Lecture 3

§ 2.8 Least Squares

$y_1, \cdots, y_p$ observations

use predict $\Delta = f(y_1, \cdots, y_p)$

Idea: find f

f ∞ dim

— $y_1, \cdots, y_p$

— restrict dim of search

parametrized class

$$g(y_1, \cdots, y_p, \theta)$$

$\downarrow$ restrict dim   predictor

distance

$$d(U, V) = E(U-V)^2$$

$$\min_\theta E(X - g(y, \theta))^2$$

Linear Least Squares (LLS)

$$g(y_1, \cdots, y_p, \theta) = \theta^T y = \sum_{i=1}^p \theta_i y_i$$

$$D(\theta) = E(X - \theta^T y)^2 = E(X^2 - 2\theta^T xy + \theta^T y y^T \theta) = E(X^2) - 2\theta^T E(Xy) + \theta^T E(yy^T)\theta$$

Thm: $a^T y$ is LLS $\iff$ a satisfies $[E(yy^T)]a = E(xy)$

want to predict $x = f(y_1, \cdots, y_p)$

Proof: $a^T y$ is LLS $\Rightarrow a$ minimizes D

$\Rightarrow \nabla D = 0$

$\downarrow$

$$-2E(XY) + 2[E(YY^T)] \cdot a = 0$$

$\Rightarrow E(XY) = E(yy^T)a$

So $E(Xy) = E(yy^T)a \Rightarrow D(a)$ is mimimized. ∎

$$yy^T = \begin{pmatrix} y_1^2 & y_1 y_2 & \cdots & y_1 y_p \\ y_2 y_1 & & & \\ \vdots & & & \\ y_p y_1 & \cdots & & y_p^2 \end{pmatrix}$$

$$E(yy^T) = \begin{pmatrix} E(y_1^2) & & \cdots & \\ \vdots & E(y_2^2) & & \\ \vdots & & \ddots & \\ & & & E(y_p^2) \end{pmatrix}$$

Expectation termwise

$$D(\theta) = E(x-\theta^T y)^2 = E(x - a^T y + (a-\theta)^T y)^2$$
$$= \underbrace{E((x-a^T y)^2}_{=D(a)} + 2(x-a^T y)(a-\theta)^T y + ((a-\theta^T)y)^2)$$

$$D(\theta) = D(a) + 2E((x-a^T y)(a-\theta)^T y) + E((a-\theta)^T y)^2$$

$E((x-a^T y)(a-\theta)^T y) = 0$ (we know $(x^T y) = y^T x$)
since $E(x-a^T y)y^T(a-\theta) = 0$

so $D(\theta) = D(a) + \underbrace{E((a-\theta)^T y)^2}_{} \geq 0$

so $D(\theta) \geq D(a)$

So a is a min of D.

b/c $E(xy^T) - a^T E(yy^T)(a-\theta)$

But $[E(yy^T)a = E(xy)]^T \Rightarrow a^T E(yy^T) = E(xy^T)$

Ex: Show $g(x, a_0, a) = a_0 + a^T y$ is LS

$\Longleftrightarrow E(x) = a_0 + a^T E(y)$ where a solves

$$E(y-E(y))(y-E(y))^T)a = E[(x-E(x))(y-E(y))]$$

$$Var(y) \cdot a = Cov.(x,y)$$

Eg : x, y scalar
$g(y, a_0, a) = a_0 + ay$
where
$Var(y) \cdot a = cov(x,y)$
$a_0 = E(x) - aE(y)$

Best predictor of x
$$g(y, a_0, a) = E(x) + \frac{Cov(x,y)}{Var(y)}(y - E(y))$$

Ex.

- x a person's weight
- y height
- $E(x) = 60$ kg std dev $= \sqrt{Var X} = 20$kg
- $E(y) = 165$ cm $\sqrt{Var y} = 10$ cm

$$\text{Correlation}(x,y) = \frac{Cov(x,y)}{\sqrt{Var X \, Var y}} = 0.7$$

$$\frac{Cov(x,y)}{Var(y)} = 1.4 \qquad g(y) = 60 + 1.4(y - 165)$$

someone's 175cm $\longrightarrow$ expected weight is $g(175) = 60 + 1.4(175-165) = 74$ kg

