Bi-variate Statistics

Focus on continuous RV's

X ~ cont. random variable

Y ~ cont. random variable

Joint Probability Distribution

- Joint CDF

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

- Joint pdf

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

Interdependence between RV's

Three Levels of Interdependence

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

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)

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

Conditional Probability

Conditional pdf of X given Y=y

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

Conditional pdf of Y given X=x

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

Uncond. Mean & Variance

Unconditional mean of Y

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

Unconditional variance of Y

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

Conditional Mean & Variance

Conditional mean of Y given X=x

$$\mu_{Y|X} = \int_{-\infty}^{\infty} yH(y|x)dy$$

in term of x
but not y

Conditional variance of Y given X=x

$$\sigma_{Y|X}^2 = \int_{-\infty}^{\infty} (y - \mu_{Y|X})^2 H(y|x)dy$$

in term of x
but not y

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

Mean-independence (2)

Note that

$\boxed{\text{Y is mean-independent from X}}$ $\not\Longleftrightarrow$ $\boxed{\text{X is mean-independent from Y}}$

Direction matters.

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

Variance-independence (2)

Note that

Y is variance-independent from X $\not\Longleftrightarrow$ X is variance-independent from Y

Direction matters.

Population Covariance (1)

Definition

$$\sigma_{XY} = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$= \iint (x-\mu_X)(y-\mu_Y) f(x,y)\,dxdy$$

a constant

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

Population Covariance (3)

Magnitude of Covarinace

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

σ_{XY} is in Baht-Kilogram

Population Correlation (1)

- Definition

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

- Sign of Correlation

  – same as that of Covariance

Population Correlation (2)

Magnitude of Correlation

always bounded between -1 and 1

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

Unit of Correlation

no unit

comparable between populations

Population Correlation (3)

Interpretation of Correlation

ρ_{XY} =+1 ==> If a variable is above or below its mean, the other will be above or below its own mean with certainty

ρ_{XY} =-1 ==> If a variable is above or below its mean, the other will be below or above its own mean with certainty

ρ_{XY} = 0 ==> If a variable is deviated from its mean, the other will be expected at its mean

Sample Covariance

s_{XY} is an estimator for σ_{XY}

Required paired sample

Estimator

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

Paired Sample of Size *n*

i	X_i	Y_i
1	X_1	Y_1
2	X_2	Y_2
:	:	:
:	:	:
n	X_n	Y_n

Sample Correlation (1)

r_{XY} is an estimator of ρ_{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

