Lec 4/12

Wednesday, April 12, 2017 15:02

Regression & correlation

§ 14.1

Let $f(x,y)$ be joint density of r.v.s X and Y.

① determine conditional density of Y given $X = x$.

② Conditional mean $\underbrace{\mu_{Y|x} = \mathbb{E}(Y|x) = \int_{-\infty}^{\infty} y\, w(y|x)\, dy}_{\text{regression equation of } Y \text{ on } X}$.

Remarks: 1 can define regression of Y on X_1 and X_2 (or $X_1, \dots, X_n$).

$$\mu_{Y|x_1,x_2} = \mathbb{E}(Y|x_1,x_2) = \int_{-\infty}^{\infty} y\, w(y|x_1,x_2)\, dy$$

2: in the discrete case, we can replace pdf w/ pmf. / Int → sum.

3: regression eqn of Y on X, $\mu_{Y|x}$, is often used for prediction of Y when we observe only X.

Ex: Let $(X_1, X_2, X_3) \sim$ pdf $f(x_1, x_2, x_3) = \begin{cases} (x_1 + x_2)e^{-x_3} & \text{for } x_1, x_2 \in (0,1),\ x_3 > 0 \\ 0 & \text{otherwise.} \end{cases}$

find $\mu_{X_2|x_1,x_3}$

Sol: marginal density of X_1, X_3.

$$m(x_1, x_3) = \begin{cases} \int_0^1 (x_1+x_2)e^{-x_3}\, dx_2 & x_1 \in (0,1),\ x_3 > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

$$= \begin{cases} (x_1 + \frac{1}{2})e^{-x_3} & x_1 \in (0,1),\ x_3 > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

$$\mu_{X_2|x_1,x_3} = \mathbb{E}(X_2|x_1,x_3) = \int^1 x_2 \cdot \frac{f(x_1,x_2,x_3)}{\dots}\, dx_2$$

$$\mu_{X_2|x_1,x_3} = \mathbb{E}(X_2 \mid x_1, x_3) = \int_0^1 x_2 \cdot \frac{f(x_1,x_2,x_3)}{m(x_1,x_3)}\, dx_2$$

$$= \int_0^1 x_2 \, \frac{x_1 + x_2}{x_1 + \frac{1}{2}} \, dx_2$$

$$= \frac{1}{x_1 + \frac{1}{2}} \int_0^1 (x_1 x_2 + x_2^2)\, dx_2$$

$$= \frac{\frac{x_1}{2} + \frac{1}{3}}{x_1 + \frac{1}{2}} = \frac{x_1 + \frac{2}{3}}{2x_1 + 1}.$$

§14.2 Linear Regression.

$\mu_{Y|x}$ is a function of x. In particular, when $\mu_{Y|x} = \alpha + \beta x$ is linear, we call it "linear regression". This is good.

Theorem 14.1: If $\mu_{Y|x}$ is linear, then:

$$\mu_{Y|x} = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

where $\mu_2 = \mathbb{E}Y$ $\quad \mu_1 = \mathbb{E}X$ $\quad \sigma_2^2 = \mathrm{Var}(Y)$ $\quad \sigma_1^2 = \mathrm{Var}(X)$

$$\rho = \underbrace{\frac{\sigma_{12}}{\sigma_1 \sigma_2}}_{\text{correlation coefficient}} \qquad \sigma_{12} \to \mathrm{cov}(X,Y)$$

