Lec 10 3/7/18 Math 3926

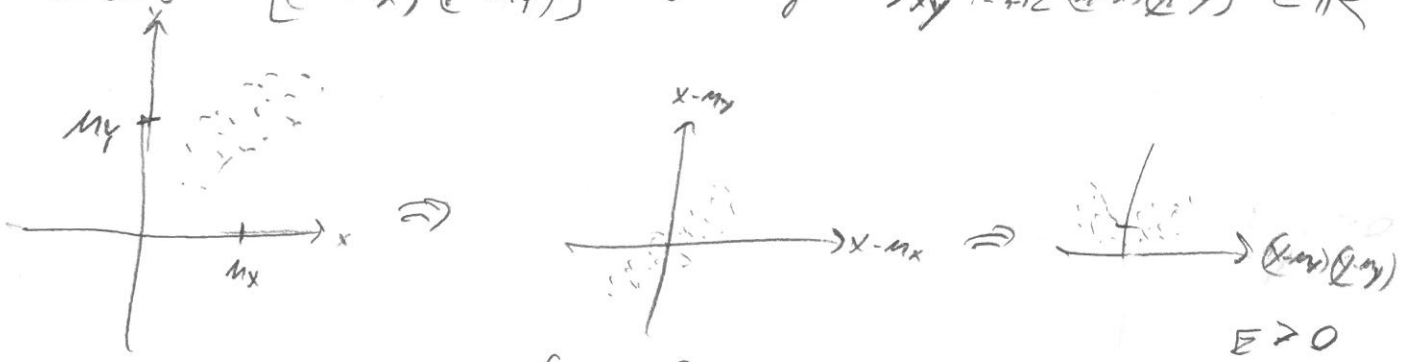
In Math 241 r.v.'s X, Y were said to be dependent if knowing the value of one affects the distribution of the other:

$$P(Y|X=x) \neq P(Y)$$

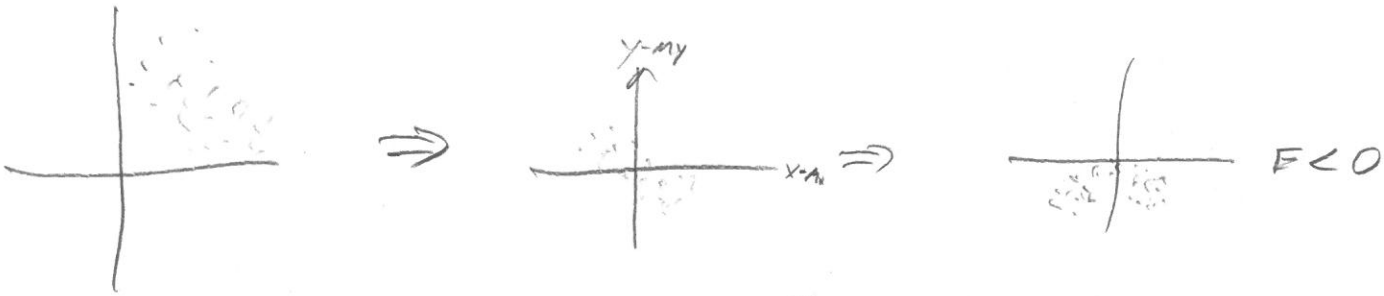
In data science if knowing a predictor x allows you to know something about y we say X, Y are "associated".

Recall covariance

$$Cov(X,Y) := E\left[(X-\mu_X)(Y-\mu_Y)\right] \quad \text{estimated by} \quad S_{xy} := \frac{1}{n-1}\sum (x_i-\bar{x})(y_i-\bar{y}) \in \mathbb{R}$$



$E(XY)$ here?



the sign of the cov. is important. It indicates if $\oplus$ if $x\uparrow \Rightarrow y\uparrow$

If $\ominus$ if $x\uparrow \Rightarrow y\downarrow$

Recall $:= Corr(X,Y) := \frac{Cov(X,Y)}{SE(X)SE(Y)} \in [-1,1]$ standardized! Convenient!

est by

$$r = \frac{S_{xy}}{S_x S_y} \in [-1,1]$$

we say X,y are pos. corr if $r>0$ which means $x\uparrow \Rightarrow y\uparrow$

we say " " neg. " " " $r<0$ " " $x\uparrow \Rightarrow y\downarrow$

" " " not " " " $r=0$ " " $x\uparrow \Rightarrow y$ unchanged

Correlation really is "linear correlation" as it appears in LS! It is a type of association

Previously $y \subseteq \mathbb{R}$, $p=1$, $\mathcal{H} = \{\vec{w}\cdot\vec{x} = w_0 + w_1 x : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}\}$

Now let $p=2 \Rightarrow \mathcal{H} = \{w_0 + w_1 x_1 + w_2 x_2 : \vec{w} \in \mathbb{R}^3\}$

for any $\vec{w}$, $(x_i - \hat{y}_i)^2$

$$SSE = \sum_{\langle \vec{x}_i, y_i\rangle \in \mathbb{D}} \left(y_i - (w_0 + w_1 x_{i1} + w_2 x_{i2})\right)^2 = \dots \text{long}$$

Take $\frac{\partial}{\partial w_0}[SSE] \overset{set}{=} 0$, $\frac{\partial}{\partial w_1}[SSE] \overset{set}{=} 0$, $\frac{\partial}{\partial w_2}[SSE] \overset{set}{=} 0$

Can we solve this any better?

