maybe?

Lecture 14: Ridge Regression

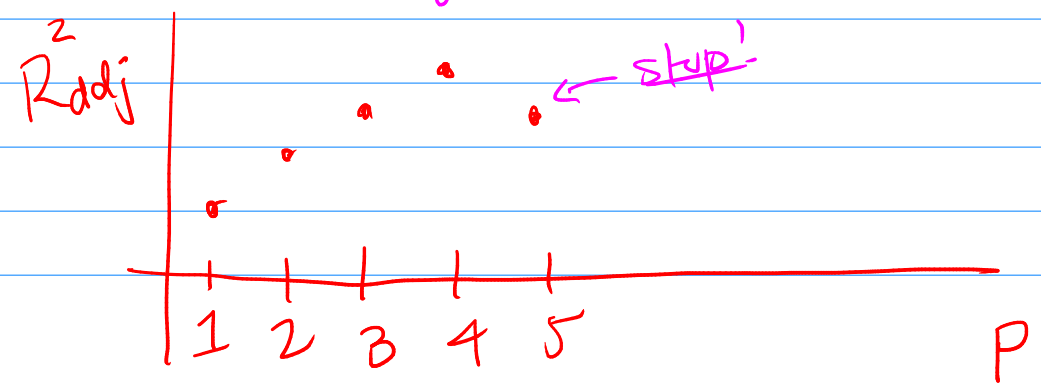
Ideally: If I have P potential covariates then I could calc. a penalized metric on all potential models and choose model w/ best metric.

Problem: generally there are 2^P possible models

Use a greedy approach

1. Forward Stepwise Selection
   - i) start w/ model just containing intercept
   - ii) add best variable to model → choose var. that improves metric most
   - iii) Calc. my penalized metric
   - iv) Loop to step ii)

Stop when my penalized metric gets worse



(2) Backwards Stepwise Selection

→ starts w/ a model w/ all vars
→ removes one at a time

Trying all possible subsets is a very flexible model building process

Might be better off w/ a more constrained stepwise search.

(think flexibility and Bias-Var tradeoff)

Ridge Regression

Q: Can we deal with ill-conditioning in a continuous way?

Recall for OLS (ordinary least squares)

$$L(\beta) = RSS(\beta) = \|y - X\beta\|^2$$

and

$$\hat{\beta}^{(OLS)} = \underset{\beta}{\operatorname{argmin}}\, L(\beta)$$

problem is that if some variables are highly co-linear (highly correlated) the assoc. $\hat{\beta}$s tend to blow up ($\pm\infty$)

Ridge regression solves this by penalizing large βs

$$\hat{\beta}^{(ridge)} = \arg\min_{\beta} \left[L(\beta) + \lambda \|\beta\|^2 \right]$$

penalty

$\lambda \geq 0$ = regularization strength

by adding $\lambda\|\beta\|^2$ we penalize/avoid choosy $\hat{\beta}$ w/ really large components

$\lambda = 0 \Rightarrow$ OLS estimates

$\lambda \to \infty \Rightarrow \hat{\beta} \to 0$

(1) Typically: ignore β_0

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

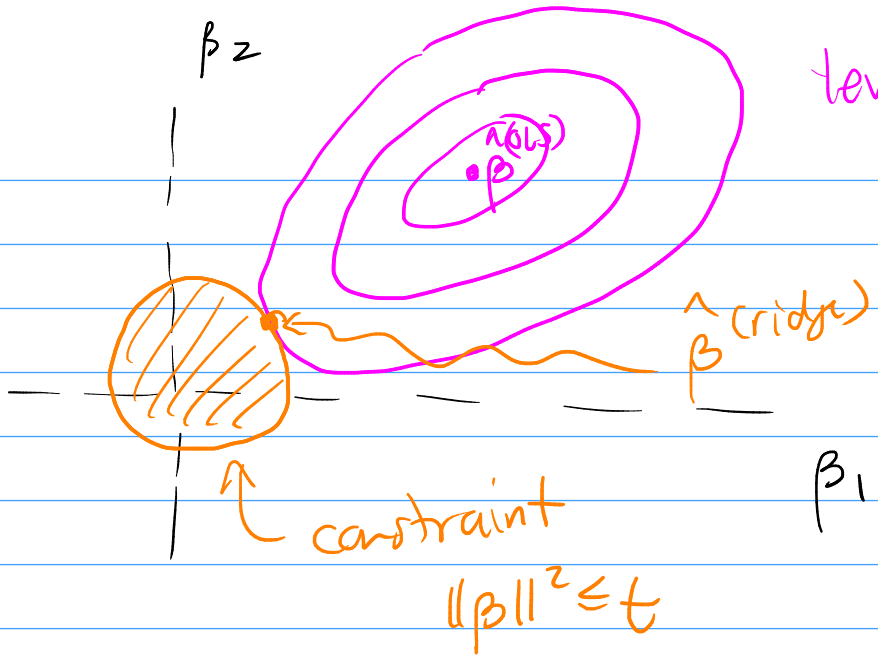
(2) Typically: standardize Xs so that βs are comparable

(3) Typically: choose λ using k-validation

Second interpretation: Ridge is equivalent to

$$\hat{\beta}^{(ridge)} = \arg\min_{\beta} L(\beta) \quad \text{s.t.} \quad \|\beta\|^2 \leq t$$

1-1 corresp. w/ λ



How do I get $\hat{\beta}^{(ridge)}$. Because penalty $\lambda\|\beta\|^2$ is quadratic, there is a closed form soln for $\hat{\beta}^{(ridge)}$.

OLS: had to solve $(X^TX)\beta = X^Ty$ ← may not be invertible

to get $\hat{\beta}^{(ols)} = (X^TX)^{-1}X^Ty$

Ridge: if we take derv. and set to zero we get

$(X^TX + \lambda I)\beta = X^Ty$ ← $\lambda > 0$ this is always invertible

so $\hat{\beta}^{(ridge)} = (X^TX+\lambda I)^{-1}X^Ty$

So I've replaced X^TX w/ $X^TX+\lambda I$

For OLS conditioning depended on $K(X^TX)$

Ridge " " $K(X^TX+\lambda I)$