Lecture 3: Linear Regression

If $\underset{\sim}{x} = (x^{(1)}, \dots, x^{(P)})$

and $\beta = (\beta^{(0)}, \dots, \beta^{(P+1)})$

and $\underline{x} = (1, x^{(1)}, \dots, x^{(P)})$

LR proposes the model

$$Y = f_{\beta}(\underset{\sim}{x}) = \underline{x}^{T}\beta = \beta^{(0)} + \sum_{j=1}^{P} x^{(j)}\beta^{(j)}$$

Looking at least-squares linear regression (OLS)

$$L(y_n, f_{\beta}(\underset{\sim}{x}_n)) = (y_n - f_{\beta}(x_n))^2$$

So if we use ERM to get $\hat{f}$

$$\hat{f} = \underset{f \in F}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} L(y_n, f_{\beta}(\underset{\sim}{x}_n))$$

$$= \underset{f \in F}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} (y_n - \underline{x}^{T}\beta)^2$$

Notice: to determine f — I simply need to determine β

So equivalently $\hat{f} = f_{\hat{\beta}}$ i.e. $\hat{f}(\underset{\sim}{x}) = \underline{x}^{T}\hat{\beta}$

where $\hat{\beta}$ minimizes ER:

$$\hat{f} = f_{\hat{\beta}} \quad \text{where} \quad \hat{\beta} = \underset{\beta}{\operatorname{argmin}} \underbrace{\frac{1}{N} \sum_{n=1}^{N} (y_n - \underline{x}^{T}\beta)^2}$$

$$\overset{\downarrow}{=\arg\min_{\beta}} \; \overset{\overset{\Downarrow}{\propto}}{\underbrace{\sum_n (y_n - \underline{x}_n^T \beta)^2}_{RSS(\beta)}}$$

So equiv.

$$\hat{f}(\underline{x}) = \underline{x}^T \hat{\beta} \quad \text{whe} \quad \hat{\beta}^{(OLS)} = \arg\min_{\beta} RSS(\beta)$$

So ERM = Ordinary Least Squares Regression

Practically, How do we find $\hat{\beta}^{(OLS)}$?

Let X be our design matrix so that

$$X = \begin{bmatrix} \vdots \\ - \underline{x}_n^T - \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & & x_1^{(P)} \\ \vdots & x_2^{(1)} & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_N^{(1)} & x_N^{(2)} & & x_N^{(P)} \end{bmatrix}_{N \times (P+1)}$$

and $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^{N \times 1}$

then I want to minimize $RSS(\beta)$

$$RSS(\beta) = \sum_{n=1}^{N} (y_n - \underline{x}^T \beta)^2$$

$$= \| \overset{N \times 1}{\overline{\underset{N \times 1}{y} - \underbrace{\underset{N \times (P+1)}{X} \; \underset{(P+1) \times 1}{\beta}}_{N \times 1}}} \|^2$$

$\|y - X\beta\|^2$

$$\hookrightarrow y - X\beta = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \overset{X}{\begin{bmatrix} — \underline{x_1}^T — \\ \vdots \\ — \underline{x_N}^T — \end{bmatrix}} \beta$$

$$= \underset{N \times 1}{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}} - \underset{N \times 1}{\begin{bmatrix} — \underline{x_1}^T \beta — \\ \vdots \\ — \underline{x_N}^T \beta — \end{bmatrix}}$$

$\|a\| = \sqrt{\sum_i a_i^2}$

$\|a\|^2 = \sum_i a_i^2$

$$= \begin{bmatrix} y_1 - \underline{x_1}^T \beta \\ y_2 - \underline{x_2}^T \beta \\ \vdots \\ y_N - \underline{x_N}^T \beta \end{bmatrix}$$

So $\|y - X\beta\|^2 = \sum_n (y_n - \underline{x_n}^T \beta)^2 = RSS(\beta)$.

So $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \underbrace{\|y - X\beta\|^2}_{RSS : \mathbb{R}^{P+1} \to \mathbb{R}} \leftarrow$

How do I find the minimizer of RSS?

