significant. (1)

(1) Enough is Enough! Handling Multicollinearity in Regression Analysis

2-8. Penalized Cost Function – Biased estimator

– Unbiased 하다?



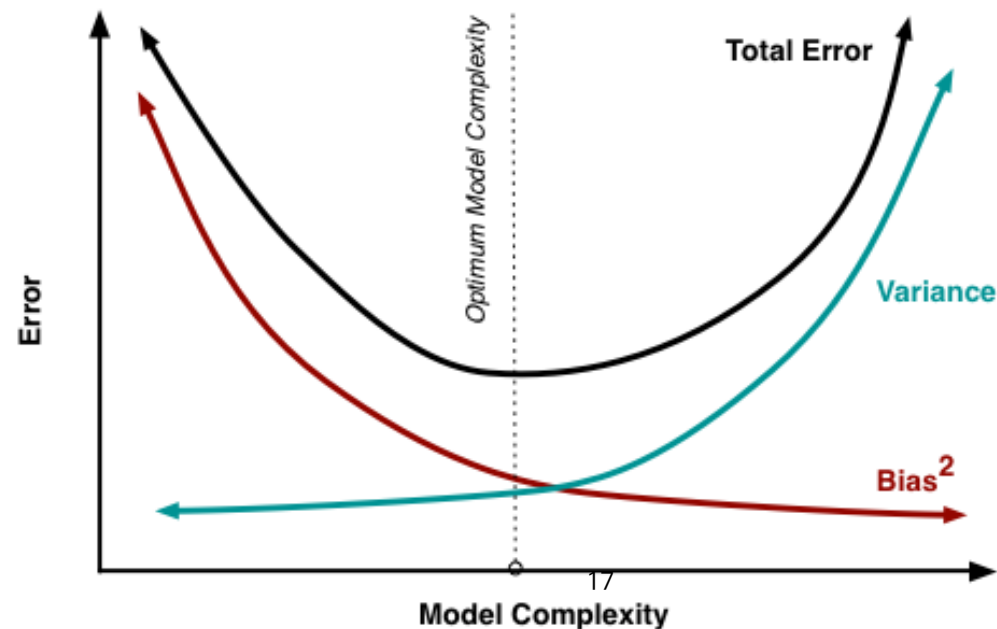
$$\widehat{\theta}_{ridge} = \frac{X^T Y}{X^T X + \lambda} = [I + \lambda(X^T X)^{-1}]^{-1} \widehat{\theta}_{ls}$$

$$\therefore \widehat{\theta}_{ridge} = \frac{\widehat{\theta}_{ls}}{(I_p + \lambda R^{-1})}$$

$$\begin{aligned} E(\widehat{\theta}_{ridge}) &= E\{(I_p + \lambda \theta^{-1})^{-1} \widehat{\theta}_{ls}\} \\ &= E(ax) = aE(x) \\ &= (I_p + \lambda \theta^{-1})^{-1} E(\widehat{\theta}_{ls}) \\ &= (I_p + \lambda \theta^{-1})^{-1} \theta \\ \text{if, } \lambda &\neq 0, \\ &\neq \theta \end{aligned}$$

2-9. Penalized Cost Function - Advantages

- The purpose of Tolerant Methods is **to reduce the sensitivity of regression parameters to multicollinearity**. Since, multicollinearity can result in large and opposite signed estimator values for correlated predictors, **a penalty function is imposed to keep the value of predictors below a pre-specified value**.
- Ridge regression attacks the multicollinearity **by reducing the apparent magnitude of the correlations**. **Biased estimation** is used to attain **a substantial reduction in variance** with an accompanied increase in stability of the regression coefficients.



Cost Function은 꼭 제곱의 형태의 Penalty Function만 가능한가?

기존의 Cost function에 어떠한 Penalty function을 주느냐에 따라
Penalty Regression의 종류가 있겠다!

3 Penalized Regression

Penalized Regression 종류

– Ridge/Lasso/Elastic-net

3-1. Ridge Regression & Lasso Regression

Ridge Regression

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

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

$$\lambda(\theta_1^2 + \theta_2^2 + \theta_3^2 + \dots)$$

Lasso Regression

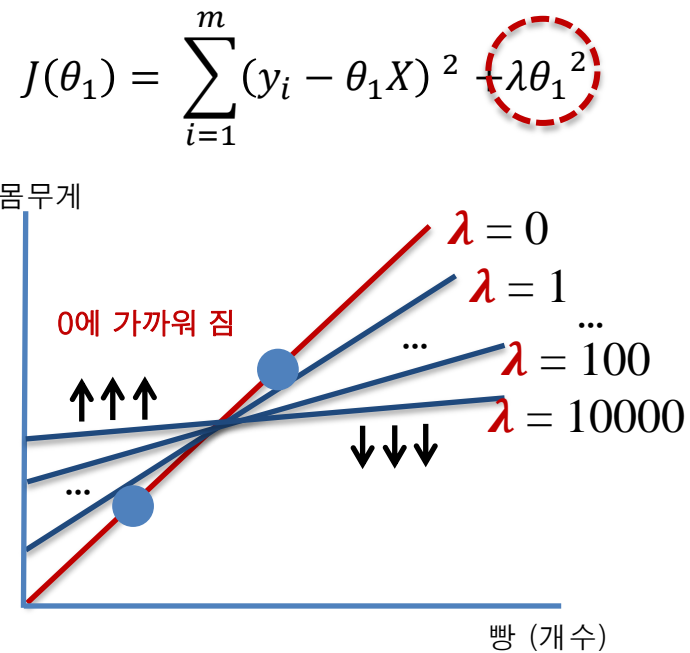
$$J(\theta_1) = \sum_{i=1}^m (y_i - \theta_1 X)^2 + \lambda |\theta_1|$$

