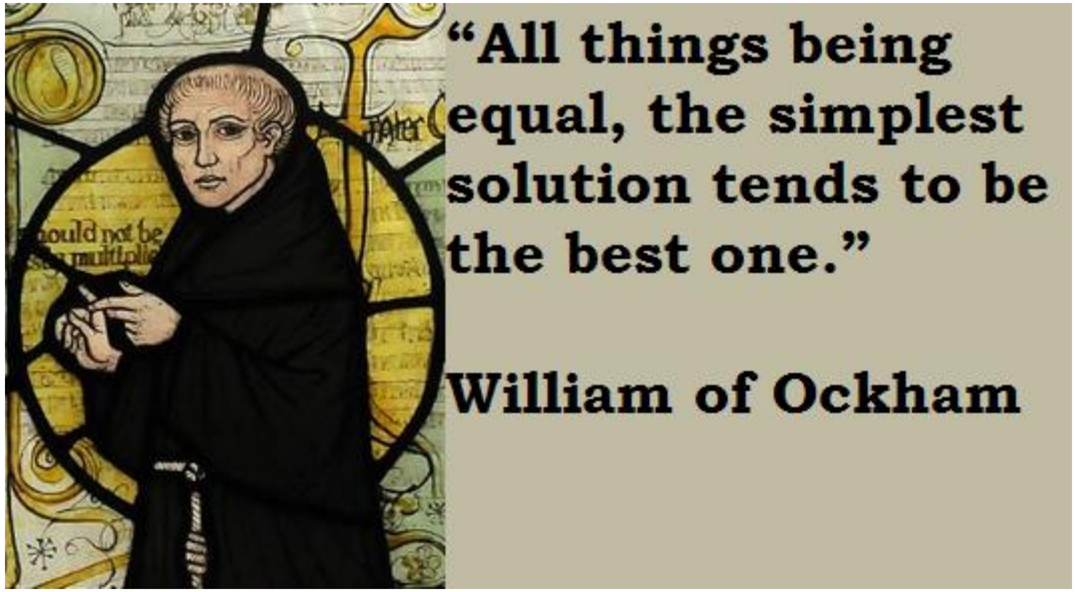


Dimensionality Reduction

Pilsung Kang
School of Industrial Management Engineering
Korea University

AGENDA

Revisit MLR

- Multiple Linear Regression
  - ✓ Formulation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

  - ✓ Objective function (should be minimized)

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{d}\hat{\beta}_j x_{ij}\right)^2$$

Revisit Logistic Regression

- Logistic Regression
  - ✓ Formulation

$$log(Odds) = log\left(\frac{p}{1-p}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots + \hat{\beta}_d x_d$$

  - ✓ Objective function (should be minimized)

$$-\sum_{i=1}^{n}\left(y_i \log\left(\frac{1}{1+\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}\right) + (1-y_i)\log\left(\frac{\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}{1+\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}\right)\right)$$

Ridge Regression

- Ridge Linear Regression

$$\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{d}\hat{\beta}_j x_{ij}\right)^2+\lambda\sum_{j=1}^{d}\hat{\beta}_j^2$$

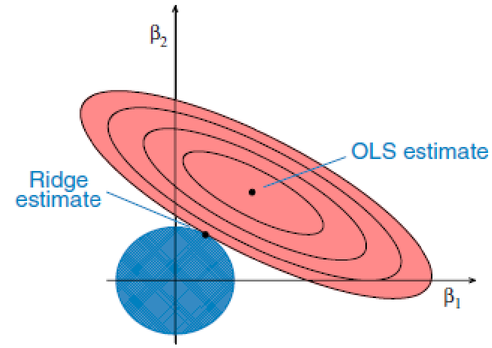
- Ridge Logistic Regression

$$-\sum_{i=1}^{n}\left(y_i\log\left(\frac{1}{1+\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}\right)+(1-y_i)\log\left(\frac{\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}{1+\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}\right)\right)+\lambda\sum_{j=1}^{d}\hat{\beta}_j^2$$

Ridge Regression

- Ridge (Logistic) Regression
  - ✓ Add L_2 nom penalty for the objective function

