Prediction, Linear Regression, Bivariate Normal Distribution.

(X, Y).

Looking at $\phi(X)$ which best predicts Y

$\Rightarrow$ minimizing $\underbrace{E[(Y-\phi(X))^2]}_{\text{squared error}}$ over ϕ.

Saw last time that $\phi_{opt}(X) = E[Y|X]$

Leads to a natural representation:

$$Y = \mu(X) + \varepsilon$$

where $\mu(X) = E[Y|X]$ regression function

and

ε (Residual) $= Y - E(Y|X)$.

and note $\underline{E[\varepsilon|X]} = 0$.

Statisticians try to model $\mu(X)$ via tractable classes of functions.

Typically we optimize

$\mathbb{E}[(Y - \phi(X))^2]$ over all ϕ belonging

to $\mathcal{F}_0 \subsetneq \mathcal{F}$ all measurable functions of X

$$\phi_{opt, \mathcal{F}_0} = \underset{\phi \in \mathcal{F}_0}{\operatorname{argmin}} \mathbb{E}[(Y - \phi(X))^2]$$

if $\phi_{opt}(X) = \mathbb{E}[Y|X] \in \mathcal{F}_0$, then

of course $\phi_{opt, \mathcal{F}_0}(X) = \phi_{opt}(X)$

$\mathbb{E}[(Y - \phi_{opt, \mathcal{F}_0}(X))^2]$ is small is

~~good~~ enough. (usually).

$\mathcal{F}_0$ = class of linear functions of X

i.e $\{\alpha + \beta X : (\alpha, \beta) \in \mathbb{R}^2\}$

$$\mathbb{E}[Y|X]_{\text{postulated}} = \alpha + \beta X$$

(working model)

We're going to predict Y using linear functions.

Case I: $E[Y|X] = \alpha_0 + \beta_0 X$ for some (α_0, β_0).

Then $Y = \alpha_0 + \beta_0 X + \varepsilon$, with $E(\varepsilon|X) = 0$

(Linear Model)

Case II: $E[Y|X]$ is NOT a linear function of X, say it's $1 + X + X^2$.

Nevertheless seek to find:

$$\underset{(\alpha,\beta)}{\operatorname{argmin}} \; \mathbb{E}\left[(Y - \alpha - \beta X)^2\right] \longleftarrow$$

Notation: $\mu_X = EX$, $\mu_Y = EY$

To be differentiated from $\mu(X) = E[Y|X]$

$\sigma_{XY} = \mathrm{Cov}(X,Y)$

$\sigma_X^2 = \mathrm{Var}(X)$

$\sigma_Y^2 = \mathrm{Var}(Y)$

$\rho_{XY} = \mathrm{Corrn}(X,Y)$

Set $\nabla_\alpha E[(Y-\alpha-\beta X)^2] = 0 \leftarrow$ (i)

$\nabla_\beta E[(Y-\alpha-\beta X)^2] = 0 \Leftarrow$ (ii)

(i) $\Rightarrow E[Y - \alpha - \beta X] \quad 0$

(ii) $\Rightarrow E[X(Y - \alpha - \beta X)] = 0$.

From (i) : $\mu_Y = \alpha + \beta\mu_X$

From (ii) $E(XY) = \alpha\mu_X + \beta\, \mathbb{E}(X^2)$

$\alpha = \mu_Y - \beta\mu_X$.

So: $E[XY] = (\mu_Y - \beta\mu_X)\mu_X + \beta\, \mathbb{E}X^2$

$= \mu_Y\mu_X + \beta(\underbrace{\mathbb{E}(X^2) - \mu_X^2}_{Var(X)})$

$$\Rightarrow \beta = \frac{\mathbb{E}(XY) - \mu_X\mu_Y}{Var(X)}$$

$$= \frac{\sigma_{XY}}{Var(X)} = \rho_{XY} \cdot \frac{\sigma_Y}{\sigma_X}$$

$\beta_0 = \rho_{XY}\dfrac{\sigma_Y}{\sigma_X}$, $\alpha_0 = \mu_Y - \beta_0\mu_X$.

$\phi_{opt, linear}(x) = \alpha_0 + \beta_0 X$.

$Y = \phi_{opt, linear}(x) + \tilde{\varepsilon}$.

$\tilde{\varepsilon}$ (pseudo residual)

$= Y - \alpha_0 - \beta_0 X$.

Note ~~that~~ $\mathbb{E}[\tilde{\varepsilon} | X] \neq 0$ when

$\mathbb{E}[Y|X]$ is NOT linear.

$$\mathbb{E}[\tilde{\varepsilon} | X] = \underline{E[Y|X]} \; \underline{- \alpha_0 - \beta_0 X}$$

$\neq 0$.

$\underline{\mathbb{E}[\tilde{\varepsilon}] = 0}$, $\underline{Cov(X, \tilde{\varepsilon}) = 0}$

Check.

We're still interested to see how our WORKING linear model (typically INCORRECT) did.