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

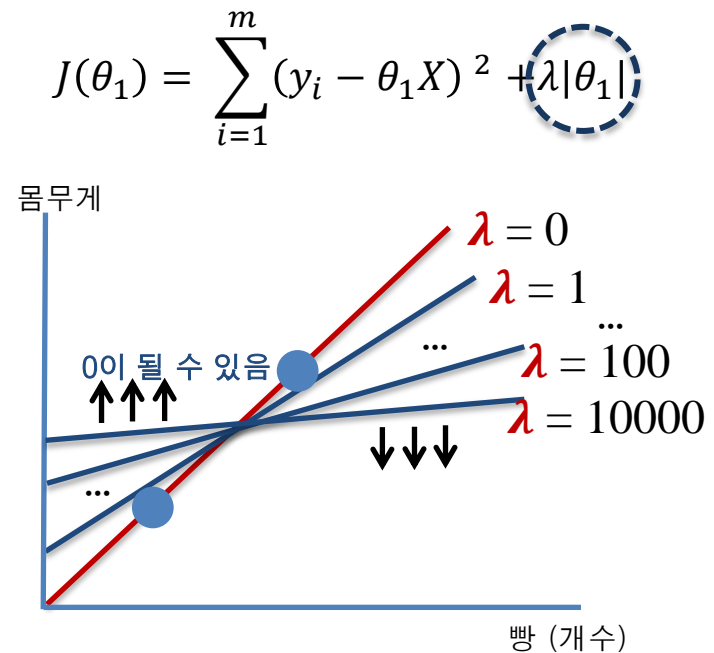
$$\lambda(|\theta_1| + |\theta_2| + |\theta_3| + \dots)$$

3-2. Ridge Regression & Lasso Regression - Difference

Ridge Regression



Lasso Regression



Ridge Regression can only shrink the slope **asymptotically close to 0**

while Lasso Regression can shrink the slope **all the way to 0**. 19

왜...? Ridge Regression의 Penalty는 0에 가까워지고
Lasso Regression의 Penalty는 0이 될 수 있는가?

3-3. Ridge Regression & Lasso Regression

Ridge Regression

$$\begin{aligned}RSS(\theta_1) &= \sum_{i=1}^m (Y_i - \theta_1 X)^2 + \lambda \theta_1^2 \\&= (Y - \theta_1 X)^T (Y - \theta_1 X) + \lambda \theta_1^2 \\&= Y^T Y - \theta_1 X^T Y - \theta_1 Y^T X + \theta_1^2 X^T X + \lambda \theta_1^2\end{aligned}$$

θ_1 에 대해 미분

$$0 = -2X^T Y + 2\theta_1 X^T X + 2\lambda \theta_1$$

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

Lasso Regression

$$\begin{aligned}RSS(\theta_1) &= \sum_{i=1}^m (Y_i - \theta_1 X)^2 + 2\lambda |\theta_1| \quad \text{let, } \theta_1 \geq 0 \\&= (Y - \theta_1 X)^T (Y - \theta_1 X) + 2\lambda \theta_1 \\&= Y^T Y - \theta_1 X^T Y - \theta_1 Y^T X + \theta_1^2 X^T X + 2\lambda \theta_1\end{aligned}$$

θ_1 에 대해 미분

$$0 = -2X^T Y + 2\theta_1 X^T X + 2\lambda$$

$$\therefore \hat{\theta} = \frac{X^T Y - \lambda}{X^T X} \rightarrow \lambda = X^T Y, \hat{\beta} = 0$$

Ridge Regression can only shrink the slope asymptotically close to 0
while **Lasso Regression** can shrink the slope all the way to 0. 20