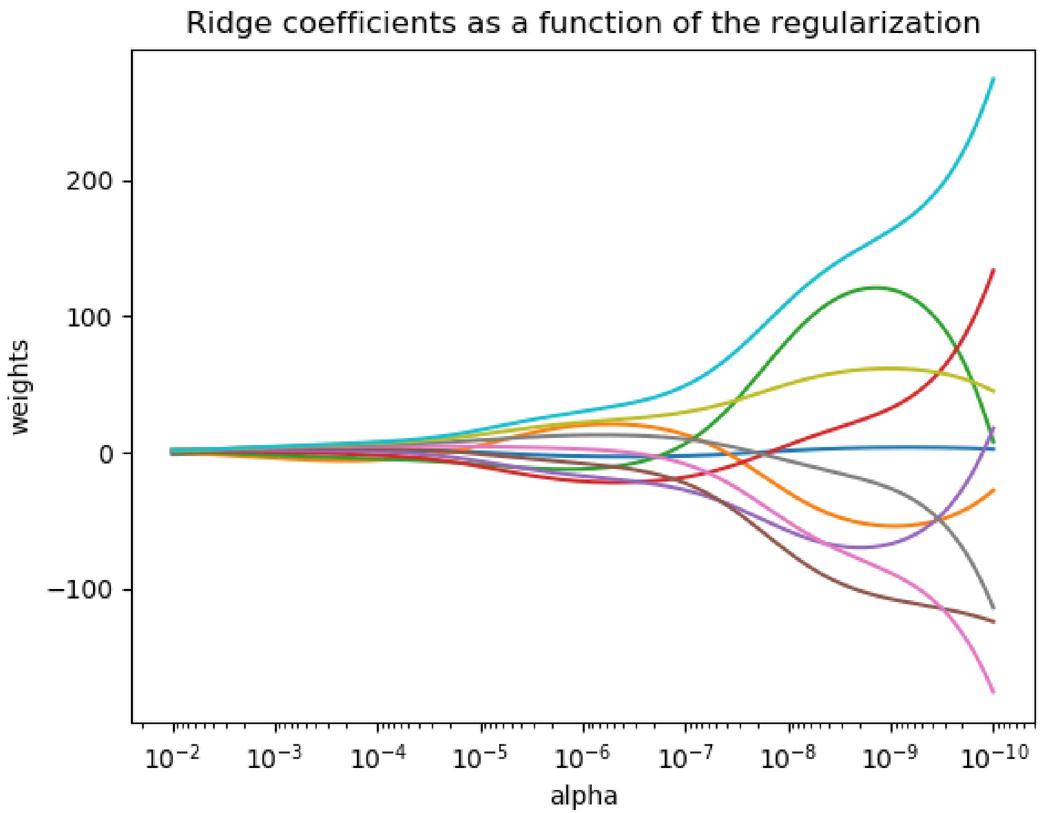
$$\lambda \sum_{j=1}^{d} \hat{\beta}_j^2$$



  - ✓ Properties
    - If two models have the same performance, smaller regression coefficients are preferred
    - Regression coefficients can be very small, but hard to make them exactly 0 ➔ not for variable selection
    - Work well when input variables have high correlations

Ridge Regression

- Ridge (Logistic) Regression
  - ✓ Example of estimated regression coefficients according to different λ



http://scikit-learn.org/stable/auto_examples/linear_model/plot_ridge_path.html#sphx-glr-auto-examples-linear-model-plot-ridge-path-py

LASSO

- LASSO: Least Absolute Shrinkage and Selection Operator
  - ✓ Multiple Linear Regression

$$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{d}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{d}|\hat{\beta}_j|$$

  - ✓ Logistic Regression

$$-\sum_{i=1}^{n}\left(y_i \log\left(\frac{1}{1+\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}\right) + (1-y_i)\log\left(\frac{\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}{1+\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}\right)\right) + \lambda\sum_{j=1}^{d}|\hat{\beta}_j|$$

LASSO

- LASSO: Least Absolute Shrinkage and Selection Operator
  - ✓ Ridge gives L_2 norm penalty while LASSO gives L_1 norm penalty
  - ✓ Can make the coefficients of irrelevant variables 0 → can do variable selection
  - ✓ The number of selected variables (variables with non-zero coefficients) vary according to λ

