REGULARIZATION

BEST SUBSET SELECTION

p FEATURES, LINEAR REGRESSION

# FEATURES	# MODELS	MODELS
p	1	$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$
$p-1$	$\binom{p}{2}$	TOO MANY TO LIST
$\vdots$	$\vdots$	
2	$\binom{p}{2}$	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, $Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$, ...
1	p	$Y = \beta_0 + \beta_1 x_1 + \varepsilon$, $Y = \beta_0 + \beta_2 x_2 + \varepsilon$, ..., $Y = \beta_0 + \beta_p x_p + \varepsilon$
0	1	$Y = \beta_0$
	2^p	

To save on computation → "Search"

Step Direction + Metric

Backwards → AIC → Default in R's step( )

Forwards

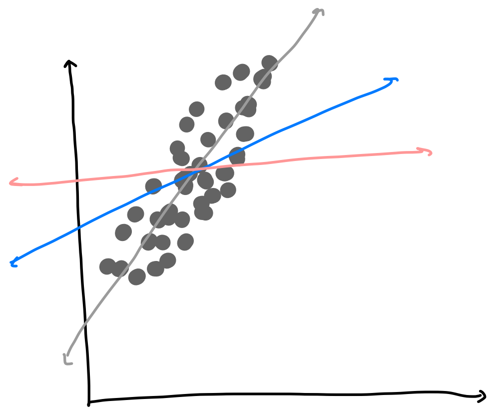
Both

BIC

Cp

Etc

Related to LOOCV asymptotically



TRUE MODEL $Y = 2 + 5x + \varepsilon$

BIAS

$$\min_{B_0, B_1} \sum_{i=1}^{n} \left(y_i - (B_0 + B_1 x_i) \right)^2 \longrightarrow \hat{B}_1 = 5.2$$

$$\min_{B_0, B_1} \sum_{i=1}^{n} \left(y_i - (B_0 + B_1 x_i) \right)^2 \quad \text{SUBJECT TO} \quad |B_1| \le 3 \longrightarrow \hat{B}_1 = 3$$

$$\min_{B_0, B_1} \sum_{i=1}^{n} \left(y_i - (B_0 + B_1 x_i) \right)^2 \quad \text{SUBJECT TO} \quad |B_1| \le 0 \longrightarrow \hat{B}_1 = 0$$

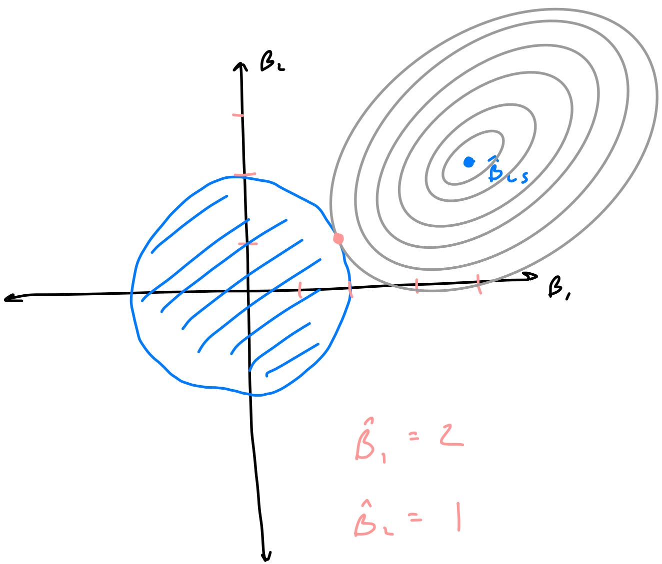
VAR

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\hat{\beta}_1 = 4, \quad \hat{\beta}_2 = 2$$

$$\sum \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})\right)^2$$

SUBJECT TO $\beta_1^2 + \beta_2^2 \leq 4$

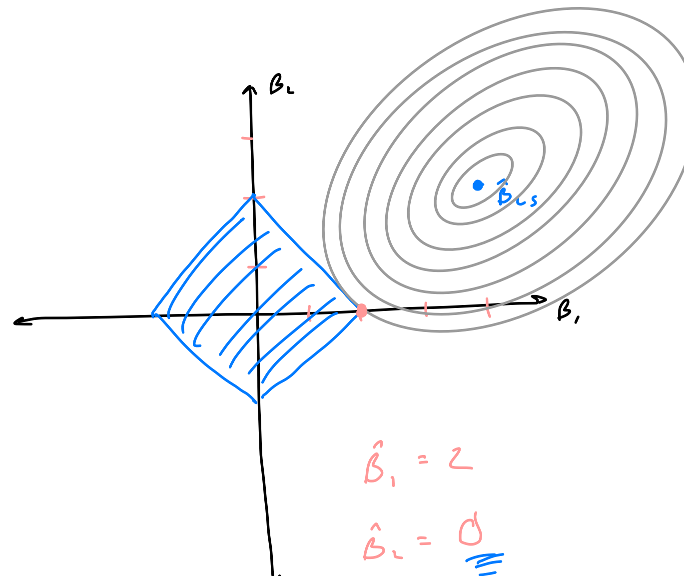


$\hat{\beta}_1 = 2$

$\hat{\beta}_2 = 1$

$$\sum \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})\right)^2$$

SUBJECT TO $|\beta_1| + |\beta_2| \leq 2$



$\hat{\beta}_1 = 2$

$\hat{\beta}_2 = 0$

LS

CONSTRAINT

$$\min \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 \le s$$

?

"BUDGET"

RIDGE

$$\min \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \le s$$

LASSO

$$\min \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \ne 0) \le s$$

BEST SUBSET SELECTION

RIDGE AND LASSO ARE GREAT WHEN p IS LARGE

LASSO DOES SELECTION!

$$\min \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le s$$

$s = 0 \longrightarrow \beta_0$

$s = \infty \longrightarrow$ OLS

$s \longleftrightarrow \lambda$

$$\min \left[\underbrace{\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2}_{\text{ERROR}} + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{PENALTY}} \right]$$

TUNING

TRADEOFF

$\lambda = 0 \longrightarrow$ OLS

$\lambda = \infty \longrightarrow \beta_0$

AS LS ERROR $\downarrow$, PENALTY $\uarrow$

A note about scaling

y	x	x^*
⋮	900,000	-1
⋮	⋮	⋮
⋮	1,000,000	0
⋮	⋮	⋮
⋮	1,100,000	1

$$x^* = \frac{x - \bar{x}}{SD[x]}$$

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad \rightarrow \quad \hat{\beta}_1 = 0.001$$

$$Y = \beta_0 + \beta_1 x^* + \varepsilon \quad \rightarrow \quad \hat{\beta}_1 = 1000$$

BIG EFFECT ON

$$\sum_{j=1}^{p} \beta_j^2$$

R

glmnet::cv.glmnet → FINDS "GOOD" λ VALUE WITH CV

alpha = 1 ⟹ LASSO DEFAULT

alpha = 0 ⟹ RIDGE

INTERNALLY HANDLES SCALING!

ALSO WORKS WITH LOGISTIC REGRESSION !