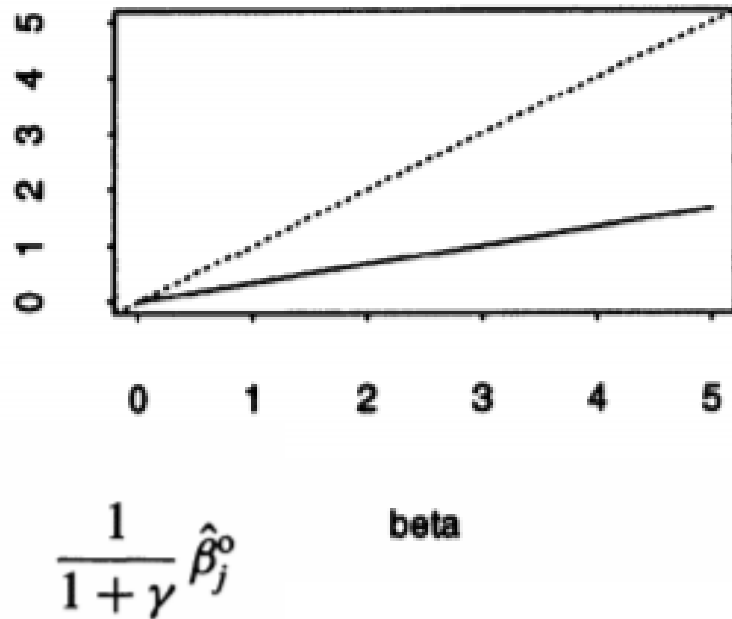
Ridge Regression : 모든 변수의 관계 식

Lasso Regression : 일부 변수의 관계 식

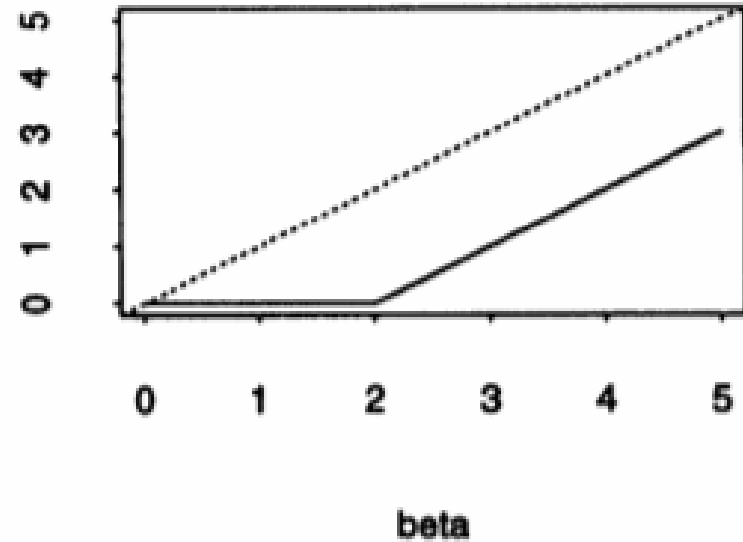
3-3. Ridge Regression & Lasso Regression

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^o)(|\hat{\beta}_j^o| - \gamma)^+ \quad \text{제약식}$$

Ridge Regression



Lasso Regression



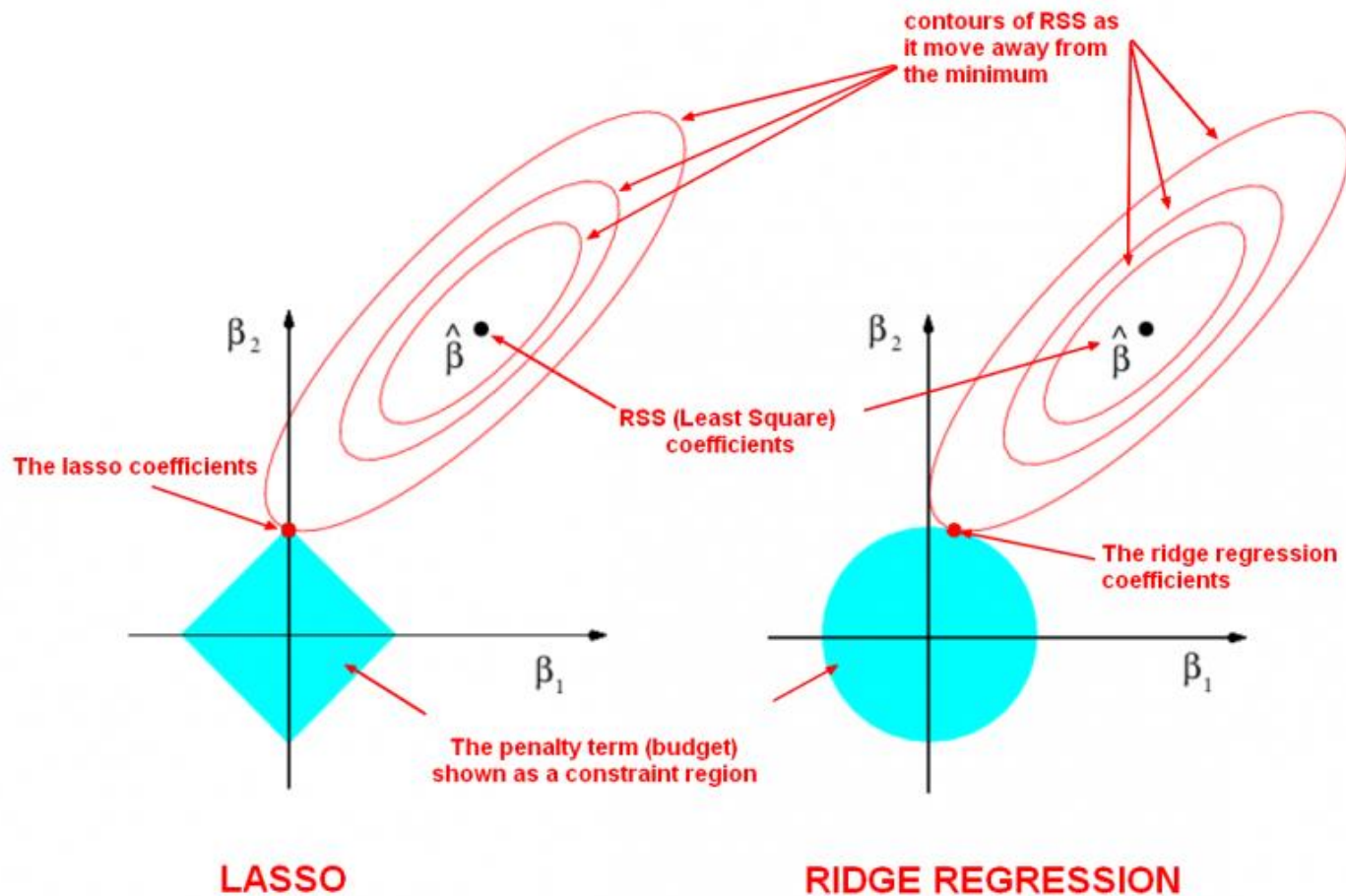
Ridge Regression can only shrink the slope asymptotically close to 0 while Lasso Regression can shrink the slope all the way to 0. 20