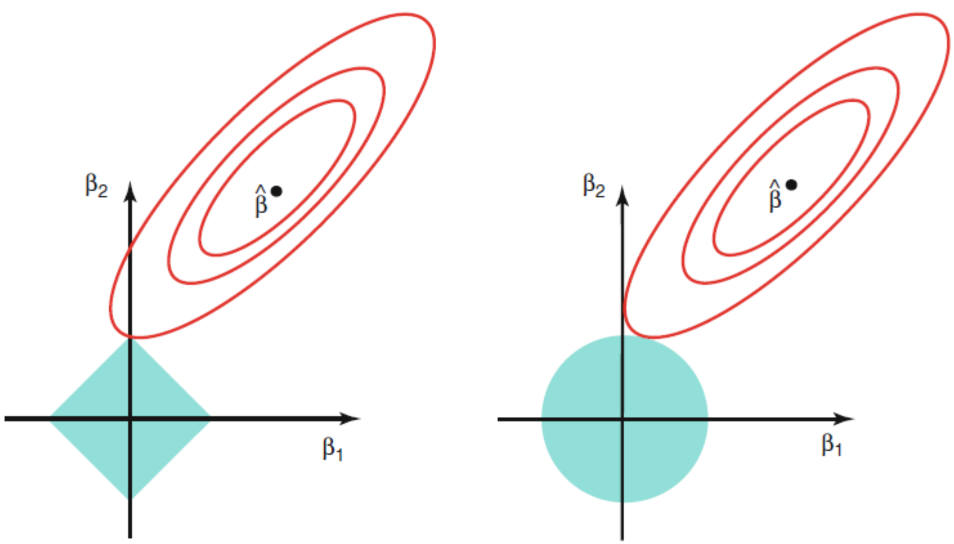
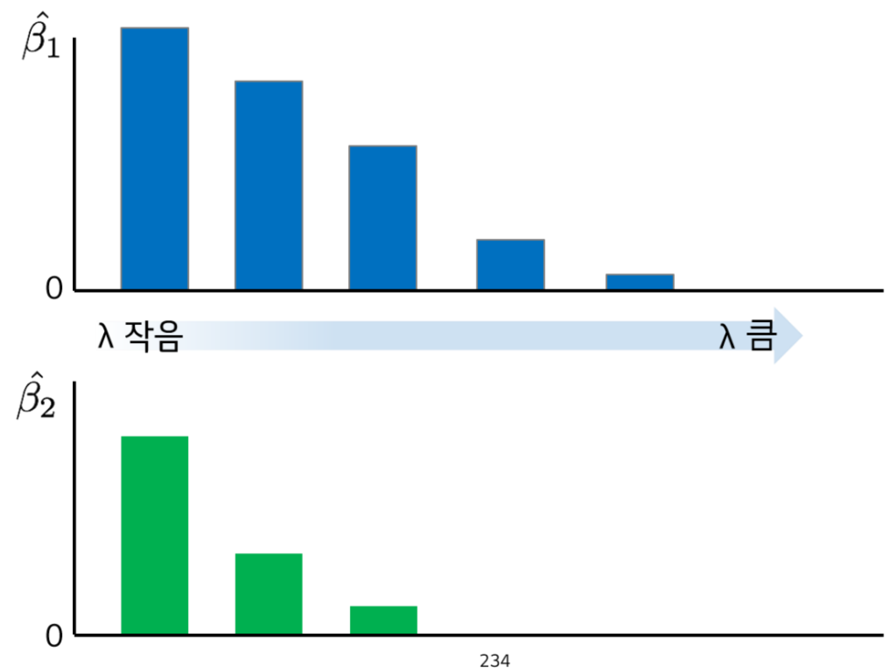
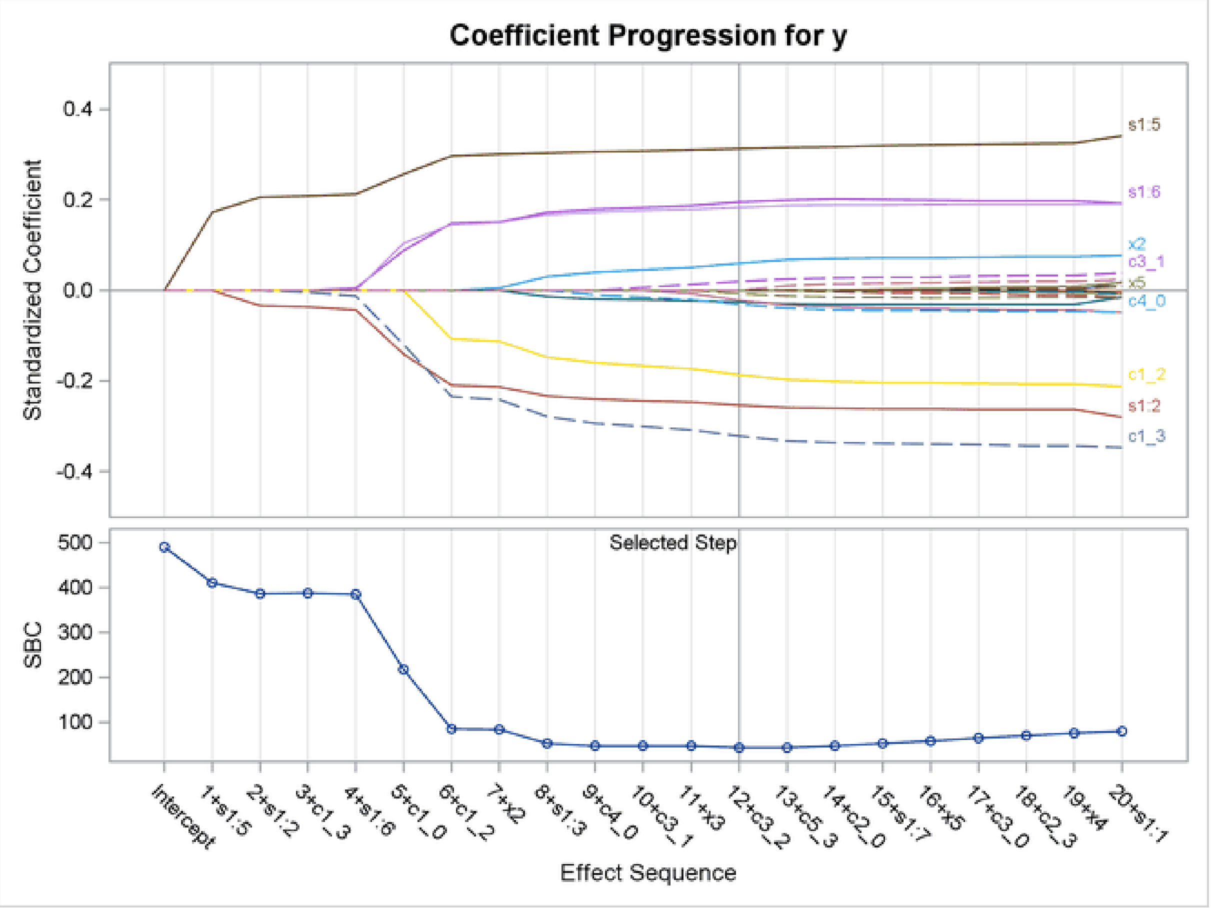


