Lecture 4: More Linear Regression + KNN

Generically I can encode a K-level factor using K-1 dummy vars.

Ex. Variable = Hogwarts House = {G, R, H, S}

$$\text{data} = \begin{bmatrix} \cdots & H & \cdots \\ & R & \\ & G & \\ & H & \\ & S & \\ & \vdots & \end{bmatrix} \Rightarrow X = \begin{bmatrix} \cdots & \begin{matrix} H & R & G \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ & \vdots & \end{matrix} & \cdots \end{bmatrix}$$

How do I interpret coefs?

Coef assoc. w/ H dummy is "holding other vars. constant, what is predicted diff btwn H and S"

Fitting issues:

Recall that I get $\hat{\beta}$ by setting $\frac{\partial RSS}{\partial \beta} = 0$

which yields my normal eqns:

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

IF $X^T X$ is invertible, then $\hat{\beta} = (X^T X)^{-1} X^T y$

when can this fail?

P or P+1

X^TX isn't invertible $\iff$ rank(X) < # cols of X

pf:

$\Leftarrow$ Assume rank(X) < # cols.

Then $\exists v \neq 0$ where $Xv = 0$.

Thus $X^TXv = 0$.

So $\exists v \neq 0$ when $(X^TX)v = 0$

i.e. rank(X^TX) < # cols of X^TX i.e. its not invertible.

$\Rightarrow$ If X^TX isn't invertible then $\exists v \neq 0$ where

$$X^TXv = 0.$$

for a, $\|a\|^2 = a^Ta$

thus $0 = v^T0 = v^TX^TXv = (Xv)^T(Xv)$

$= \|Xv\|^2$

So $\|Xv\|^2 = 0 \Rightarrow \|Xv\| = 0$ can only happen if $Xv = 0$

$\exists v \neq 0$ where $Xv = 0$, i.e. X is rank deficient.

$\|a\| = 0 \iff a = 0$

When does this happen in reality?

(1) Accidentially include a var. twice in my design

$$X = \begin{bmatrix} 1 & 1 & 1 \\ \vdots & 2 & 2 \\ \vdots & 3 & 3 \\ 1 & 4 & 4 \end{bmatrix}$$

(2) If one of my vars is a LC of the others.

$$X = \begin{bmatrix} & M & F \\ 1 & 0 & 1 \\ \vdots & 1 & 0 \\ \vdots & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Intercept = M + F

(3) If # cols of X > # obs.

P > N

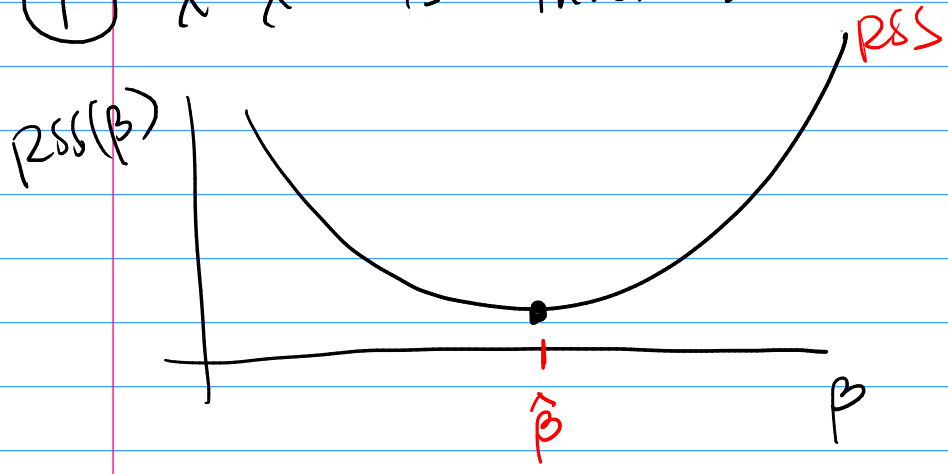
Ex. I measure 50,000 genes for each of 30 patients