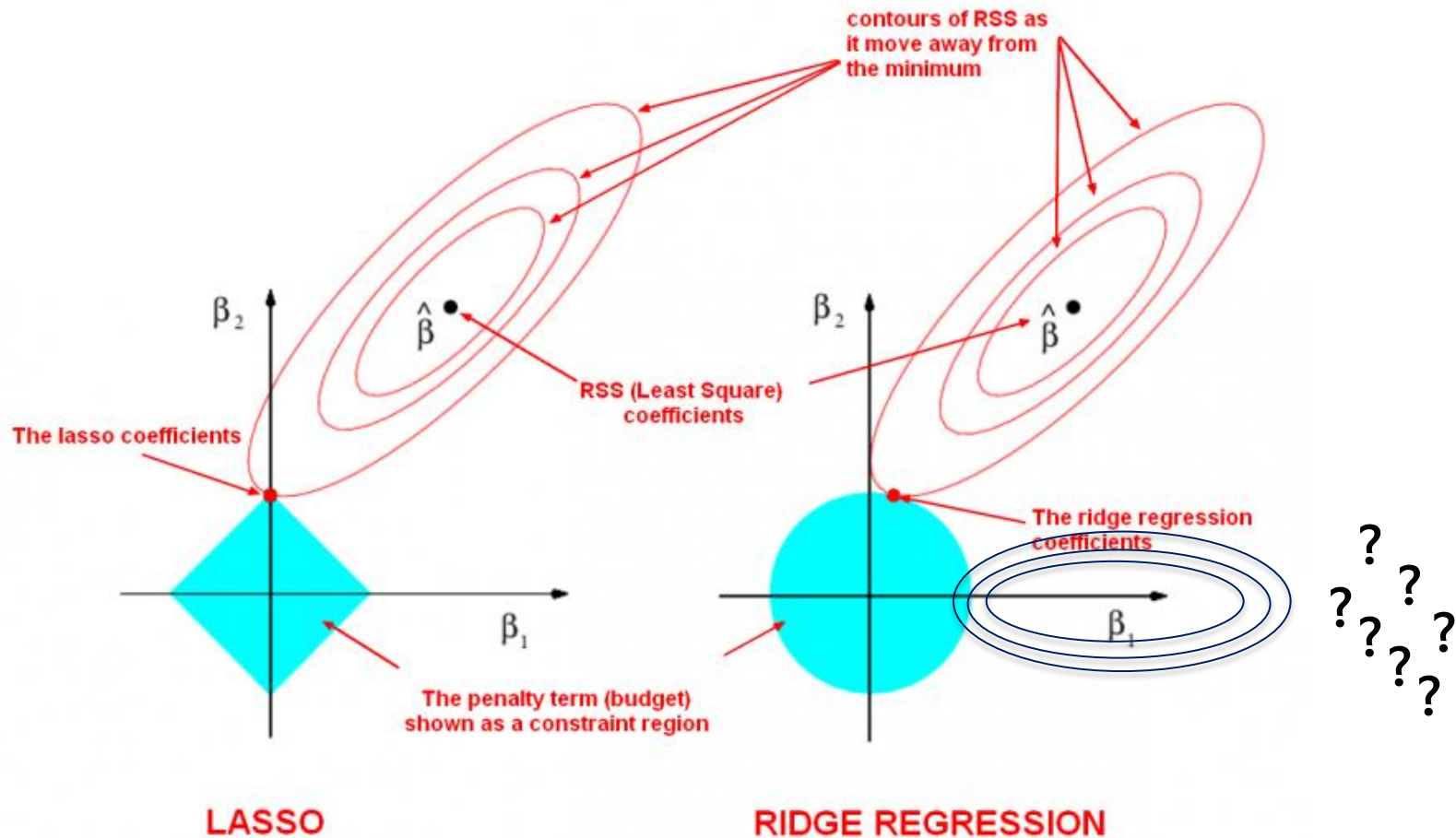
Ridge Regression : 모든 변수의 관계 식

Lasso Regression : 일부 변수의 관계 식

3-4. Ridge Regression & Lasso Regression



3-4. Ridge Regression & Lasso Regression



3-5. Ridge Regression & Lasso Regression - Example

몸무게 = intercept + β_1 빵 먹은 날 + β_2 초코 빵 + β_3 빵 개수 + β_4 식전 빵 유무 + β_5 오늘 날씨 + β_6 미드 본 횟수

Ridge Regression

$$\lambda(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2 + \beta_5^2 + \beta_6^2)$$

the larger we make λ

Shrink little bit ...

