Lecture 13: Variable Selection

All the procedures so far have worked w/ a fixed set of features

May want to select the "best" set of features.

Why?

1. prediction accuracy
2. interpretation
3. model may be ill-conditioned ($P \gg N$)

Back to LS OLS regression:

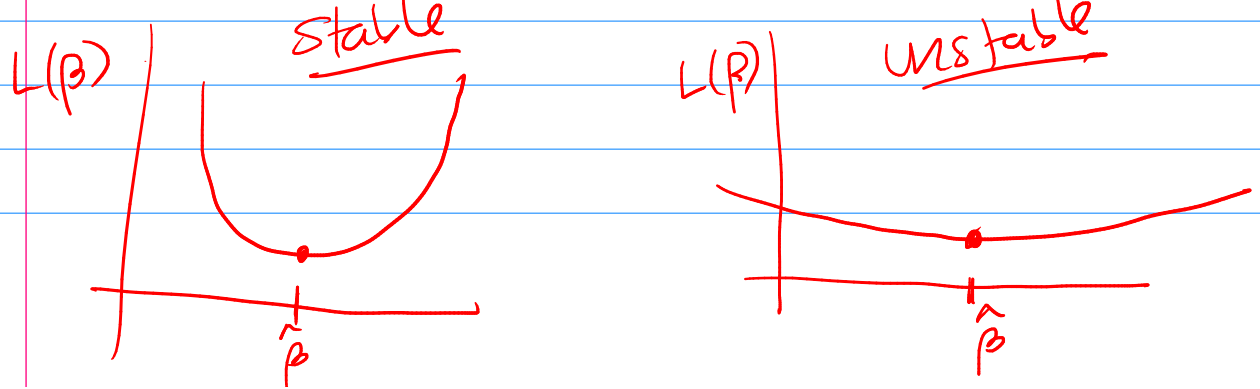
recall that $\hat{\beta}$ comes from solving "normal" eqns

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

to get $\hat{\beta}$ we multiply by $(X^T X)^{-1}$ to get

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

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



If X^TX isn't invertible, $\hat{\beta}$ not unique.

Simple illustration

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

What happens if $X_1 \approx X_2$ (highly correlated)

then

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_1 + \cdots$$

$$\approx \beta_0 + (\beta_1 + \beta_2) X_1 + \cdots$$

If $\beta_1 = 5$, $\beta_2 = 7$ (sum $= 12$)

then basically as good to have $\beta_1 = 0$, $\beta_2 = 12$ (sum $= 12$)

or $\beta_1 = -100$, $\beta_2 = 112$

What tends to happen is that $\beta_1, \beta_2 \to \pm\infty$

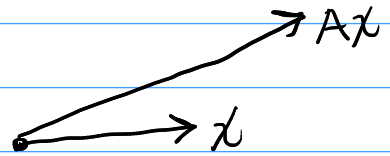
The stability depends on inverting X^TX

i.e. the condition number of X^TX.

Condition Number:

For A ($n \times n$ mtx) let

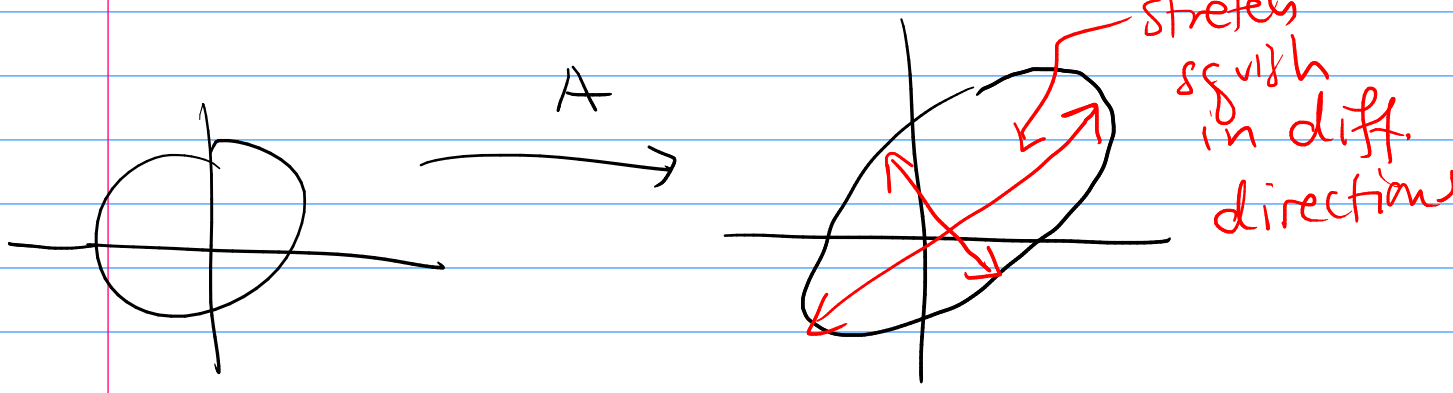
$$M = \max_x \frac{\|Ax\|}{\|x\|} = \max_{x:\|x\|=1} \|Ax\|$$