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \le s$ *and* $\beta_1^2 + \beta_2^2 \le s$*, while the red ellipses are the contours of the RSS.*



LASSO

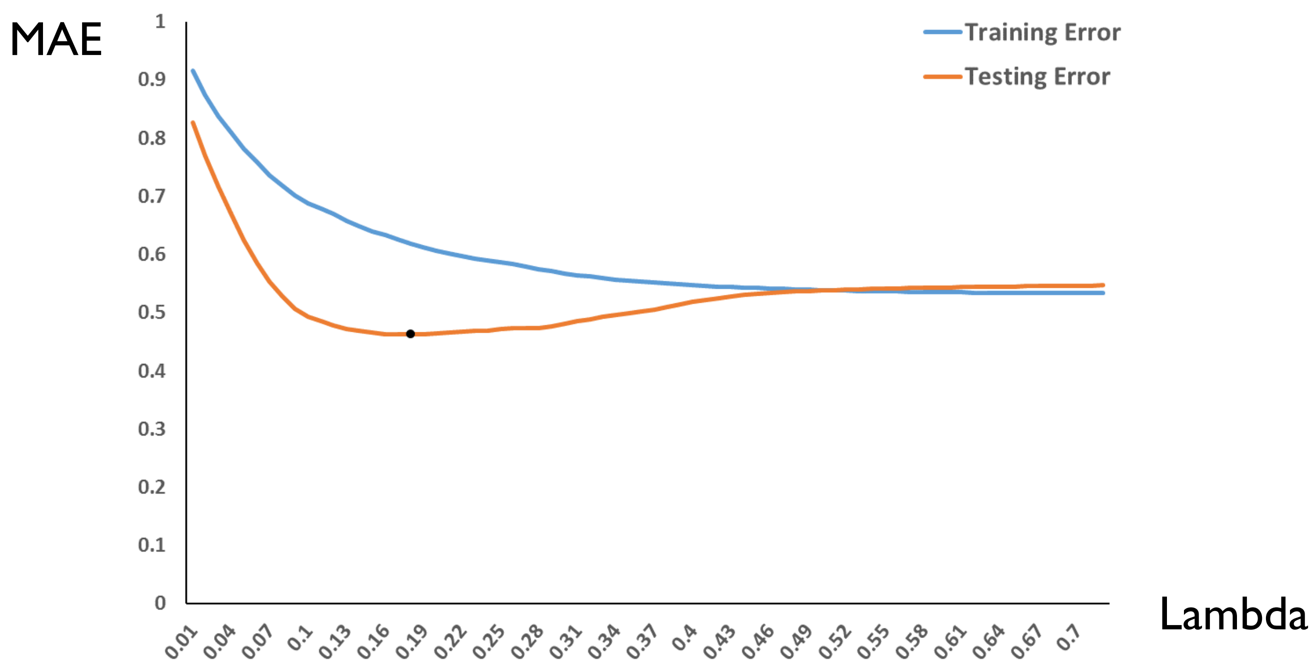
- LASSO: Least Absolute Shrinkage and Selection Operator
  - ✓ Example of estimated regression coefficients according to different λ



http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_glmselect_examples08.htm

LASSO

- LASSO: Least Absolute Shrinkage and Selection Operator
  - ✓ determine the best λ with the highest regression performance



  - ✓ Limitation: Both variable selection and regression performance degenerate if variables are highly correlated

Elastic Net

- Elastic Net
  - ✓ Can have advantages of both Ridge (considering correlation between variables) and LASSO (variable selection ability)
  - ✓ Multiple Linear Regression

