Lecture 9

Statistical Data Analysis and Trend Analysis of Environmental Radiation Monitoring

Follow-up Training Course 2025

Environmental Radioactivity Monitoring

16th October 2025

OHKURA Takehisa
Japan Atomic Energy Agency, JAEA

Outline

1. Sampling and Estimate

2. Statistic analysis 1: Summary of Regression Analysis (Singe Regression Analysis)

3. Statistic analysis 2: Summary of Time Series Analysis

0. Introduction

Introduction

➢ *Generic Objectives of Environmental Monitoring*

1. to assess individual radiation doses of residents surrounding nuclear facilities, either in emergency or in normal situation.
2. to monitor radiation doses that are below the annual dose limits set by the national regulations and guidelines.
3. to obtain accurate information of radiation sources in order to take appropriate protective measures and environment monitoring programmes.

Dose limits for the public by law (in Japan)

Dose Type	Dose Limit
Effective Dose	1 mSv / year
Equivalent Dose	50 mSv / year (skin)
	15 mSv / year (crystalline lens)

ICRP Publication 43 ~Principles of Monitoring for the Radiation Protection of the Population~

➢ *Objectives of Routine Environmental Monitoring in Japan*

~ Under Operational Situation ~

1. to estimate and evaluate radiation doses of local residents.
2. to collect information of radioactive materials accumulated in the environment.
3. to detect the unanticipated release of radioactive materials and radiation from nuclear facilities at the early stage, and to evaluate their possible impacts on the environment.
4. to prepare the system of environmental radiation monitoring for abnormal and emergency situations.

"Guidelines for Environmental Radiation Monitoring" by Japan Nuclear Safety Commission
"Guidelines for Nuclear Emergency Countermeasures – Supplemental material for Normal Environmental Monitoring -" by Japan Nuclear Regulation Authority

◆ Environmental monitoring is aimed at obtaining measurement values.

● Observation, Measurement

☐ Merit : This is true!!

- If ensure that suitable method and techniques (analysis, measurement, calibration, routine inspection, etc.)
- Measurement value: uncertainty is associated, evaluation on the basis of unified strategy

☐ Demerit : There is no spatial or temporal continuity, and spatial or temporal variability, measurement values are not necessarily representative.

◆ Environmental monitoring is aimed at obtaining representative values.

➢ Representativeness means that the sample should reflect the conditions in the environment from which it is taken.

Environmental and Source Monitoring for Purposes of Radiation Protection, IAEA Safety Standards, Safety Guide, No. RS-G-1.8

- Estimate by a numerical simulation model

  Can complement spatial or temporal gap which observation/measurement cannot cover.

- Suitable sampling techniques/strategies

  Enable us to estimate representative values from samples.

Aims of environmental monitoring

→ Aims of statistical analysis of environmental data

Estimate representative value of obtained data by statistical method

➢ *Evaluation of Effect of facility-induced radiation*

- Points to keep in mind when evaluating from environmental radiation measurement results

  - ✓ Environmental Factors Affecting Radiation Monitoring

  - ✓ Distinguish between “facility-induced radiation” and “natural radiation, nuclear test-derived, past controlled releases from facilities, accidental releases, etc”.

  - ✓ Understanding the range of normal fluctuations and trends in fluctuations

➢ *Detecting abnormal values and normal range of variation*

- One objective of environmental monitoring

  → **detecting the unanticipated situation at the early stage**

Criteria is necessary in order to determine whether measurement value is abnormal value of within normal range.

"Variation range on normal situation"

"Guidelines for Nuclear Emergency Countermeasures – Supplemental material for Normal Environmental Monitoring -" by Japan Nuclear Regulation Authority

“Variation range on normal situation”

○Establishing “**Upper-level value** of variation range on normal situation”

→ In case of excessing **the value** → abnormal value

Criteria 1

Mean ± (3 × standard deviation[σ]) for measurement values in the past several years

→ “**Variation range on normal situation**”

✓ Indispensable condition is that statistical distribution of measurement values is depending on (or regarded as) “normal distribution or log-normal distribution”.

Criteria 2

Maximum vales for measurement values in the past several years of since the measurement has started

is set as **upper-level value**

“Guidelines for Nuclear Emergency Countermeasures – Supplemental material for Normal Environmental Monitoring -” by Japan Nuclear Regulation Authority

➢ *Evaluation of Effect of facility-induced radiation*

- Points to keep in mind when evaluating from environmental radiation measurement results

  - ✓ Environmental Factors Affecting Radiation Monitoring

  - ✓ Distinguish between "facility-induced radiation" and "natural radiation, nuclear test-derived, past controlled releases from facilities, accidental releases, etc".

  - ✓ Understanding the range of normal fluctuations and trends in fluctuations

It is useful to understand "baseline trend", "cycle" and associated irregular variance to the trend.

Outline

1. Sampling and Estimate

2. Statistic analysis 1: Summary of Regression Analysis
   (Singe Regression Analysis)

3. Statistic analysis 2: Summary of Time Series Analysis

1.1 Sampling techniques

Sampling techniques

In general, sampling techniques are divided in five, as the followings

- Simple random sampling
- Cluster sampling
- Stratified sampling
- Multi-stage sampling
- Systematic sampling

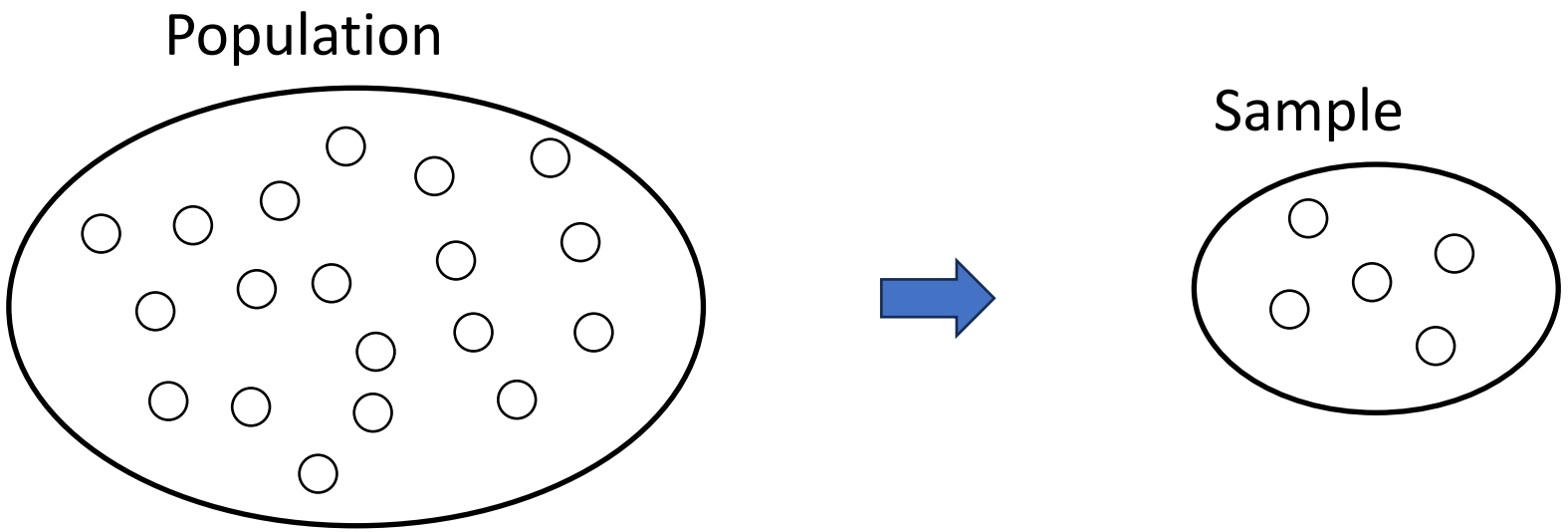
Table 1.1-1 Sampling Techniques For Environmental Monitoring

Sampling technique	Description	Comment
Judgemental sampling	Sample is taken on the judgement of the sampling person	Increased probability of biased sampling; representativeness cannot be quantified; accuracy cannot be quantified
Simple random sampling	Any sample has the same probability of being included	Provides representativeness; problems may arise with inhomogeneous terrain
Stratified sampling	The sample in its entirety is divided into parts that are known to be more homogeneous; simple random sampling is then applied to the remaining subdivisions	Requires knowledge of the inhomogeneity of the entire sample; may lead to bias if the fractions of the samples are not properly estimated
Systematic sampling	Starting from a randomly selected point, sampling follows a strict predefined sampling grid	In comparison with random sampling, easier to implement in practice; spatial contamination patterns may be overlooked

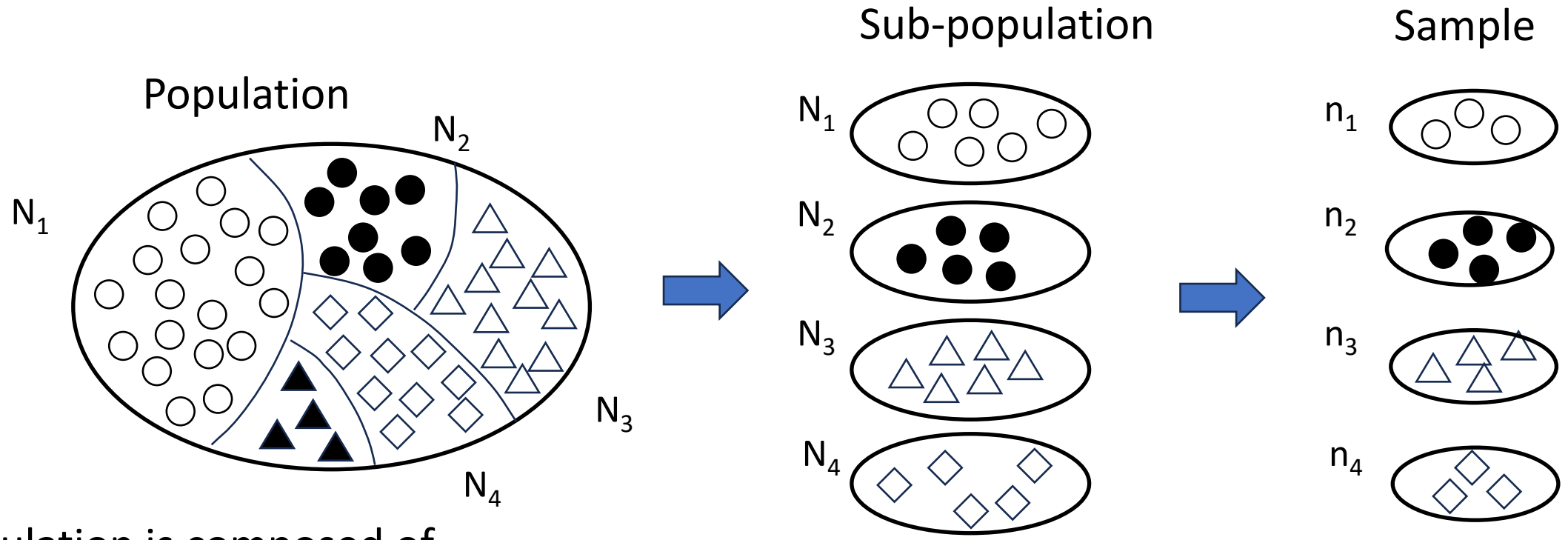
Cited in TABLE 5. SAMPLING TECHNIQUES FOR ENVIRONMENTAL MONITORING, IAEA, No. RS-G-1.8

Environmental and Source Monitoring for Purposes of Radiation Protection, IAEA Safety Standards, Safety Guide, No. RS-G-1.8

Simple random sampling



Stratified sampling



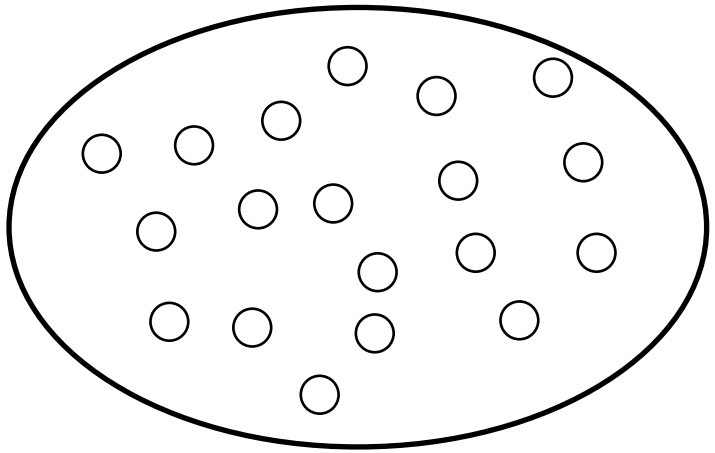
The population is composed of strata that are heterogeneous to each other.

N, n are sample size.

Sub-populations are set so that variation among each sub-population is as large as possible and the variation within each sub-population is as small as possible.

Systematic sampling

Population



Numbering the population

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ・・・

Sample

× ○ × × ○ × × ○ × × ○ × × ○ ×

Sampling at regular interval

This case, starting with No. 2 and sampling at interval 3.

→ Sampling No. (2 + 3(k-1)). (k=1, 2,3 4, ・ ・ ・)

1. 2 Estimation

◆Point estimation and interval estimation

- **Point estimation**

  A method to estimate population parameters such as population mean, population variance with a single value

- **Interval estimation**

  A method to estimate population parameters, such as population mean, population variance with some range of values

Point estimation

◆ Estimator

Quantity estimated as the population (parameter, not measurable in reality) of a probability distribution based on actual sample data measured in reality

The following equation, which calculate the mean of the value$x_i (i = 1, 2, 3, \cdots)$ obtained by n trials, used in the point estimate of the population mean, is itself an "estimator".

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

◆ Consistency

- When we estimate the parameters of a population, such as the population mean or the population variance, based on an estimator, the estimate must be accurate.
- The law of large numbers was that as the sample size n increases, the sample mean approaches the population mean.
- This property, in which the estimator gradually approaches the true parameters as n increases, is called "**consistency**".

◆Unbiasedness

- When we guess the parameters of a population, such as the population mean or the population variance, based on an estimator, it is meaningless if the estimation deviates significantly from the true parameters.

- In other words, the expected value of the estimator must meet the parameter.
- This property is called **unbiasedness**.

- Expressing unbiasedness in terms of the estimator $\widehat{\theta}$ and the true parameter θ the following equation is obtained.

$$E(\hat{\theta}) = \theta$$

- It indicates that "the expected value of $\hat{\theta}$ is θ regardless of the value of n."
- In other words, the outliers of the estimator are not biased (the outliers are the same above and below) when n is small or large.

◆ Unbiased estimator

- Unbiasedness means that the calculated estimates are, on average, free of bias, independent of sample size, and estimators that satisfy this property are called "unbiased estimators”.

- When the expected value of an estimator measured from a sample is equal to that of the population, the estimator is called an unbiased estimator.

  - For example, the sample mean is an unbiased estimator because the expected value of the sample mean is equal to the population mean.

◆ Unbiased estimator of mean

As sample mean is $\bar{x}$,

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} \times n\mu = \mu$$

then, since $E(\bar{x}) = \mu$, unbiased estimator of population mean $\hat{\mu}$ is sample mean $\bar{x}$.

$$\hat{\mu} = \bar{x}$$

and,

$$V(\bar{x}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} V(x_i) = \frac{\sigma^2}{n}$$

the above property is also valid.

◆ Unbiased estimator of variance

- Since the unbiased estimator of the mean was the sample mean, it seems that the unbiased estimator of the variance could also be expressed as the following variance obtained from the sample mean,

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

In fact, it doesn't.

- The unbiased estimator of the population variance $\hat{\sigma}^2$ is expressed by the following equation.

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\hat{\sigma}^2$ that satisfies this equation is called **unbiased variance**.

- This is a very important concept because the sample unbiased variance is the estimate of the population variance used in the analysis when the population variance is unknown.

Appendix

To confirm $E(s^2) \neq \sigma^2$

Let's calculate $E(s^2)$ in reality

$$E(s^2) = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= \frac{1}{n}E\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= \frac{1}{n}E\left(\sum_{i=1}^{n}[(x_i - \mu) - (\bar{x} - \mu)]^2\right)$$

$$= \frac{1}{n}E\left(\sum_{i=1}^{n}(x_i - \mu)^2 - 2\sum_{i=1}^{n}\big((x_i - \mu)\big)(\bar{x} - \mu) + \sum_{i=1}^{n}(\bar{x} - \mu)^2\right)$$

$$= \frac{1}{n}E\left(\sum_{i=1}^{n}(x_i - \mu)^2 - 2n(\bar{x} - \mu)^2 + n(\bar{x} - \mu)^2\right)$$

$$= \frac{1}{n}E\left(\sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}E[(x_i - \mu)^2] - E[(\bar{x} - \mu)^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}V(x_i) - V(\bar{x})$$

$$= \sigma^2 - \frac{\sigma^2}{n}$$

$$= \frac{n-1}{n}\sigma^2$$

Then, due to $E(s^2) \neq \sigma^2$, It can be confirmed that the variance obtained from the sample mean is not an unbiased estimator of the population variance σ^2.

Appendix

Because,

$$E(s^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

To obtain the population variance from this equation, the equation is transformed as follows.

$$E\left(\frac{n}{n-1}s^2\right) = \sigma^2$$

This means that the parentheses ( ) on the left-hand side indicate the unbiased estimator $\hat{\sigma}^2$ of the population estimator σ^2.

