

# Bi-variate Statistics

Focus on continuous RV's

$X \sim$  cont. random variable

$Y \sim$  cont. random variable

# Interdependence between RV's

Three Levels of Interdependence

- Stochastic Interdependence
- Mean-dependence, variance-dependence
- Correlation

## Joint Probability Distribution

- Joint CDF

$$F(x, y) \equiv \Pr(X \leq x, Y \leq y)$$

- Joint pdf

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}$$

## Stochastic Independence

$X$  and  $Y$  are stochastically independent if  $f(x, y) = g(x)h(y)$  for all  $x, y$ .

Otherwise, they are stochastically interdependent.

No direction.

## Unconditional Probability

Marginal pdf of X (unconditional on y)

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{not in term of } y$$

Marginal pdf of Y (unconditional on x)

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad \text{not in term of } x$$

(c) Pongsa Pornchaiwiseskul,

5

## Uncond. Mean & Variance

Unconditional mean of Y

$$\mu_Y = \int_{-\infty}^{\infty} yh(y) dy \quad \text{a constant}$$

Unconditional variance of Y

$$\sigma_Y^2 = \int_{-\infty}^{\infty} (y - \mu_Y)^2 h(y) dy \quad \text{a constant}$$

(c) Pongsa Pornchaiwiseskul,

7

## Conditional Probability

Conditional pdf of X given Y=y

$$G(x|y) = \frac{f(x, y)}{h(y)} \quad \text{in term of } (x, y)$$

Conditional pdf of Y given X=x

$$H(y|x) = \frac{f(x, y)}{g(x)} \quad \text{in term of } (x, y)$$

(c) Pongsa Pornchaiwiseskul,

6

## Conditional Mean & Variance

Conditional mean of Y given X=x

$$\mu_{Y|X} = \int_{-\infty}^{\infty} yH(y|x) dy \quad \begin{array}{l} \text{in term of } x \\ \text{but not } y \end{array}$$

Conditional variance of Y given X=x

$$\sigma_{Y|X}^2 = \int_{-\infty}^{\infty} (y - \mu_{Y|X})^2 H(y|x) dy \quad \begin{array}{l} \text{in term of } x \\ \text{but not } y \end{array}$$

(c) Pongsa Pornchaiwiseskul,

8

## Mean-independence (1)

Y is mean-independent from X if

$\mu_{Y|X}$  is constant or does not depend on the value of X

X is mean-independent from Y if

$\mu_{X|Y}$  is constant or does not depend on the value of Y

## Variance-independence (1)

Y is variance-independent from X if

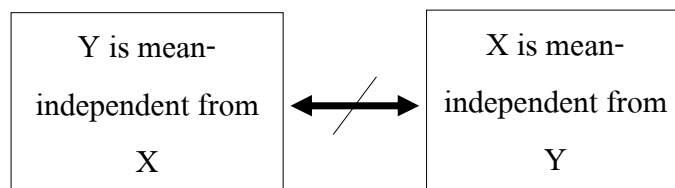
$\sigma_{Y|X}$  is constant or does not depend on the value of X

X is variance-independent from Y if

$\sigma_{X|Y}$  is constant or does not depend on the value of Y

## Mean-independence (2)

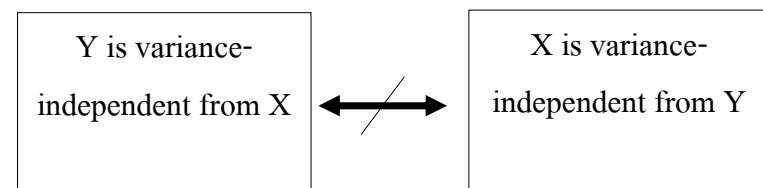
Note that



Direction matters.

## Variance-independence (2)

Note that



Direction matters.

