<u>Situation:</u> $\underset{N\times 1}{Y} = \overset{N\times p}{X}\overset{p\times 1}{\beta} + \overset{N\times 1}{\varepsilon}$ $\quad E(\varepsilon)=0, \ Var(\varepsilon)=\sigma^2 I$

when $\underset{p\times p}{X^TX = I}$ $\quad$ (and also $(X^TX)^{-1} = I$)

(as before we have centered x's and y, so β_0 is not needed)

1) <u>Least squares</u> ($\overset{LS}{OLS}$)

$$\hat\beta_{LS} = (X^TX)^{-1}X^TY = X^TY$$

$$E(\hat\beta_{LS}) = X^T E(Y) = \underbrace{X^TX}_{I}\beta = \beta \qquad \text{(we knew that)}$$

$$Cov(\hat\beta_{LS}) = X^T Cov(Y) X = (X^TX)\sigma^2 I = \sigma^2 I$$

$$MSE(\hat\beta_{LS}) = (E(\underbrace{\hat\beta_{LS}) - \beta}_{0})^T (E(\hat\beta_{LS}) - \beta) + tr(\underbrace{Var(\hat\beta_{LS})}_{\underset{p\times p}{\sigma^2 I}})$$

general formula

$$= p\sigma^2$$

2) <u>Ridge regression</u>

$$\hat\beta(\lambda) = (\underbrace{X^TX}_{I} + \lambda I)^{-1}X^TY = \underset{p\times p}{\begin{bmatrix} 1+\lambda & & 0 \\ & \ddots & \\ 0 & & 1+\lambda \end{bmatrix}^{-1}} \underset{p\times N \ N\times 1}{X^TY}$$

$$= \left(\frac{1}{1+\lambda}\right) \cdot \underset{p\times 1}{X^TY}$$

$$= (1+\lambda)^{-1}\hat\beta_{LS}$$

Can use results for $\hat\beta_{LS}$ to get E, Cov, MSE for $\hat\beta(\lambda)$

$$E(\hat\beta(\lambda)) = (1+\lambda)^{-1}\underbrace{E(\hat\beta_{LS})}_{\beta} = (1+\lambda)^{-1}\beta$$

$$\mathrm{Cov}(\hat{\beta}(\lambda)) = (1+\lambda)^{-1}\,\mathrm{Cov}(\hat{\beta}_{LS})\,(1+\lambda)^{-1}$$

$$= (1+\lambda)^{-2} \cdot \sigma^2 I = \frac{\sigma^2}{(1+\lambda)^2} \underset{p\times p}{I}$$

$$\mathrm{MSE}(\hat{\beta}(\lambda)) = \left((1+\lambda)^{-1}\beta - \beta\right)^T\left((1+\lambda)^{-1}\beta - \beta\right)$$

$$+ \mathrm{tr}\left((1+\lambda)^{-2}\sigma^2 \underset{p\times p}{I}\right)$$

$$= \underbrace{\left(\frac{1}{1+\lambda} - 1\right)}_{\frac{1-1-\lambda}{1+\lambda}} \beta^T\beta \left(\frac{1}{1+\lambda} - 1\right) + (1+\lambda)^{-2} p \sigma^2$$

$$= \underline{\lambda^2 (1+\lambda)^{-2} \beta^T\beta + p(1+\lambda)^{-2}\sigma^2}$$

When is $\mathrm{MSE}(\hat{\beta}(\lambda))$ minimal wrt λ?

$$\frac{\lambda^2}{(1+\lambda)^2}$$

$$\frac{d\,\mathrm{MSE}(\hat{\beta}(\lambda))}{d\lambda} = \frac{2\lambda \cdot (1+\lambda)^2 - \lambda^2 \cdot 2(1+\lambda)}{(1+\lambda)^4} \beta^T\beta$$

$$- 2p(1+\lambda)^{-3}\sigma^2 = 0$$

$$\frac{2\lambda\left((1+\lambda) - \lambda\right)}{(1+\lambda)^3}\beta^T\beta - \frac{2p\sigma^2}{(1+\lambda)^3} = 0$$

$$\lambda\beta^T\beta - p\sigma^2 = 0 \Leftrightarrow \underline{\underline{\lambda = \frac{p\sigma^2}{\beta^T\beta}}}$$

This means that $\lambda_{opt} = \sigma^2 \cdot \frac{p}{\beta^T\beta}$

Is this a minimum? Yes - could have checked!

See plots in L7.html under solutions.

Theobald 1974: There exists $\lambda > 0$ such that $MSE(\hat{\beta}(\lambda)) < MSE(\hat{\beta}(0))$

Proof in WNvW show $\lambda < 2\sigma^2 (\beta^T\beta)^{-1}$ is sufficient for

$MSE(\hat{\beta}(\lambda)) < MSE(\hat{\beta}(0))$, but for us the optimal λ is larger than

this upper limit.