From this, the unbiased variance u^2 with respect to the population variance is expressed as the following equation,

$$u^2 = \hat{\sigma}^2 = \frac{n}{n-1}s^2$$

$$= \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

From the above, it is verified that unbiased estimator $\hat{\sigma}^2$ of the population variance is expressed as the following equation,

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Appendix

To confirm whether expectation of the unbiased variance is σ^2, Let's calculate $E(\hat{\sigma}^2)$ in reality.

$$E(\hat{\sigma}^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n}[(x_i - \mu) - (\bar{x} - \mu)]^2\right)$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(x_i - \mu)^2 - 2\sum_{i=1}^{n}\big((x_i - \mu)\big)(\bar{x} - \mu) + \sum_{i=1}^{n}(\bar{x} - \mu)^2\right)$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(x_i - \mu)^2 - 2n(\bar{x} - \mu)^2 + n(\bar{x} - \mu)^2\right)$$

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 - n\frac{\sigma^2}{n}\right)$$

$$= \frac{1}{n-1}(n-1)\sigma^2$$

$$= \sigma^2$$

Then, since $E(\hat{\sigma}^2) = \sigma^2$ can be verified, it can be confirmed that the obtained variance is the unbiased estimator of the population variance σ^2.

◆ Experience of unbiased variance

Appendix

It is shown that the unbiased estimator of the population variance $\hat{\sigma}^2$ is expressed

not by $$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

but by the followings,

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Here, it is confirmed that variance divided by (n-1) is unbiased estimator, not by n, experimentally

➢ Experiment 1

Table 1.2-1 is 500-serie data set $\{x_1, x_2, \cdots, x_{10}\}$. Each series consists of 10 numbers produced by uniform random number between $[0, 1]$.

➢ Experiment 2

Table 1.2-2 is 500-series data set $\{x_1, x_2, \cdots, x_{10}\}$. Each series consists of 10 numbers produced by normal random number on $\mathrm{N}(0, 5^2)$.

➢ Experiment 1

Table 1-1 is 500-serie data set $\{x_1, x_2, \cdots, x_{10}\}$. Each series consists of 10 numbers produced by uniform random number between $[0, 1]$.

For each series,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is calculated, and compare with variance of uniform distribution (theoretical value)

variance of uniform distribution $\sigma^2 = {}^1/_{12} = 0.08333$ $(\sigma = 0.28868)$

Appendix

Table 1.2-1　Uniform random number between [0, 1] (10 trials)

n	1	2	3	4	5	6	7	8	9	10	s^2	S^2
Series (1)	0.098025	0.407616	0.097986	0.207633	0.037446	0.188794	0.47152	0.568171	0.650698	0.854852	0.075694	0.068124
Series (2)	0.488397	0.039892	0.658855	0.600485	0.754915	0.394106	0.9169	0.379349	0.706264	0.58258	0.059318	0.053387
Series (3)	0.984103	0.473377	0.972474	0.931397	0.568638	0.325457	0.258968	0.099538	0.207376	0.201434	0.119196	0.107276
・ ・ ・												
Series (500)	0.601781	0.742323	0.145295	0.498098	0.282957	0.128412	0.048815	0.299301	0.90575	0.545944	0.080718	0.072646

Mean of the 500 series of $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is 0.083164527.

Mean of the 500 series of $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is 0.074848074.

According to the above, s^2 is closer to the (theoretical) variance of the uniform distribution (=0.08333).

Experiment 2

Table 1.2-2 is 500-series data set $\{x_1, x_2, \cdots, x_{10}\}$. Each series consists of 10 numbers produced by normal random number on $\mathrm{N}(0, 5^2)$.

For each series,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

is calculated, and compare with variance of normal distribution (=5^2=25)

Appendix

Table 1.2-2 Normal random number (N(0, 5^2)) (10 trials)

n	1	2	3	4	5	6	7	8	9	10	s^2	S^2
Series (1)	-7.27489	-3.91004	-6.61902	-1.56849	-11.0902	-6.60541	-0.2081	0.030154	-4.84266	-5.88393	12.12075	10.90868
Series (2)	8.457031	2.932561	-4.08836	-7.36171	-14.9562	3.699649	4.76861	-5.25169	1.373052	-1.96064	47.67886	42.91097
Series (3)	-4.17528	-2.04682	3.928549	1.821708	-5.8745	-0.32389	-1.55149	-9.39743	0.598407	-0.0438	15.21646	13.69482
・ ・ ・												
Series (500)	-3.06523	7.300844	0.114641	5.964053	4.725016	-1.18933	-12.2817	6.628699	4.82325	3.159174	35.9852	32.38668

Mean of the 500 series of $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is 25.55971

Mean of the 500 series of $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is 23.00374

According to the above, s^2 is closer to the variance of the normal distribution (=5^2=25).

Interval estimation

For example, when the distribution followed by the population can be assumed to be a normal distribution, the method of estimating the parameters, such as the population mean expressed in a certain interval, using values obtained from a sample.

This interval is called "**confidence interval**"

➢ The interval estimation of the population mean is calculated differently when the population variance is known or unknown.

- **When the population variance is known**

  Using the value of the population variance σ^2, calculate a confidence interval using **standard normal distribution**

- **When the population variance is not known**

  Using the value of the unbiased variance s^2, calculate a confidence interval using **t-distribution**

✓ However, since there are few situations in reality where the population mean is not known but only the population variance is known, the t-distribution method is usually used for interval estimation of the population mean.

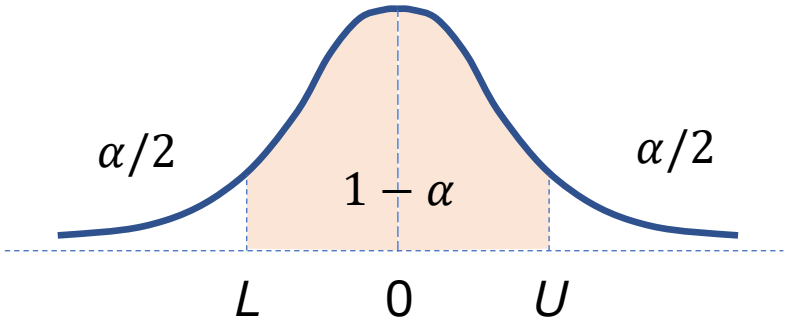
◆ Interval estimation of population mean

Point estimation:θ is estimated as a single value $\hat{\theta}$.

Interval estimation: estimation using the concept of probability for θ

Interval estimation is a method that guarantees that the probability that the true population (parameter) value θ falls into some interval (L, U) is greater than or equal to 1-α, where α is the probability that θ does not fall into the interval.

$$P(L \leq \theta \leq U) \geq 1 - \alpha$$



L: Lower confidence limit

U: Upper confidence limit

1-α: Confidence coefficient

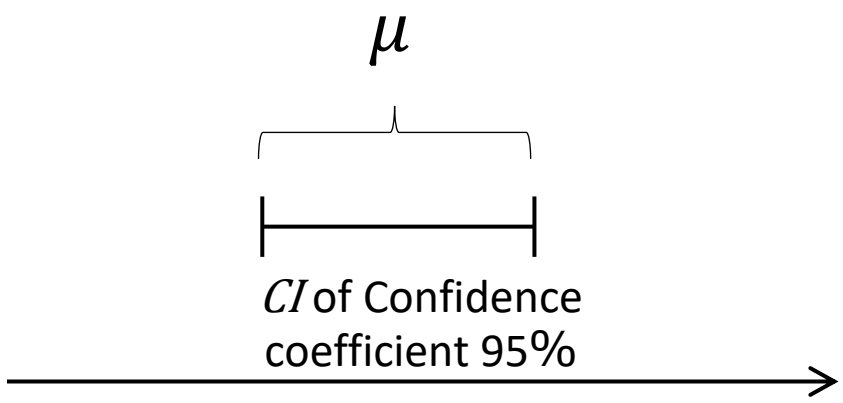
Interval [L, U]: 100%(1-α) confidence interval

- For 1-α, select an appropriate value depending on the purpose. It is usually set to 0.99 or 0.95.
- Confidence intervals are obtained from the sample distribution of $\hat{\theta}$.

NOTE !!

◆ Meaning of Interval Estimation

Wrong



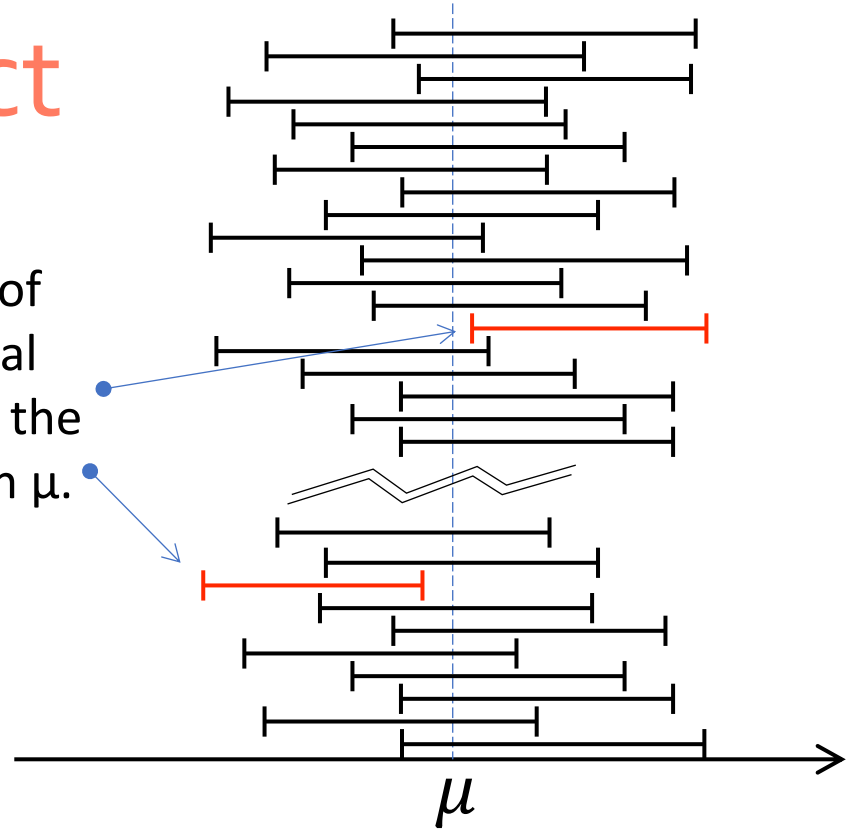
"*CI* of Confidence coefficient 95%" dose **NOT** mean that

probability that the population mean μ exists within the 95% confidence interval is 95%.

CI: Confidence Interval

Correct

Five times out of 100, the interval does not include the population mean μ.



"*CI* of Confidence coefficient 95%" means that when the confidence intervals is calculated, 95% include the population mean μ within that interval.

When 100 times the sample data is used to estimate a confidence interval, 95 of those times include the population mean μ within that interval.

Outline

1. Sampling and Estimate

2. Statistic analysis 1: Summary of Regression Analysis (Singe Regression Analysis)

3. Statistic analysis 2: Summary of Time Series Analysis

◆ Regression analysis

- “Regression” is an expression by an equation for a response variable (=y) using an explanatory variable (=x).

- This equation is called “Regression Equation”.
- To find “Regression Equation” is called “Regression Analysis”.

- In case, single explanatory variable is used,

  ➢ Singe regression analysis

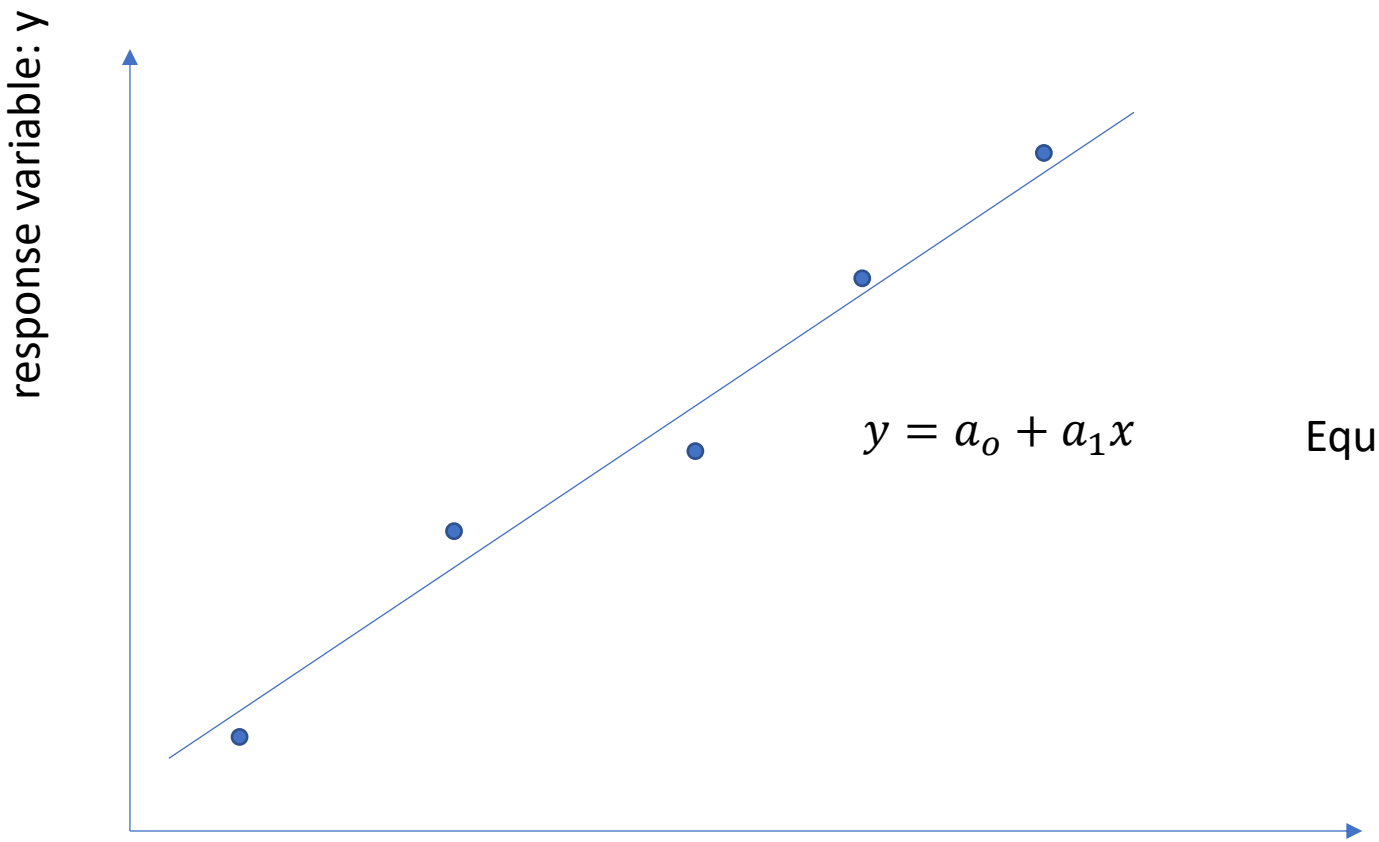
- In case, multiple explanatory variables are used,

