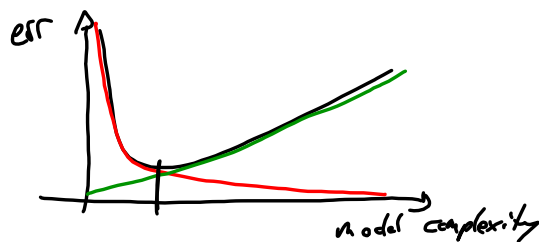


PENALIZED REGRESSION



LS $\rightarrow$ Best
 Linear
 Unbiased
 Estimator

IDEA: trade some bias in order to reduce the variance, and consequently get a smaller prediction error.

↑
of penalized regression

$$RSS(\beta) + \lambda \text{penalty}(\beta)$$

RIDGE REGRESSION

Consider a model $y = X\beta + \epsilon$

Ridge regression estimates the regression coefficients β by minimizing

$$\hat{\beta}_{\text{ridge}} = \arg \min_{\beta} \sum_{i=1}^n (y_i - \beta_0 - x_{ij}\beta_j)^2 + \underbrace{\lambda \sum_{j=1}^p \beta_j^2}_{\text{penalty}}$$

$\rightarrow$ shrinkage: $\hat{\beta}_{\text{ridge}}$ are closer to 0 w.r.t $\hat{\beta}_{\text{OLS}}$

$$\hat{\beta}_{\text{ridge}} = \arg \min_{\beta} \|y - X\beta\|_2^2 + \lambda \beta^T \beta$$

$$1 \quad \frac{\partial}{\partial \beta} (\|y - X\beta\|_2^2 + \lambda \beta^T \beta)$$

$$-2X^T(y - X\beta) + 2\lambda\beta = 0$$

$$X\beta + X^T X \beta = X^T y$$

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

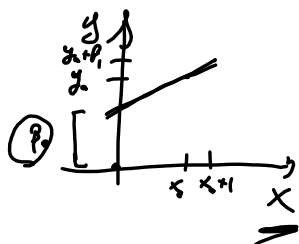
ridge estimator $\lambda I = \begin{pmatrix} \lambda & & 0 \\ & \lambda & \\ 0 & & \lambda \end{pmatrix}$

The hyperparameter (or tuning parameter) λ controls how much penalty I want to add.

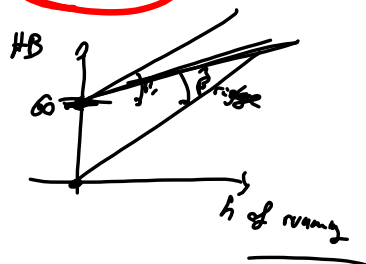
smaller λ , more complex model $\rightarrow \lambda = 0 \rightarrow \hat{\beta}_{\text{ridge}} = \hat{\beta}_{\text{OLS}}$

larger λ , simpler model $\rightarrow \lambda = +\infty \rightarrow \hat{\beta}_{\text{ridge}} = \hat{\beta}_0$

The penalty affects only the regression coefficients, not the intercept, $\beta_0 = \bar{y}$. In order to have a "fair penalization", we need to standardize the X .



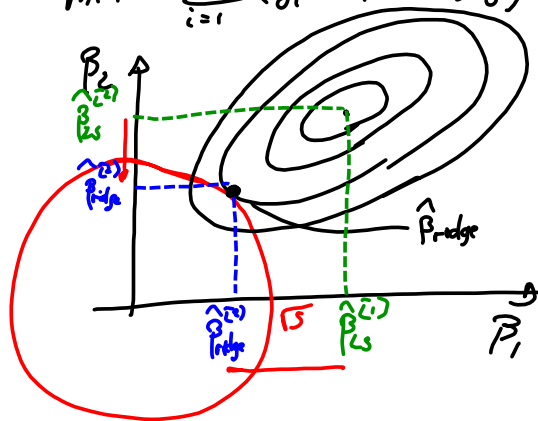
$$y = \beta_0 + \beta_1 x + \epsilon$$



penalized regression can also be seen as constrained optimization

$$\min \sum_{i=1}^n (y_i - \beta_0 - x_{ij}\beta_j)^2 \quad \text{subject to } \sum_{j=1}^p \beta_j^2 \leq S$$

$$\beta = (\beta_0, \beta_1)$$



Note that, due to correlation

$$\lambda_a > \lambda_b \Rightarrow |\hat{\beta}_j(\lambda_a)| < |\hat{\beta}_j(\lambda_b)|$$

$$\frac{\text{Bias}}{E}[\hat{\beta}_{LS}] = \beta \quad \text{LS is an unbiased estimator}$$

$$\begin{aligned} E[\hat{\beta}_{\text{bridge}}] &= E[(X^T X + \lambda I)^{-1} X^T y] \\ &= E[(X^T X + \lambda I)^{-1} (X^T X) \underbrace{(X^T X)^{-1} X^T y}_{\hat{\beta}_{LS}}] \end{aligned}$$

