Lecture 3 - More Linear Regression

Variable transformations

Regression: $\boxed{Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j}$

we want to estimate these βs

actually have: $\{(y_n, x_n)\}_{n=1}^{N}$

$x_n = (x_{n1}, x_{n2}, \dots, x_{np})$

$$\hat{\beta} = \arg\min_{\beta} L(\beta)$$

argument that minimizes ...

$$L(\beta) = \sum_{n=1}^{N} \Big(y_n - \beta_0 - \sum_{j=1}^{p} \beta_j x_{nj}\Big)^2$$

If we cleverly encode xs as lgs:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

design matrix

$$X = \begin{bmatrix} 1 & — & x_1 & — \\ \vdots & — & x_2 & — \\ \vdots & & \vdots & \\ 1 & — & x_N & — \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{N1} & \cdots & \cdots & x_{Np} \end{bmatrix}}_{\text{design matrix}}$$

then

$$\boxed{L(\beta) = \|y - X\beta\|^2}$$

$$\begin{bmatrix} \beta_0 + \sum_j \beta_j X_{1j} \\ \beta_0 + \sum_j \beta_j X_{2j} \\ \vdots \\ \beta_0 + \sum_j \beta_j X_{Nj} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & & \ddots & \\ \vdots & & & \\ 1 & x_{N1} & \cdots & x_{Np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$(p+1)$

$X_{N \times (p+1)}$ β_{p+1}

$\boxed{\text{Linear Model: } Y \approx X\beta}$

Model:

linear in βs

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

$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$ = proj. of y onto $\mathrm{Col}(X)$.

OG Regression: $Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j = \boxed{X\beta}$ lin. in β

Is this regression? $\boxed{Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j^2}$ Yup.

linear in β

$Y = X^* \beta$

$\exists X^*$ where we can write the above model in this way

$$\begin{bmatrix} 1 & | & | & & | \\ \vdots & X_1^2 & X_2^2 & \cdots & X_p^2 \\ 1 & | & | & & | \end{bmatrix} = X^*$$

$$\| $$

$$\begin{bmatrix} 1 & x_{11}^2 & x_{12}^2 & \cdots & x_{1p}^2 \\ \vdots & & \vdots & & \\ 1 & x_{N1}^2 & \cdots & & x_{Np}^2 \end{bmatrix}$$

What about this? No. Not linear in βs

$Y = \beta_0^2 + \sum_{j=1}^{p} X_j \beta_j^2$ ← Squaring βs

$$Y = \beta_0^2 + \sum_{j=1} X_j \beta_j^2$$ ← Squaring βs

No way to write
$Y = X\beta$

Other examples:

→ $Y = \beta_0 + \beta_1 X_1^2 + \beta_2 \log(X_2) + \beta_3 X_3^5$

this is regression

$\boxed{Y = X\beta}$

$$\begin{bmatrix} | & | & | & | \\ 1 & X_1^2 & \log(X_2) & X_3^5 \\ | & | & | & | \end{bmatrix} = X = \text{design mtx}$$

In general any transformation of the covariates is allowed.

Ex. $Y = X_1 X_2 \beta_1 + X_3^2 \beta_2$

$\boxed{Y = X\beta}$

$$X = \begin{bmatrix} | & | \\ X_1 X_2 & X_3^2 \\ | & | \end{bmatrix}$$

For all of these

design mtx X

For all of these

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

~~How do we get $\hat{\beta}$?~~ Fitting issues

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

Recall: $a \in \mathbb{R}^N$

$\|a\| = \sqrt{a^T a}$

$\|a\|^2 = a^T a$

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

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

Calc I says to get $\hat{\beta}$

we look at $\frac{\partial L}{\partial \beta} = 0$

Calc III says use gradient

$$\frac{\partial L}{\partial \beta} = \text{gradient} = \boxed{2(y - X\beta)^T(-X)}$$

$$\underbrace{(1 \times N)\ \ (N \times (p+1))}_{(p+1)\text{-vector}}$$

So Look at where $\frac{\partial L}{\partial \beta} = 0$

$$\cancel{2}(y - X\beta)^T(\cancel{-}X) = 0$$

$$\Rightarrow y^T X - \beta^T X^T X = 0$$

$$\Rightarrow \boxed{X^T y = X^T X \beta} \quad \text{Normal equation}$$

$$\Rightarrow \boxed{X^T y = X^T X \beta}$$ Normal equation

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

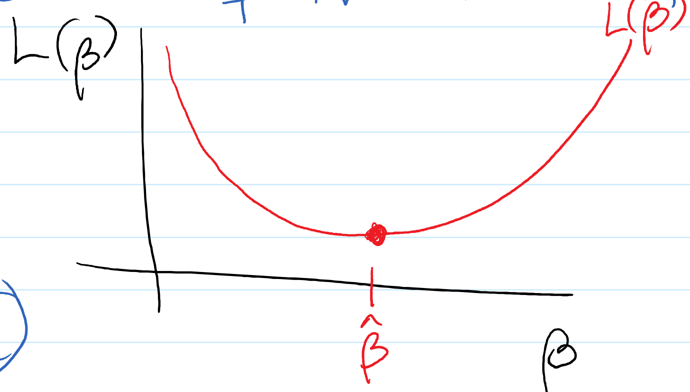
→ This works as long as $X^T X$ is invertible.