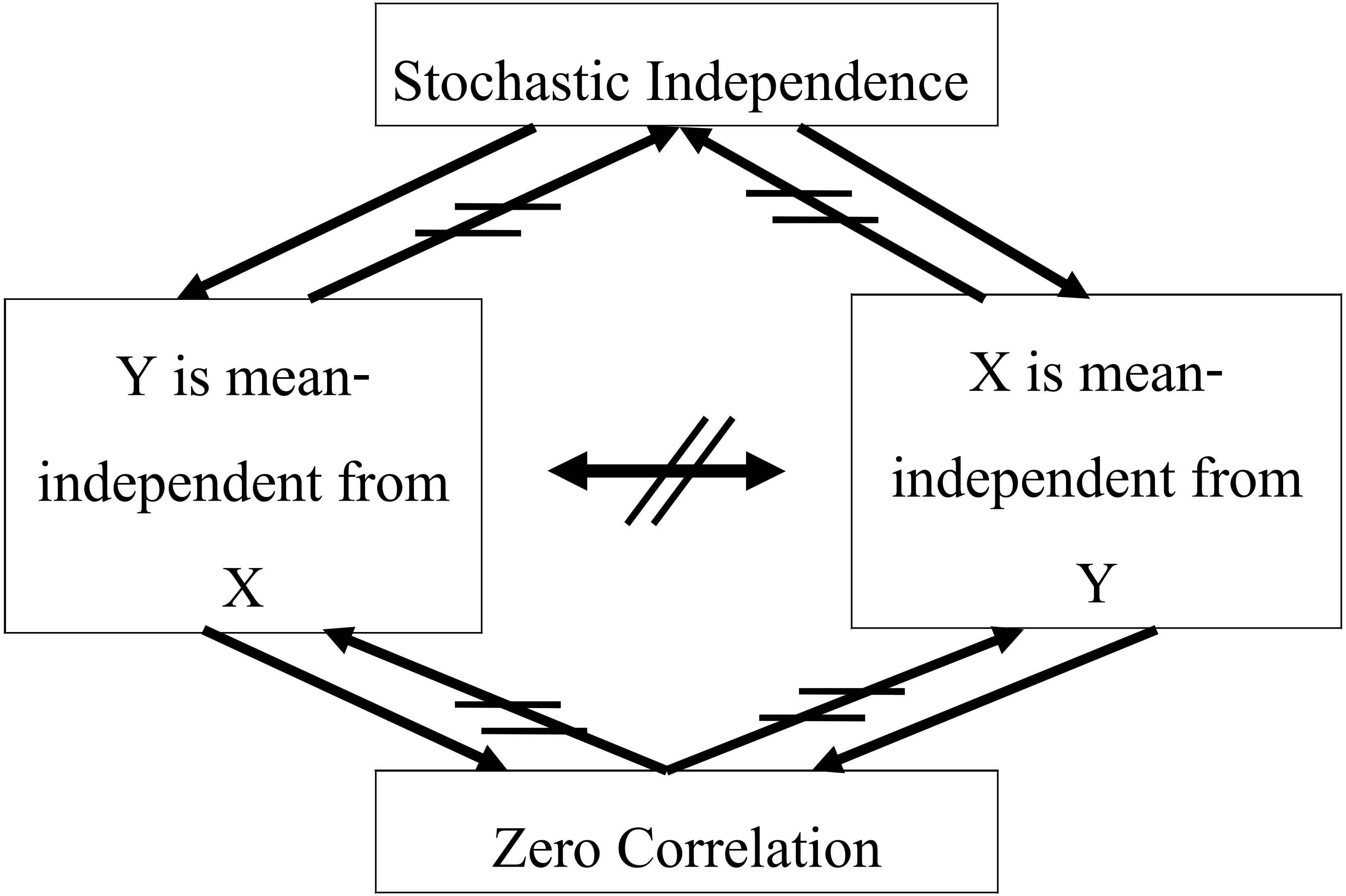
$$-1 \leq r_{XY} \leq +1$$

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

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.

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

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

One-way ANOVA (2)

Assumptions

- Variance-independence or $\sigma^2_{Y|X} = \sigma^2$

  where σ^2 is a positive parameter.

- Given $X=x_i$, $Y|x_i \sim N(\mu_i, \sigma^2)$

One-way ANOVA (3)

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

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

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

where μ_0 is the common mean.

One-way ANOVA (4)

Define

$\overline{Y}_i$ = sample conditional mean of Y given X=x_i

= estimator of μ_i

Calculated from a sub-sample of size n_i

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

One-way ANOVA (5)

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

= estimator of σ^2

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_i\right)^2}{\sum_{i=1}^{m}(n_i - 1)}$$

One-way ANOVA (6)

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

= estimator of μ_0 if H_0 is true

Calculated from all the sub-samples with

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

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

One-way ANOVA (7)

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

= also estimator of σ^2 if H_0 is true

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

One-way ANOVA (8)