= 0에 가까워짐
never be equal to 0

Lasso Regression

$$\lambda(|\beta_1| + |\beta_2| + |\beta_3| + |\beta_4| + |\beta_5| + |\beta_6|)$$

increase the value for λ

Shrink little bit ...

= 0으로 수렴함
go all the way to 0

3-6. Ridge Regression & Lasso Regression

Ridge Regression

Ridge Regression :

[Prediction] + **[Bias Variance Trade-off]**

유익한 변수가 많은 경우, Ridge Regression이 유용

Lasso Regression

Ridge Regression :

[Prediction] + [Bias Variance Trade-off]

+ **[Feature Selection]**

변수 선택의 측면에서는, LASSO Regression이 유용

독립변수가 많고, 어떤 변수가 유의한 의미를 가질 지 배경지식도 없고
변수를 일일이 하나씩 볼 수 없다면…?

3-7. Elastic Net Regression

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

Ridge Regression + Lasso Regression



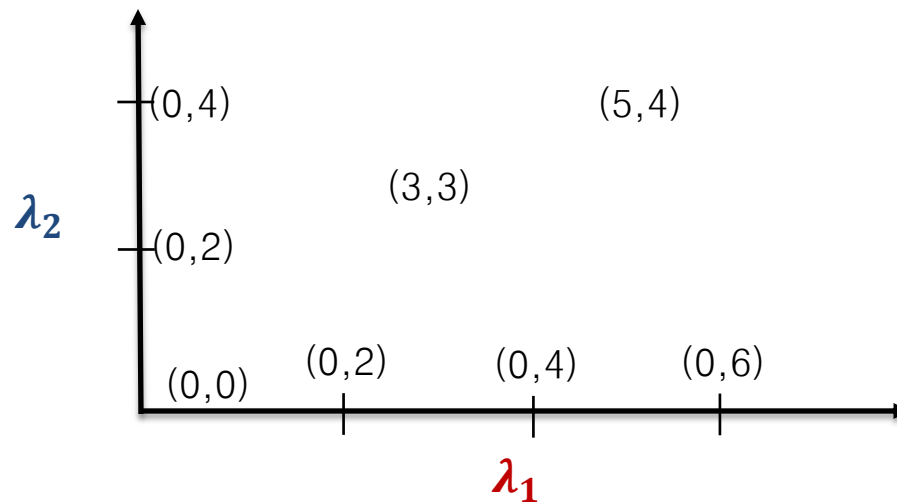
$$J(\theta_1) = \sum_{i=1}^m (y_i - \theta_1 X)^2 + \lambda_1 \theta_1^2 + \lambda_2 |\theta_1|$$

**The hybrid Elastic-Net Regression is especially good at
dealing with situations when there are correlations between parameters. (1)**

3-8. Elastic Net Regression

$$J(\theta_1) = \sum_{i=1}^m (y_i - \theta_1 X)^2 + \lambda_1 \theta_1^2 + \lambda_2 |\theta_1|$$

Cross Validation on different combinations of λ_1 and λ_2



Ridge, Lasso, Elastic-net Regression만 잘 쓰면 되겠네?

It seems safe to conclude that the lasso is an oracle procedure for simultaneously achieving consistent variable selection and optimal estimation (prediction). **However**, there are also **solid arguments against the lasso oracle statement**.

Fan and Li (2001) showed that the lasso can perform automatic variable selection because the l_1 penalty is singular at the origin. On the other hand, the lasso shrinkage produces biased estimates for the large coefficients, and thus it could be suboptimal in terms of estimation risk. Fan and Li conjectured that **the oracle properties** do not hold for the lasso.

Meinshausen and Bühlmann (2004) also showed the conflict of optimal prediction and consistent variable selection in the lasso. They **proved that the optimal λ for prediction gives inconsistent variable selection results; in fact, many noise features are included in the predictive model**. This conflict can be easily understood by considering an orthogonal design model (Leng, Lin, and Wahba 2004)

4 Extended Penalized Regression

Extended Penalized Regression

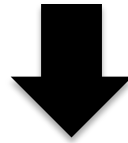
- Adaptive Lasso
- Group Lasso
- Fused Lasso

4-1. Adaptive Lasso Regression – Oracle Properties

Consistent in variable selection \neq Consistent in parameter estimation

An oracle procedure is one that has the following oracle properties:

- Identifies the right subset of true variables; and
- Has optimal estimation rate.



Consistent in variable selection **and Consistent in parameter estimation**

4-2. Adaptive Lasso Regression

Lasso Regression

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

모든 변수에 동일한 Penalty 를 부여

Adaptive Lasso Regression

수식 summation 으로 변경

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

$$\text{with, } \widehat{W} = \frac{1}{(|\hat{\beta}_j^*|)^\gamma}$$

$\hat{\beta}_j^*$ = usually Ridge Regression

γ = Thresholding functions

변수에 따라 상이한 Penalty 를 부여

✓ Oracle Properties

The adaptive lasso yields consistent estimates of the parameters while retaining the attractive convexity property of the lasso.