Prob. LS

know E, xy, yy^T

$\Rightarrow E(yy^T), E(xy)$

$\Rightarrow a, g(y,a)$

Stat. LS

don't know E

instead, know sample of size n: $(x_1, y^1), \cdots (x_n, y^n)$

from E

$$\hat{D}(\theta) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta^T y^i)^2}_{\text{if } n \text{ is big, } \approx D(\theta)} \xrightarrow{LLN} E[(X-\theta^T y)^2]$$

if $\hat{a}$ minimizes $\hat{D}$, hope $\hat{a} \approx a$

Applications

– **Inequalities**

$$\mathbb{1}(X \geq a) = \begin{cases} 1, & \text{if } x \geq a \\ 0, & \text{o.w.} \end{cases}$$

$\downarrow$

Indicator function

Sps $X \geq 0$, $a > 0$

① $a > x \Rightarrow$ then $\mathbb{1}(X \geq a) = 0 = 0 \cdot (\frac{x}{a}) = (\frac{x}{a})\mathbb{1}(X \geq a) \leq X/a$

② $a \leq x \Rightarrow \mathbb{1}(X \geq a) \leq (\frac{x}{a})\mathbb{1}(X \geq a) \Rightarrow \mathbb{1}(X \geq a) \leq X/a \Rightarrow P(X \geq a) = E(\mathbb{1}(X \geq a)) \leq E(\frac{X}{a})$

Markov's Inequality

given X, $E(X)$, what's $P(|X - E(X)| > x)$?

if $Var(X)$ exists $\quad = P(|X - E(X)|^2 > x^2) \leq \frac{E(|X-E(X)|^2)}{x^2} = \frac{Var(X)}{x^2}$

$\Rightarrow$ Chebychev's Inequality $P(|X - E(X)| > x) \leq \frac{Var(X)}{x^2}$

if $Var(X)$ exists

could have said: $P(|X - E(X)| > x) \overset{\text{by Markov}}{\leq} \frac{E|X - E(X)|}{x} \leq \frac{\sqrt{Var(X)}}{x}$

$E(X) \leq E(X^2)^{\frac{1}{2}}$

Application

if $Var(X)=0$, $P(X=E(X))=1$

(Proof): $\forall K\in\mathbb{N}$, $P(|X-E(X)|>1/K)\leq K^2 Var(X)=0$

$1/K\to 0$, $|X-E(X)|>1/K$, $|X-E(X)|>1/(K+1)$

$y_K=\mathbb{1}(|X-E(X)|>1/K)\longrightarrow \mathbb{1}(|X-E(X)|>0)$

$P(|X-E(X)|>1/K)\longrightarrow P(|X-E(X)|>0)\leftarrow 0$

By Axiom 5 $\Rightarrow P(|X-E(X)|>0)=\lim_{K\to\infty} P(|X-E(X)|>0)=0$

$P(X=E(X))=1-P(X\neq E(X))=1-P(|X-E(X)|>0)=0=1-0=1$

Thm: Cauchy-Schwartz:
X,Y are 2 r.v.'s . then $(E(XY))^2\leq E(X^2)E(Y^2)$

Corollary: $E|X-E(X)|^2=E[(1)\cdot|X-EX|]^2$

by C-S $\leq(E\,1^2)(E(X-EX)^2)\Rightarrow E|X-EX|\leq[Var(X)]^{\frac{1}{2}}$

"=" $\Leftrightarrow E[(C_1X+C_2Y)^2]=0$, $C_1,C_2\neq 0$

Proof: take $Z=(X,Y)^T$, $\forall C\in\mathbb{Z}^2$

$\forall C$. YZ $E(C^TZ)^2\geq 0$ $\rightarrow$ $=C^T(E\,ZZ^T)C$

$$\begin{pmatrix} Ex^2 & Exy \\ Exy & Ey^2 \end{pmatrix}\triangleq u$$

u is positive,
by def,
$\Rightarrow \det(u)\geq 0$
$Ex^2\cdot Ey^2-(E(xy))^2$

In LS, we used $d(u,v)=E(u-v)^2$
To measure , closeness
- If $d(u,v)$ is small, in what way are they close:
$Var(u-v)=E(u-v)^2-(E(u-v))^2\leq E(u-v)^2=d(u,v)$

If $d(u,v)=0$, then $Var(u-v)=0\Rightarrow P(U-V=E(u-v))=1\Rightarrow P(u=v)=1$
$E(u-v)^2=Ec^2=c^2=d(u,v)$

by chebychev : $P(|u-v-E(u-v)|>x)\leq Var(u-v)/x^2\leq\frac{d(u,v)}{x^2}$
markov

if $d(u,v)\leq\varepsilon^2$ (small), then $P(|u-v-E(u-v)|>x)\leq(\frac{\varepsilon}{x})^2$

$|E(u-v)|\leq\varepsilon$

C-S : $E(u-v)\leq(d(u,v))^{\frac{1}{2}}$