= max amt. A stretches a unit vector x

$$m = \min_x \frac{\|Ax\|}{\|x\|} = \min_{\|x\|=1} \|Ax\|$$

= min. amt. A stretches a unit vector.



The condition number $K(A) = \frac{M}{m}$

Imagine solving $Az = b$

$A(z+\delta_z) = (b+\delta_b)$

perturb b by δ_b

my soln is perturbed to $z+\delta_z$

& I have $Az = b$ and $A\delta_z = \delta_b$

notice: $\|b\| = \|Az\| \le M\|z\|$ (by defn) ①

$\|\delta_b\| = \|A\delta_z\| \ge m\|\delta_z\|$ ②

$\left[\frac{\|Az\|}{\|z\|} \le M\right]$

$\left[\frac{\|A\delta_z\|}{\|\delta_z\|} \ge m\right]$

rearrange ① $\frac{M}{\|b\|} \ge \frac{1}{\|z\|}$

② $\frac{\|\delta_b\|}{m} \ge \|\delta_z\|$

multiply

$$\underbrace{\frac{M}{m}}_{K(A)} \underbrace{\frac{\|\delta_b\|}{\|b\|}}_{\text{rel. size of } \delta_b} \ge \underbrace{\frac{\|\delta_z\|}{\|z\|}}_{\text{rel. size of } \delta_z}$$

If I perturb b by δ_b, rel. change in z may be up to $K(A)$ times rel. change in b.

If $K(A)$ is small then system is stable, b/c rel. large change in b produce small changes in z

and vice-versa.

Notice that if A isn't invertible then

$$m = 0$$

so $K(A) = \infty$

Fact: $K(A) = \dfrac{\sigma_{max}}{\sigma_{min}}$ ← larger sing. val.; ← smallest sing. val.

pf: $A = UDV^T$

$$M = \max_{\|x\|=1} \|Ax\| = \max_{\|x\|=1} \|UDV^T x\|$$

claim: multiplying by orthog mtx doesn't change length of vector

$\|q\| = \sqrt{q^T q}$ →

$$\|Qx\| = \sqrt{(Qx)^T(Qx)} = \sqrt{x^T Q^T Q x} = \sqrt{x^T x} = \|x\|$$

$$= \max_{\|y\|=1} \|UDy\|$$

$$= \max_{\|y\|=1} \|Dy\| \qquad D = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix}$$

$\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \cdots$

$$= \max_{\|y\|=1} \sqrt{(\sigma_1 y_1)^2 + (\sigma_2 y_2)^2 + \cdots + (\sigma_n y_n)^2}$$

choose $y = (1, 0, 0, 0, \ldots 0)$

$$= \sqrt{\sigma_1^2} = \sigma_1 = \sigma_{max}$$

Similarly $m = \sigma_{min}$

Why do we care?

For regression solve $(X^TX)\beta = X^Ty$

and so the stability of system depends on

$K(X^TX)$.