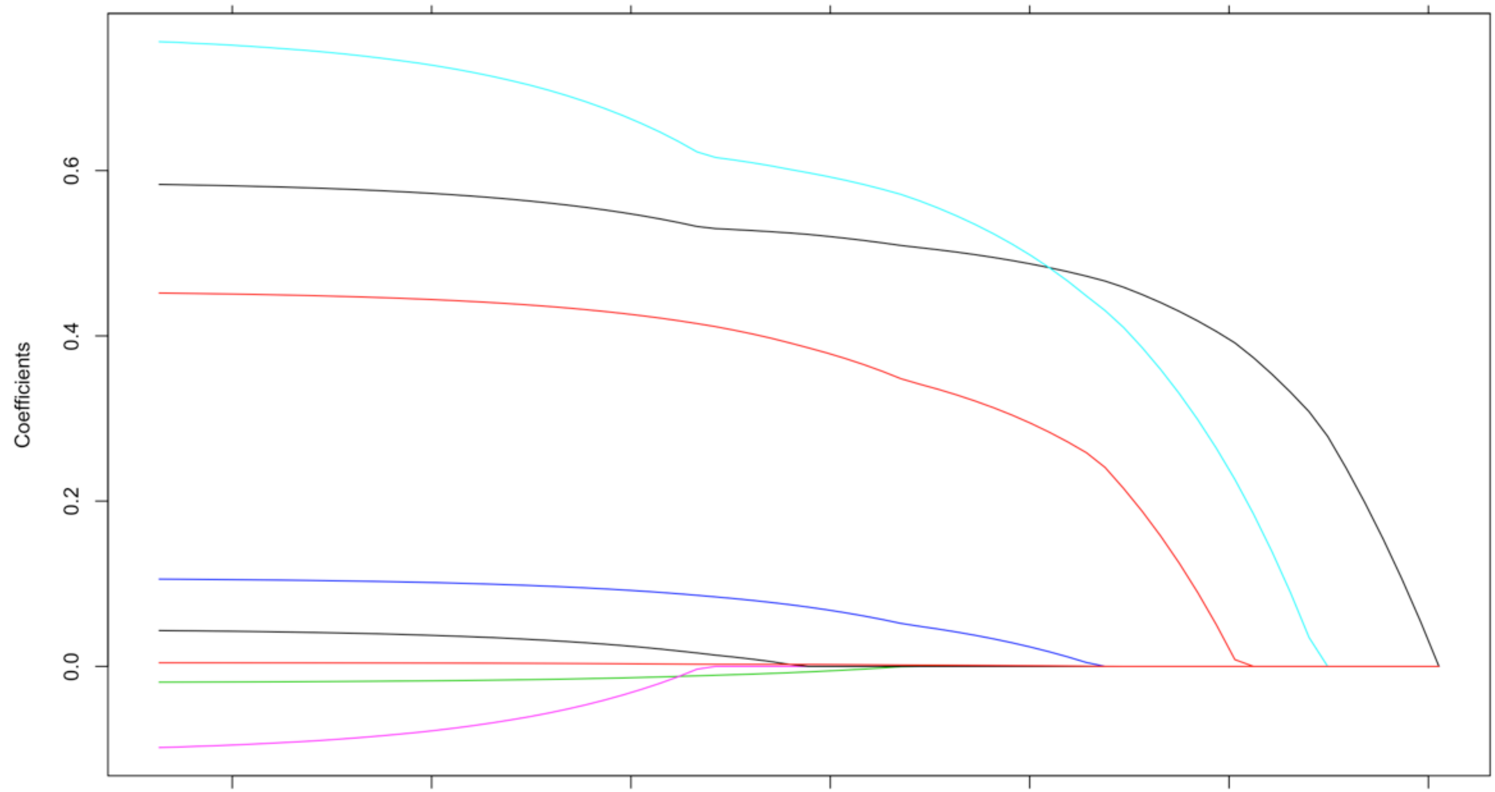
$$\frac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{d}\hat{\beta}_j x_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{d}|\hat{\beta}_j| + \lambda_2 \sum_{j=1}^{d}\hat{\beta}_j^2$$

  - ✓ Logistic Regression

$$-\sum_{i=1}^{n}\left(y_i \log\left(\frac{1}{1+\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}\right) + (1-y_i)\log\left(\frac{\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}{1+\exp(-\sum_{j=0}^{d}\hat{\beta}_j x_j)}\right)\right)$$

$$+\lambda_1 \sum_{j=1}^{d}|\hat{\beta}_j| + \lambda_2 \sum_{j=1}^{d}\hat{\beta}_j^2$$