  ➢ Multiple regression analysis

2.1 Singe regression analysis

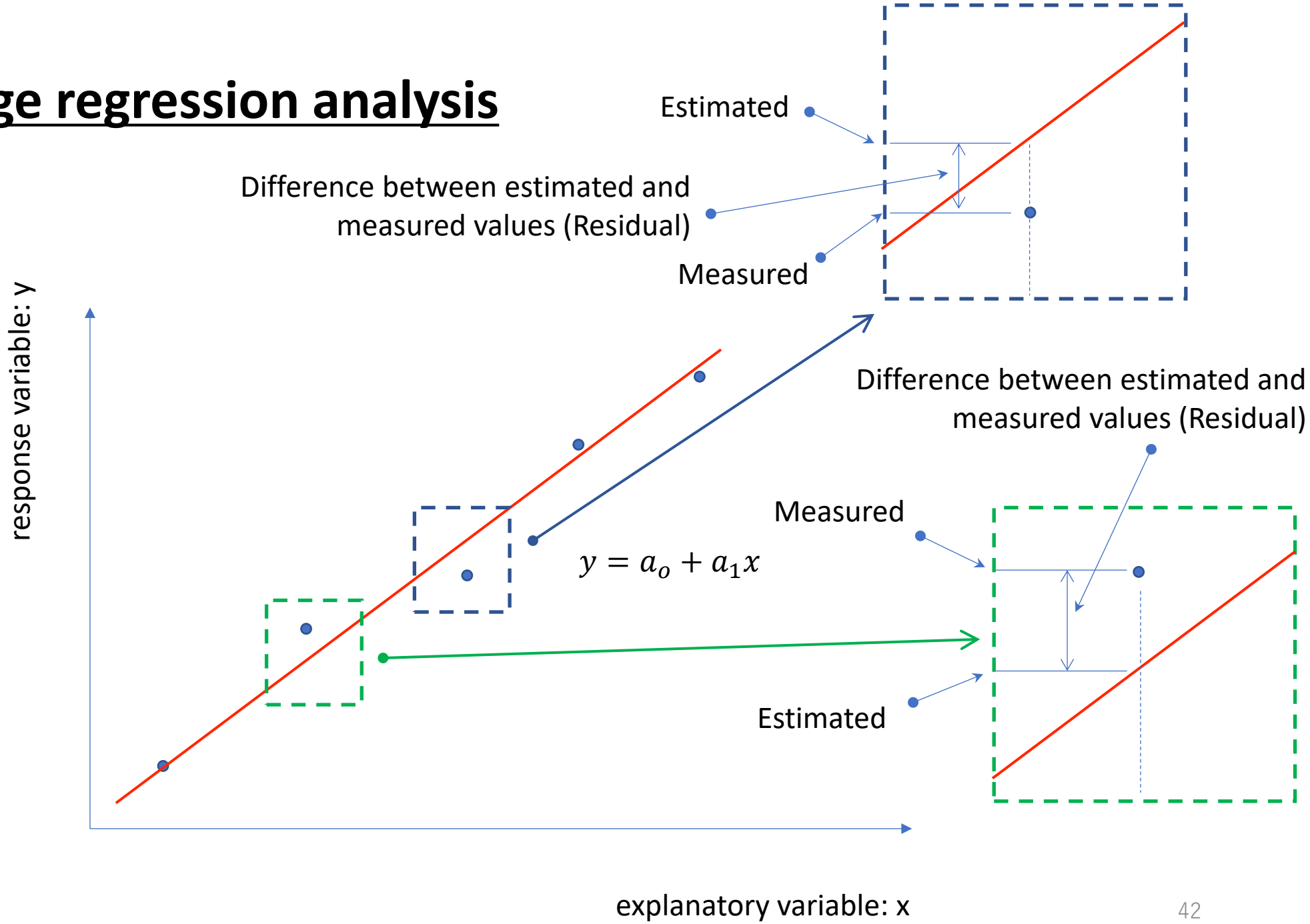
Least Square Method

Singe regression analysis

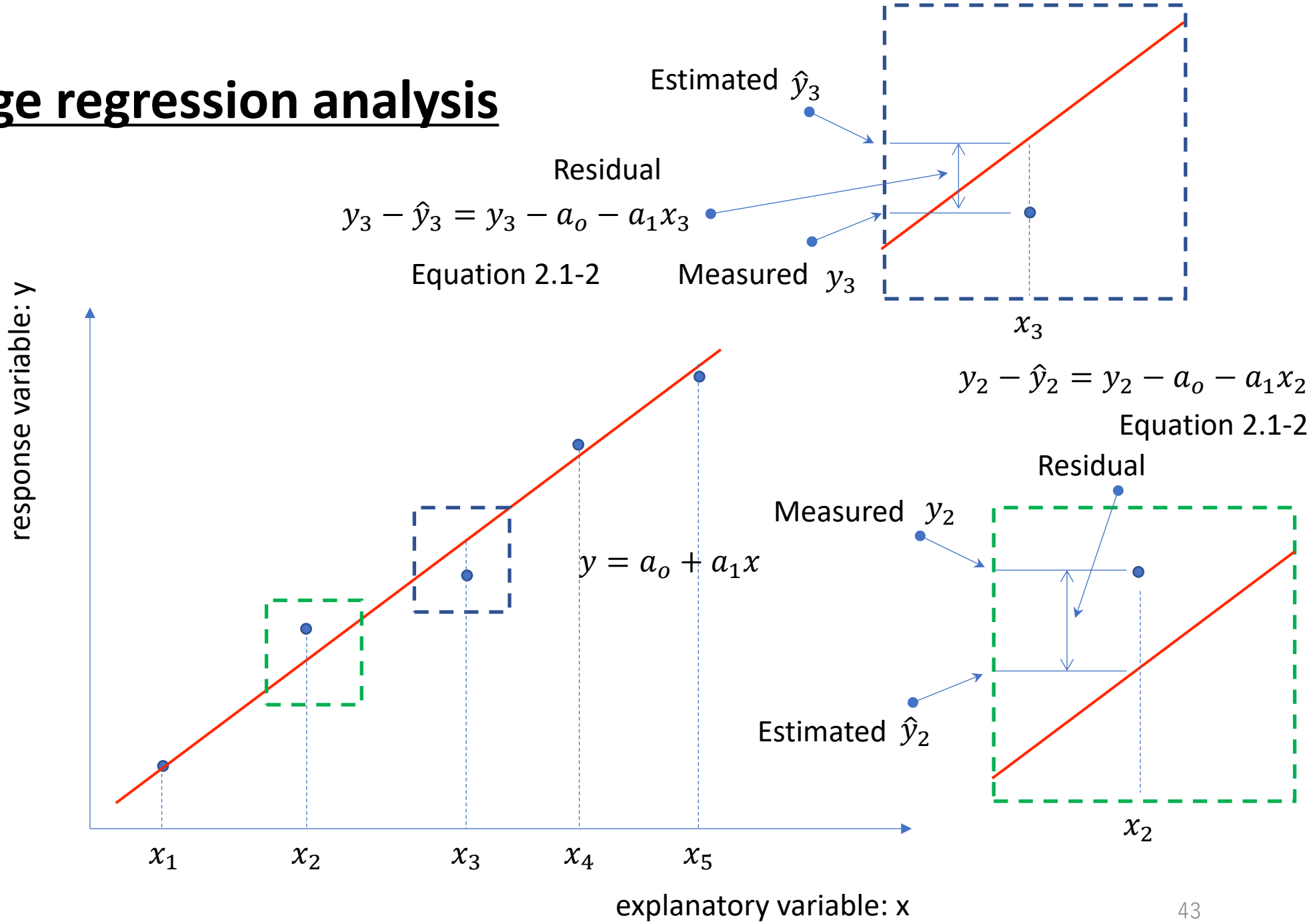


$$y = a_o + a_1 x$$ Equation 2.1-1

Singe regression analysis



Singe regression analysis



Singe regression analysis

Sum of residual of all data

$$+\Big)\begin{array}{l}(y_1-\hat{y}_1)^2=(y_1-a_0-a_1x_1)^2\\(y_2-\hat{y}_2)^2=(y_2-a_0-a_1x_2)^2\\(y_3-\hat{y}_3)^2=(y_3-a_0-a_1x_3)^2\\(y_4-\hat{y}_4)^2=(y_4-a_0-a_1x_4)^2\\(y_5-\hat{y}_5)^2=(y_5-a_0-a_1x_5)^2\end{array}$$

$$\sum_i^5 (y_i-a_0-a_1x_i)^2$$

Find the coefficients a_0 and a_1 that minimize the above.

Least Square Method

Practice of single regression analysis

i	x	y
1	2	5
2	4	5
3	6	7
4	7	8
5	8	7
6	9	10
MEAN	6	7

➢ Single regression analysis with n data

Regression equation

$$y = a_o + a_1 x$$ Equation 2.1-3

Data structure of i-th measured value

$$y_{(i)} = a_o + a_1 x_{(i)} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$ Equation 2.1-4

Estimated value of i-th data

$$\hat{y}_{(i)} = \hat{a}_0 + \hat{a}_1 x_{(i)}$$ Equation 2.1-5

Residual: difference between i-th measurement value and estimated value by regression equation

$$e_{(i)} = y_{(i)} - \hat{y}_{(i)} = y_{(i)} - (\hat{a}_0 + \hat{a}_1 x_{(i)})$$ Equation 2.1-6

Sum square of residual

$$S_e = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} \{y_{(i)} - (\hat{a}_0 + \hat{a}_1 x_{(i)})\}^2$$ Equation 2.1-7

To find $\hat{a}_0$, $\hat{a}_1$ to minimize the above.

Partial differentiation of S_e with $\hat{a}_0$, $\hat{a}_1$ respectively, and take these to zero

$$\frac{\partial S_e}{\partial \hat{a}_0} = -2 \sum_{i=1}^{n} \left(y_{(i)} - \hat{a}_0 - \hat{a}_1 x_{(i)}\right) = 0$$ Equation 2.1-8

$$\frac{\partial S_e}{\partial \hat{a}_1} = -2 \sum_{i=1}^{n} x_{(i)} \left(y_{(i)} - \hat{a}_0 - \hat{a}_1 x_{(i)}\right) = 0$$ Equation 2.1-9

organized these equations,

$$n\hat{a}_0 + \hat{a}_1 \sum x_{(i)} = \sum y_{(i)}$$ Equation 2.1-10

$$\hat{a}_0 \sum x_{(i)} + \hat{a}_1 \sum x_{(i)}^2 = \sum x_{(i)} y_{(i)}$$ Equation 2.1-11

These equation 2.1-10 to 2.1-11 are simultaneous equations of $\hat{a}_0$, $\hat{a}_1$. These are called normal equation.

From these,

$$\hat{a}_0 = \frac{\sum y_{(i)}}{n} - \hat{a}_1 \frac{\sum x_{(i)}}{n}$$

$$= \bar{y} - \hat{a}_1 \bar{x}$$

Equation 2.1-12

$$\bar{y} = \hat{a}_0 + \hat{a}_1 \bar{x}$$

Equation 2.1-13

$\bar{x}$; mean of $x_{(i)}$
$\bar{y}$; mean of $y_{(i)}$

The estimated regression equation passes through the point $(\bar{x}, \bar{y})$.

$$\hat{a}_0 = \frac{\sum y_{(i)}}{n} - \hat{a}_1 \frac{\sum x_{(i)}}{n}$$ Equation 2.1-12

$$\hat{a}_0 \sum x_{(i)} + \hat{a}_1 \sum x_{(i)}^2 = \sum x_{(i)} y_{(i)}$$ Equation 2.1-11

Substituting Equation 2.1-12 into Equation 2.1-11

$$\left(\frac{\sum y_{(i)}}{n} - \hat{a}_1 \frac{\sum x_{(i)}}{n} \right) \sum x_{(i)} + \hat{a}_1 \sum x_{(i)}^2 = \sum x_{(i)} y_{(i)}$$ Equation 2.1-14

Organize equation 2.1-14

$$\hat{a}_1\left(\sum x_{(i)}^2 - \frac{\left(\sum x_{(i)}\right)^2}{n}\right) = \sum x_{(i)}y_{(i)} - \frac{\left(\sum x_{(i)}\right)\left(\sum y_{(i)}\right)}{n}$$

Equation 2.1-15

Term [ ] of equation 2.1-15

$$S_{xx} = \sum_{i=1}^{n}\left(x_{(i)} - \bar{x}\right)^2 = \sum x_{(i)}^2 - \frac{\left(\sum x_{(i)}\right)^2}{n}$$

Equation 2.1-16

$$S_{yy} = \sum_{i=1}^{n}\left(y_{(i)} - \bar{y}\right)^2 = \sum y_{(i)}^2 - \frac{\left(\sum y_{(i)}\right)^2}{n}$$

Equation 2.1-17

$$S_{xy} = \sum_{i=1}^{n}\left(x_{(i)} - \bar{x}\right)\left(y_{(i)} - \bar{y}\right) = \sum x_{(i)}y_{(i)} - \frac{\left(\sum x_{(i)}\right)\left(\sum y_{(i)}\right)}{n}$$

Equation 2.1-18

Substituting these equation 2.1-16 and 2.1-18 into equation 2.1-15

$$\hat{a}_1 S_{xx} = S_{xy}$$ Equation 2.1-19

is obtained. Then,

$$\hat{a}_1 = \frac{S_{xy}}{S_{xx}}$$ Equation 2.1-20

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}$$

$$= \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$ Equation 2.1-21

Practice of single regression analysis

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	5	2-6	5-7	16	4	8
2	4	5	4-6	5-7	4	4	4
3	6	7	6-6	7-7	0	0	0
4	7	8	7-6	8-7	1	1	1
5	8	7	8-6	7-7	4	0	0
6	9	10	9-6	10-7	9	9	9
SUM	36	42	0	0	S_{xx}=34	S_{yy}=18	S_{xy}=22
MEAN	$\bar{x}$=6	$\bar{y}$=7					

Single regression equation

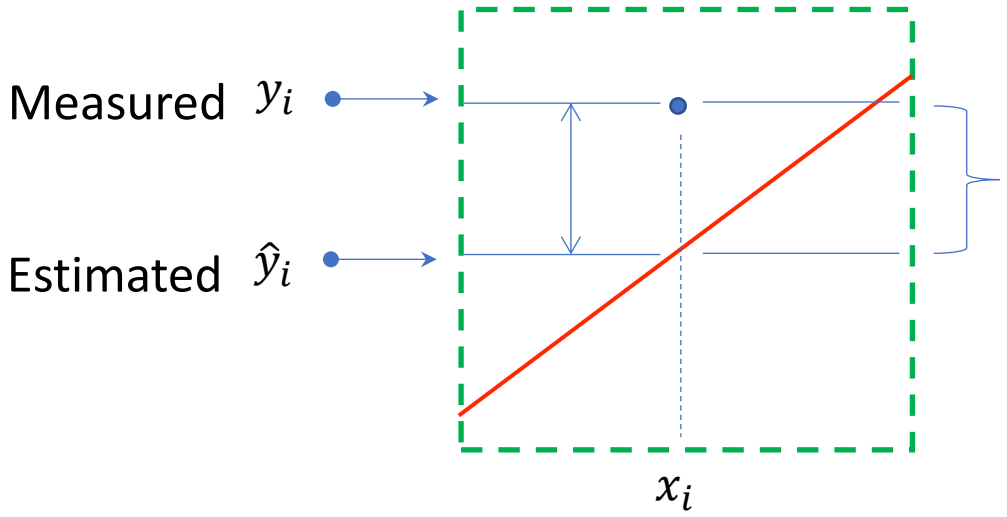
$$y = a_o + a_1 x = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} + \frac{S_{xy}}{S_{xx}}x$$

$$\hat{a}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \qquad \hat{a}_1 = \frac{S_{xy}}{S_{xx}}$$

$$y = 3.118 + 0.647x$$

◆ Validation of the accuracy of single regression analysis

- Coefficient of Determination, Correlation Coefficient



e_i Residual (difference between estimated value $\hat{y}_i$ and measured value y_i)
$= y_i - \hat{y}_i = y_i - a_o - a_1 x_i$

i	x_i	y_i	$(y_i-\bar{y})^2$	$\hat{y}_i$	$(\hat{y}_i-\bar{y})^2$	$e_i=y_i-\hat{y}_i$	e_i^2
1	2	5	$(5-7)^2=4$	4.41176	6.69896	0.58824	0.34602
2	4	5	4	5.70588	1.67474	-0.70588	0.49827
3	6	7	0	7.00000	0	0	0
4	7	8	1	7.64706	0.41869	0.35294	0.12457
5	8	7	0	8.29412	1.67474	-1.29412	1.67474
6	9	10	9	8.94118	3.76817	1.05882	1.12111
SUM	36	42	$S_{yy}=18$		$S_R=14.23529$		$S_e=3.76471$
MEAN	$\bar{x}=6$	$\bar{y}=7$					

Sum square of residual of y

$$S_T = S_{yy} = \sum (y_i - \bar{y})^2$$

Equation 2.1-22

Sum square of regression

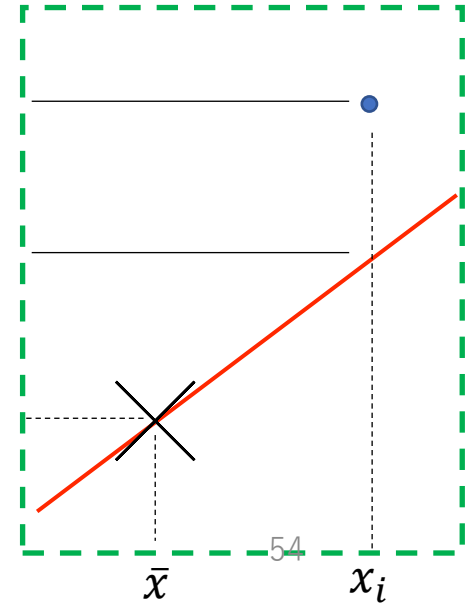
$$S_R = \sum (\hat{y}_i - \bar{y})^2$$

Equation 2.1-23

Sum square of residual of error

$$S_e = \sum (y_i - \hat{y}_i)^2$$

Equation 2.1-24



● Coefficient of Determination

Let's see relationship between S_T, S_R and S_e

$$S_T = S_R + S_e \qquad \text{Equation 2.1-25}$$

$$18 = 14.23529 + 3.76471$$

Divide both sides of this equation 2.1-25 by S_T

$$1 = \frac{S_R}{S_T} + \frac{S_e}{S_T} \qquad \text{Equation 2.1-26}$$

Since the regression analysis aims to reduce the sum of squares of the errors,

$$\frac{S_R}{S_T} = \frac{\text{Sum square of regression}}{\text{Sum square of measurements}} \qquad \text{Equation 2.1-27}$$

The closer A is to 1, the "better the regression equation fits".

$$R^2 = \frac{S_R}{S_T}$$

Equation 2.1-28

is defined as “Coefficient of Determination” (or “contribution ratio”)

The contribution ratio explains the proportion of the variation in y that is due to the variation in the regression.

● Correlation Coefficient

✓ For the obtained regression equation to be valid, the measured value y_i and the estimated value $\hat{y}_i$ should fit better.

Correlation Coefficient of $(y_i, \hat{y}_i)$

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

Equation 2.1-29

is calculated and can be used for evaluation of regression equation.

✓ There is the following relationship between “Correlation Coefficient” and “Coefficient of Determination”,

“Square of Correlation Coefficient” = “Coefficient of Determination”

Probability distribution for single regression analysis

Probability distribution of estimator of single regression analysis $\hat{a}_0 + \hat{a}_1 x_1$ is expressed by the following,

$$\hat{a}_0 + \hat{a}_1 x_1 \quad \sim N\left(a_o + a_1 x_1, \quad \left\{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right\}\sigma^2\right) \qquad \text{Equation 2.1-30}$$

where,

$$S_{xx} = \sum_i (x_i - \bar{x})^2$$

Appendix

◆Analysis of Variance for Single Regression Analysis

Methods for confirming the significance of regression equations

There is “Test of a single regression by ANOVA (ANalysis Of VAriance)”.

Variation factor	Sum square	Degree of freedom	Mean square	F-value
Variation by regression	$S_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$\phi_R = 1$	$V_R = \frac{S_R}{\phi_R}$	$F_0 = \frac{V_R}{V_e}$
Variation by residual	$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$\phi_e = n - 2$	$V_e = \frac{S_e}{\phi_e}$	
Total variation	$S_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$\phi_T = n - 1$		

$$S_T = S_R + S_e$$

Equation 2.1-25

Appendix

This ANOVA tests

null hypothesis H_0: Regression equations are not useful for estimation.

,using a test statistic F_0.