We can get a large $K(X^TX)$ i.e. ill-conditioned X^TX for a couple reasons

Ex. X^TX not invertible b/c one var is a LC of the others

Ex. one var is approx. a LC of others

Ex. $P > N$ then X^TX is not invertible and $K(X^TX) = \infty$

E.g. X measures gene expression in $N = 30$ patients across $P = 20{,}000$ genes.

How do we deal with this?

Today: (1) Variable selection.

Next time: (2) Shrinkage (Ridge/LASSO)

(3) Dimensionality Reduction (PCA)

Variable Selection

Goal: pick a subset of important variables and just use this subset

Q: How do I define "important"

Two approaches:

(1) Use an individual metric for each var. and use subset w/ best metric

e.g. p-values for each var.

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

(2) Calc. a metric for groups of vars and use group w/ best metric.

Careful: Looking at train metrics can be misleading.

e.g. $RSS_{train} \downarrow$ as $p \uparrow$

Solns: (1) use some test metric (e.g. k-val.)

(2) use a penalized in-sample metric

Ex. Adjusted R^2

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

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

Ex. Mallow's C_p

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

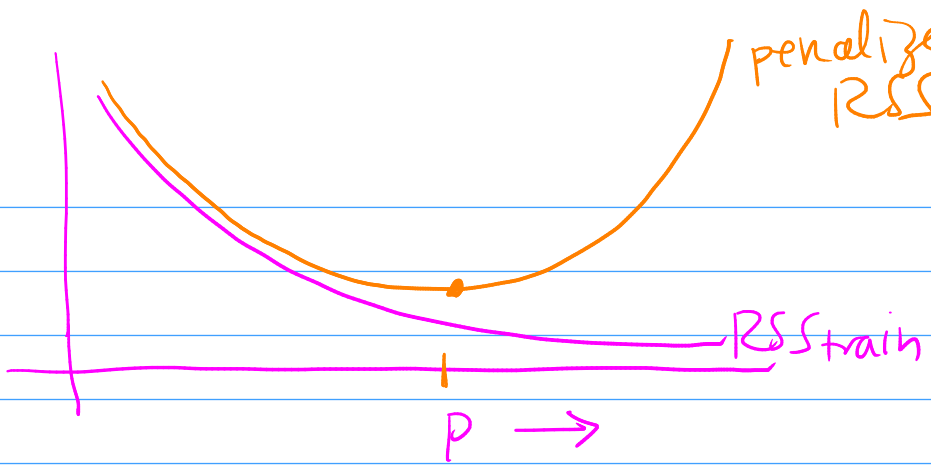
penalty increases w/ P

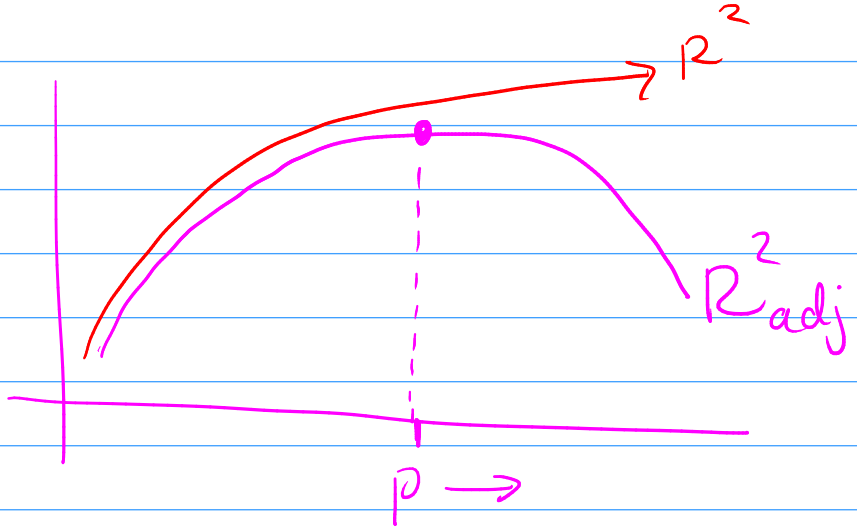
Ex. AIC

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

Ex. BIC

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


penalized RSS
RSS train
p →


R^2
R^2_{adj}
p →