Proof: $\mu_{Y|x} = \alpha + \beta x$

(1)

$$\Rightarrow \int y \, w(y|x)\, dy = \alpha + \beta x$$

$\cdot\, g(x)$

$$\Rightarrow \int y \, \underbrace{w(y|x)\, g(x)}_{f(x,y)}\, dy = (\alpha + \beta x)\, g(x)$$

marginal density on x

$$\Rightarrow \iint y \, f(x,y)\, dy\, dx = \int (\alpha + \beta x)\, g(x)\, dx$$

$$\Rightarrow \quad \mathbb{E}Y = \alpha \int g(x)\,dx + \beta \int x g(x)\,dx$$

$$= \alpha + \beta\, \mathbb{E}X$$

$$\Rightarrow \quad \mu_2 = \alpha + \beta \mu_1 \qquad (\star_1)$$

② $\star_1 \xrightarrow{\cdot\, x g(x)} \int xy\, f(x,y)\,dy = (\alpha x + \beta x^2)\, g(x)$

$$\Rightarrow \iint xy\, f(x,y)\,dy\,dx = \int (\alpha x + \beta x^2)\, g(x)\,dx$$

$$\Rightarrow \quad \mathbb{E}(XY) = \alpha\, \mathbb{E}(X) + \beta \int x^2 g(x)\,dx$$

$$= \alpha\, \mathbb{E}(X) + \beta\left(\mathrm{Var}(X) + (\mathbb{E}X)^2\right)$$

$$\Rightarrow \quad \sigma_{12} + \mu_1\mu_2 = \alpha\mu_1 + \beta(\sigma_1^2 + \mu_1^2) \qquad (\star_2)$$

Now solve the system $(\star_1, \star_2)$

$$\Rightarrow \quad \alpha = \mu_1 - \rho\frac{\sigma_2}{\sigma_1}\mu_1$$

$$\beta = \rho\frac{\sigma_2}{\sigma_1}$$

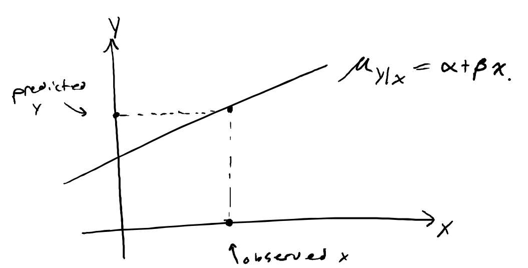
HW: if Regression of Y on X is linear, is the regression of X on Y also linear?

Remarks: ① $\rho = 0 \Leftrightarrow \mu_{Y|x}$ is constant (does not depend on x).

Ex 14.9

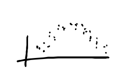
when $\rho = 0$, X, Y are uncorrelated, but not necessarily independent.

② $-1 \le \rho \le 1$ and is slope of regression line

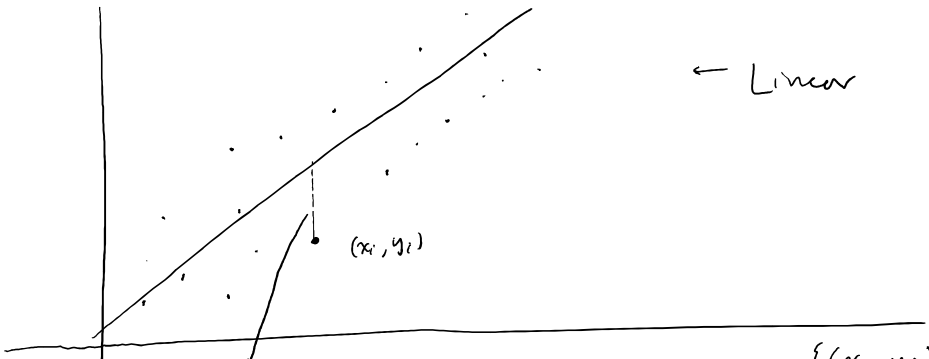


§14.3 Method of Least Squares.

When only paired data is given. "curve fitting."



not linear



← Linear

$e_i = y_i - (\alpha + \beta x_i)$ for $i = 1, \dots, n$. $\{(x_i, y_i)\}_{i=1}^{n}$ is a set of paired data.

The least-squares estimate $\hat{\alpha}^*$ and $\hat{\beta}^*$ minimize $\sum_{i=1}^{n} e_i^2$.