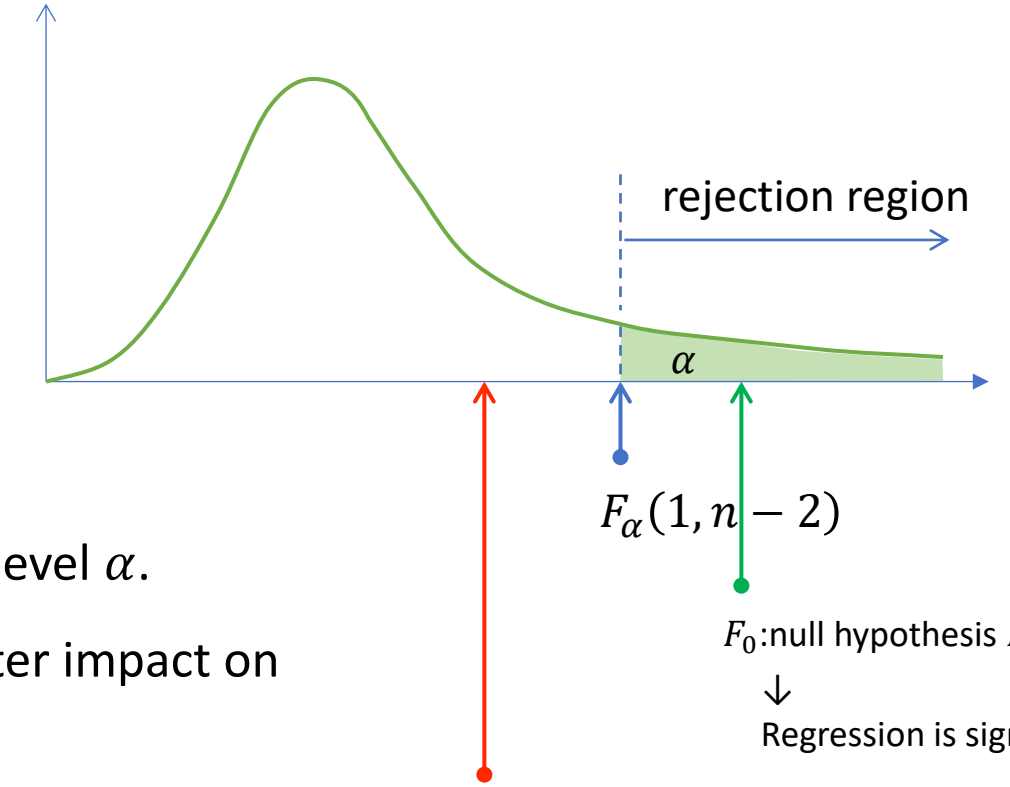
Therefore, "**rejecting the null hypothesis H_0**" is meaningful in this test.

Since this test statistic follows F-distribution of degree of freedom$(1,\ n-2)$,

if

$$F_0 \geq F_\alpha(1, n-2)$$

The null hypothesis H_0 is rejected by significance level α.



This means that "Variation by regression has a greater impact on total variation than variation by residuals."

⇓

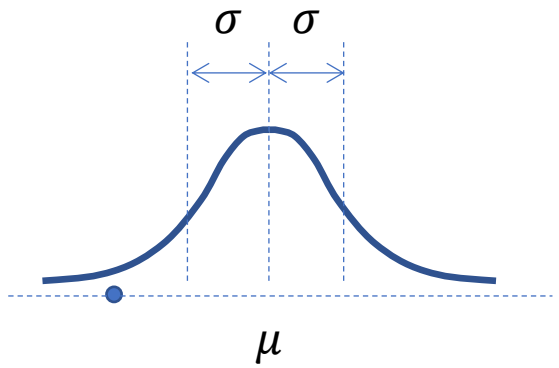
variation by regression >> variation by residuals

⇓

i.e., the single regression equation is significant.

Maximum likelihood estimation

Difference between Maximum Likelihood Estimation and Least Square Method



μ Expectation

σ Standard deviation

- **Least Square Method**

  Estimate mean to minimize the error of the estimated value between the sample data

- **Maximum Likelihood Estimation**

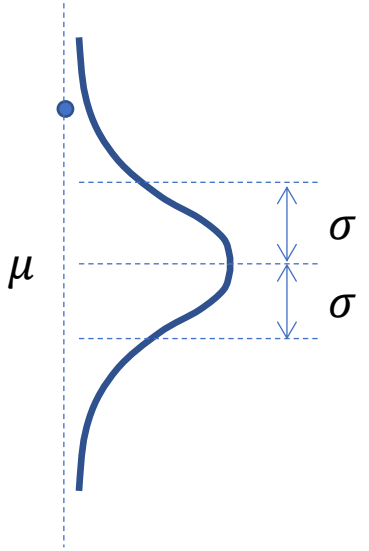
  Estimate mean to maximize the probability (likelihood) obtained from the sample data

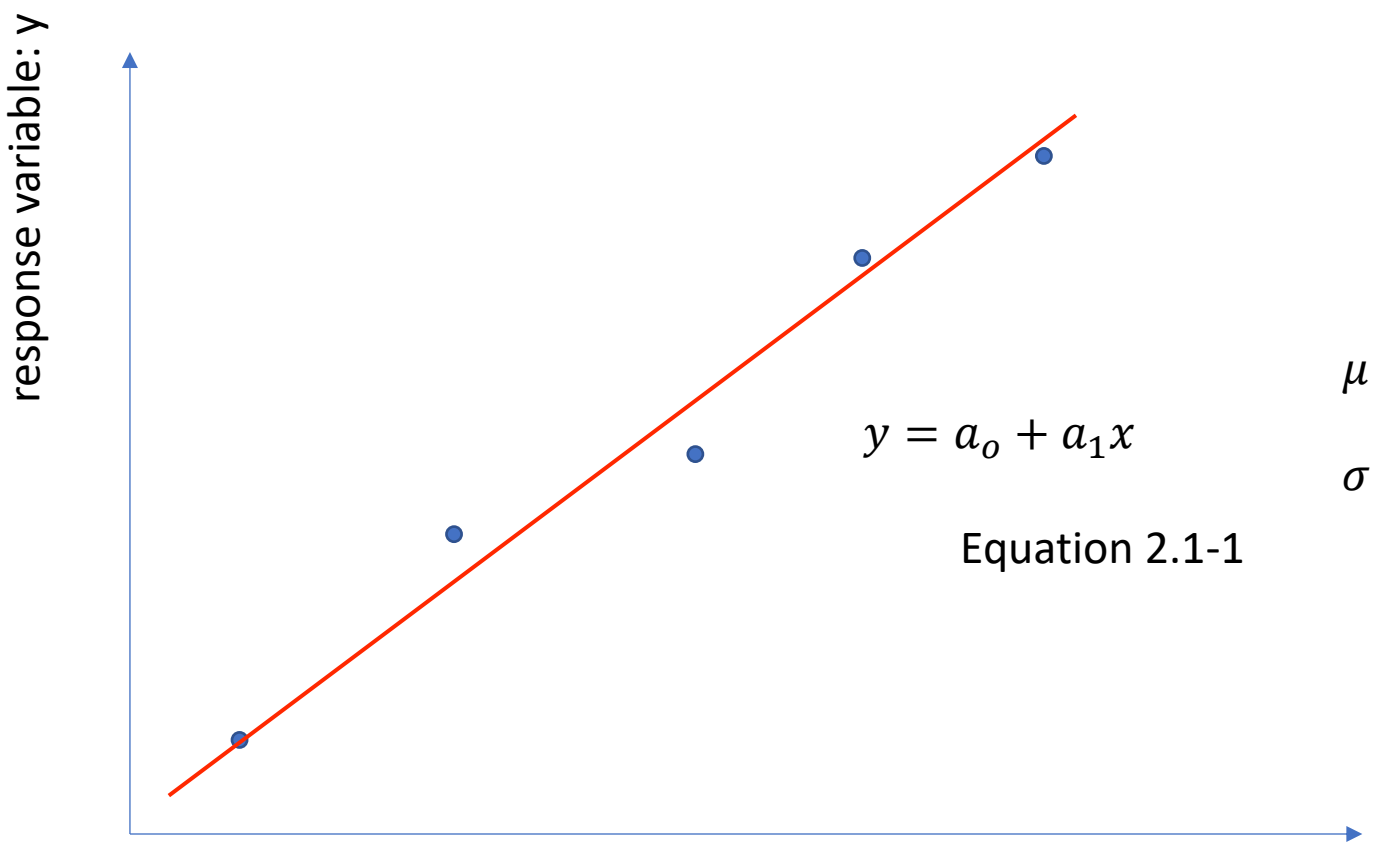
Appendix

Singe regression analysis

- Assumptions of Regression Analysis:

  Each data has a probability distribution.





$$y = a_o + a_1 x$$

Equation 2.1-1

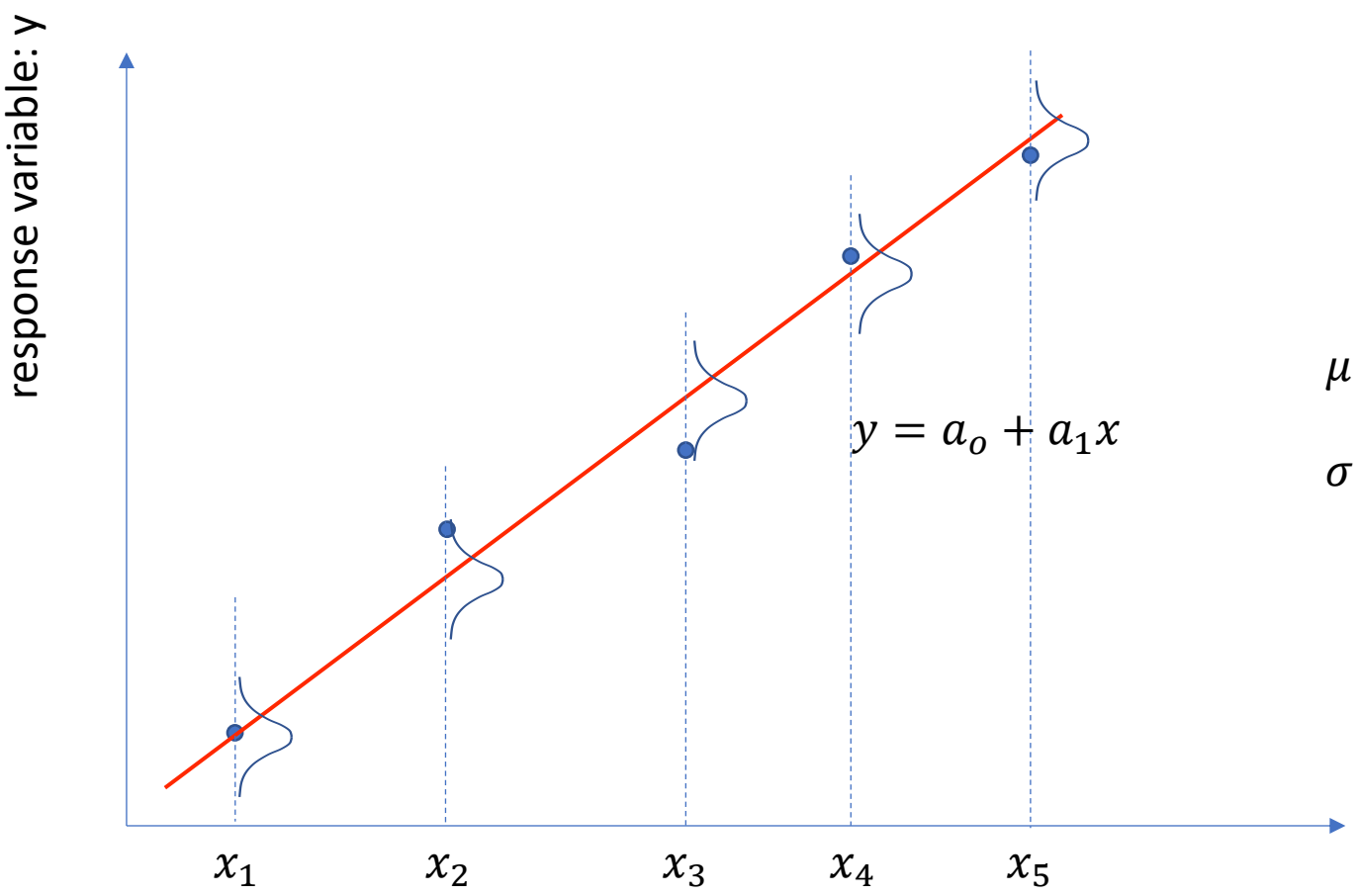
μ Expectation

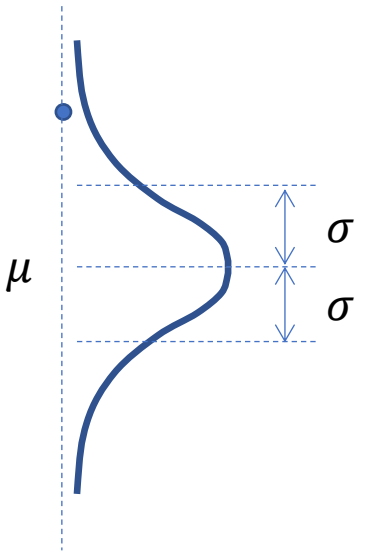
σ Standard deviation

Appendix

Each data has a probability distribution.

- Assuming that the probability distribution is normal and has a common variance σ^2,





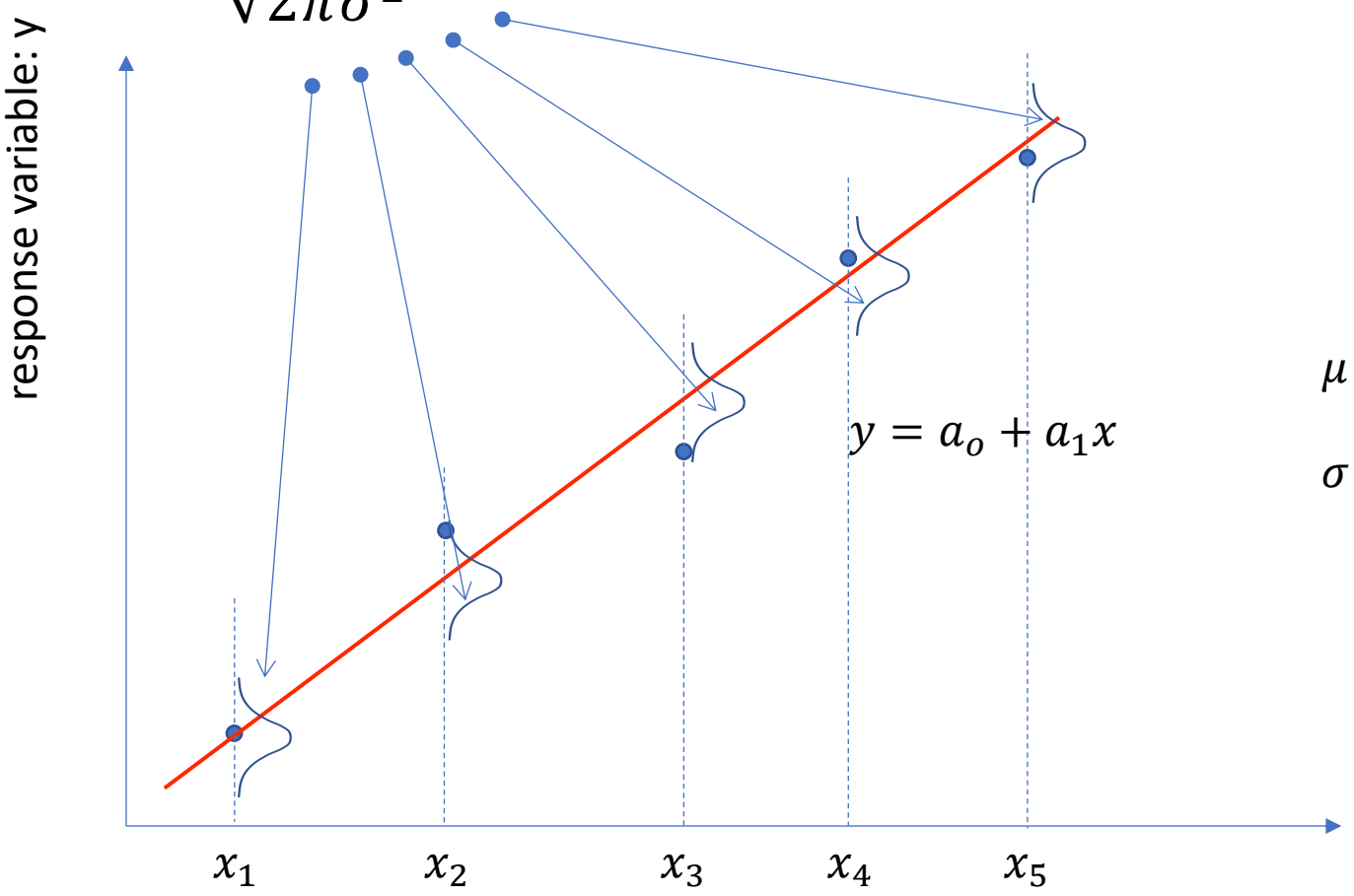
μ Expectation

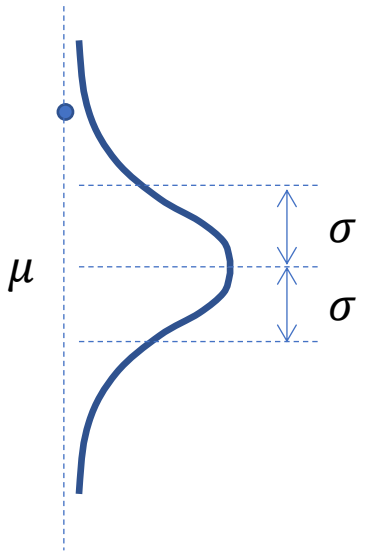
σ Standard deviation

Appendix

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{(y_i - a_o - a_1 x_i)^2}{2\sigma^2}\right\}}$$