## Population Covariance (1)

### Definition

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$
$$= \iint (x - \mu_X)(y - \mu_Y)f(x, y)dx dy$$

a constant

(c) Pongsa Pornchaiwiseskul,

13

## Population Covariance (3)

### Magnitude of Covariance

unbounded

depends on the units of both RV's

### Unit of covariance

= unit of X times unit of Y

e.g., X is in Baht and Y is in Kilogram

$\sigma_{XY}$  is in Baht-Kilogram

(c) Pongsa Pornchaiwiseskul,

15

## Population Covariance (2)

### Sign of Covariance

Positive ==> if one RV is above or below its mean, the other RV tends to be also above or below its mean

Negative ==> if one RV is above or below its mean, the other RV tends to be below or above its mean

(c) Pongsa Pornchaiwiseskul,

14

## Population Correlation (1)

### • Definition

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

### • Sign of Correlation

—same as that of Covariance

(c) Pongsa Pornchaiwiseskul,

16

## Population Correlation (2)

### Magnitude of Correlation

always bounded between -1 and 1

$$-1 \leq \rho_{XY} \leq +1$$

### Unit of Correlation

no unit

comparable between populations

(c) Pongsa Pornchaiwiseskul,

17

## Sample Covariance

$s_{XY}$  is an estimator for  $\sigma_{XY}$

Required paired sample

Estimator

$$s_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

(c) Pongsa Pornchaiwiseskul,

19

## Population Correlation (3)

### Interpretation of Correlation

$\rho_{XY} = +1 \implies$  If a variable is above or below its mean, the other will be above or below its own mean with certainty

$\rho_{XY} = -1 \implies$  If a variable is above or below its mean, the other will be below or above its own mean with certainty

$\rho_{XY} = 0 \implies$  If a variable is deviated from its mean, the other will be expected at its mean

(c) Pongsa Pornchaiwiseskul,

18

## Paired Sample of Size $n$

$i$	$X_i$	$Y_i$
1	$X_1$	$Y_1$
2	$X_2$	$Y_2$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$n$	$X_n$	$Y_n$

(c) Pongsa Pornchaiwiseskul,

20

## Sample Correlation (1)

$r_{XY}$  is an estimator of  $\rho_{XY}$

Definition 
$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

Sign of sample Correlation

same as that of sample Covariance

## Sample Correlation (2)

Magnitude of Sample Correlation

same as population correlation

always bounded between -1 and 1

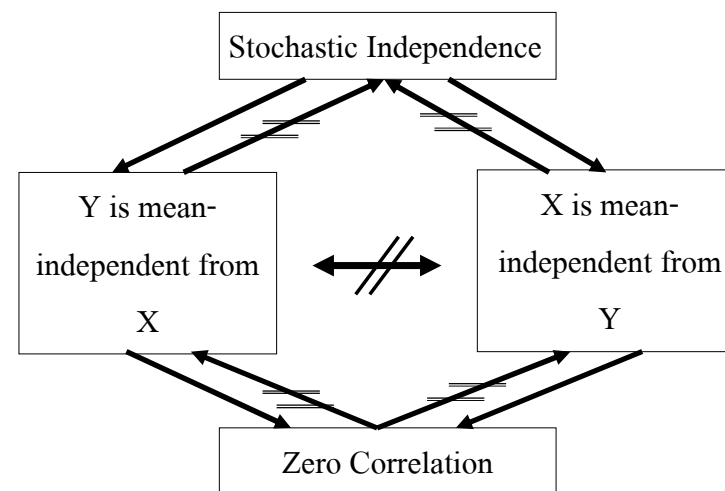
$$-1 \leq r_{XY} \leq +1$$

Unit of sample Correlation

no unit

comparable between data sets

## Hierarchy of Independence



## Tests for Independence

- Test for Stochastic Independence
- Test for Zero Correlation
- Test for Mean-independence

## Test for Stochastic Independence

- non-parametric test
- generally included in fundamental Statistics textbooks
- require a match-paired sample
- generate a frequency table
- compare observed (actual) frequencies with expected (if independent) frequencies
- Chi-square test

(c) Pongsa Pornchaiwiseskul,

25

## Tests for Mean-independence

- Analysis of Variance (ANOVA)

assume **no** functional form of the conditional mean

- Regression Analysis

assume a functional form of the conditional mean

(c) Pongsa Pornchaiwiseskul,

27

## Test for Zero Correlation

$$H_0 : \rho_{XY} = 0$$

$$H_1 : \rho_{XY} \neq 0$$

Theorem

$$t_{cal} = \frac{r_{XY}}{\sqrt{\frac{1-r_{XY}^2}{n-2}}} \sim t(n-2)$$

Perform a Two-sided test.