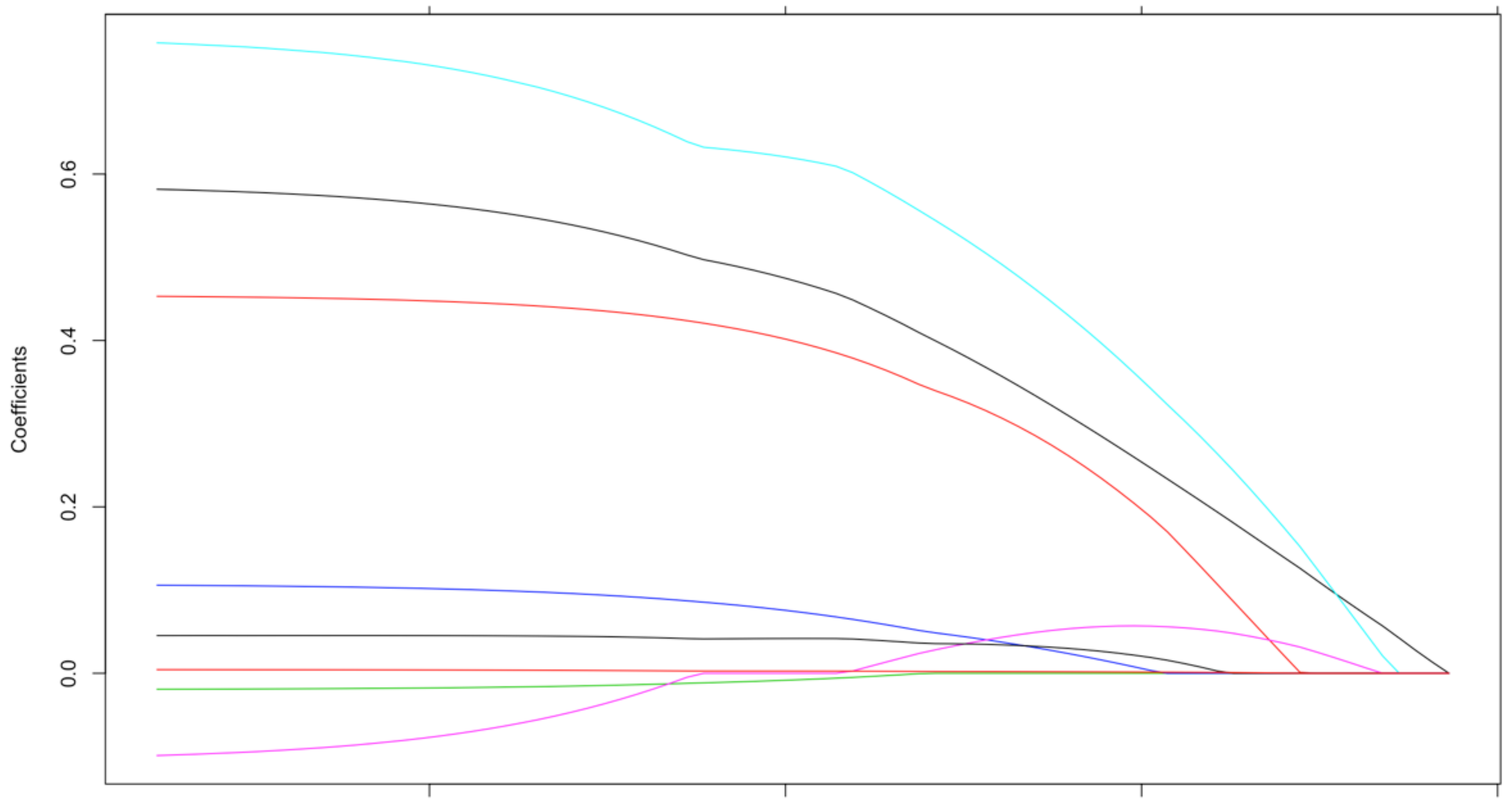
Elastic Net



λ_1 Increases

Number of variable decreases

Elastic Net

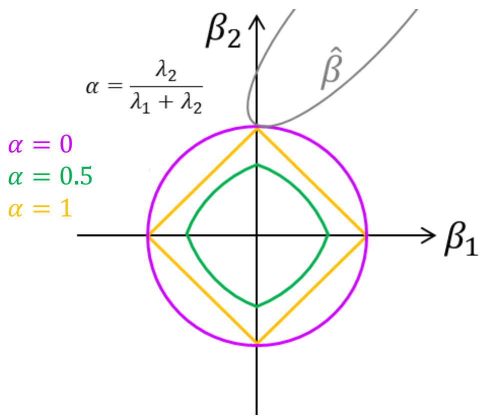


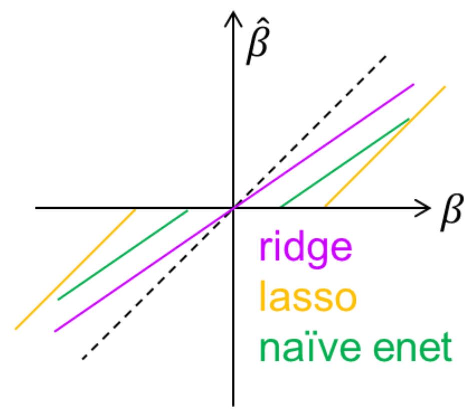
λ_2 Increases

Little impact on variable selection

Empirical Study

- Compare four variable selection methods and three shrinkage methods
  - ✓ Variable selection: Forward selection, Backward elimination, Stepwise selection, GA
  - ✓ Shrinkage: Ridge, Lasso, Elasitic Net





Ridge	$\hat{\beta} = \min_{\beta} \lvert Y - X\beta \rvert^2 + \lambda_1 \lvert \beta \rvert^2$	shrinkage
Lasso	$\hat{\beta} = \min_{\beta} \lvert Y - X\beta \rvert^2 + \lambda_2 \lvert \beta \rvert^1$	shrinkage, variable selection
Elastic net	$\hat{\beta} = \min_{\beta} \lvert Y - X\beta \rvert^2 + \lambda_2 \lvert \beta \rvert^1 + \lambda_1 \lvert \beta \rvert^2$	shrinkage, variable selection, grouping effect

