29. Ridge Regression, part 2

November 21, 2025 10:00 AM

Ridge Regression: $\hat{\beta}^R_\lambda = \underset{\tilde{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ \underbrace{\sum_{i=1}^n (Y_i - X_i^T \tilde{\beta})^2}_{\text{Squared Error}} + \lambda \underbrace{\sum_{j=1}^p \tilde{\beta}_j^2}_{\text{penalty on using "big" values for } \tilde{\beta}} \right\}$

$$\hat{\beta}^R_\lambda = (X^TX + \lambda I_p)^{-1} X^T Y$$

λI_p is "stablizing" the inverse of X^TX

- if we have Multicolinearity then $(X^TX)^{-1}$ is unstable leading to a large variance in $\hat{\beta}$
- if $p > n$ then $(X^TX)^{-1}$ Does not exist.

However $(X^TX + \lambda I_p)^{-1}$ always exists for any $\lambda > 0$.

$\Rightarrow$ Reduce variance and stop us from overfitting

Theorem: there exists a $\lambda > 0$, small enough, s.t.

$$MSE(\hat{\beta}^R_\lambda) < MSE(\hat{\beta})$$

Ridge < Least Squares

Proof: • $MSE(\hat{\beta}) = Var(\hat{\beta})$ Unbiased!

$= \sigma^2 (X^TX)^{-1}$ For simplicity, we assume this exists.

• $MSE(\hat{\beta}^R_\lambda) := E[(\hat{\beta}^R_\lambda - \beta)(\hat{\beta}^R_\lambda - \beta)^T]$

$= Var(\hat{\beta}^R_\lambda) + bias(\hat{\beta}^R_\lambda)\, bias(\hat{\beta}^R_\lambda)^T$

Here we use an outer product $\hat{\beta}\hat{\beta}^T$ a matrix instead of inner product $\hat{\beta}^T\hat{\beta} \in \mathbb{R}$

Note: $tr(\hat{\beta}\hat{\beta}^T) = \hat{\beta}^T\hat{\beta}$

• $bias(\hat{\beta}^R_\lambda) = E\hat{\beta}^R_\lambda - \beta$

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

$= (X^TX + \lambda I_p)^{-1} X^T\, EY - \beta$

$= (X^TX + \lambda I_p)^{-1} X^T X \beta - \beta$

$= [\underbrace{(X^TX + \lambda I_p)^{-1} X^T X}_{< I} - I] \beta$

$= [(X^TX + \lambda I_p)^{-1} X^TX - (X^TX)^{-1}(X^TX)] \beta$

$= [(X^TX + \lambda I_p)^{-1} - (X^TX)^{-1}] (X^TX) \beta$

Point: Bias is negative wrt β i.e. if $\beta_i > 0$

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

Point: Bias is negative wrt β. i.e. if $\beta_i > 0$ then the $\text{bias}(\hat{\beta}^R_{\lambda i}) < 0$ and if $\beta_i < 0$ then the $\text{bias}(\hat{\beta}^R_{\lambda i}) > 0$

$\Rightarrow$ Shrinkage estimator as the bias is always point towards zero.

$$\text{Var}(\hat{\beta}^R_\lambda) = \text{Var}\left((X^T X + \lambda I_p)^{-1} X^T Y \right)$$

$$= \left[(X^T X + \lambda I_p)^{-1} X^T \right] \text{Var}(Y) \left[(X^T X + \lambda I_p)^{-1} X^T \right]^T$$

$$= \left[(X^T X + \lambda I_p)^{-1} X^T \right] (\sigma^2 I_n) \left[X \underbrace{(X^T X + \lambda I_p)^{-1}}_{\text{Symmetric}} \right]$$

$$= \sigma^2 (X^T X + \lambda I_p)^{-1} (X^T X) (X^T X + \lambda I_p)^{-1}$$

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \cancel{(X^T X)} \cancel{(X^T X)}^{-1}$$

Multiplied by $(X^T X)^{-1}(X^T X)$ and grouped terms together

$$= \sigma^2 (X^T X)^{-1} \left[(X^T X)(X^T X + \lambda I_p)^{-1} \right]^2$$

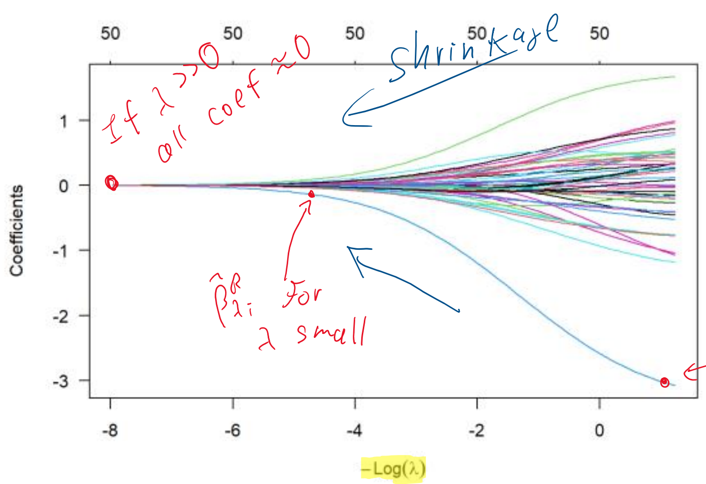
Claim: $< I_p$ then we have reduced the variance.

Simple Case: $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

Now, $(X^T X)(X^T X + \lambda I_p)^{-1} = \dfrac{\sum x_i^2}{\sum x_i^2 + \lambda} < 1$

```
# Run a ridge regression
md.ridge = glmnet( xx,yy,alpha=0 )
plot( md.ridge,las=1 )
```

Output from "glm net"



Each line corresponds to a $\hat{\beta}^R_{\lambda, i}$

If $\lambda \approx 0$ then $\tilde{\beta}^R_\lambda \approx \hat{\beta}$

$\hat{\beta}^R_{\lambda i}$ for λ big

Interpolating between $\hat{\beta}$, 0/1, and 0

Next time: go from Ridge to Lasso estimator

Next time: go from Ridge to lasso estimator

$\lambda \sum \tilde{\beta}_i^2$ $\lambda \sum |\beta_i|$

lasso = Least Absolute Shrinkage + Selection Operator

$|\beta|$ $\downarrow 0$ set $= 0$

↳ Do variable selection, but still maintains a convex optimization problem.