MV Calc Problem: get $\underbrace{\text{derivative}}_{\text{gradient}}$ and set equal to zero

Turns Out :

gradient of RSS(β) w.r.t. $\beta = \frac{\partial RSS}{\partial \beta} = \underbrace{-2\overbrace{(y - X\beta)^T}^{1 \times N} \overbrace{X}^{N \times (P+1)}}_{1 \times (P+1) \text{ row vector}}$

So if $\frac{\partial RSS}{\partial \beta} = -2(y - X\beta)^T X$

then Calc 3 says set equal to zero

$$\frac{\partial RSS}{\partial \beta} = -2(y - X\beta)^T X = 0$$

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

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

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

If $\underbrace{\underset{p+1\,\times N\,\times p+1}{X^T X}}_{(p+1)\times(p+1)}$ is invertible the

$$(X^T X)^{-1} X^T y = \overset{I}{\cancel{(X^T X)^{-1} X^T X}}\beta$$

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

So $\hat{f}(\underset{\sim}{x}) = \underline{x}^T \hat{\beta}$ when $\hat{\beta} = (X^T X)^{-1} X^T y$

$$= \underline{x}^T (X^T X)^{-1} X^T y$$

Consider predictions on training data

$$\hat{y}_n = \hat{f}(\underset{\sim}{x}_n) = \underline{x}_n^T \underbrace{(X^T X)^{-1} X^T y}_{\hat{\beta}}$$

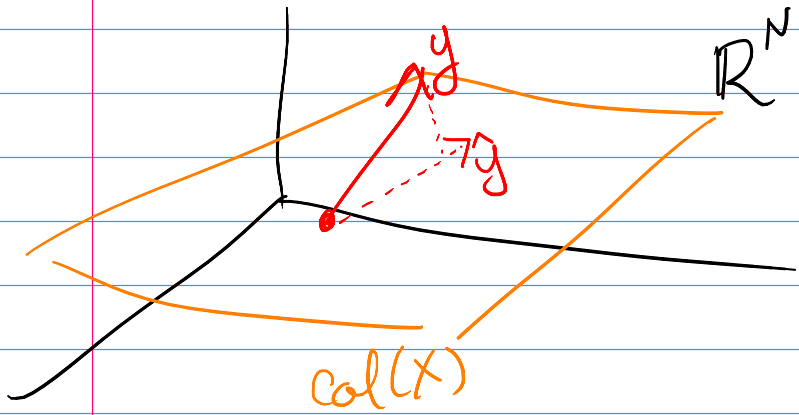
$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} \text{---}\ \underline{x}_n^T \hat{\beta}\ \text{---} \end{bmatrix} = X\hat{\beta}$$

For training data

$\hat{y} = X\hat{\beta}$

notice then that

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{\text{projection mtx onto Col}(X)}y = \text{proj. of } y \text{ onto Col}(X)$$



How flexible is OLS?

Is this regression?

$$Y = \beta^{(0)} + \sum_{j=1}^{p} \beta^{(j)} X^{(j)2} \ ?$$

Yes. This is still linear in βs.

lets just change the design

$$\underline{X} = (1, X^{(1)2}, X^{(2)2}, \dots, X^{(p)2})$$

then $Y = \underline{X}^T\beta$

↑ this is still a linear operation wrt β

So as before

$\hat{\beta} = (X^TX)^{-1}X^Ty$ but now I'm using a slightly different X

Ex. What about this:

$$Y = f(\underset{\sim}{x}) = \underline{\beta^{(0)}} + \underline{\beta^{(1)}} x^{(1)2} + \underline{\beta}^{(2)} \log(x^{(2)}) + \underline{\beta^{(3)}} \sin(x^{(1)} x^{(2)}) \ ?$$

Is this linear regression?

Yes. All we need to do is change the design

$$\underset{\sim}{x} = (1, x^{(1)2}, \log(x^{(2)}), \sin(x^{(1)}x^{(2)}))$$

So if

$$X = \begin{bmatrix} 1 & | & | & | \\ \vdots & x^{(1)2} & \log(x^{(2)}) & \sin(x^{(1)}x^{(2)}) \\ 1 & | & | & | \end{bmatrix}$$

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

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

Generically we still have an OLS method in any basis expansion so long as the bases don't depend on βs:

$$x^{(j)} = \varphi_j(\underset{\sim}{x})$$ for arb. φ_j doesn't depend on βs

then

$$X = \begin{bmatrix} | & | & \\ \varphi_1(\underset{\sim}{x}) & \varphi_2(\underset{\sim}{x}) & \cdots \\ | & | & \end{bmatrix}$$ and $\hat{\beta} = (X^T X)^{-1} X^T y$

What about categorical vars? (factors in R)

e.g. race, color, gender etc.