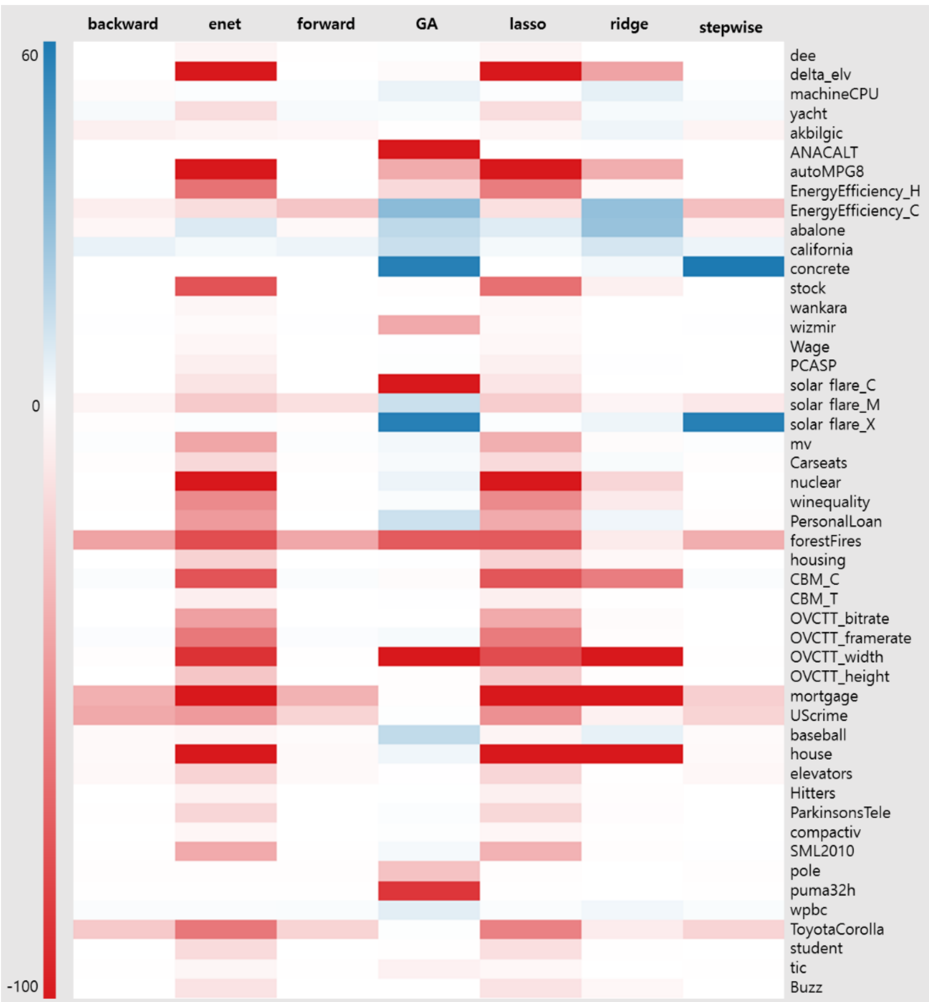
Empirical Study

- Data sets: 49 regression data sets

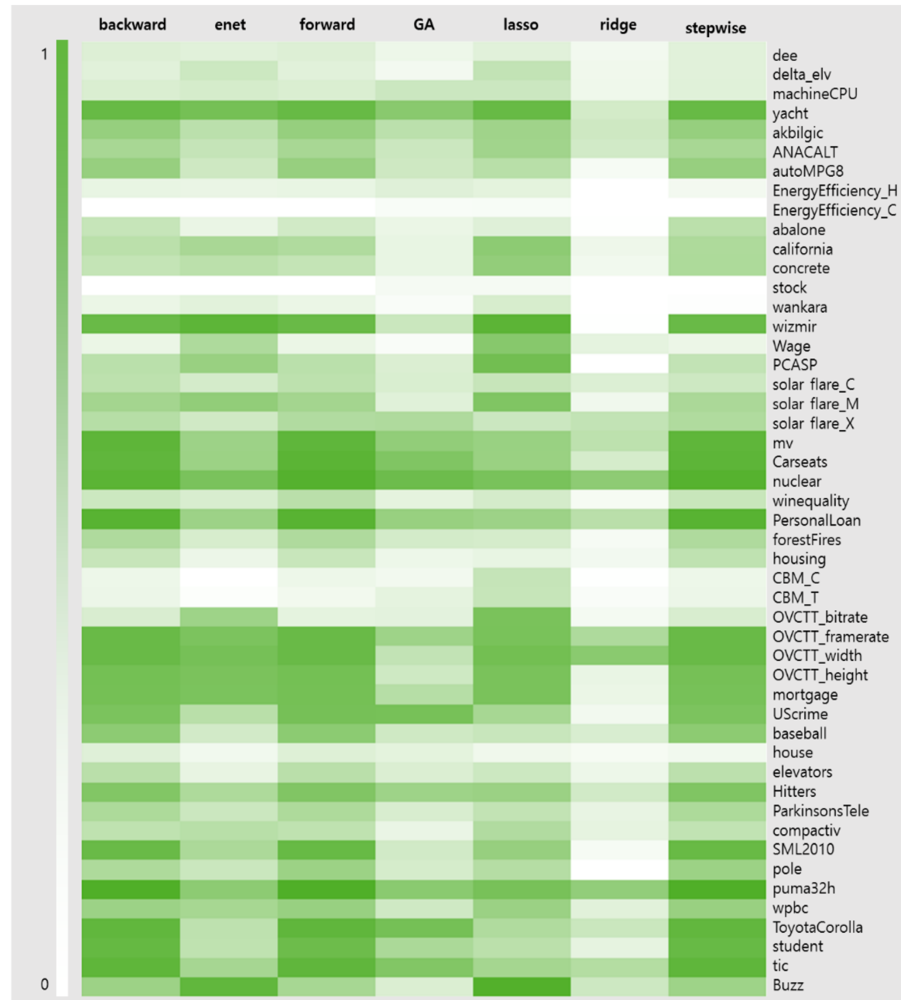
Dataset	source	records	variables	Dataset	source	records	variables
abalone	KEEL	4,177	9	OVCTT_bitrate	UCI	68,784	16
akbilgic	UCI	536	8	OVCTT_framerate	UCI	68,784	16
ANACALT	KEEL	4,052	8	OVCTT_height	UCI	68,784	16
autoMPG8	KEEL	392	8	OVCTT_width	UCI	68,784	16
baseball	KEEL	336	17	ParkinsonsTele	UCI	5,875	21
Buzz	UCI	28,179	95	PersonalLoan	etc.	2,500	13
california	KEEL	20,640	9	PCASP	UCI	45,730	10
Carseats	R	400	11	pole	KEEL	14,998	27
CBM_C	UCI	11,934	15	puma32h	KEEL	4,124	33
CBM_T	UCI	11,934	15	SML2010	UCI	4,137	24
compactiv	KEEL	8,192	22	solar flare_C	UCI	323	11
concrete	KEEL	1,030	9	solar flare_M	UCI	323	11
dee	KEEL	365	7	solar flare_X	UCI	323	11
delta_elv	KEEL	9,517	7	stock	KEEL	950	10
elevators	KEEL	16,599	19	student	UCI	382	51
EnergyEfficiency_C	UCI	768	9	tic	KEEL	9,822	86
EnergyEfficiency_H	UCI	768	9	ToyotaCorolla	etc.	1,436	34
forestFires	KEEL	517	13	UScrime	R	47	16
Hitters	R	263	20	Wage	R	3,000	10
house	KEEL	22,784	17	wankara	KEEL	1,609	10
housing	UCI	506	14	winequality	UCI	6,497	12
machineCPU	KEEL	209	7	wizmir	KEEL	1,461	10
mortgage	KEEL	1,049	16	wpbc	UCI	194	34
mv	KEEL	40,768	11	yacht	UCI	308	7
nuclear	R	32	11				