(4) This can "approx." happen if some vars are almost a LC of each other

Ex. two vars. are highly correlated

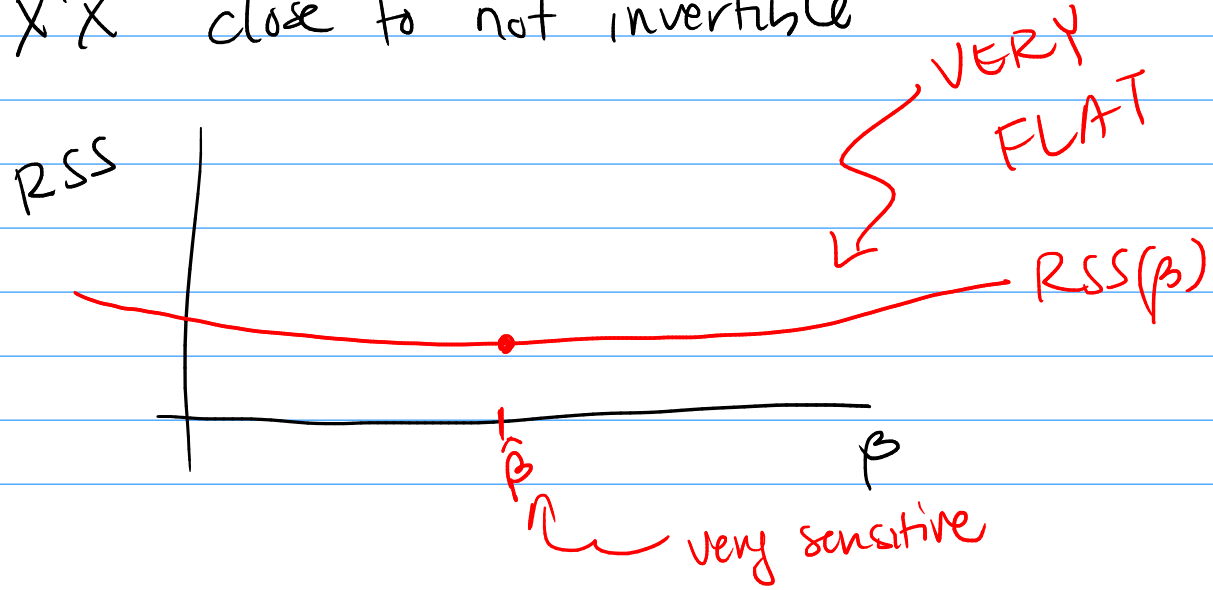
What RSS(β) looks like in various situations

(1) X^TX is invertible



(2) X^TX not invertible

(3) X^TX close to not invertible

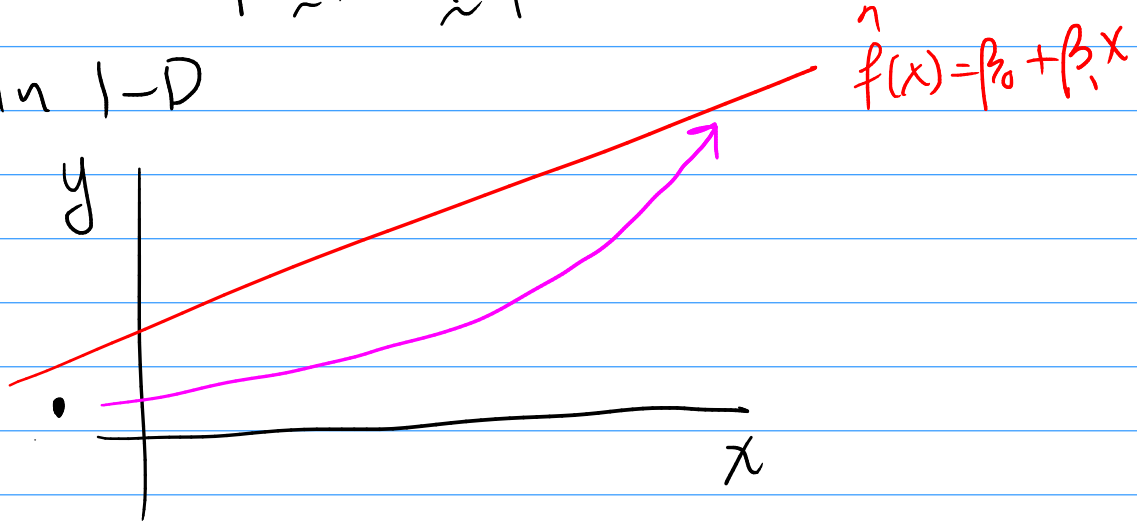


KNN Regression (K nearest neighbors regression)

For LR we have a really strong global assumption about the form of f

$$f(\underset{\sim}{x}) = \underset{\sim}{x}^T \beta$$

e.g. in 1-D



Predictions follow linear fn in entire space

Training data affects fit very far away

Benefit: strong global assumption makes $\hat{f}$ practical to find (opt. β over a p-dimil space)

KNN makes a weaker local assumption abat the form of $\hat{f}$

where $\hat{f}(\underset{\sim}{x})$ only depend on nearby training pts

In particular $\hat{f}(\underset{\sim}{x}) = \frac{1}{K} \sum_{n \in N_k(\underset{\sim}{x})} y_n$ = avg y_n over K nearest train pts to $\underset{\sim}{x}$

$N_K(x)$ = K neighborhood of $\underset{\sim}{x}$
= indicies of K nearest traing pts to $\underset{\sim}{x}$