(c) Pongsa Pornchaiwiseskul,

26

## One-way ANOVA (1)

- Question: Is Y mean-independent from X?
- assume no functional form of the conditional mean of Y given X
- ideal for X with discrete values
- group the Y by the values of independent variable X
- test for the variation between groups

(c) Pongsa Pornchaiwiseskul,

28

## One-way ANOVA (2)

### Assumptions

- Variance-independence or  $\sigma_{Y|X}^2 = \sigma^2$   
where  $\sigma^2$  is a positive parameter.
- Given  $X=x_i$ ,  $Y|x_i \sim N(\mu_i, \sigma^2)$

## One-way ANOVA (4)

### Define

$\bar{Y}_i$  = sample conditional mean of Y given  $X=x_i$   
= estimator of  $\mu_i$

Calculated from a sub-sample of size  $n_i$

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

## One-way ANOVA (3)

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_m$$

$$H_1 : \mu_1 \neq \mu_2 \neq \dots \neq \mu_m$$

If  $H_0$  is true,  $\mu_1 = \mu_2 = \dots = \mu_m = \mu_0$

where  $\mu_0$  is the common mean.

## One-way ANOVA (5)

$\hat{\sigma}^2$  = within-group variation or pooled variance of Y

= estimator of  $\sigma^2$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{\sum_{i=1}^m (n_i - 1)}$$

## One-way ANOVA (6)

$\bar{Y}$  = sample unconditional mean of Y

= estimator of  $\mu_0$  if  $H_0$  is true

Calculated from all the sub-samples with

total sample size =  $n_1 + n_2 + \dots + n_m$

$$\bar{Y} = \left( \sum_{i=1}^m \sum_{j=1}^{n_i} Y_{ij} \right) / \left( \sum_{i=1}^m n_i \right)$$

(c) Pongsa Pornchaiwiseskul,

33

## One-way ANOVA (8)

Theorem

$$(n-m) \frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-m) \text{ where } n = \sum_{i=1}^m n_i$$

$$(m-1) \frac{\widehat{\widehat{\sigma^2}}}{\sigma^2} \sim \chi^2(m-1)$$

(c) Pongsa Pornchaiwiseskul,

35

## One-way ANOVA (7)

$\widehat{\widehat{\sigma^2}}$  = between-group variation

= also estimator of  $\sigma^2$  if  $H_0$  is true

$$\widehat{\widehat{\sigma^2}} = \frac{\sum_{i=1}^m n_i (\bar{Y}_i - \bar{Y})^2}{m-1}$$

(c) Pongsa Pornchaiwiseskul,

34

## One-way ANOVA (9)

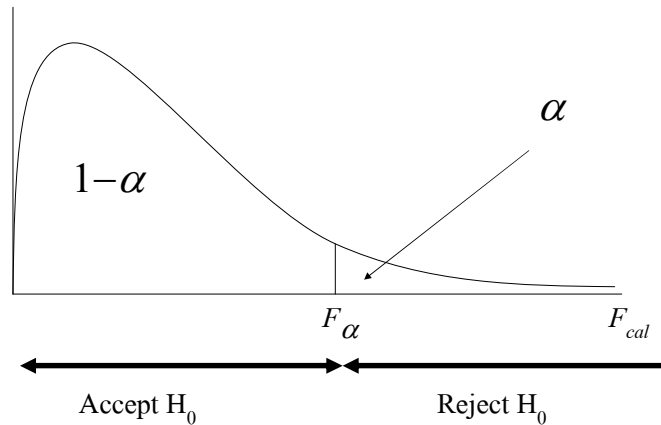
$$F_{cal} = \left( \frac{(m-1) \frac{\widehat{\widehat{\sigma^2}}}{\sigma^2}}{m-1} \right) / \left( \frac{(n-m) \frac{\widehat{\sigma^2}}{\sigma^2}}{n-m} \right)$$

$$= \frac{\widehat{\widehat{\sigma^2}}}{\widehat{\sigma^2}} \sim F(m-1, n-m)$$

(c) Pongsa Pornchaiwiseskul,

36

## One-way ANOVA (10)



(c) Pongsa Pornchaiwiseskul,

37

## Regression Analysis (1)

- assume a functional form of the conditional mean up to its parameters
- parameters are unknown
- estimate parameters
- test the parameters

