Lecture 10 -- Variable Selection

Tuesday, October 1, 2024 3:29 PM

So far procedures have had a fixed set of vars.

May want to select "best" set:

Why? (1) prediction accuracy: bias and variance

(2) interpretation

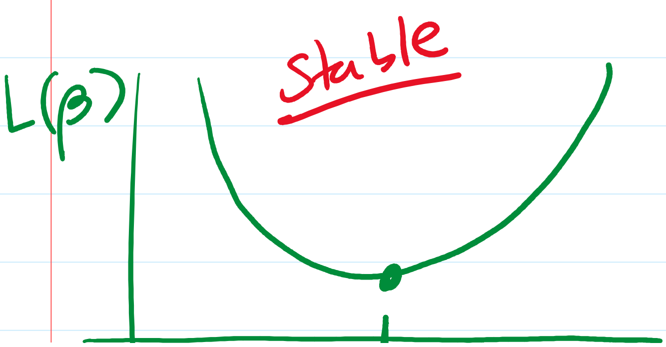
---

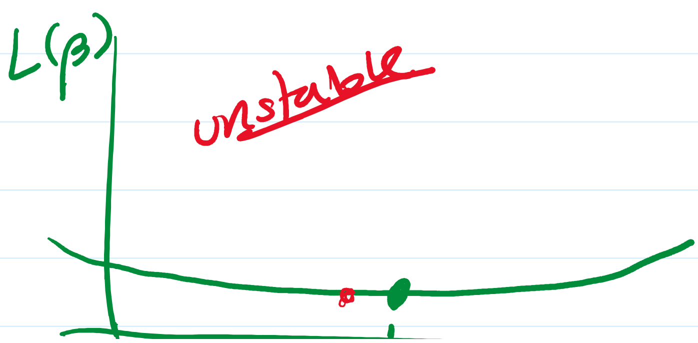
Back to OLS

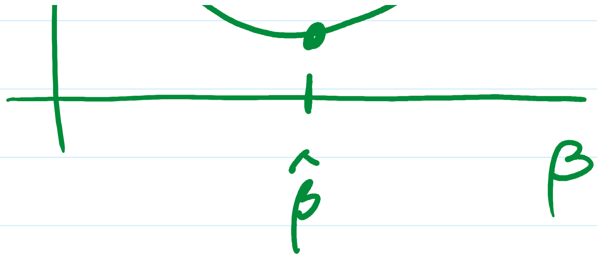
Recall that $\hat{\beta}$ was obtained by solving

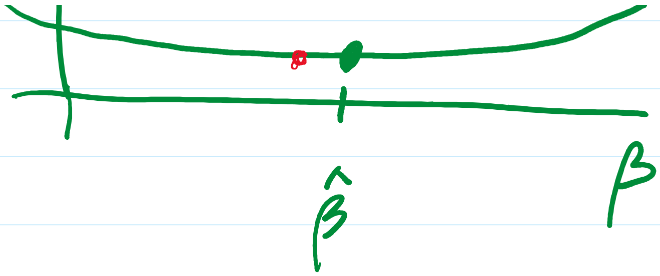
$$X^T X \beta = X^T y.$$

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






## Condition Number

For a linear system $Az = b$ the stability of soln deps on the condition number of $A$ :

$$K(A).$$

Basically: measures how sens. soln. is to small perturbations in $A$ or $b$.

Fact: $K(A) = \dfrac{\sigma_{max}(A)}{\sigma_{min}(A)}$ ← largest sing. val. of $A$

← smallest.

Large $K(A)$ = very sensitive (ill-conditioned)

Small $K(A)$ = insensitive (well-cond.)

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

---

why do I care?

Want to solve $\underbrace{(X^T X)}_{A}\underbrace{\beta}_{z} = \underbrace{X^T y}_{b}$.

then the stability of $\hat{\beta}$ depends on $K(X^T X)$.

---

Why do we get a large $K(X^T X)$?

(1) If $P < N$ but one var. is (approx) a lin. comb. of others

(2) If $P > N$ then $K(X^T X) = \infty$

<u>e.g.</u> $X$ meas. $P = 20{,}000$ genes among $N = 30$ patients.

---

How can we deal w/ this?

1. Variable selection
2. Shrinkage

---

Goal of 1 is to pick some subset of important vars to use.

Approach #1 calc. some importance metric for each var and then only keep those w/ best value.

e.g. calc. p-val. for each $\hat{\beta}_j$ and only retain those w/ low p-vals.

problem: perf. of one var may depend on others.

- - - - - - - - - - - -

Approach #2 : calc. metric for groups of

Approach #2 : calc. metric for groups of vars and choose group w/ best val.

problem: w/ p vars I have $2^p$ possible subsets to try.

---

Possible metrics for approach #2

Know: don't look at traing metric alone

as $p \uparrow$ the traing metric $\downarrow$.

Soln! (1) use some val/testg data or cross-val etc.

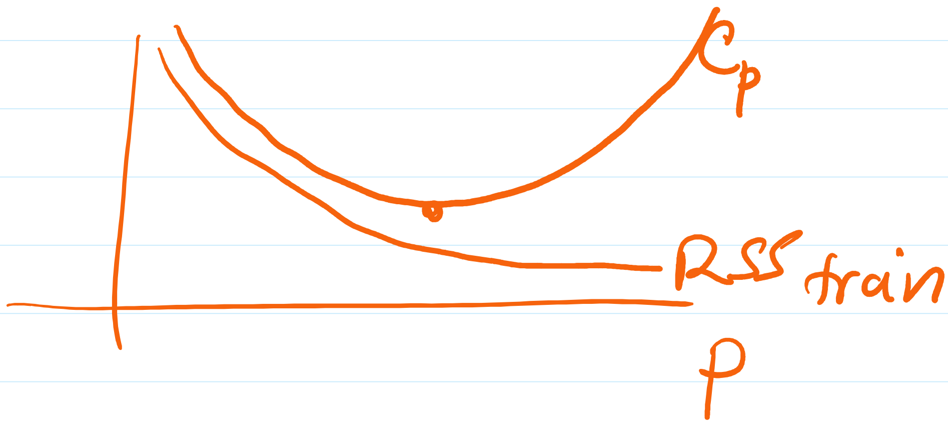
(2) penalized traing metric (classic)

$$RSS_{train} = \sum_n (y_n - \hat{y}_n)^2$$

Ex. Mallow's $C_p$

## Ex. Mallow's $C_p$

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




## Ex. AIC

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

## Ex. BIC

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

---

## Ex. Adjusted $R^2$

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

---

**Problem:** have $2^P$ models to check

Problem: have $2^p$ models to check

Use a greedy approach

## Forward Selection

(i) start w/ model w/ just intercept

(ii) add var. to model that improves metric most
(dec. AIC or inc. Adj. $R^2$)

(iii) repeat (ii) until my metric stops improving.

Can I deal w/ ill-conditioned problems in a continuous way? <u>Shrinkage</u>.

## Ridge Regression

For OLS we minimize

$$L(\beta) = RSS(\beta) = \|y - X\beta\|_2^2$$

and we let

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

---

If my $X^TX$ is ill conditioned
(some of my vars are highly cor)

and then my elements of $\hat{\beta}$ tend to get really large ($\to \pm\infty$)

<u>Ex.</u> $Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

but $X_1 \approx X_2$ say $\hat{\beta}_1 = 5$

but $X_1 \approx X_2$

say $\hat{\beta}_1 = 5$

$\hat{\beta}_2 = 7$

then $Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_1$

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

basically as good as

$\hat{\beta}_1 = 5000 \quad \hat{\beta}_2 = -4988$