Look at the squared error.

$$\mathbb{E}\left[(Y - \alpha_0 - \beta_0 X)^2\right]$$

$$= \mathbb{E}\left[(Y - \mu_Y) - \beta_0 (X - \mu_X)\right]^2$$

$$= \underline{\mathbb{E}\left[(Y-\mu_Y)^2\right]} - 2\beta_0 \, \underline{\mathbb{E}\left[(Y-\mu_Y)(X-\mu_X)\right]}$$

$$+ \beta_0^2 \, \underline{\mathbb{E}\left[(X-\mu_X)^2\right]}$$

$$\boxed{\beta_0 = \frac{\sigma_{XY}}{\sigma_X^2}}$$

$$= \sigma_Y^2 - 2\beta_0 \sigma_{XY} + \beta_0^2 \cdot \sigma_X^2$$

$$= \sigma_Y^2 - 2\frac{\sigma_{XY}^2}{\sigma_X^2} + \frac{\sigma_{XY}^2}{\sigma_X^4} \cdot \sigma_X^2$$

$$= \sigma_Y^2 - \frac{\sigma_{XY}^2}{\sigma_X^2}$$

Greatest reduction when $|\rho|$ near 1

$$= \sigma_Y^2 \left[1 - \frac{\sigma_{XY}^2}{\sigma_X^2 \cdot \sigma_Y^2}\right]$$

$$= \underline{\underline{\sigma_Y^2}} \underline{\left[1 - \rho_{XY}^2\right]}$$

- if you predicted Y by μ_Y squared error is σ_Y^2.

Correlation therefore captures the goodness of the linear fit.

$\rho_{XY} \simeq 0$ means linear functions of X are terrible for predicting Y.

But X itself may NOT be a terrible predictor provided you use the right sort of function to predict.

~~look at the~~ example:

$Y = X^2 + \varepsilon, \quad X \sim \text{Unif}[-1, 1]$

ε is independent of X

$\varepsilon \sim N[0, \sigma^2]$

$E[Y \mid X] = X^2$

X is useful but not if you restrict to the linear model.

Linear Regression and the Bivariate Normal Distribution:

Classical Linear Regression Model (Signal + Noise model):

$$Y = \alpha_0 + \beta_0 X + \varepsilon$$

(X, ε) are mutually independent.

$\mathbb{E}\varepsilon = 0$. Note: $\mathbb{E}[\varepsilon | X] = 0$

Further assume normal distributions

$$X \sim N(\mu_X, \sigma_X^2)$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$E[Y|X] = \alpha_0 + \beta_0 X.$$

$$Y|X \sim N[(\alpha_0 + \beta_0 X), \sigma^2]$$

$$X \sim N(\mu_X, \sigma_X^2)$$

Deviate from the notes a bit.

$f_{X,Y}(x,y)$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right]$$

$$\times \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}\left(y-\alpha_0-\beta_0 x\right)^2\right]$$

$$= \frac{1}{2\pi\sigma_X\sigma} \exp\left[-\frac{1}{2\sigma_X^2}(x-\mu_X)^2 - \frac{1}{2\sigma^2}(y-\alpha_0-\beta_0 x)^2\right]$$

$\uparrow$

Write this in terms of the moment-parameters $(\mu_X, \mu_Y, \sigma_{XY}, \sigma_X^2, \sigma_Y^2)$

$$\beta_0 = \frac{\sigma_{XY}}{\sigma_X^2} = \rho \cdot \frac{\sigma_Y}{\sigma_X}$$

$\mu_Y = \alpha_0 + \beta_0 \mu_X$, so $\alpha_0 = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X$

Substitute the expressions in terms of the moments

Algebra

$\Rightarrow f(x,y)$

$$= \frac{1}{2\pi\sigma_x \underline{\underline{\sigma}}} \exp\Big[-\frac{1}{2\sigma_x^2}(x-\mu_x)^2$$

$$- \frac{1}{2\sigma^2}\Big[y-\Big(\mu_y+\rho\cdot\frac{\sigma_y}{\sigma_x}(x-\mu_x)\Big)\Big]^2\Big]$$

$*$

$$\sigma_y^2 = Var(Y) = E[\underline{Var(Y|X)}] + Var[E(Y|X)]$$

$$= \sigma^2 + Var(\alpha_0 + \beta_0 X)$$

$$= \sigma^2 + \beta_0^2\sigma_x^2.$$

$$\Rightarrow \sigma^2 = \sigma_y^2\left(1 - \frac{\beta_0^2\sigma_x^2}{\sigma_y^2}\right)$$

$*$

$$= \sigma_y^2\left(1 - \frac{\sigma_{xy}^2}{\sigma_x^4\sigma_y^2}\sigma_x^2\right)$$

$$= \sigma_y^2(1-\rho_{xy}^2) \leftarrow$$

Replacing σ^2 by its form in terms of the moment parameters:

$$f(x,y)$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{(X-\mu_X)^2}{2\sigma_X^2} - \frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\left(\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right)\right]^2\right]$$

look at the term within BIG [ ]:

$$= -\frac{1}{2(1-\rho^2)}\left[(1-\rho^2)\left(\frac{X-\mu_X}{\sigma_X}\right)^2 + \frac{1}{\sigma_Y^2}\left[(Y-\mu_Y)-\rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X)\right]^2\right]$$

Then just expand square and combine terms to get:

$$-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]$$

This finally puts the joint density in the standardized form:

$Y = \alpha_0 + \beta_0 X + \varepsilon$

$$f(x,y) = \boxed{\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}}\exp\left[-T\right] \longleftarrow$$

$$T = \frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right\}$$

$$\mathrm{BVN}\left[\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}\right]$$

We'll see ~~that~~ a reverse linear model representation also holds in this case:

I.E:

$$X = \tilde{\alpha}_0 + \tilde{\beta}_0 Y + \tilde{\varepsilon}$$

$\tilde{\varepsilon} \perp\!\!\!\perp Y$ and $\tilde{\varepsilon} \sim N[0, \tilde{\sigma}^2]$

$$\tilde{\sigma}^2 = \sigma_X^2 (1 - \rho_{XY}^2)$$

<u>Exercise</u>: Use COVT in 2 dimensions to show that:

$$(U, V) \equiv \left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y} \right)$$

$$\sim BVN[0, 0, 1, 1, \rho]$$

$$f_{U,V}(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)\right]$$