Lecture 3: Linear Regression

Why?

1. classic method – very well studied
2. Simple (good)
3. powerful
4. basis of more complicated methods

Setup: For LR we assume we have some design (input variables):

$$\underset{\sim}{X} = (X^{(1)}, \dots, X^{(p)}) \in \mathbb{R}^p \leftarrow p = \# \text{ vars}$$

and assoc. coef.

$$\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$$

then we assume a model of the form

$$Y = f(\underset{\sim}{X}) = \underset{\sim}{X}^T \beta$$
$$= \sum_{j=1}^{p} \beta_j X^{(j)}$$

Then to learn $\hat{f}$ we simply need to learn some estimates $\hat{\beta}$ of β.

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

But wait, where's the intercept?

Typically people write the model as

$$\hat{f}(\underset{\sim}{X}) = \hat{\beta}_0 + \hat{\beta}_1 X^{(1)} + \hat{\beta}_2 X^{(2)} + \cdots$$

Doesn't really matter, hidden in design.

Use $\underset{\sim}{X}$ in two ways:

1. vars. we measure
2. input to a ML algo (design)

Need not be the same, could transform

$\underline{\text{input}}$: X^2, $\log(X)$, $X^{(1)} \cdot X^{(2)}$, ...

In this course we typ. use 2 so $\underset{\sim}{X}$ is whatever we input after whatever transf.

To get an intercept, just assume first var in $\underset{\sim}{X}$ is always $\underline{1}$

$$\underset{\sim}{X} = (1, \text{rest of vars})$$

then

$$\underset{\sim}{X}^T \hat{\beta} = (1, X^{(1)}, \ldots, X^{(p)})^T (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$$

$$= \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X^{(j)} .$$

How do I learn good values for $\hat{\beta}$?

I want to choose them so that

$$Y \approx \underset{\sim}{X}^T \hat{\beta}$$

Linear $\boxed{\text{Least-Squares}}$ Regression

Residual Sums of Squares

Find $\hat{\beta}$ as value for β that minimizes RSS

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

pred. val. for n^{th} training resp. using β

val. of resp. for n^{th} training sample

$\approx$ prediction err. in training data

So what we do is

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \; RSS(\beta)$$

$=$ value that minimizes $RSS(\beta)$

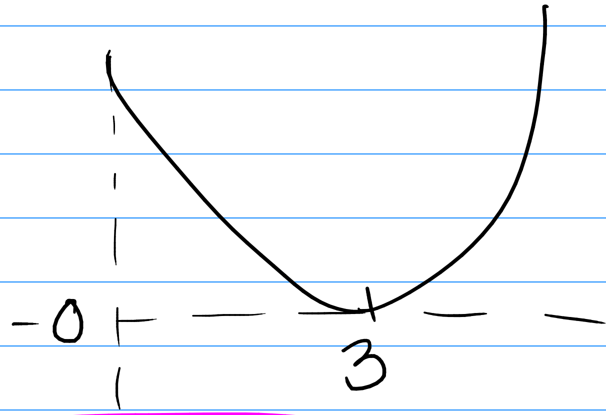
Aside:

$$\min_x f(x) = \text{minimum value of } f$$

$$\arg\min_x f(x) = \text{value at which } f \text{ takes on min}$$

$$\min_x (x-3)^2 = 0$$

$$\arg\min_x (x-3)^2 = 3$$



How do we get $\hat{\beta}$?

mtx w/ rows that are training x_ns

Let X be our design matrix

$$X = \begin{bmatrix} — \; \underset{\sim}{x}_n^T \; — \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(p)} \\ \vdots & & \ddots & & \\ 1 & x_N^{(1)} & \cdots & \cdots & x_N^{(p)} \end{bmatrix}$$

$N \times (p+1)$

and $y = (y_1, \dots, y_N)^T \in \mathbb{R}^N$

then

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

$$= \| \underbrace{y}_{N \times 1} - \underbrace{X}_{N \times (p+1)} \underbrace{\beta}_{(p+1) \times 1} \|^2$$

$$y - X\beta = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \text{---} \; \underset{\sim}{x}_1^T \; \text{---} \\ \vdots \\ \text{---} \; \underset{\sim}{x}_N^T \; \text{---} \end{bmatrix} \beta$$

$$= \begin{bmatrix} y_1 - \underset{\sim}{x}_1^T\beta \\ \vdots \\ y_N - \underset{\sim}{x}_N^T\beta \end{bmatrix}$$

so $$\|y - X\beta\|^2 = \sum_{n=1}^{N} (y_n - \underset{\sim}{x}_n^T\beta)^2 = RSS(\beta).$$

So $$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \underbrace{\|y - X\beta\|^2}_{\text{function of } \beta}$$

A calculus 3 problem. All I need to do is set derivative to zero.

$$\text{gradient} = \frac{\partial RSS}{\partial \beta} = \underbrace{-2\underbrace{(y - X\beta)^T}_{N\times 1} \underbrace{X}_{N\times(p+1)}}_{1\times(p+1)}$$

p+1 vector of partial derivs

ok, so set equal to zero

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

$$\Leftrightarrow y^T X - (X\beta)^T X = 0$$

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

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

$$\Leftrightarrow \boxed{(X^TX)\beta = X^Ty}$$

Normal Equation

If X^TX is invertible then can multiply both sides by $(X^TX)^{-1}$ to get

$$\boxed{\hat{\beta} = (X^TX)^{-1}X^Ty}$$

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

$$= \underset{\sim}{x}^T(X^TX)^{-1}X^Ty$$

Consider predictions for training data: $\hat{y}_n = \hat{f}(\underset{\sim}{x}_n)$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} — \underset{\sim}{x}_n^T\hat{\beta} — \end{bmatrix} = \begin{bmatrix} — \underset{\sim}{x}_n — \end{bmatrix}\hat{\beta}$$