Equation 2.1-31





μ Expectation

σ Standard deviation

Appendix

Multiplying the probability of each parameter

"Likelihood function"

is generally expressed as the following equation,

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$ Equation 2.1-32

Θ： Parameter of probability distribution (in case normal distribution, mean and standard deviation)

x： Parameter we wants (in case normal distribution, mean and standard deviation)

$$L(\mu, \sigma^2|x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2)$$ Equation 2.1-33

Appendix

$$L(\mu,\sigma^2|x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\left\{-\frac{(y_i-a_o-a_1x_i)^2}{2\sigma^2}\right\}}$$

$$=\left(\frac{1}{\sqrt{2\pi}}\right)^n\frac{1}{\sigma^n}exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-a_o-a_1x_i)^2\right\}$$

Equation 2.1-34

$$L(\mu,\sigma^2|x_1,x_2,x_3,x_4,x_5)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\left\{-\frac{(y_1-a_o-a_1x_1)^2}{2\sigma^2}\right\}}\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{\left\{-\frac{(y_2-a_o-a_1x_2)^2}{2\sigma^2}\right\}}\times$$

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{\left\{-\frac{(y_3-a_o-a_1x_3)^2}{2\sigma^2}\right\}}\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{\left\{-\frac{(y_4-a_o-a_1x_4)^2}{2\sigma^2}\right\}}\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{\left\{-\frac{(y_5-a_o-a_1x_5)^2}{2\sigma^2}\right\}}$$

$$=\left(\frac{1}{\sqrt{2\pi}}\right)^5\frac{1}{\sigma^5}exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{5}(y_i-a_o-a_1x_i)^2\right\}$$

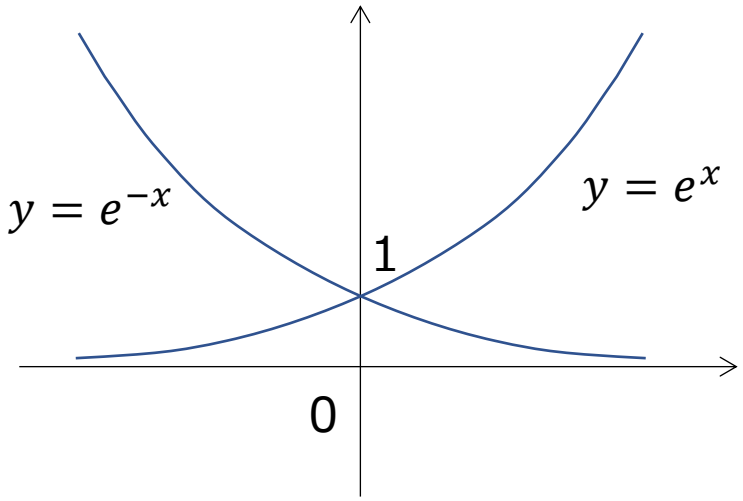
Equation 2.1-35

Appendix

$$L(\mu, \sigma^2 | x_1, x_2, x_3, x_4, x_5) = \left(\frac{1}{\sqrt{2\pi}}\right)^5 \frac{1}{\sigma^5} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{5}(y_i - a_o - a_1 x_i)^2\right\}$$

- To maximize likelihood function L,

  A part of [ ] in right-hand side can be minimize,



- By the way,

  A part of [ ] in right-hand side

  is the same with

  functions aiming at minimization on “Least-squares method”

  In other words,

  the least squares method does the same thing
  as assuming a normal distribution in the likelihood method.

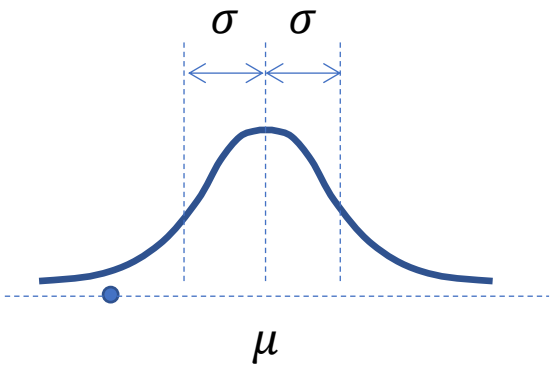
◆ Difference between Maximum Likelihood Estimation and Least Square Method

- **Least Square Method**

  Estimate mean to minimize the error of the estimated value between the sample data

- **Maximum Likelihood Estimation**

  Estimate mean to maximize the probability (likelihood) obtained from the sample data



μ Expectation

σ Standard deviation

- the least squares method does the same thing, as assuming a normal distribution in the likelihood method.
- The likelihood method can be used for any probability distributions other than the normal distribution.

Outline

1. Sampling and Estimate

2. Statistic analysis 1: Summary of Regression Analysis (Singe Regression Analysis)

3. Statistic analysis 2: Summary of Time Series Analysis

Time Series Analysis

- Time series data:

  Data that records the results of measurements and observations made in the order of time for phenomenon that varies with time.

- Objectives of time series analysis:

  Statistical analysis of the behavior of time-series data to understand the characteristics of data variation, interrupt phenomena, and predict future variation.

◆ Preparation of data

- Legitimization
- Difference
- Ratio (compared to previous data, Year-to-year comparison)
- Moving Average
- Geometric Mean

◆ Decompose time series data into several variation

Trend : T

Variations that represent basic long-term movements.
There is a case that Circular variation is not separated. This is called Trend-cycle variation (TC).

Cycle variation : C

A variation that repeats a cycle with a relatively long period of time. Sometimes this is included in “trend”.

Seasonal variation : S

A variation that repeats a cycle with a period of one year.

Irregular variation : I

Variation other than those listed above, which are not regular.
Sometimes variations due to various environmental dynamic factors is included in irregular variation.

$y_t = TC_t + S_t + I_t$ ※ Here, “cycle variation” is included in “trend”.

◆ Procedure of time series analysis

- Extract “trend” TC_t
- Extract “seasonal variation” S_t
- Extract “irregular variation” I_t

◆ How to extract "Trend" TC_t at time t

- Moving average method
- exponential smoothing method

Moving average method

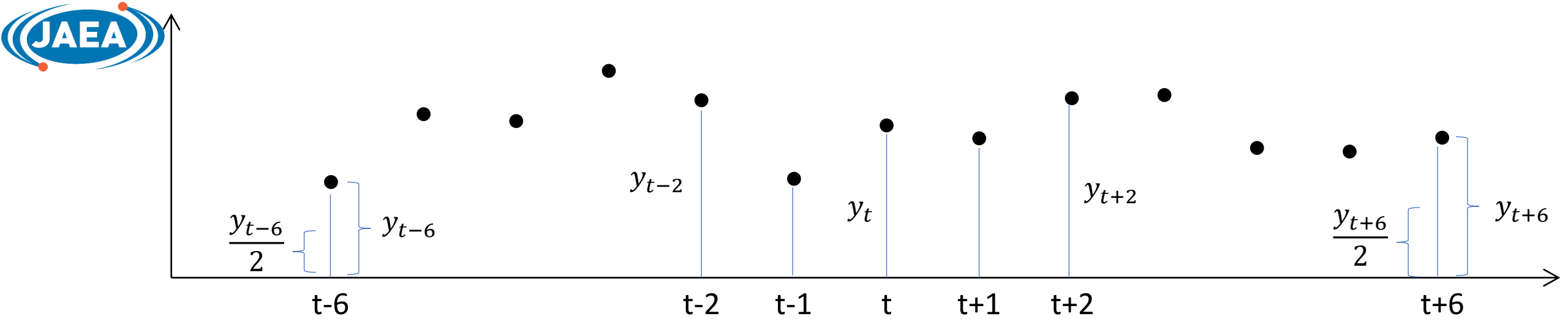
A value at time t obtained by moving average : $\widehat{TC_t}$

Mean of 2k+1 values ka value at time t and from before point k to after point

$$\{y_s | y_{t-k}, y_{t-k+1}, y_{t-k+2}, \cdots, y_{t-1}, y_t, y_{t+1}, \cdots, y_{t+k-2}, y_{t+k-1}, y_{t+k}\}$$

$$\widehat{TC_t} = \sum_{s=t-k}^{t+k} \frac{y_s}{2k+1}$$ (2k+1) moving average

➢ In general, using an m-term moving average removes components that cycle at period m.



Moving average: For data expected to cycle in 12 periods, a 12-term moving average is used.

If the number of terms is even, as in the case of 12 terms, the (2k+1)-term moving average cannot be applied.

Assuming k=6, using data set $\{y_s | y_{t-6}, y_{t-5}, \cdots, y_{t-1}, y_t, y_{t+1}, \cdots, y_{t+5}, y_{t+6}\}$, $\widehat{TC_t}$ is obtained.

In this case, as initial term $\frac{y_{t-6}}{2}$ instead of y_{t-6}

, as the final term $\frac{y_{t+6}}{2}$ instead of y_{t+6}

$$\widehat{TC_t} = \left(\frac{y_{t-6}}{2} + y_{t-5} + \cdots + y_{t+5} + \frac{y_{t+6}}{2}\right) \Big/ 12$$

◆ How to extract “seasonal variation” St

By subtracting the trend TC_t from the original time series data

$$w_t = y_t - \widehat{TC_t}$$

It includes seasonal and irregular variation.

Calculate “monthly mean : $\overline{w}(m)_t = \frac{\sum_{i=0}^{n} w_{t+12i}}{n} \left(= (\overline{w}(m)_{(t\pm 12k)}\right)$ “ of $\{w_t\}$

Subtract those averages from the monthly average to obtain the seasonal variation $\widehat{S_t}$, in order to set the average to 0

$$\widehat{S_t} = \overline{w}(m)_t - \sum_{l=0}^{11} \overline{w}(m)_{(t+l)}$$

Seasonal variation for the same month of each year have the same value.

◆ How to extract “irregular variation” It

Irregular variation $\widehat{I_t}$ is obtained as residual $\hat{I} = y_t - \widehat{TC_t} - \widehat{S_t}$.

◆ Auto-correlation

Appendix

Since time-series data deals with periods, correlations between time series that shift the time points of the same series are important.

Original time series data: $\{y_t\}$

Time series data shifted from the original time series data by τ : $\{y_{t+\tau}\}$

The autocorrelation coefficient : correlation coefficient between $\{y_t\}$ and $\{y_{t+\tau}\}$

Calculate correlation between $x = y_t = y(t)$ and $y = y_{t+\tau} = y(t+\tau)$.

Here, let x(t) be the amount of time-dependent random variation,

Mean of the product of two variations separated by τ time : $C(t,\tau) = \frac{1}{N}\sum_{t=1}^{N} y(t)y(t+\tau) = E[y(t)y(t+\tau)]$

N: the size of a data series

Statistical function defined as the above equation is called “auto-correlation function”.

Here, τ is called “lag”, this means time delay or gap.

Appendix

Mean of $x = y_t = y(t)$ and $y = y_{t+\tau} = y(t+\tau)$ is defined as μ .

Function of lag τ as autocovariance function is calculated as

$$C(\tau) = E[(y(t) - \mu)(y(t+\tau) - \mu)\,]$$

Using this, “autocorrelation function” is obtained as the following,

$$r_\tau = \frac{C(\tau)}{C(0)}$$

Where, denominator C(0) means covariance (that is variance) at τ=0,

This serves to normalize the autocovariance of numerator to [-1,1].

- ✓ Note that “the mean and variance of y_t“ and “the covariance between $\{y_t, y_{t+\tau}\}$ ” are assumed to be independent of time (stationary stochastic process).

Example

Trend analysis of environmental monitoring data

Provided data:

1. Cs-137 Fall out in Fukushima, Ibaraki and Miyagi area after the Fukushima Daiichi-Nuclear Power Station Accident - time series data -

Practice 1 Time series analysis of Cs-137 fall out data.

-> Estimation of variation range during normal situation.

Practice 1

◆Variation range during normal situation

Future prediction: Estimated from past achieved data

- Fitting approximate function to “trend variation”

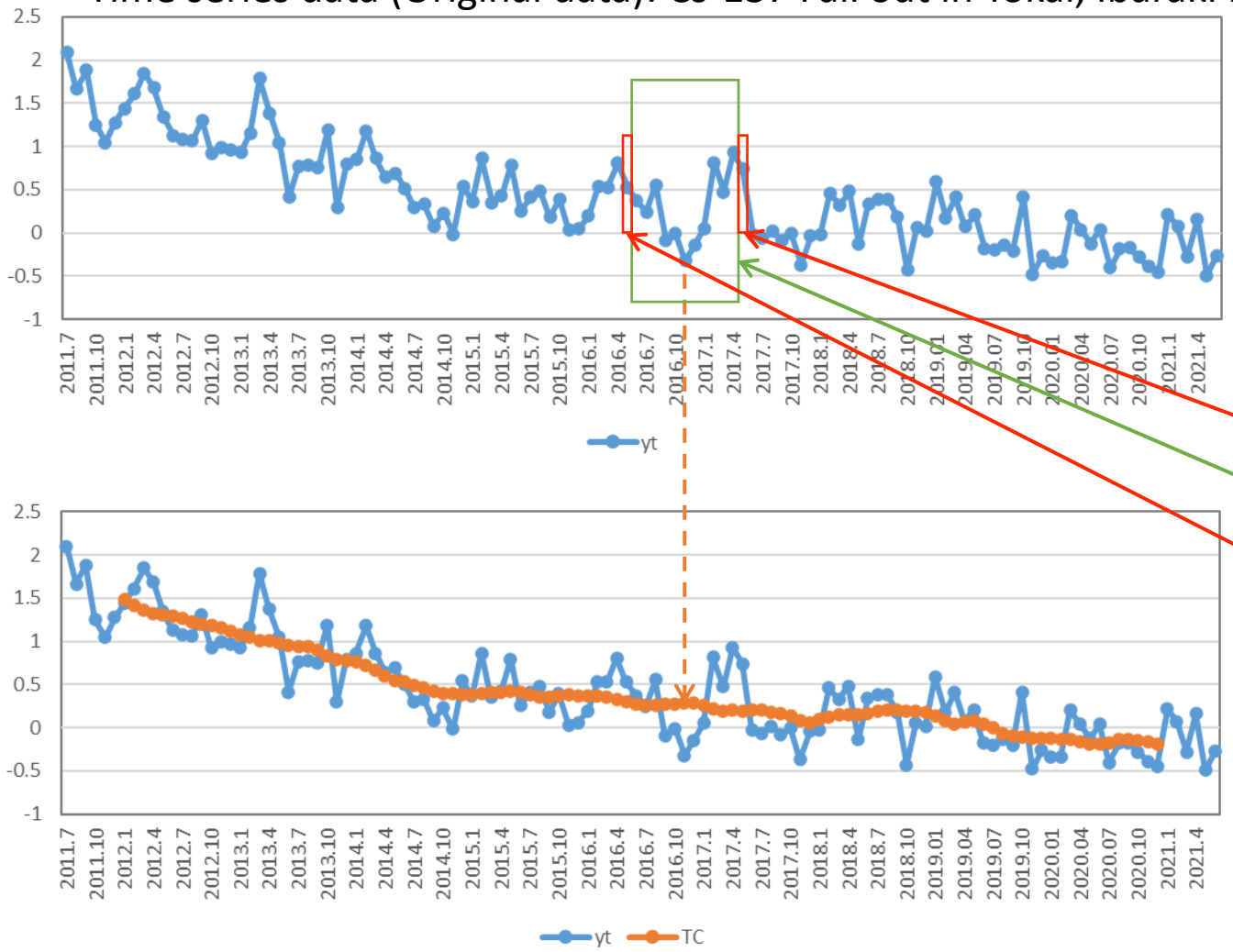
  If necessary, divided into several interval

- Variance of “Irregular variation”

Example)

STEP 1 Extract “Trend variation”

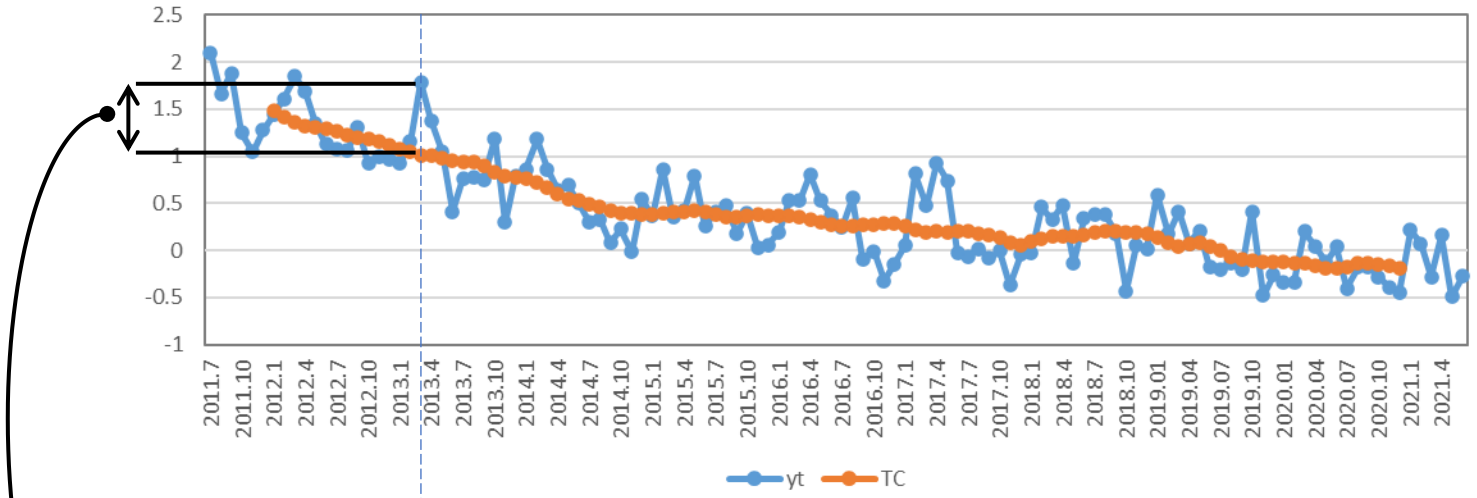
Time series data (Original data): Cs-137 Fall out in Tokai, Ibaraki from July 2011 to June 2021