How do I do something like

$$Y = \beta^{(0)} + \beta^{(1)} \overset{X^{(1)}}{\underline{\text{Gender}}} \; ?$$

Can do this using dummy variable encoding.

$$\underline{x} = (1, 0) \text{ if female} \qquad X^{(1)} = \begin{cases} 0 & \text{fem} \\ 1 & \text{male} \end{cases}$$

$$\underline{x} = (1, 1) \text{ if male}$$

then my design mtx X will look somethg like

$$X = \begin{array}{c} \begin{bmatrix} \underline{1} & 0 \\ \vdots & 1 \\ \vdots & \vdots \\ \vdots & 0 \\ \underline{1} & 1 \end{bmatrix} \end{array} \begin{array}{l} \text{females} \\ \text{males} \\ \\ \\ \\ \end{array}$$

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

How do I interpret $\hat{\beta}$?

$\underline{\text{Ex.}}$ $(\hat{\beta}^{(0)}, \hat{\beta}^{(1)})$ interpret?

If gender = F then $Y \approx \hat{\beta}^{(0)} + \hat{\beta}^{(1)} \cdot 0 = \hat{\beta}^{(0)}$

gender = M then $Y \approx \hat{\beta}^{(0)} + \hat{\beta}^{(1)} \cdot 1$

So $\hat{\beta}^{(0)}$ as the typ. val. for Y (w/ other vars fixed) for a female

$$\hat{\beta}^{(1)} = \underbrace{\hat{\beta}^{(0)} + \hat{\beta}^{(1)}}_{\text{typ val for M}} - \underbrace{\hat{\beta}^{(0)}}_{\text{typ val for F}} =$$ contrast = diff. btwn typ. vals for M and F

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

Hogwarts House

4-level

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

How do I interpret $\hat{\beta}$?

Holing other vars. constant what is the diff in typ. val for Y btwn ___ and S

Fitting Issues

Recall that $\frac{\partial RSS}{\partial \beta} = 0$ yielded "normal equations"

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

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

When can this fail? when X^TX isn't invertible.

→ $\boxed{X^TX \text{ isn't invertible} \iff \text{rank}(X) < \#\text{cols } X}$

⇐ $\text{rank}(X) < \#\text{cols } X$ then $\exists\, v \neq 0$ s.t. $Xv = 0$.

Consequently $X^TXv = X^T0 = 0$

So $\exists\, v \neq 0$ s.t. $(X^TX)v = 0$ i.e. X^TX is rank defficient so its not invertible.

⇒ If X^TX isn't invertible then $\exists\, v \neq 0$ s.t.

$X^TXv = 0$ $\qquad \|u\| = \sqrt{u^Tu}$

hence $0 = v^T0 = v^TX^TXv = (Xv)^T(Xv) = \|Xv\|^2$

So $\exists\, v \neq 0$ when $\|Xv\| = 0$ hence $Xv = 0$

$(\|a\| = 0 \Leftrightarrow a = 0)$

So X is rank difficient ie. $\text{rank}(X) < \#\text{cols}$.

---

when does this happen in reality?

① If I accidentically incl. a var. twice in design

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

lin. dep. ⇒ X rank def.

this is like saying

$$\hat{Y} = \hat{\beta}^{(0)} + \hat{\beta}^{(1)} X^{(1)} + \hat{\beta}^{(2)} X^{(1)}$$

e.g. $= 1 + 5X^{(1)} + 7X^{(1)}$

this has exactly same preds as

$$= 1 + \underbrace{3X^{(1)} + 9X^{(1)}}_{12X^{(1)}}$$

or any model where sum of $\hat{\beta}^{(1)} + \hat{\beta}^{(2)} = 12$

2 If # cols of $X > N$

if X is $N \times P+1$

and $P+1 > N$

then $\text{rank}(X) \leq \min\{\#\text{rows}, \#\text{cols}\} = N < P+1$