개별적으로 따로 따로 하나씩 다른 Penalty를 주기 보다,
주요한 영향을 미치는 변수들만 묶어서 한번에 선택과 집중을 하면 안되나?

4-3. Group Lasso Regression

Structured Sparsity Method

Structured = '구조'

Sparsity Method = '변수 선택'

→ 변수들 간에 유사한 구조 (그룹)이 있는 경우(사전지정), 해당 그룹에 따라 변수를 선택하는 Regression
같은 군집에 속한 모든 변수를 함께 선택

$$\min_{\beta \in \mathbb{R}^p} \left(\|y - \sum_{\ell=1}^L \mathbf{X}_{\ell} \beta_{\ell}\|_2^2 + \lambda \sum_{\ell=1}^L \sqrt{p_{\ell}} \|\beta_{\ell}\|_2 \right)$$



$$\|\mathbf{p} - \mathbf{q}\| = \sqrt{(\mathbf{p} - \mathbf{q}) \cdot (\mathbf{p} - \mathbf{q})} = \sqrt{\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 - 2\mathbf{p} \cdot \mathbf{q}}$$

4-4. Group Lasso Regression

Structured Sparsity Method

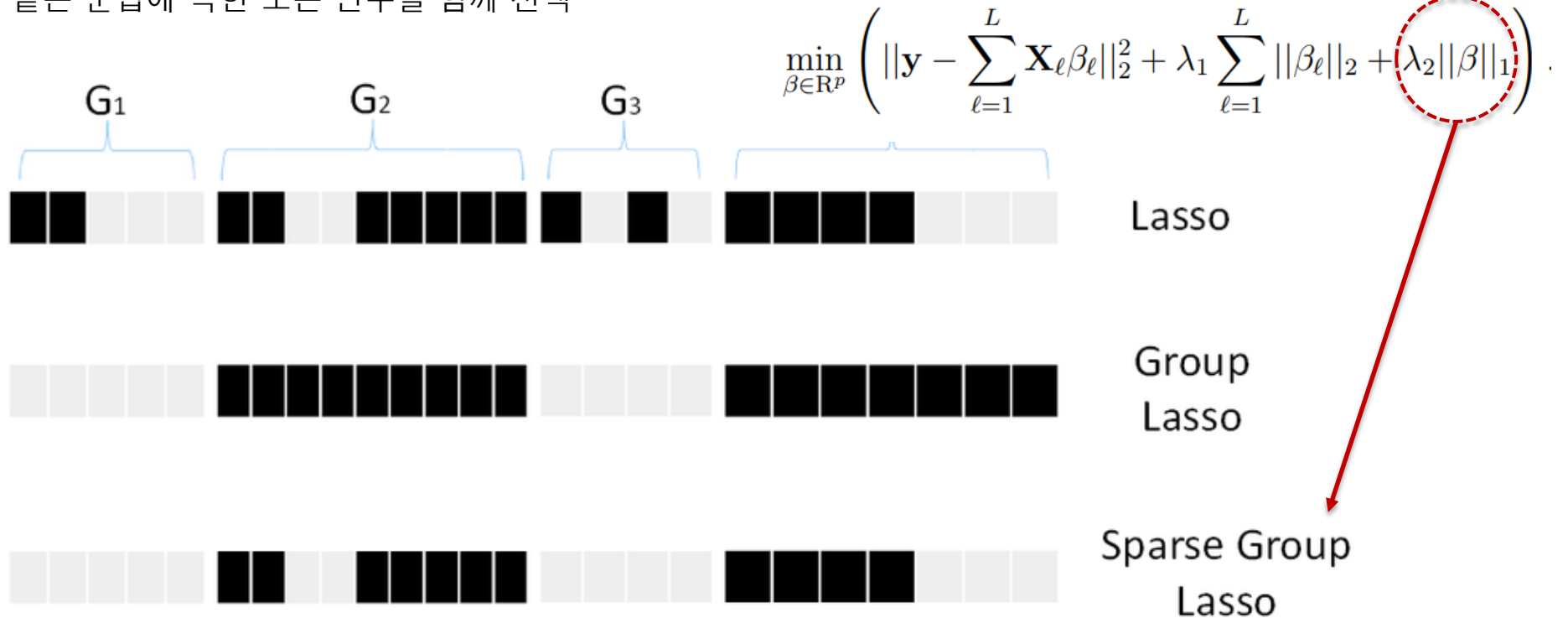
Structured = '구조'

Sparsity Method = '변수 선택'

