Lecture 12: Variable Selection

So far procedures have a fixed set of vars.

May want to select a "best" set

Why? (1) prediction accuracy

(2) interpretation

(3) model may be ill-conditioned

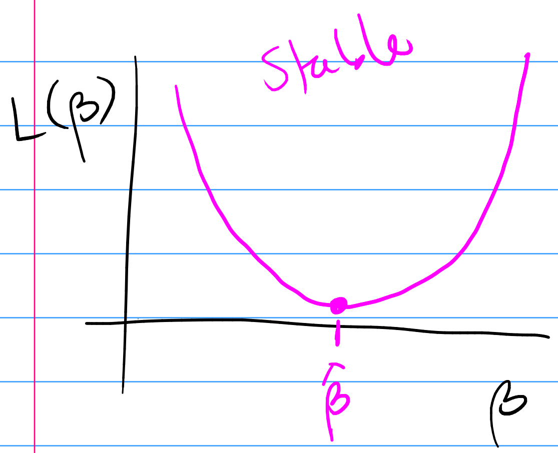
(OLS regression: $P > N$)

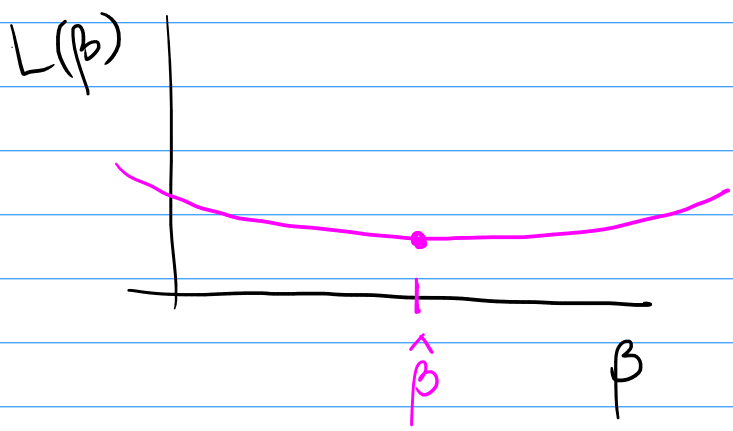
Back to OLS

Recall that $\hat{\beta}$ is obtained by solving

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

the stability of $\hat{\beta}$ depends on inverting $X^T X$





Condition Number

For a linear system $Az = b$

the stability of the solution depends on

the condition Number of A denoted $K(A)$

↳ how sensitive is solution to changes in A or b.

Fact: $K(A) = \dfrac{\sigma_{max}(A)}{\sigma_{min}(A)}$ ← largest singular value of A

smallest sing. val.

Large $K(A)$ = very sensitive linear system

Small $K(A)$ = not so sensitive

$K(A) = \infty \iff A$ isn't invertible.

Why do we care?

Want to solve $\underbrace{(X^TX)}_{A}\underbrace{\beta}_{z} = \underbrace{X^Ty}_{b}$

The stability of $\hat{\beta} = (X^TX)^{-1}X^Ty$ depends on $K(X^TX)$

If $K(X^TX)$ is large then the regression is ill-conditioned — very sensitive to X and Y

Several causes:

Ex, one var is (approx.) a LC of others

(vars. highly correlated)

Ex, $P > N$ (so $K(X^TX) = \infty$)

e.g. X measures $P = 20K$ genes for $N = 30$ patients

How to deal w/ this?

1. Variable selection
2. Shrinkage

Goal of 1 is to pick some subset of "important" variables to use.

how do we define?

how do we pick?

Two approaches:

(1) use some individual metric for each var and choose vars w/ best metric.

e.g. calc. p-val. for each var.

and use set of vars w/ smallest p-vals.

potential problem: performance of one var. may depend on others.

(2) calc. some metric on groups of vars - choose group w/ best metric.

potential problem: w/ P vars I have 2^P possible subsets

Careful not to look at training metric

$RSS_{train} \downarrow$ as $P \uparrow$

Solns: (1) X-val. (comp. expensive)

(2) penalized training metric (classic)

Ex. Adjusted R^2:

$$R^2_{adj.} = 1 - \frac{N-1}{N-P-1}(1-R^2)$$

as $P\uparrow$ eventually $R^2_{adj}\downarrow$

Ex. Mallow's C_p

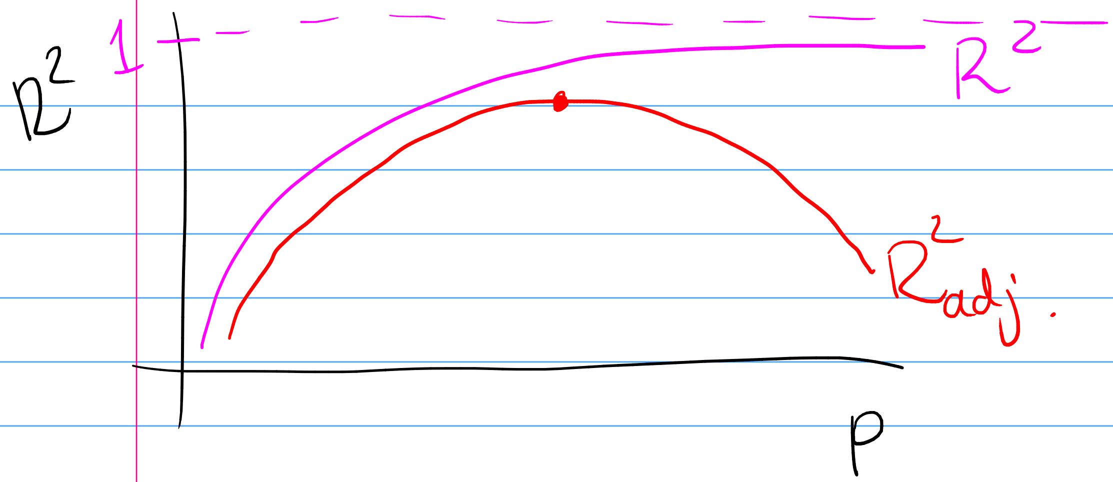
$$C_p = \frac{1}{N}(RSS_{train} + \underbrace{2P\hat{\sigma}^2}_{\text{penalty}})$$

Ex. AIC

$$AIC = \frac{1}{N\hat{\sigma}^2}(RSS_{train} + 2P\hat{\sigma}^2)$$

Ex. BIC

$$BIC = \frac{1}{N}(RSS_{train} + \log(N)P\hat{\sigma}^2)$$



Problem is that potentially have 2^p models to consider

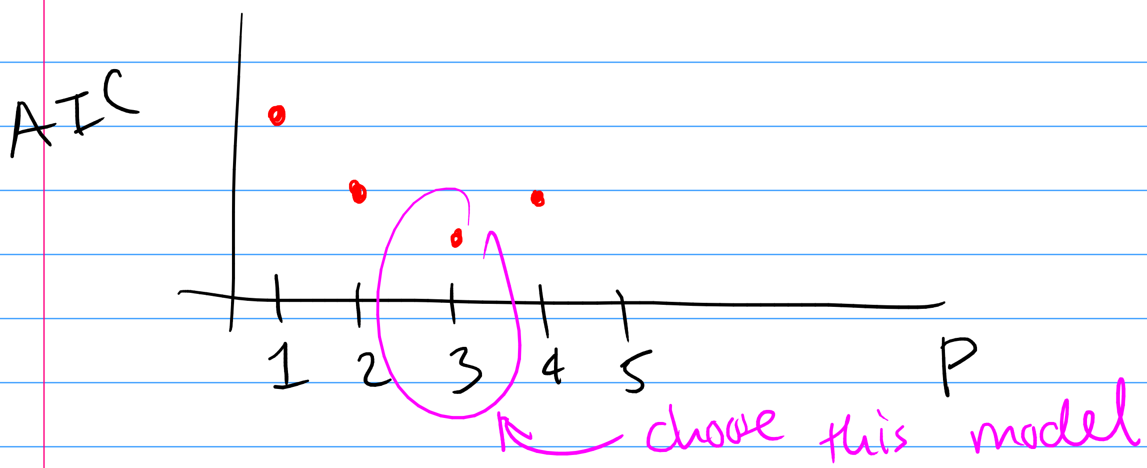
Use a greedy approach:

1. Forward Selection:

   i. start w/ model w/ just intercept

   ii. add variable that improves metric the most (inc. R^2_{adj} or dec. AIC)

   iii. repeat step ii until my metric gets worse



(2) Backwards Selection

→ start w/ model w/ all vars
→ removes them one-at-a-time.

Can I deal w/ ill-conditioning in a more continuous way?

Ridge Regression

For OLS we minimize

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

i.e.

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

If some of my vars are highly correlated then the assoc. coef. tend to blow-up ($\pm\infty$)

Ex. $Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ say $\hat{\beta}_1 = 5$, $\hat{\beta}_2 = 7$

if $X_1 \approx X_2$ then model is

$$Y \approx \hat{\beta}_0 + (\underbrace{\hat{\beta}_1 + \hat{\beta}_2}_{12}) X_1$$

equiv. as good is $\hat{\beta}_1 = 5000$, $\hat{\beta}_2 = -4988$

Ridge regression penalizes optimization to avoid large $\hat{\beta}$s.

$$\hat{\beta}^{(ridge)} = \underset{\beta}{\operatorname{argmin}}\ L(\beta) + \lambda\|\beta\|^2$$

$\lambda \geq 0$

By adding an $\lambda\|\beta\|^2$ if elements of β get too large this penalty becomes large.

If $\lambda = 0$ we get OLS $\hat{\beta}$.

as $\lambda \to \infty$ we get $\hat{\beta}^{(ridge)} \to 0$

(Typically exclude β_0 from penalty)

Often we standardize vars before ridge.

Also typically choose λ via X-val.