(c) Pongsa Pornchaiwiseskul,

39

## One-way ANOVA (11)

Mean-independence if  $H_0$  has  
been accepted

Do an ANOVA exercise in Excel

(c) Pongsa Pornchaiwiseskul,

38

## Regression Analysis (2)

Focus on the mean-dependence and  
variance-dependence of Y on X

In general,

$$\mu_{Y|X} = m(x) \quad \text{a function of } x$$

$$\sigma_{Y|X}^2 = v(x) \quad \text{a function of } x$$

(c) Pongsa Pornchaiwiseskul,

40

# Simple Linear Regression

## Assumptions

- 1)  $\mu_{Y|X} = \beta_1 + \beta_2 x$  a linear function of  $x$
  - 2)  $\sigma^2_{Y|X} = \sigma^2$  variance-independent
  - 3)  $Y|X \sim N(\beta_1 + \beta_2 X, \sigma^2)$
- $\beta_1, \beta_2$  and  $\sigma^2$  are unknown parameters

# Simple Linear Regression Model (1)

The model based upon the assumptions

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i$$

where  $i$  = index of the observation

$\epsilon_i$  = identical and independent normal error term

$$\epsilon_i \sim N(0, \sigma^2) \text{ for all } i=1, \dots, n$$

# Simple Linear Regression Model (2)

$X_i$  is given or non-random but  $Y_i$  or  $\epsilon_i$  is randomly sampled.

It is also required that  $\sum_{i=1}^n (X_i - \bar{X})^2 > 0$

Why? Will see.

# Estimator of $\beta_1$ and $\beta_2$

## Ordinary Least Square(OLS) Method

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

The “hat” about the parameter symbol denotes the estimator for the parameter.

## Estimator of $\sigma^2$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \{Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)\}^2}{n-2}$$

Why  $n-2$ ?

## Properties of OLS Estimators (2)

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \sigma^2 \left( -\frac{\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

## Properties of OLS Estimators (1)

$$E(\hat{\beta}_2) = \beta_2 \quad \text{and} \quad V(\hat{\beta}_2) = \sigma^2 \left( \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

$$E(\hat{\beta}_1) = \beta_1 \quad \text{and} \quad V(\hat{\beta}_1) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

## Properties of OLS Estimators (3)

$$\hat{\beta}_2 \sim N \left( \beta_2, \sigma^2 \left( \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \right)$$

$$\hat{\beta}_1 \sim N \left( \beta_1, \sigma^2 \left( \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \right)$$

## Properties of OLS Estimators (4)

$$(n-2) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

$$E(\hat{\sigma}^2) = \sigma^2 \quad \text{and} \quad V(\hat{\sigma}^2) = \frac{2\sigma^4}{n-2}$$

(c) Pongsa Pornchaiwiseskul,

49

## Statistical Inference (1)

(1- $\alpha$ )100% Confidence Interval for  $\beta_2$

$$= \hat{\beta}_2 \pm t_{\frac{\alpha}{2}} se(\hat{\beta}_2)$$

where  $se(\hat{\beta}_2) = \sqrt{\hat{\sigma}^2 \left( \frac{1}{\sum (X_i - \bar{X})^2} \right)}$

(c) Pongsa Pornchaiwiseskul,

51

## Properties of OLS Estimators (5)

$$t_{cal} = \frac{\frac{\hat{\beta}_2 - \beta_2}{\sqrt{\sigma^2 \frac{1}{\sum (X_i - \bar{X})^2}}}}{\sqrt{\frac{(n-2) \frac{\hat{\sigma}^2}{\sigma^2}}{n-2}}} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\hat{\sigma}^2 \frac{1}{\sum (X_i - \bar{X})^2}}} \sim t(n-2)$$

(c) Pongsa Pornchaiwiseskul,

50

## Statistical Inference (2)

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

Perform a two-tailed  $t$ -test

$$t_{cal} = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} \sim t(n-2)$$

(c) Pongsa Pornchaiwiseskul,

52

## Statistical Inference (3)

Y has Mean-independence from X at significant level of  $\alpha$  if the  $(1-\alpha)100\%$  Confidence Interval for  $\beta_2$  covers zero or  $H_0 : \beta_2=0$  has been accepted at significant level  $\alpha$