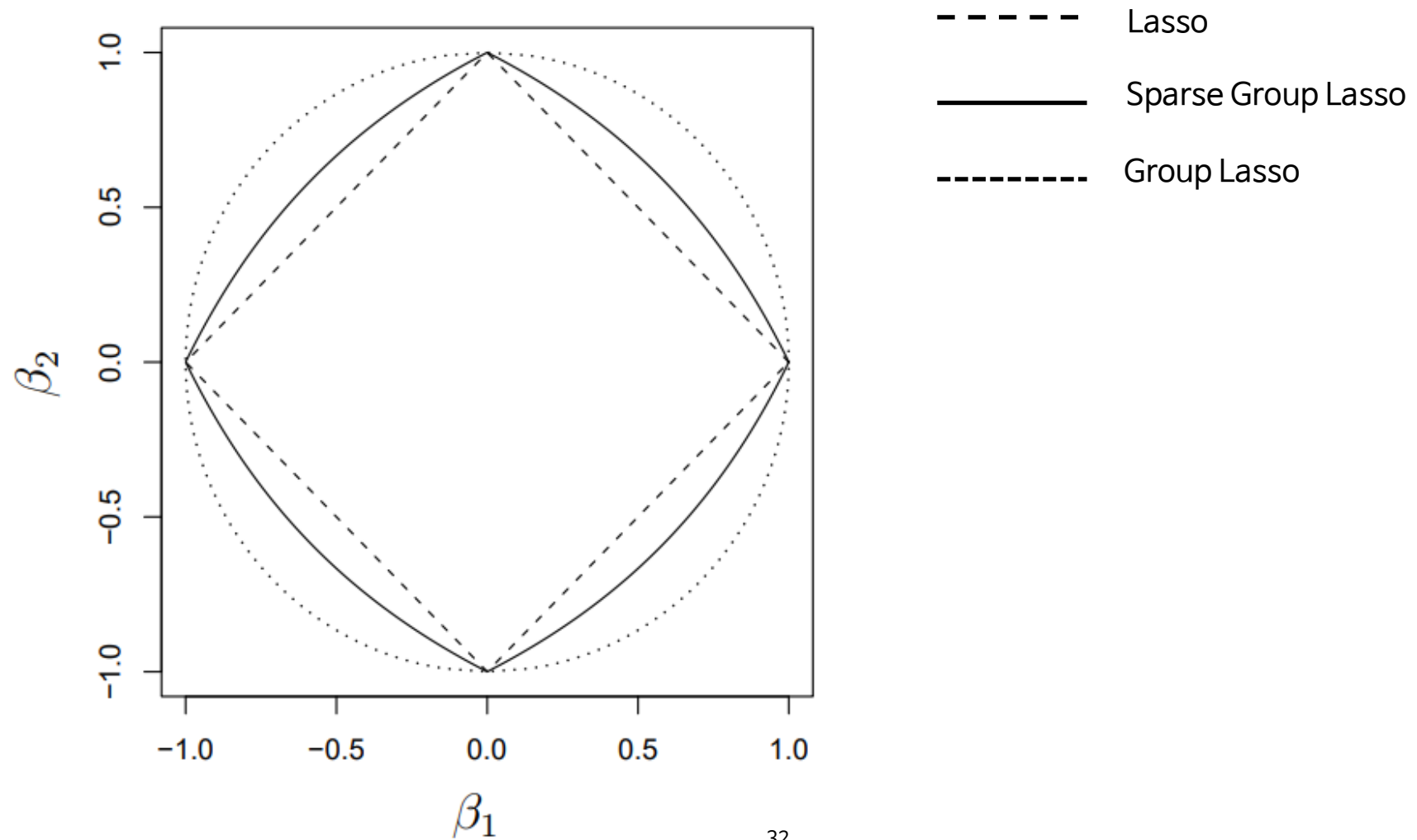
→ 변수들 간에 유사한 구조 (그룹)이 있는 경우, 해당 그룹에 따라 변수를 선택하는 Regression