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

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

$$= W(\lambda) E[\hat{\beta}_{LS}] = W(\lambda) \beta$$

symmetric
unbiased only if $W(\lambda) = I$,
i.e., $\lambda = 0$

$$\frac{\text{Variance}}{\text{Var}}(\hat{\beta}_{\text{bridge}}) = \text{Var}(W(\lambda) \hat{\beta}_{LS}) = W(\lambda) \text{Var}(\hat{\beta}_{LS}) W(\lambda) = \sigma^2 W(\lambda) (X^T X)^{-1} W(\lambda)$$

$$\begin{aligned} \text{Var}(\hat{\beta}_{LS}) - \text{Var}(\hat{\beta}_{\text{bridge}}) &= \sigma^2 [(X^T X)^{-1} - W(\lambda) (X^T X)^{-1} W(\lambda)] \\ &= \sigma^2 W(\lambda) [(I + \lambda (X^T X)^{-1}) (X^T X)^{-1} (I + \lambda (X^T X)^{-1}) - (X^T X)^{-1}] W(\lambda) \\ &= \sigma^2 W(\lambda) [\cancel{(X^T X)^{-1}} + \lambda (X^T X)^{-2} + \lambda (X^T X)^{-2} + \lambda^2 (X^T X)^{-3} - \cancel{(X^T X)^{-1}}] W(\lambda) \\ &= \sigma^2 W(\lambda) [2\lambda (X^T X)^{-2} + \lambda^2 (X^T X)^{-3}] W(\lambda) \neq 0 \end{aligned}$$

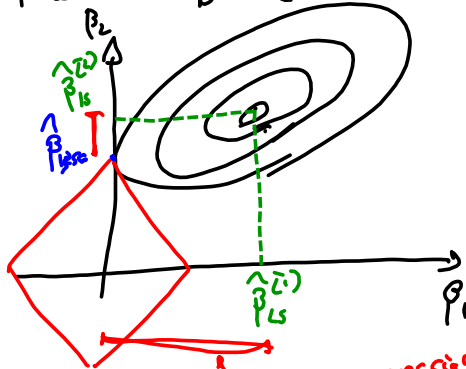
$$\Rightarrow \text{Var}(\hat{\beta}_{LS}) \geq \text{Var}(\hat{\beta}_{\text{bridge}})$$

② λ is often computed by resampling technique like cross validation or bootstrapping

LASSO (Least Angle Shrinkage and Selection Operator)

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

$$\hat{\beta}_{\text{Lasso}} = \arg \min_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - x_{ij} \beta_j)^2 \right\} \text{ subject to } \sum_{j=1}^p |\beta_j| \leq s$$



Some regression coefficients are forced to be 0, i.e., the related variable excluded from the model

sparser = with less covariates

main advantage of Lasso: sparser models

main disadvantage of Lasso: we do not have a closed form for the estimator
 $\Rightarrow$ must be computed numerically

penalized regression (BRIDGE REGRESSION)

$$\sum_{i=1}^n (y_i - \beta_0 - x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|^q$$

$q=0$ subset selection

$q=1$ Lasso

$q=2$ ridge regression

Lasso has the advantages of variable selection
 ridge regression is much better in handling the correlation among the covariates

We would like to have both the advantages of Lasso and ridge:

$1 < q < 2$ does not really work, we lose the properties of the Lasso

$\rightarrow$ alternative that works: elastic net

$$\hat{\beta}_{\text{elnet}} = \arg \min_{\alpha} \left\{ \sum_{i=1}^n (y - \beta_0 - x_{ij} \beta_j)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\}$$

$\swarrow$ L_1 loss $\swarrow$ L_2 loss

$$+ \lambda \sum_{j=1}^p [\alpha |\beta_j| + (1-\alpha) \beta_j^2]$$

$$\lambda \geq 0, \quad 0 \leq \alpha \leq 1$$

$\uparrow$ ridge $\uparrow$ Lasso