Empirical Study

Error Rate Improvement

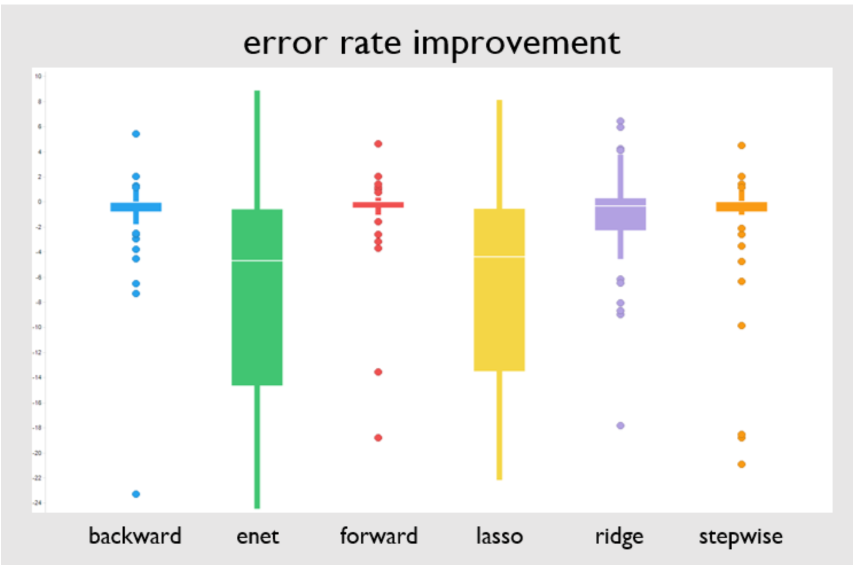


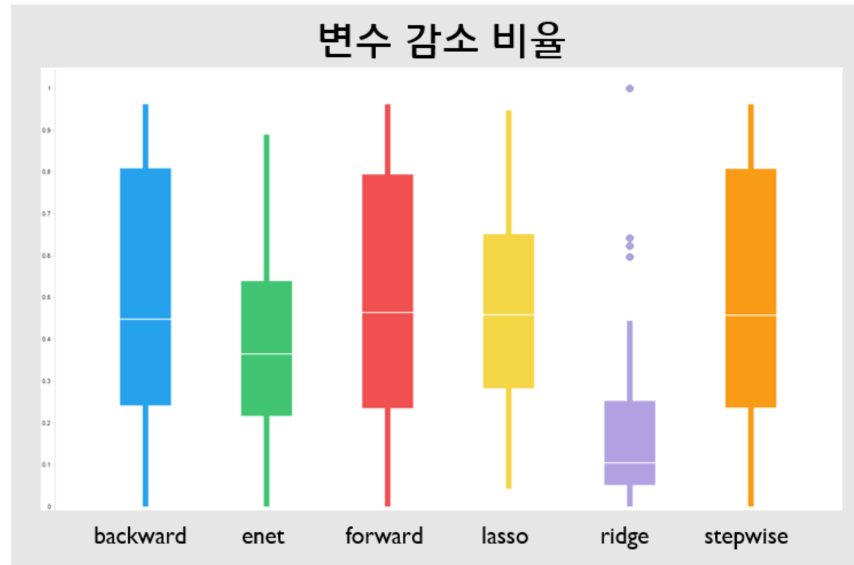
Variable Reduction Ratio



Empirical Study

- Performance comparison





변수선택 방법	예측 정확도	변수 감소율	계산 효율성
Forward	4	4	1
Backward	3	3	2
Stepwise	2	2	6
Ridge	1	6	5
Lasso	6	1	3
Elastic Net	5	5	4


ANY
questions?