→ Fail if $X^T X$ isn't invertible.

When can that happen?

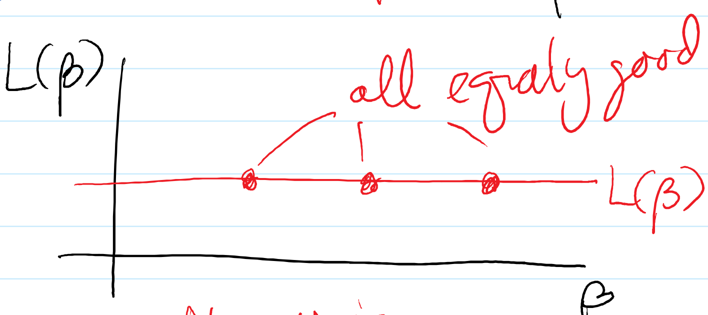
→ If any of our cols of X are linear combinations of any other.

1 ↳ If $N <$ # cols of X



If $X^T X$ is invertible

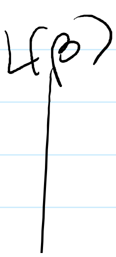
2



If $X^T X$ isn't invertible

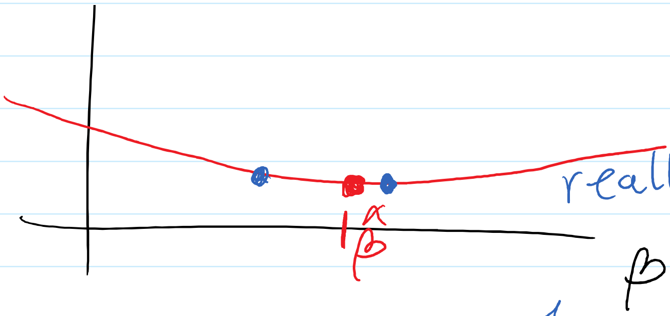
No uniq minimum

3



$X^T X$ is close to not invertible

$X^T X$ is
not invertible



really shallow loss.

The fit will be very sensitive

Categorical Variables

So far we need numeric variable.

What if I have non-numeric?

- → race
- → color
- → gender
- → ...

What I would like to do: $\boxed{Y = \beta_0 + \beta_1 \text{gender}}$

$$X = \begin{bmatrix} 1 & M \\ \vdots & M \\ \vdots & F \\ 1 & F \\ & \vdots \end{bmatrix}$$ ← ain't gonna work

Use a dummy variable:

encode $M = 0$ and $F = 1$

$$X = \begin{bmatrix} 1 & 0 \\ \vdots & 0 \\ \vdots & 1 \\ 1 & 1 \\ & \vdots \end{bmatrix} \rightsquigarrow \boxed{Y = \begin{cases} \beta_0 & \text{if gender} = M \\ \beta_0 + \beta_1 & \text{if gender} = F \end{cases}}$$

more generally we can encode multi-level categorical variables using dummies

If I have a K-level factor I use K-1 dummy variables

Ex,

$$\text{data} = \begin{bmatrix} \cdots & \text{Hogwarts House} & \cdots \\ & H & \\ & R & \\ & R & \\ & S & \\ & G & \\ & G & \end{bmatrix} = \begin{array}{c} \begin{array}{ccc} HG & RG & SG \end{array} \\ \begin{bmatrix} \cdots & 1 & 0 & 0 & \cdots \\ & 0 & 1 & 0 & \\ \cdots & 0 & 1 & 0 & \cdots \\ & 0 & 0 & 1 & \\ & 0 & 0 & 0 & \\ & 0 & 0 & 0 & \end{bmatrix} \end{array}$$

4-level factor

use 3 dummy vars using G as a base level

the resulting model

binary 0/1

$$Y = \beta_0 + \beta_1 \underset{1}{HG} + \beta_2 \underset{0}{RG} + \beta_3 \underset{0}{SG} + \cdots$$

$$= \begin{cases} \beta_0 + \beta_1 & \text{if } H \\ \beta_0 + \beta_2 & R \\ \beta_0 + \beta_3 & S \\ \beta_0 & G \end{cases}$$