## Statistical Inference (4)

$$H_0 : \beta_2 = 0.6$$

$$H_1 : \beta_2 \neq 0.6$$

Perform a two-tailed  $t$ -test

$$t_{cal} = \frac{\hat{\beta}_2 - 0.6}{se(\hat{\beta}_2)} \sim t(n-2)$$

## Statistical Inference (5)

$$H_0 : \sigma^2 = 4$$

$$H_1 : \sigma^2 > 4$$

Perform a one-tailed Chi-square-test

$$\chi_{cal}^2 = (n-2) \frac{\hat{\sigma}^2}{4} \sim \chi^2(n-2)$$

## Central Limit Theorem for $\hat{\beta}_1, \hat{\beta}_2$

- Violation of normal distribution of  $\mathcal{E}$  but still  $E(\mathcal{E}) = 0$  and  $V(\mathcal{E}) = \sigma^2$  and still  $\mathcal{E}_i$  is i.i.d.
- $\hat{\beta}_1$  and  $\hat{\beta}_2$  are approximately normal when  $n \rightarrow \infty$
- CI and hypothesis testing for  $\beta_1, \beta_2$  and  $\sigma^2$  are still acceptable if the sample is large

## Gauss-Markov Theorem (1)

Given that  $X$  is non-random,

OLS estimator  $(\hat{\beta}_1, \hat{\beta}_2)$  is BLUE

Best

Linear

Unbiased

Estimator

Note that Gauss-Markov Theorem does not requires normality assumption

## Gauss-Markov Theorem (3)

OLS estimator is unbiased

OLS estimators has the smallest variance

among linear estimators. There could be a non-linear estimator that is more efficient.

## Gauss-Markov Theorem (2)

OLS estimator is a linear estimator. E.g.,

$$\hat{\beta}_2 = \sum_{i=1}^n \frac{1}{c} (X_i - \bar{X})(Y_i - \bar{Y}) \text{ where } c = \sum_{j=1}^n (X_j - \bar{X})^2$$

$$= \sum_{i=1}^n \underbrace{\frac{1}{c} (X_i - \bar{X})}_{k_i} Y_i - \sum_{i=1}^n \frac{1}{c} (X_i - \bar{X}) \bar{Y} = \sum_{i=1}^n k_i Y_i$$

Is linear in  $Y$  (the random component). Note that  $k_i$  is non-random as it is in terms of  $X$  only.

## OLS with Random $X_i$ 's (1)

Note that non-randomness of  $X$  is assumed for all the theorems.

In general,  $X_i$ 's are random as they are also observed or recorded as they happen. They are not given as assumed.

What is their effect on the validity of OLS estimators?

## OLS with Random $X_i$ 's (2)

Fortunately, OLS estimators are still “valid” for most but not all random distributions of  $X$ . In general, OLS estimators will not be unbiased anymore. The best they can be is consistent estimators. Gauss-Markov Theorem does not apply in this case.

**Just be aware.**

## Bi-variate Normal (1)

Given that  $R$  is BVN,  $R \sim BVN(\mu, \Sigma)$

where

$$R \equiv \begin{pmatrix} X \\ Y \end{pmatrix}, \mu \equiv \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \Sigma \equiv \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{pmatrix}$$

Random Vector      Mean Vector      Variance-covariance Matrix

Note that a Bi-variate Normal(BVN) is a form of joint probability distribution of 2 RV's

## Bi-variate Normal (2)

Joint pdf of  $R$

$$f(x, y) = \frac{1}{2\pi} \left| \Sigma \right|^{-\frac{1}{2}} e^{-\frac{1}{2} r^T \Sigma^{-1} r}$$

$$-\infty < x < \infty, -\infty < y < \infty$$

where  $r^T = [x \ y]$

## Bi-variate Normal (3)

It can be proved that

- 1) conditional pdf follows the normal distribution
- 2) conditional mean is a linear function

$$E(Y | X) = \left( \mu_Y - \frac{\sigma_{XY}}{\sigma_{XX}} \mu_X \right) + \left( \frac{\sigma_{XY}}{\sigma_{XX}} \right) X$$

$$E(X | Y) = \left( \mu_X - \frac{\sigma_{XY}}{\sigma_{YY}} \mu_Y \right) + \left( \frac{\sigma_{XY}}{\sigma_{YY}} \right) Y$$