$\mathbb{D} = \langle X, \vec{y}\rangle$ where $X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \in \mathbb{R}^{n \times 3}$ ($p+1$)

where is $X\vec{w} = \begin{bmatrix} w_0 + w_1 x_{11} + w_2 x_{12} \\ w_0 + w_1 x_{21} + w_2 x_{22} \\ \vdots \\ w_0 + w_1 x_{n1} + w_2 x_{n2} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} \Rightarrow \vec{\hat{y}} = X\vec{w}$

Recall $(\vec{a}+\vec{b})^T = \vec{a}^T + \vec{b}^T$

$$\Rightarrow SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (\vec{y} - \vec{\hat{y}})^T(\vec{y} - \vec{\hat{y}}) = (\vec{y}^T - \vec{\hat{y}}^T)(\vec{y} - \vec{\hat{y}})$$

$$= \vec{y}^T\vec{y} - \vec{\hat{y}}^T\vec{y} - \vec{y}^T\vec{\hat{y}} + \vec{\hat{y}}^T\vec{\hat{y}}$$

Recall $\vec{v}\cdot\vec{v} = \sum_{j=1}^{d} v_j^2$, $\vec{v} \in \mathbb{R}^d$, $\vec{v}^T\vec{v}$

Recall $\vec{a}\cdot\vec{b} = \vec{b}\cdot\vec{a}$

$$= \vec{y}^T\vec{y} - 2\vec{\hat{y}}^T\vec{y} + \vec{\hat{y}}^T\vec{\hat{y}}$$

$$= \vec{y}^T\vec{y} - 2(X\vec{w})^T\vec{y} + (X\vec{w})^T(X\vec{w})$$

Recall $(AB)^T = B^TA^T \Rightarrow = \vec{y}^T\vec{y} - 2\vec{w}^TX^T\vec{y} + \vec{w}^TX^TX\vec{w}$

Now we need $\frac{\partial}{\partial w_0}[\,]$, $\frac{\partial}{\partial w_1}[\,]$, ... $\frac{\partial}{\partial w_p}[\,]$

$\frac{\partial}{\partial \vec{z}}[\eta] = \begin{pmatrix} \frac{\partial}{\partial z_1}[\eta] \\ \vdots \\ \frac{\partial}{\partial z_n}[\eta] \end{pmatrix} = \vec{0}$ (2)

$\vec{z} \in \mathbb{R}^n$

we can imagine taking the derivative wrt to the whole vector

$$\frac{\partial SSE}{\partial \vec{w}} := \begin{bmatrix} \frac{\partial}{\partial w_0}[SSE] \\ \vdots \\ \frac{\partial}{\partial w_p}[SSE] \end{bmatrix}$$ i.e. the vector derivative

e.g. $\frac{\partial}{\partial \vec{z}}[q\vec{z}] = [q_1 c_1$

e.g. $\frac{\partial}{\partial \vec{z}}\left[\vec{z}^T \vec{b}\right] = \frac{\partial}{\partial \vec{z}}\left[c_1 b_1 + c_2 b_2 + \dots + c_n b_n\right] = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \vec{b}$

when $\vec{z} \in \mathbb{R}^n$, $\vec{b} \in \mathbb{R}^n$

e.g. $\frac{\partial}{\partial \vec{z}}[q f(\vec{z}) + g(\vec{z})]$

$= \begin{bmatrix} \frac{\partial}{\partial c_1}[q f(\vec{z}) + g(\vec{z})] \\ \vdots \\ \frac{\partial}{\partial c_n}[q f(\vec{z}) + g(\vec{z})] \end{bmatrix}$

$= \begin{bmatrix} q \frac{\partial}{\partial z_1}[f(\vec{z})] + \frac{\partial}{\partial c_1}[g(\vec{z})] \\ q \frac{\partial}{\partial c_n}[f(\vec{z})] + \frac{\partial}{\partial c_n}[g(\vec{z})] \end{bmatrix}$

$= q \frac{\partial}{\partial \vec{z}}[f(\vec{z})] + \frac{\partial}{\partial \vec{z}}[g(\vec{z})]$

e.g. $\frac{\partial}{\partial \vec{z}}\left[\vec{z}^T A \vec{z}\right]$ where $A \in \mathbb{R}^{n \times n}$ and $\vec{z} \in \mathbb{R}^n$ and A symmetric

Note $A\vec{z} = \begin{bmatrix} a_{11}c_1 + a_{12}c_2 + \dots + a_{1n}c_n \\ a_{21}c_1 + a_{22}c_2 + \dots + a_{2n}c_n \\ \vdots \\ a_{n1}c_1 + a_{n2}c_2 + \dots + a_{nn}c_n \end{bmatrix}$, $\vec{z}^T(A\vec{z}) = c_1(a_{11}c_1 + a_{12}c_2 + \dots + a_{1n}c_n)$