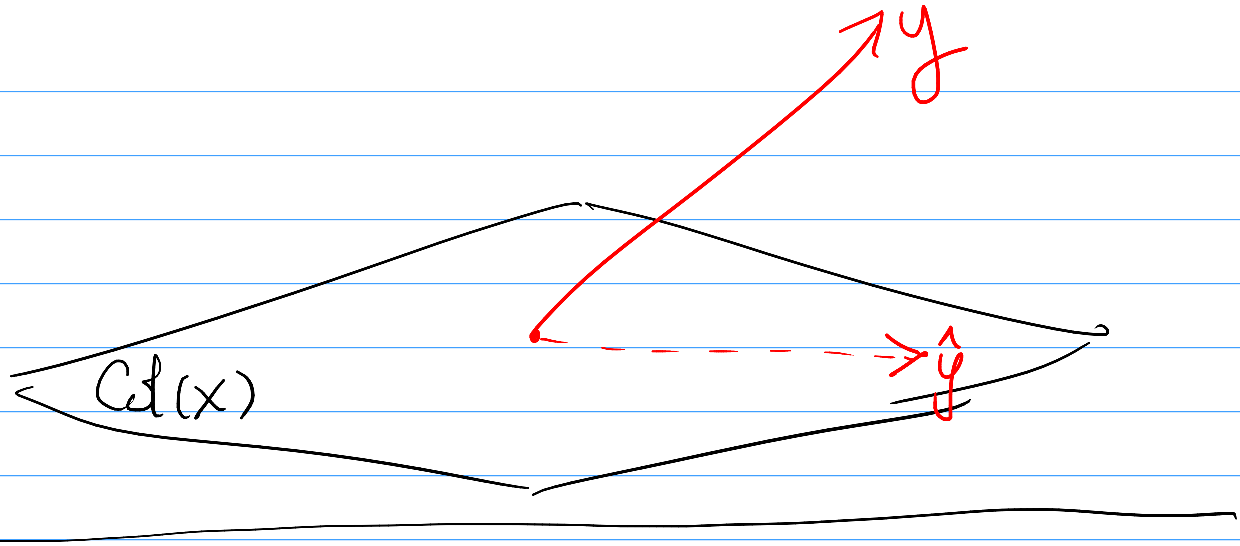
$$= X\hat{\beta}$$

$$= \underbrace{X(X^TX)^{-1}X^T}_{P_X = \text{proj. onto col}(X)}y$$

$$= P_Xy$$

P_Xv = proj. of v onto col(X)

= closest vec. to v that is a LC of cols of X



How flexible is OLS (ordinary least squares)?

Is this a regression model?

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X^{(j)2}$$

Yes. This is still linear in βs.
Just a different design.

$$Y = \underset{\sim}{x}^T \beta \qquad (1, X^{(1)2}, X^{(2)2}, \ldots,)$$

Ex.

$$Y = \beta_0 + \beta_1 X^{(1)3} + \beta_2 \sin(X^{(2)}) + \beta_3 \log(X^{(3)}) + \cdots$$

All I need to do to find $\hat{\beta}$, is change X

$$X = \begin{bmatrix} 1 & | & | & | & \\ \vdots & x^{(1)2} & \sin(x^{(2)}) & \log(x^{(3)}) & \cdots \\ 1 & | & | & | & \end{bmatrix}$$

What about categorical vars? (Factors in R)

e.g. race, gender, color, flavor, ...

How do I fit a regression of

$$Y = \beta_0 + \beta_1 (\text{Color})$$

Can do this using dummy variables.

$$\text{Color} \in \{Y, R, B\}$$

in design matrix I can encode color as two variables

$$(0,0) = R$$

$$(0,1) = B$$

Blue

$$(1,0) = Y$$

Yellow

$$\text{data} = \begin{bmatrix} R \\ Y \\ Y \\ R \\ B \end{bmatrix} \longrightarrow X = \begin{array}{c} \quad\; Y \;\; B \\ \begin{bmatrix} 1 & 0 & 0 \\ \vdots & 1 & 0 \\ \vdots & 1 & 0 \\ \vdots & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \end{array}$$

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

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 (\underbrace{\text{Yellow Dummy}}^{1 \text{ if } Y,\ 0 \text{ else}}) + \hat{\beta}_2 (\overbrace{\text{Blue dummy}}^{1 \text{ if } B,\ 0 \text{ else}})$$

$$= \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 & \text{if yellow} \\ \hat{\beta}_0 + \hat{\beta}_2 & \text{if blue} \\ \hat{\beta}_0 & \text{if red} \end{cases}$$

base

How to interpret

$$\hat{\beta}_0 + \hat{\beta}_1 - \hat{\beta}_0 = \hat{\beta}_1 = (\text{pred. } Y) - (\text{Pred. } R)$$

contrast btwn Y and R

$$\hat{\beta}_2 = \hat{\beta}_0 + \hat{\beta}_2 - \hat{\beta}_0 = \text{Contrast btwn B and R}$$