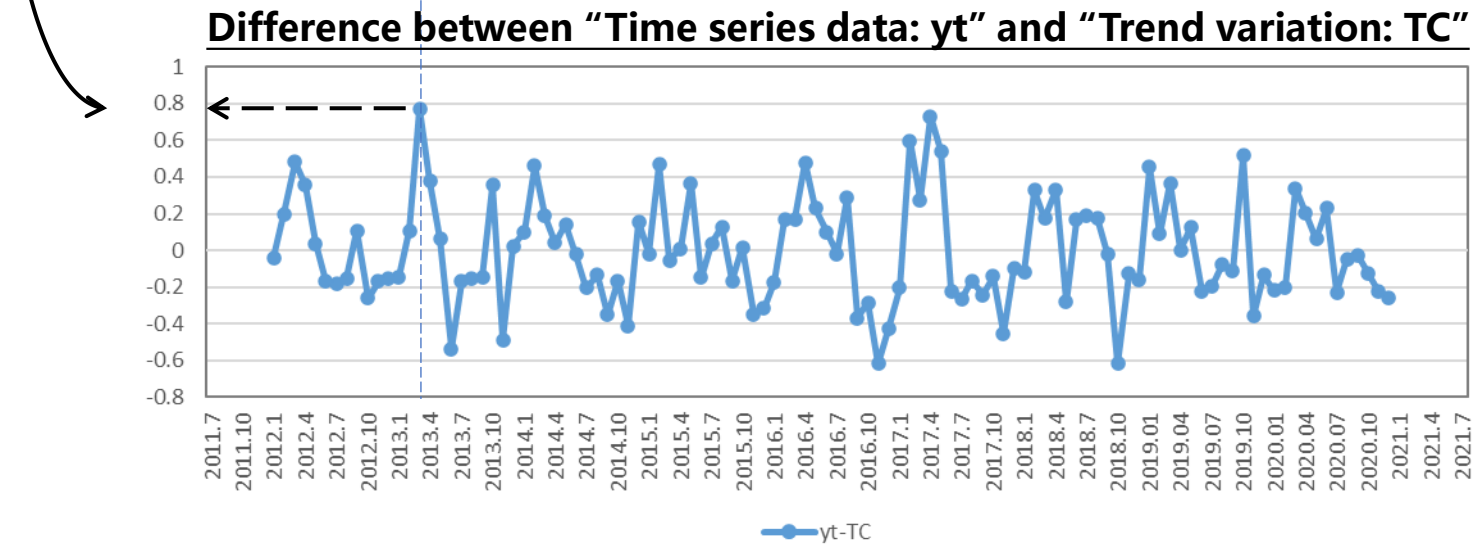


Moving average:

$$\widehat{TC_t} = \left(\frac{y_{t-6}}{2} + y_{t-5} + \cdots + y_{t+5} + \frac{y_{t+6}}{2} \right) \Big/ 12$$

STEP 2

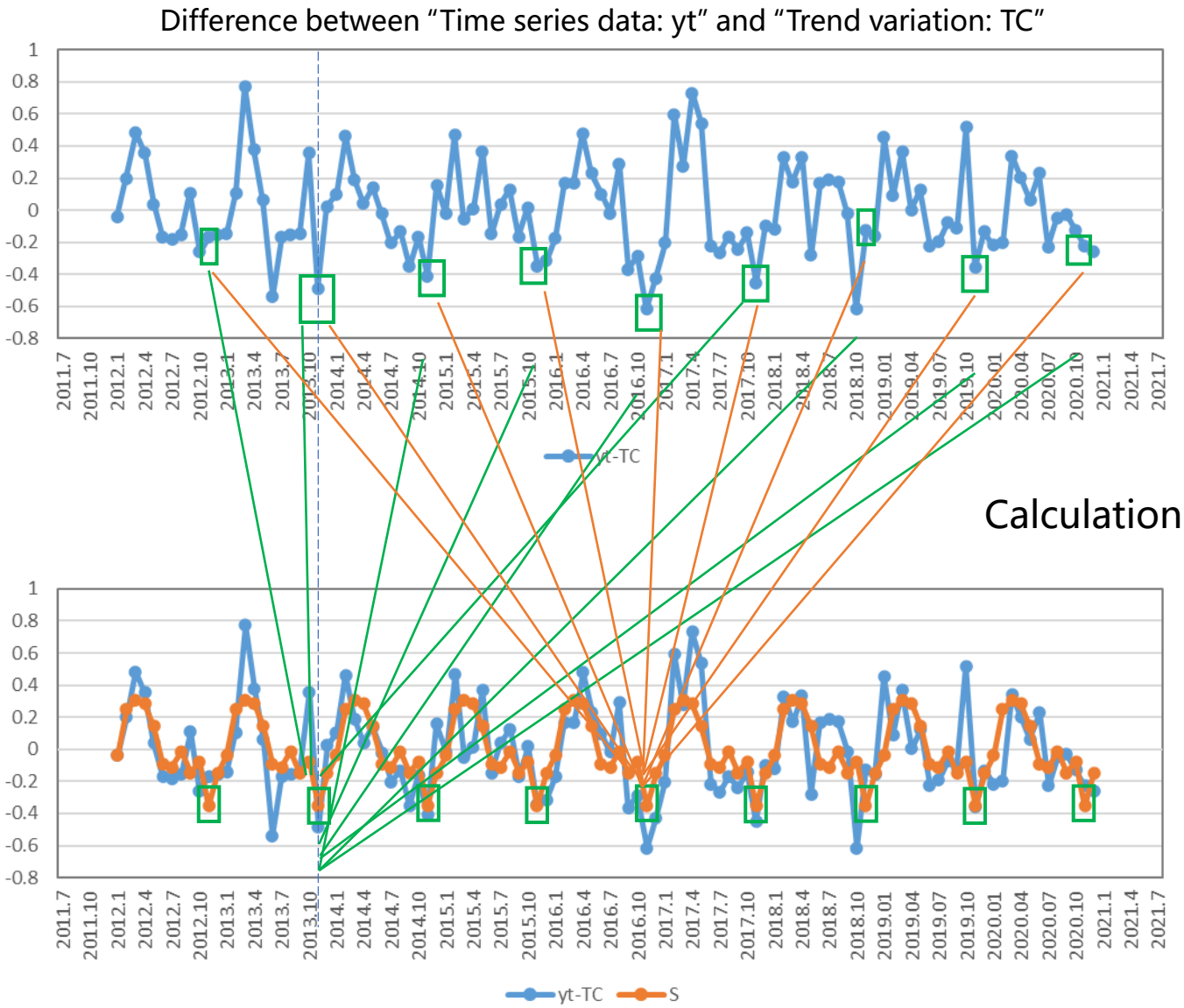




Time series data: yt

Trend variation: TC

STEP 3 Extract “Seasonal variation”



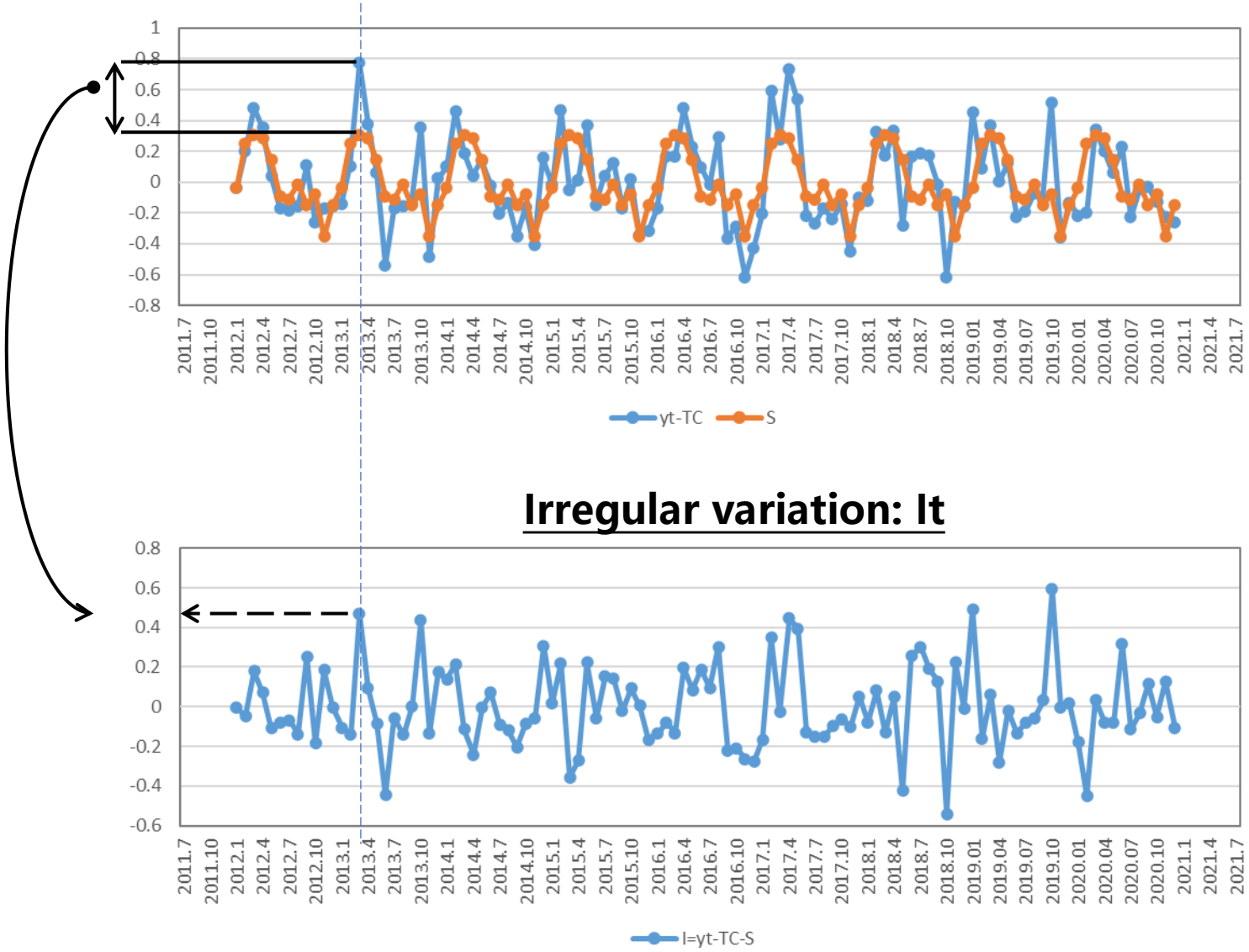
Calculation of monthly mean from each month data

Difference between “Time series data yt” and “Trend variation TC” : yt-TC

Seasonal variation: S

STEP 4

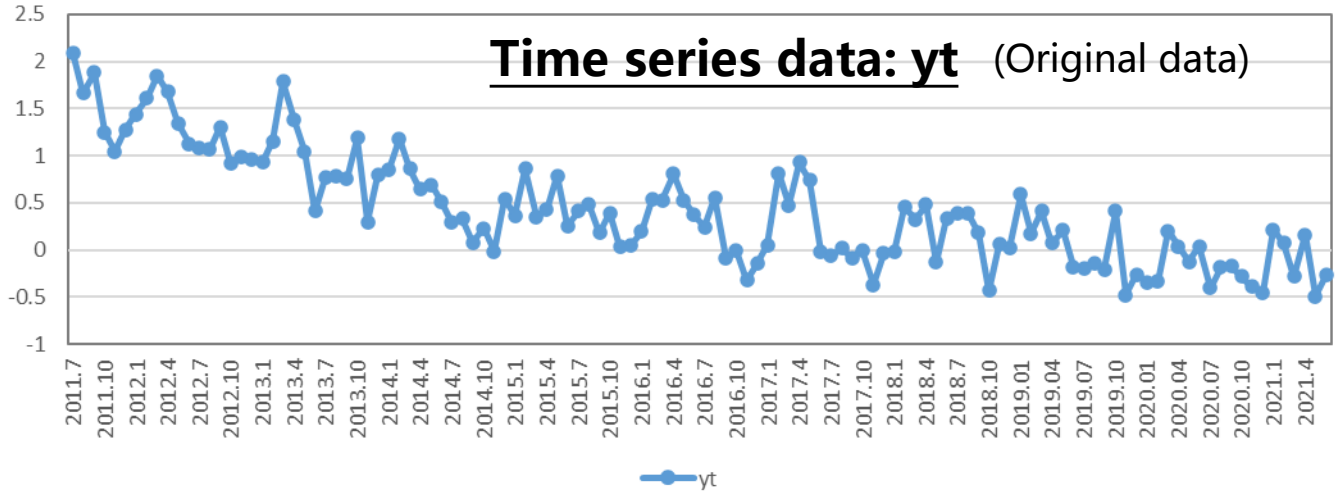
Extract “Irregular variation”



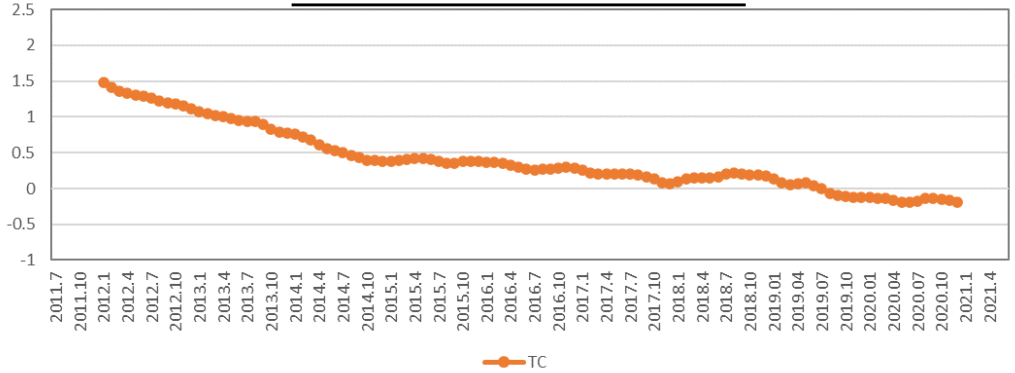
Difference between “Time series data yt” and “Trend variation TC” : yt-TC

Seasonal variation: S

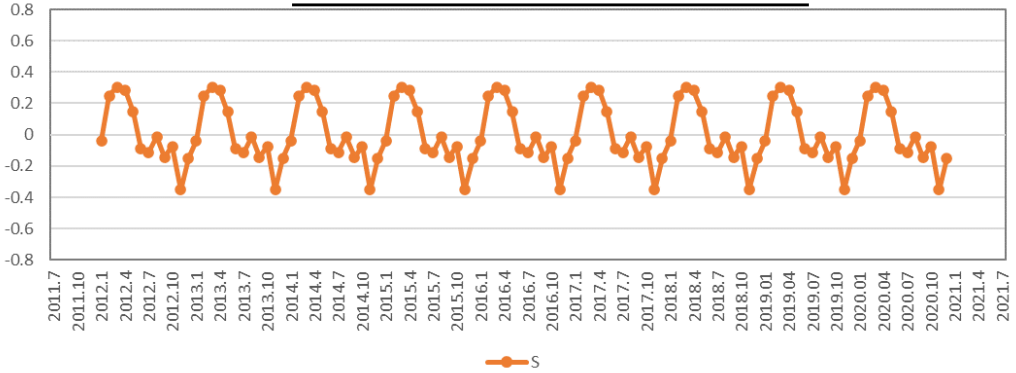
Time series data: yt (Original data)



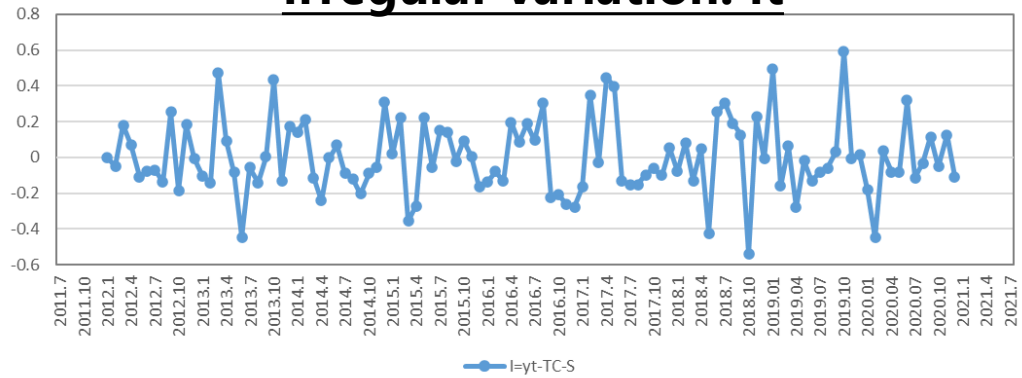
Trend variation: TC



Seasonal variation: St



Irregular variation: It



Summary

- In environmental monitoring, representative values of measurement results with statistical variability are estimated using statistical methods.
- The least-squares method is a likelihood estimation method that assumes that the probability distribution of each measurement has a normal distribution.
- Multiple regression analysis is a regression analysis in which multiple explanatory variables are used to explain the response variable. !! Note multicolinearity !!
- In multiple regression analysis, the coefficient of determination has the property that as the number of explanatory variables increases, the coefficient of determination automatically increases. → adjusted coefficient of determination
- Principal component analysis is a method that summarizes multiple quantitative explanatory variables into a smaller number of indicators called “principal component”
- Time series analysis is an analytical method that decomposes time series data into trend variation, cyclical variation, seasonal variation, irregular components, etc.
- Multiple regression analysis, principal component analysis, and time series analysis can be used in combination.
- In regression analysis, uncertainty can be estimated using the probability distribution of the estimate.