$+ c_2(a_{21}c_1 + a_{22}c_2 + \dots + a_{2n}c_n)$

$\vdots$

$+ c_n(a_{n1}c_1 + a_{n2}c_2 + \dots + a_{nn}c_n) = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}c_i c_j$

$\frac{\partial}{\partial \vec{z}}[\ ] = \begin{bmatrix} \frac{\partial}{\partial c_1}[\ ] = 2c_1 a_{11} + a_{12}c_2 + \dots + a_{1n}c_n + c_2 a_{21} + \dots + c_n a_{n1} = 2c_1 a_{11} + 2c_2 a_{12} + \dots + 2c_n a_{1n} \\ \frac{\partial}{\partial c_2}[\ ] = a_{21}c_1 + 2a_{22}c_2 + \dots + a_{2n}c_n + c_1 a_{12} + c_3 a_{32} + \dots = 2\vec{a}_{2\cdot}^T \vec{z} \\ \vdots \end{bmatrix}$

$= 2\vec{a}_{1\cdot}^T \vec{z}$

$= c_1(\vec{z}^T \vec{a}_{1\cdot}) + c_2(\vec{z}^T \vec{a}_{2\cdot}) + \dots + c_n(\vec{z}^T \vec{a}_{n\cdot})$

$= 2\begin{bmatrix} \vec{a}_{1\cdot}^T \\ \vec{a}_{2\cdot}^T \\ \vdots \\ \vec{a}_{n\cdot}^T \end{bmatrix} \vec{z} = 2A\vec{z}$

Now let's try to apply this to our problem at hand

$\frac{\partial}{\partial \vec{w}}[SSE] = \frac{\partial}{\partial \vec{w}}\left[\vec{y}^T\vec{y} - 2\vec{w}^T X^T \vec{y} + \vec{w}^T X^T X \vec{w}\right]$

$= \frac{\partial}{\partial \vec{w}}[\vec{y}^T\vec{y}] - 2\frac{\partial}{\partial \vec{w}}[\vec{w}^T X^T \vec{y}] + \frac{\partial}{\partial \vec{w}}[\vec{w}^T \underbrace{X^T X}_{\text{symmetric?}} \vec{w}]$

$= \vec{0} - 2X^T\vec{y} + 2X^T X\vec{w} \overset{\text{set}}{=} 0$

$(X^T X)^T = (X^T(X^T)^T) = X^T X$ ✓

$\Rightarrow X^T\vec{y} = X^T X\vec{w} \Rightarrow \boxed{\vec{b} = (X^T X)^{-1} X^T \vec{y}}$

LS solution!

Note: $\underset{(p+1)\times n}{(X^T)}\underset{(n\times p+1)}{(X)}$ must be invertible!

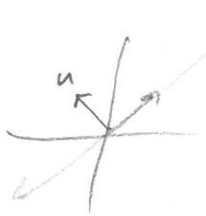
$\in \mathbb{R}^{p+1 \times p+1}$

When is $X^T X$ invertible?

Only when $\text{rank}[X] = p+1$ Proof:

Assume $\text{rank}(X) < p+1$ $\Rightarrow$ there's a non-trivial nullspace i.e. a vector $\vec{u} \neq 0 \in \mathbb{R}^{p+1}$ that can be mapped to $\vec{0}$.

i.e. $X\vec{u} = \vec{0}_n$ $\quad X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$



$V = \text{colspace}(X)$ is just this line

$\vec{u} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

But then if we use X^TX to map vector $\vec{u}$,

$$(X^TX)\vec{u} = X^T(X\vec{u}) = X^T\vec{0}_n = \vec{0}_{p+1} \qquad X\vec{u} = \vec{0}$$

this means $\vec{u} \in \text{Nullspace}(X^TX)$ and thus $\dim[\text{Nullspace}(X^TX)] > 0 \Rightarrow X^TX$ is not invertible.

What does $\text{rank}(X) = p+1$ mean? Each col is not lin. dep. on other cols.
$\Rightarrow$ No predictor information is duplicated

$$X = \begin{bmatrix} 1 & \text{Salary\_2000} & \text{Salary\_2001} & \cdots & \text{not\_salary} \\ 1 & \vdots & \vdots & & \vdots \\ \vdots & & & & \\ 1 & & & & \end{bmatrix}$$

not full rank $\Rightarrow$ delete not_salary! (or one of the yrs)

$$X = \begin{bmatrix} 1 & \text{height\_in\_feet} & \text{height\_in\_meters} \\ 1 & \vdots & \vdots \\ \vdots & & \\ 1 & & \end{bmatrix}$$

// $\Rightarrow$ // // //

What is g?

$$g(\vec{x}^a) = \vec{x}^{a\,T}\vec{b} = \begin{bmatrix} 1 & x_1^a & \cdots & x_p^a \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} = b_0 + b_1x_1^a + \cdots + b_px_p^a$$