같은 군집에 속한 모든 변수를 함께 선택

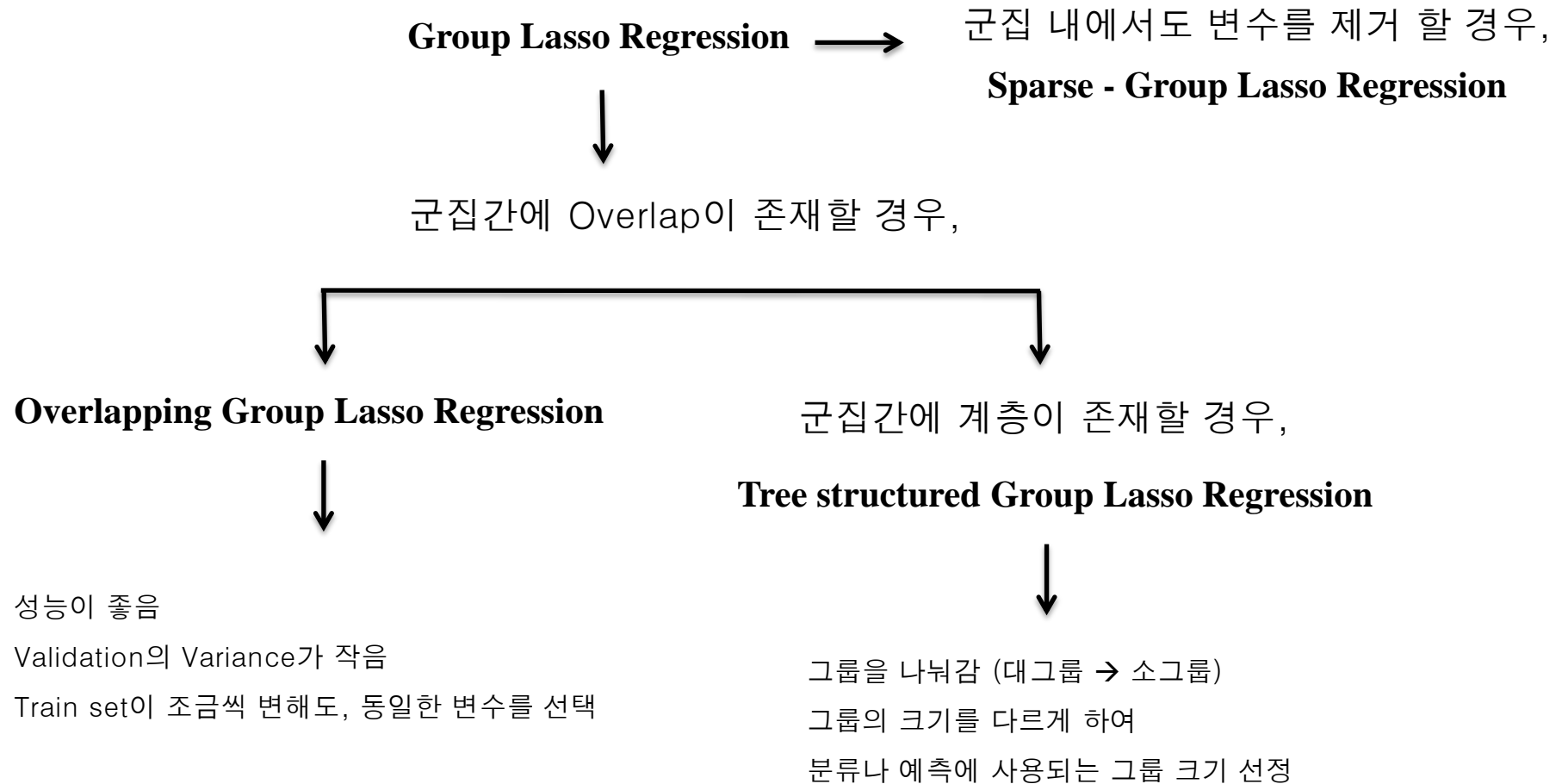
+ Dummy 변수를 Grouping 하는 데에도 유의



4-5. Group Lasso Regression



4-6. Group Lasso Regression – related lasso regression



4-7. Group Lasso Regression with example

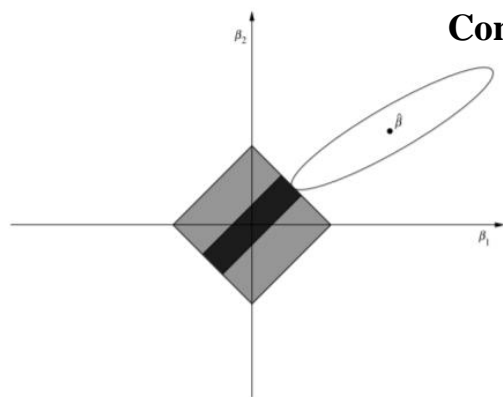
- Group Lasso를 통한 중학생의 삶의 만족도에 영향을 미치는 변수 탐색
 - : 한국 청소년 정책연구원의 KCYPS 패널 자료가 수백 개의 변수들을 제공 (338)
 - : 연속형 뿐만 아니라 범주형 변수를 함께 다루고자
 - : 338개 변수 중 유의한 15개의 변수가 선택됨

4-8. Fused Lasso Regression

변수에 순서 정보가 존재하는 경우, 유의하게 사용될 수 있음

특정한 순서 정보에 따른 회귀 계수의 차이에 Lasso Regression의 L1 penalty를 추가한 모형

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \underbrace{\lambda_1 \sum_{i=1}^p |\beta_j|}_{\text{Controls the degree of sparsity}} + \underbrace{\lambda_2 \sum_{i=2}^p |\beta_j - \beta_{j-1}|}_{\text{Controls the degree of smoothing between successive differences.}}$$



4-9. Fused Lasso Regression with example

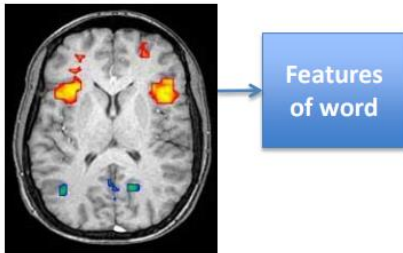
– Fused Lasso 회귀 모형 기반의 수익성에 대한 요인 연구

: 국내 21개의 대학병원의 운영 자료를 기반으로 대학병원 수익성에 영향을 미치는 요인 분석

: 연도별 변수들의 평균 변화를 반영하기 위해서 연도별 평균이 보정된 fused lasso 회귀 모형을 적합

: 종속 변수인 의료수익의료이익률에 통계적으로 유의한 영향을 주는 독립변수로 간호사당 환자수, 외래환자 ...

■ **Goal:** Predict semantic features from fMRI image



5 Python3 실습

- Adaptive Lasso
- Group Lasso
- Ridge Regression with Sentiment Analysis

5-1. Python



1. Penalized Regression을 이용한 감성 분석 (Sentiment Scoring)
2. Lasso, Group Lasso, Sparse Group Lasso를 이용한
감기진료건수 예측

참고하기 좋은 논문

- Regression Shrinkage and Selection via the Lasso By Robert Tibshirani

End of Document

jeons9677@gmail.com