Theorem

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

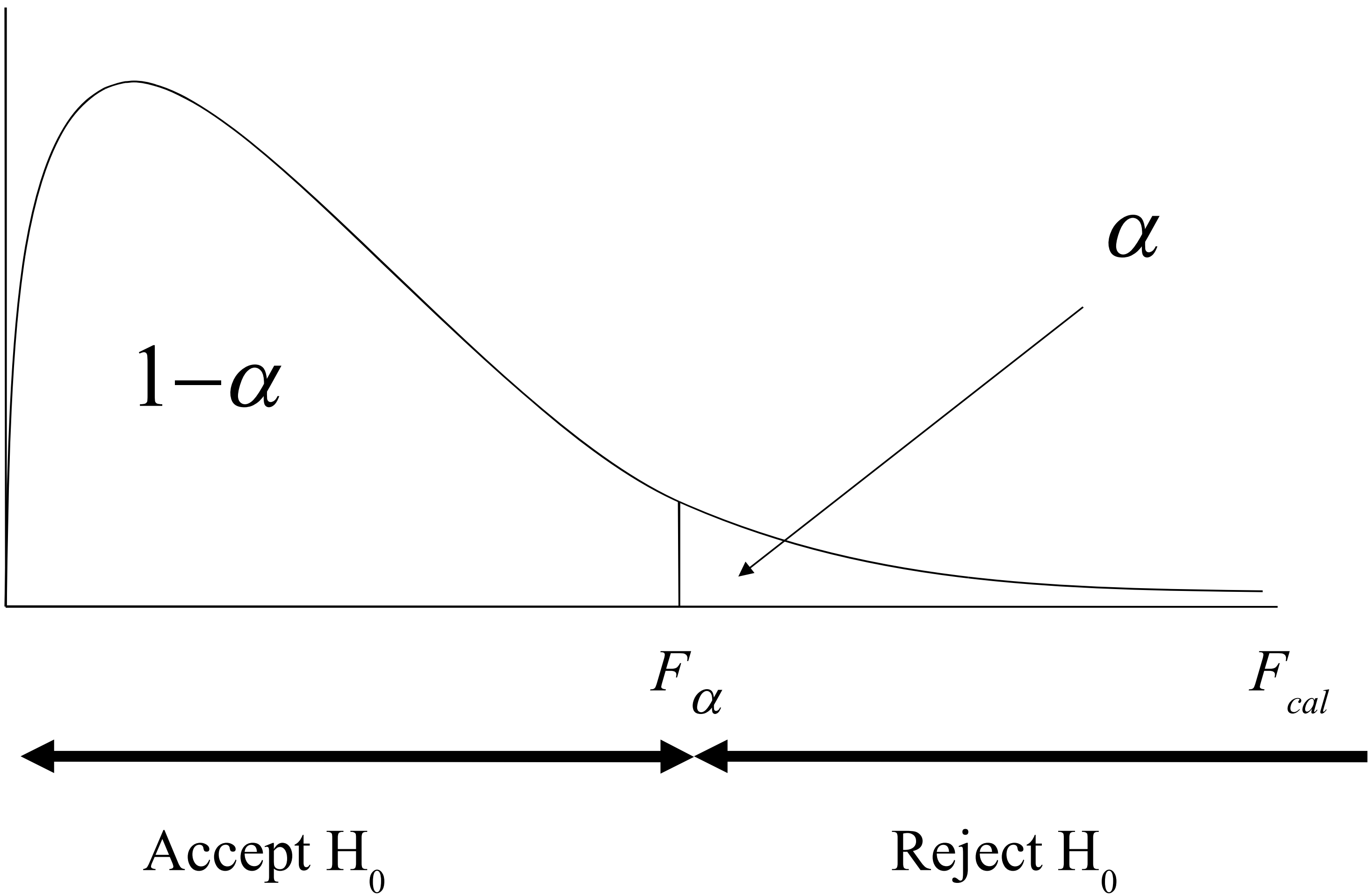
$$(m-1)\frac{\widehat{\widehat{\sigma^2}}}{\sigma^2} \sim \chi^2(m-1)$$

One-way ANOVA (9)

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

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

One-way ANOVA (10)



One-way ANOVA (11)

Mean-independence if H_0 has been accepted

Do an ANOVA exercise in Excel

Regression Analysis (1)

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

Regression Analysis (2)

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

In general,

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

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

Simple Linear Regression

<u>Assumptions</u>

1) $\mu_{Y|X} = \beta_1 + \beta_2 x$ a linear function of x

2) $\sigma^2_{Y|X} = \sigma^2$ variance-independent

3) $Y|X \sim N(\beta_1 + \beta_2 X, \sigma^2)$

β_1, β_2 and σ^2 are unknown parameters

Simple Linear Regression Model (1)

The model based upon the assumptions

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

where i = index of the observation

ε_i = identical and independent normal error term

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

Simple Linear Regression Model (2)

X_i is given or non-random but Y_i or ε_i is randomly sampled.

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

Why? Will see.

Estimator of β_1 and β_2

Ordinary Least Square(OLS) Method

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

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

The "hat" about the parameter symbol denotes the estimator for the parameter.

Estimator of σ^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 (1)

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

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

Properties of OLS Estimators (2)

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

Properties of OLS Estimators (3)

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

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

Properties of OLS Estimators (4)

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

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

Properties of OLS Estimators (5)

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

Statistical Inference (1)

(1-α)100% Confidence Interval for β_2

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

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

Statistical Inference (2)

$$\mathrm{H}_0 : \beta_2 = 0$$
$$\mathrm{H}_1 : \beta_2 \neq 0$$

Perform a two-tailed *t*-test

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

Statistical Inference (3)

Y has Mean-independence from X at significant level of α if the $(1-\alpha)100\%$ Confidence Interval for β_2 covers zero or $H_0 : \beta_2=0$ has been accepted at significant level α

Statistical Inference (4)

$$\mathrm{H}_0 : \beta_2 = 0.6$$

$$\mathrm{H}_1 : \beta_2 \neq 0.6$$

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)

$$\mathrm{H}_0 : \sigma^2 = 4$$

$$\mathrm{H}_1 : \sigma^2 > 4$$

Perform a one-tailed Chi-square-test

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

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

- Violation of normal distribution of ε
  but still E(ε) = 0 and V(ε) = σ^2
  and still ε_i is i.i.d.
- $\hat{\beta}_1$ and $\hat{\beta}_2$ are approximately normal when $n \to \infty$
- CI and hypothesis testing for β_1, β_2 and σ^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 (2)

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

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

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

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

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.