# Bi-variate Normal (4)

3) variance-independence

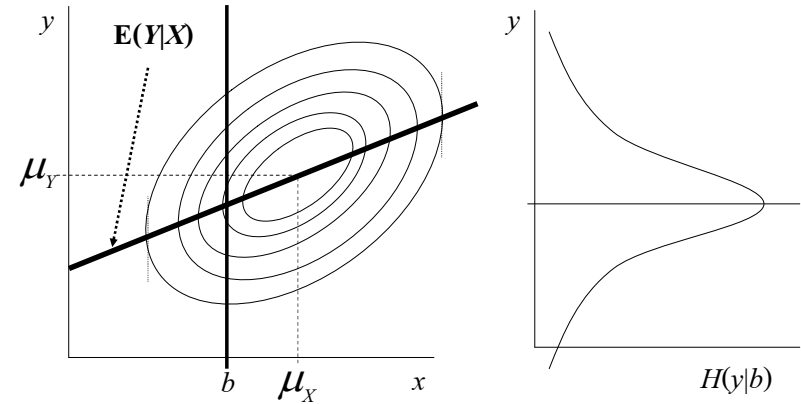
$$V(Y | X) = \sigma_{YY} - \left( \frac{\sigma_{XY}}{\sigma_{XX}} \right)^2 \sigma_{XX}$$

$$V(X | Y) = \sigma_{XX} - \left( \frac{\sigma_{XY}}{\sigma_{YY}} \right)^2 \sigma_{YY}$$

(c) Pongsa Pornchaiwiseskul,

65

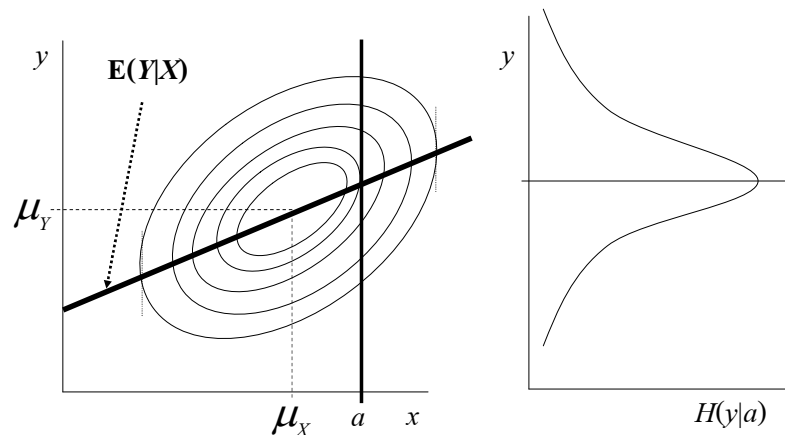
# Bi-variate Normal (6)



(c) Pongsa Pornchaiwiseskul,

67

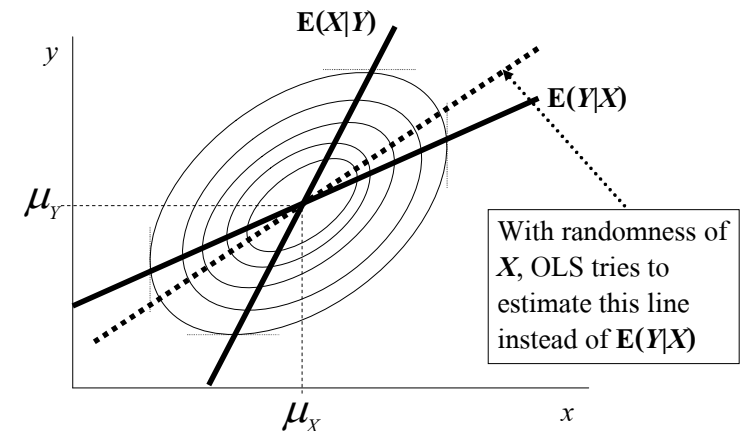
# Bi-variate Normal (5)



(c) Pongsa Pornchaiwiseskul,

66

# Bi-variate Normal (7)



(c) Pongsa Pornchaiwiseskul,

68

## OLS Estimator $\hat{\beta}_1, \hat{\beta}_2$ is BVN

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim BVN \left( \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} & -\frac{\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ -\frac{\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} & \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{pmatrix} \right)$$