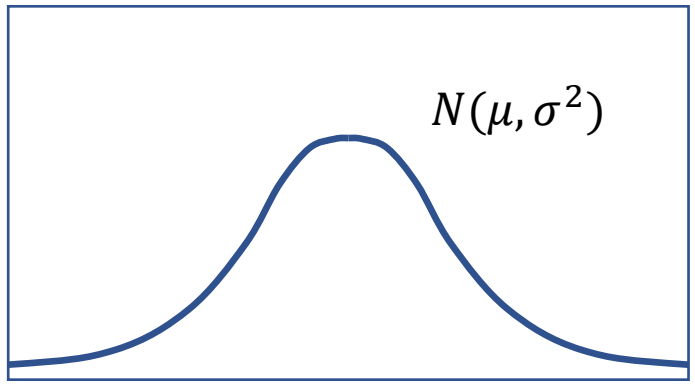
Appendix

Appendix 1

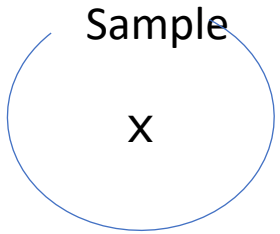
Interval estimation

Interval estimation of the population mean by sample (known population variance)

Appendix



Sampling



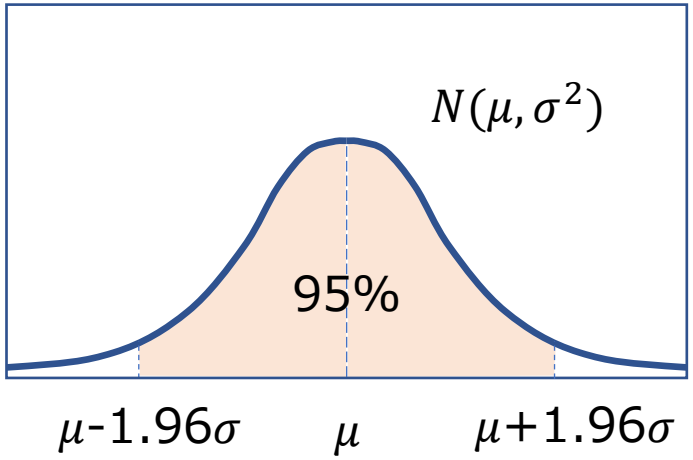
I know the population variance, but I don't know the population mean, so I want to calculate the population mean using sample data by intervals estimation.

μ: unknown

σ^2: known

Interval estimation requires a condition of what percentage confidence coefficient to use for interval estimation.

Interval estimation of the population mean μ usually at 95% confidence coefficient



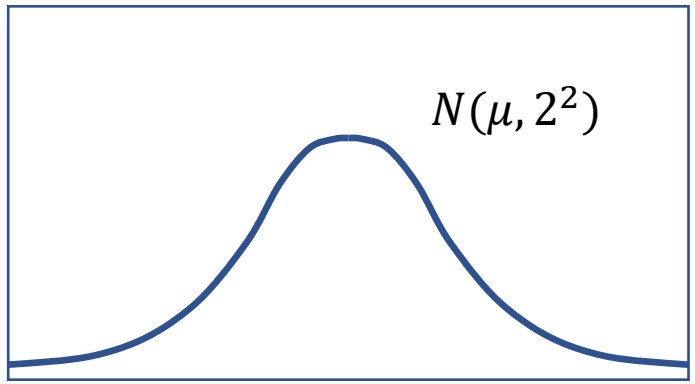
When 100 trials of sampling one sample from $N(\mu, \sigma^2)$ are performed, 95 times the following equation holds.

$$\mu - 1.96\sigma \leq x \leq \mu + 1.96\sigma$$

$$x - 1.96\sigma \leq \mu \leq x + 1.96\sigma$$

95% confidence interval for the population mean μ

Example)



Sampling a data

Sample

5

μ: unknown

σ^2 : known = 4

Interval estimation of population mean μ by 95% confidence coefficient

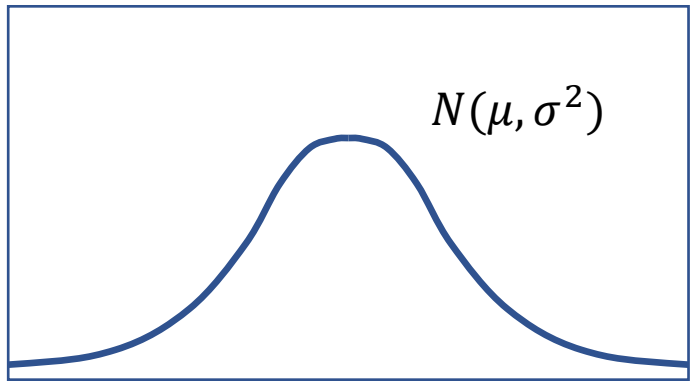
$$5 - 1.96 \times 2 \leq \mu \leq 5 + 1.96 \times 2$$

$$1.08 \leq \mu \leq 8.92$$

95% confidence interval for the population mean μ

= Interval estimation results for the population mean μ

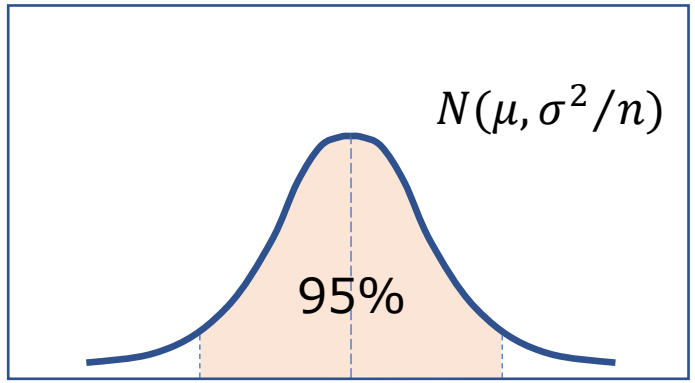
Interval estimation of the population mean with multiple samples (known population variance)



μ: unknown

σ^2: known

Mean of n data sampled



Interval estimation of the population mean μ by 95% confidence coefficient

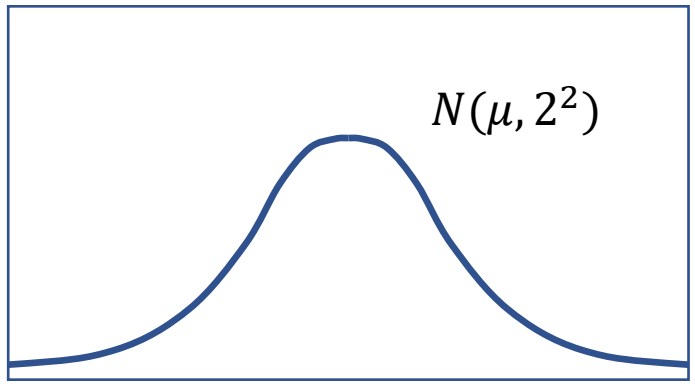
When 100 trials of sampling n samples from the population $N(\mu, \sigma^2)$ are performed, 95 times the following equation holds.

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}$$

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

95% confidence interval for the population mean μ

Example)



μ: unknown

σ^2: known = 4

Sampling some data

Sample

8.5, 9.1, 8.7, 8.9

Sample mean

$\bar{x} = 8.8$

Interval estimation of population mean μ by 95% confidence coefficient

$$8.8 - 1.96\frac{2}{\sqrt{4}} \leq \mu \leq 8.8 + 1.96\frac{2}{\sqrt{4}}$$

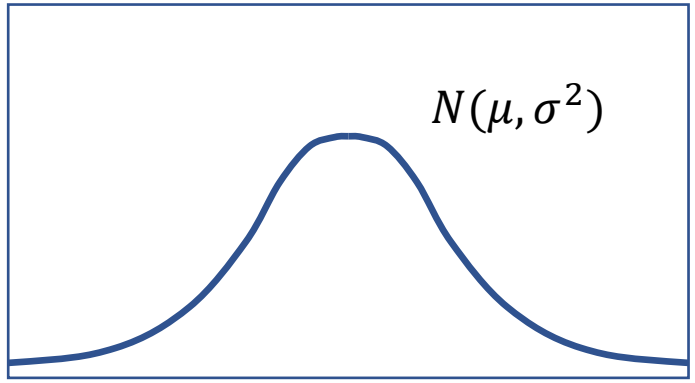
$$6.84 \leq \mu \leq 10.76$$

95% confidence interval for the population mean μ

Interval estimation of the population mean (UNKNOWN population variance)

Appendix

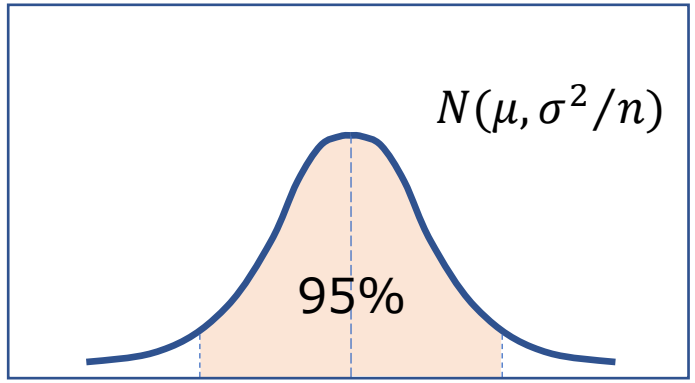
So far, it has talked that the case where the population variance of the distribution of the population is known, but in practice it is not often the case that the population variance is known.



μ: unknown

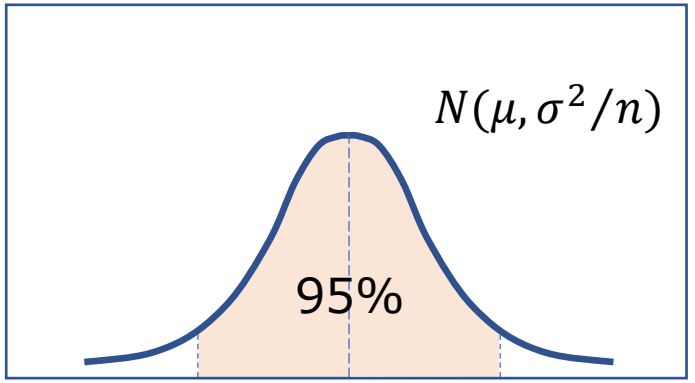
σ^2: **known**

Mean of n data sampled



Standardized

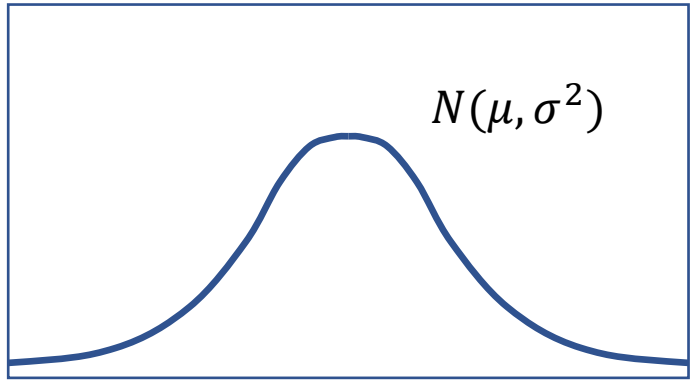
$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$$



Interval estimation of the population mean (UNKNOWN population variance)

Appendix

Substitute in an estimator of σ^2 for unknow σ^2



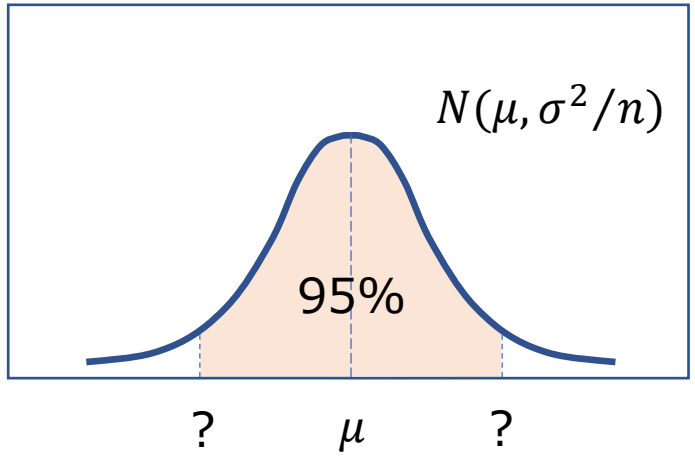
μ: unknow

σ^2: **unknown**

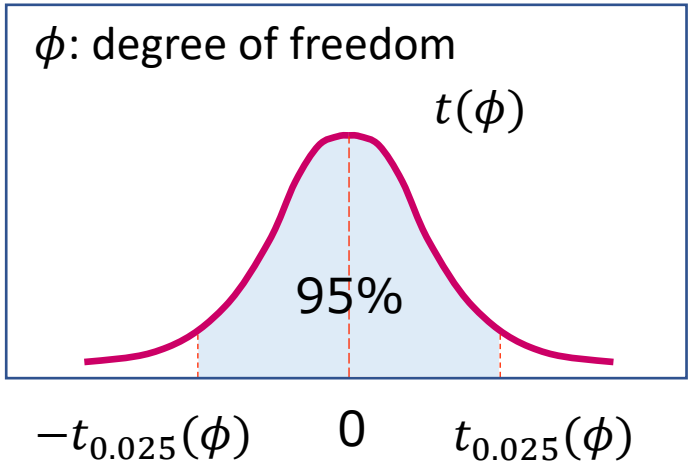
Good estimator of σ^2 is unbiased variance, --> $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

A follows a t-distribution, not a standard normal distribution. --> $t = \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}}$

Mean of n data sampled



$$t = \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}}$$



Appendix

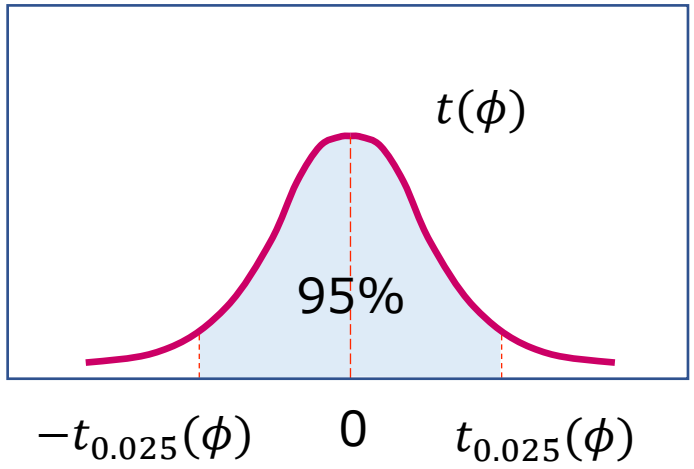
Interval estimation of population mean μ by 95% confidence coefficient

When 100 trials of sampling n samples from a population $N(\mu, \sigma^2)$ of unknown population variance are performed, 95 times the following equation holds.

$$-t_{0.025}(\phi) \le t \le t_{0.025}(\phi)$$

$$t = \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}}$$

ϕ: Degree of freedom



$$-t_{0.025}(\phi) \le \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}} \le t_{0.025}(\phi)$$

$$\mu - t_{0.025}(\phi)\sqrt{\frac{\hat{\sigma}^2}{n}} \le \bar{x} \le \mu + t_{0.025}(\phi)\sqrt{\frac{\hat{\sigma}^2}{n}}$$

$$\bar{x} - t_{0.025}(\phi)\sqrt{\frac{\hat{\sigma}^2}{n}} \le \mu \le \bar{x} + t_{0.025}(\phi)\sqrt{\frac{\hat{\sigma}^2}{n}}$$

95% confidence interval for the population mean μ

◆ Comparison between 95% confidence intervals for the population mean μ

known population variance :

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}$$

unknown population variance :

$$\bar{x} - t_{0.025}(\phi)\sqrt{\frac{\hat{\sigma}^2}{n}} \leq \mu \leq \bar{x} + t_{0.025}(\phi)\sqrt{\frac{\hat{\sigma}^2}{n}}$$

Appendix 2

Multiple regression analysis

Appendix

Multivariate analysis

A generic term for analysis on multivariate data. There are many analytical methods for multivariate analysis,

Example)

- Multiple regression analysis
- Principal component analysis
- Discriminant analysis
- Cluster analysis

Multiple Regression Analysis

Appendix

◆ Single regression analysis and Multiple regression analysis

- Single regression analysis

Explanatory variable: x → Response variable: y

Structure

$$y = a_o + a_1 x_1$$

Equation 2.2-1

- Multiple regression analysis

Explanatory variable: x_1
Explanatory variable: x_2
Explanatory variable: x_3
⋮
Explanatory variable: x_n

→ Response variable: y

Structure

$$y = a_o + a_1 x_1 + a_2 x_2 + a_3 x_3 + \cdots + a_n x_n$$

Equation 2.2-2

Appendix

◆ Main aims of multiple regression analysis

1. Identify the major factors of impact.
2. Predict the results.

◆ Procedure of multiple regression analysis

1. Preparation
   i. Consider factors that influence the response variable.
   ii. Consider the significance of obtaining multiple regression equations.
   Such as, whether the explanatory variables are correlated with the response variable
2. Obtain multiple regression equation.
3. Validation of multiple regression equation
   i. Accuracy of the estimated multiple regression equation
   ii. Statistical significance (precision of multiple regression equations, coefficients)

◆ How to find multiple regression equation

Appendix

In case, single regression analysis,

- Regarding to the following equation of relationship between a single explanatory variable x and a response variable y,

$$y = a_o + a_1 x \qquad \text{Equation 2.2-3}$$

- Residual (Difference between measured value and estimated value by single regression equation)

$$y_i - \hat{y}_i = y_i - a_o - a_1 x_i \qquad \text{Equation 2.2-4}$$

, and sum of square of each residual,

$$\sum_i^n (y_i - a_0 - a_1 x_i)^2 \qquad \text{Equation 2.2-5}$$

a_0 and a_1 is calculated to minimize the above equation 2.2-5.

The concept is the same for multiple regression analysis

Appendix

In case, multi regression analysis,

- Relational equation of m explanatory variables $\{x_1, x_2, x_3, \ldots, x_m\}$ and response variable y

$$y = a_o + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m \qquad \text{Equation 2.2-6}$$

- Residual (Difference between measured value and estimated value by single regression equation)

$$y_{(i)} - \hat{y}_{(i)} = y_{(i)} - (a_o + a_1 x_{1(i)} + a_2 x_{2(i)} + \cdots + a_m x_{m(i)}) \qquad \text{Equation 2.2-7}$$

, and sum of square of each residual,

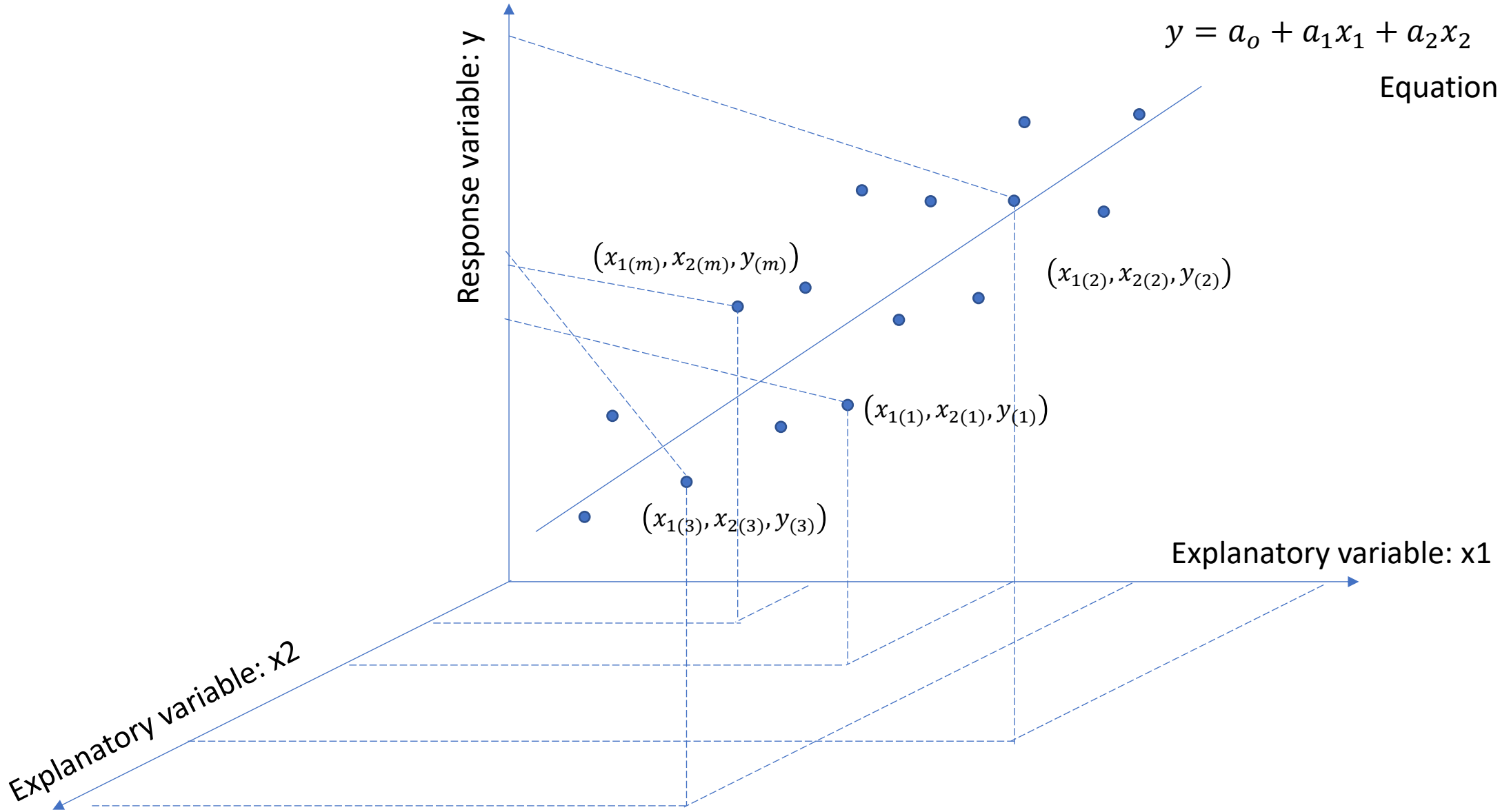
$$\sum_i^n \{y_{(i)} - (a_o + a_1 x_{1(i)} + a_2 x_{2(i)} + \cdots + a_m x_{m(i)})\}^2 \qquad \text{Equation 2.2-8}$$

$a_0, a_1, \ldots, a_m$ is calculated to minimize the above equation 2.2-8.

Case) Multi regression analysis with 2 explanatory-variables

$$y = a_o + a_1 x_1 + a_2 x_2$$

Equation 2.2-9



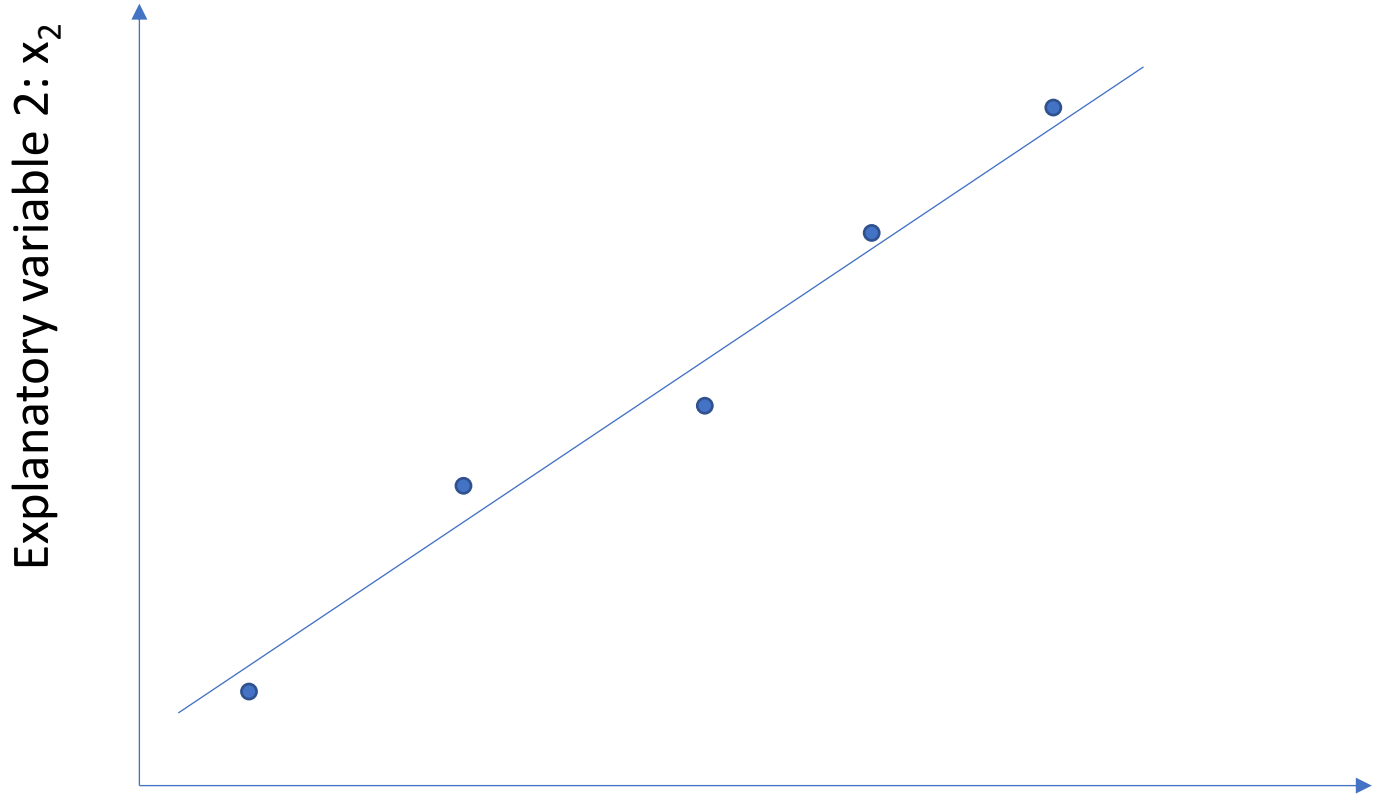
◆ Consideration of multi regression analysis

- Multicollinearity
- Effect of the number of explanatory variables on the coefficient of determination

What does situation of "multicollinearity exists" mean?

For example, when multiple regression analysis with 5 explanatory variables $\{x_1, x_2, x_3, x_4, x_5\}$, if x_1 and x_2 are under correlation,

x_2 is necessary for this multiple regression analysis, or not?



Appendix

◆ Situation of "multicollinearity exists"

$$S_{11}S_{22} - S_{12}^2 = 0 \quad \Leftrightarrow \quad \frac{S_{12}^2}{S_{11}S_{22}} = 1$$

$$\Leftrightarrow \quad r_{x1x2}^2 = \frac{S_{12}}{\sqrt{S_{11}S_{22}}} = 1$$

$$\Leftrightarrow \quad r_{x1x2} = \pm 1$$

Equation 2.2-10

When correlation coefficient of x_1 and x_2 is 1 or -1,

this means that "multicollinearity exists".

- $r_{x1x2} = \pm 1$ means that when either x_1 or x_2 is decided, the other is decided from a linear relationship.
- In other words, from the perspective of explaining the response variable y, if either x_1 or x_2 is known, information on the other is no longer necessary.
- The more easily interpretable of either x_1 or x_2 should be included in the model.

multiple correlation coefficient

✓ In multiple regression analysis, as in the case of single regression analysis, so that the regression equation obtained is valid, the measured value y_i and the estimated value $\hat{y}_i$ should fit better,

correlation coefficient of $(y_i, \hat{y}_i)$

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

Equation 2.2-11

is calculated and can be used for evaluation of regression equation.

✓ There is the following relationship between “Multiple Correlation Coefficient” and “Coefficient of Determination”,

“Square of Multiple Correlation Coefficient” = “Coefficient of Determination”

Coefficient of Determination (or Contribution Ratio)

Appendix

The relationship between the sum square of total variation S_T, the sum square of regressionS_R、 and the sum square of error S_e is expressed by the following equation, as same with the single regression analysis.

$$S_T = S_R + S_e \qquad \text{Equation 2.2-12}$$

$$1 = \frac{S_R}{S_T} + \frac{S_e}{S_T} \qquad \text{Equation 2.2-13}$$

Since the regression analysis aims to reduce the sum of squares of the errors,

$$\frac{S_R}{S_T} = \frac{\text{Sum square of regression}}{\text{Sum square of measurements}} \qquad \text{Equation 2.2-14}$$

The closer A is to 1, the "better the multiple regression equation fits".

Appendix

$$R^2 = \frac{S_R}{S_T}\left(= \frac{S_T - S_e}{S_T} = 1 - \frac{S_e}{S_T}\right)$$

Equation 2.2-15

is defined as “Coefficient of Determination” (or “contribution ratio”)

The contribution ratio explains the proportion of the variation in y that is due to the variation in the regression.

Appendix

◆ Adjusted coefficient of determination

- In multiple regression analysis, the property is that as the number of explanatory variables increases, the coefficient of determination (contribution ratio) of equation 2.2-15 automatically increases.

- In other words, when comparing “the contribution ratio when two explanatory variables are used” with “the contribution ratio when three explanatory variables are used”, the latter always has a larger contribution ratio.

- In other words, it is possible and not desirable to intentionally increase the apparent contribution ratio by adding completely meaningless explanatory variables.

- Instead of just a ratio of the sum squares as in Equation 2.2-15, A penalty that reduces the significance of the multiple regression equation is imposed,

  - This penalty is imposed by increasing the number of explanatory variables. As a result, the significance of the multiple regression equation is reduced. This penalty is imposed by adjusting using the degree of freedom

$$R^{*2} = 1 - \frac{S_e/\phi_e}{S_T/\phi_T}$$

$$= 1 - \frac{S_e/(n-p-1)}{S_T/(n-1)}$$

n: number of measured value
p: number of explanatory variable

Equation 2.2-16

This is called "Adjusted Coefficient of Determination"

  - There is the following relationship between coefficient of determination R^2 and adjusted coefficient of determination R^{*2},

$$R^{*2} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Equation 2.2-17

Double-adjusted Coefficient of Determination”

Appendix

Unfortunately, as the number of explanatory variables increases, the adjusted coefficient of determination also increases.

- Further penalized, “double-adjusted coefficient of determination”

$$R^{**2} = 1 - \frac{(n+p+1)S_e/(n-p-1)}{(n+1)S_T/(n-1)}$$

Equation 2.2-18

is also used.

- This “double-adjusted coefficient of determination” can be derived from the optimal combination of explanatory variables,

  with the coefficient of determination peaking at a certain optimal combination of explanatory variables and decreasing as more explanatory variables are added.

Appendix

◆Analysis of Variance for Multi Regression Analysis

Methods for confirming the significance of multi regression equations

There is “Test of a multi regression by ANOVA”.

ANOVA: analysis of variance

Variation factor	Sum square	Degree of freedom	Mean square	F-value
Variation by regression	$S_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$\phi_R = p$	$V_R = \frac{S_R}{\phi_R}$	$F_0 = \frac{V_R}{V_e}$
Variation by residual	$S_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$\phi_e = n - p - 1$	$V_e = \frac{S_e}{\phi_e}$	
Total variation	$S_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$\phi_T = n - 1$		

$$S_T = S_R + S_e$$

n: the number of data

p: the number of explanatory variables of regression analysis

Equation 2.2-12

Appendix

This ANOVA tests

null hypothesis H_0: Multi regression equations are not useful for estimation.

,using a test statistic F_0.

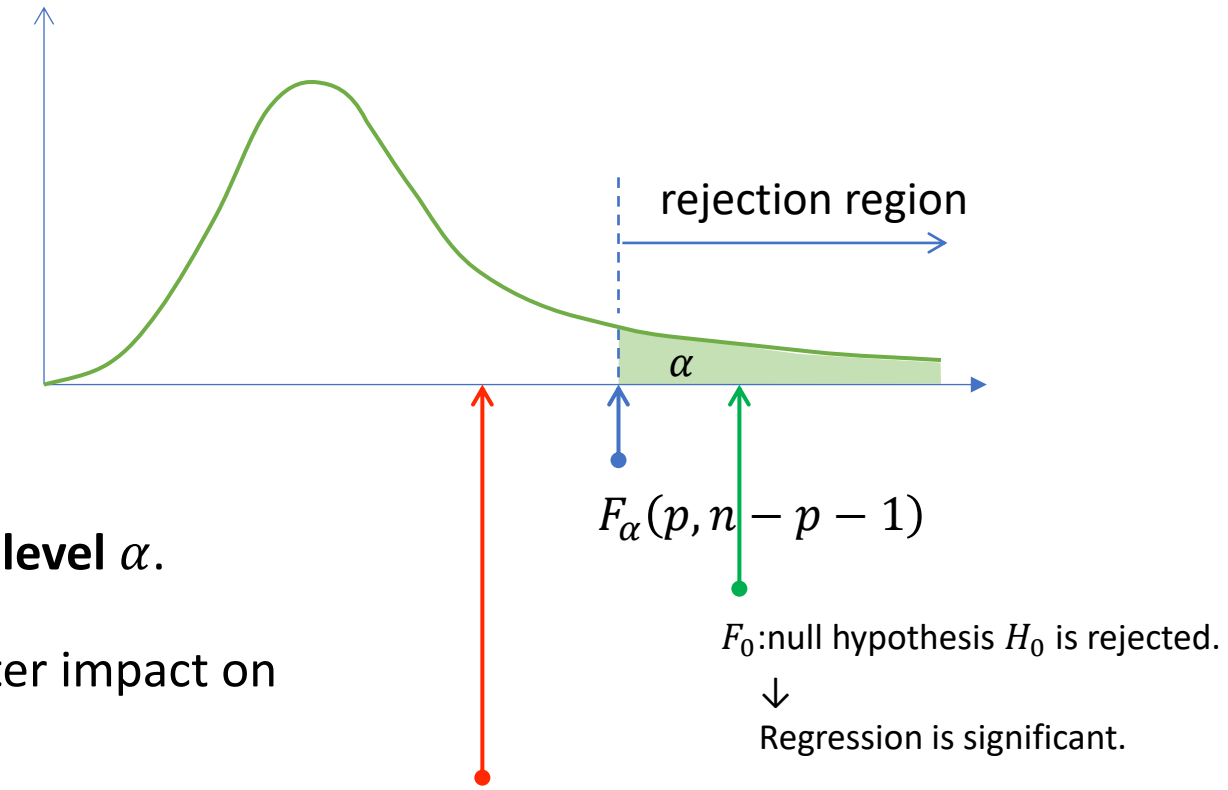
Therefore, "**rejecting the null hypothesis H_0**" is meaningful in this test.

Since this test statistic follows F-distribution of degree of freedom$(p,\ n-p-1)$,

if

$$F_0 \geq F_\alpha(p, n-p-1)$$

The **null hypothesis** H_0 is rejected by **significance level** α.



This means that "Variation by regression has a greater impact on total variation than variation by residuals."

=

variation by regression >> variation by residuals

⇓

i.e., the multi regression equation is significant.

Appendix 3

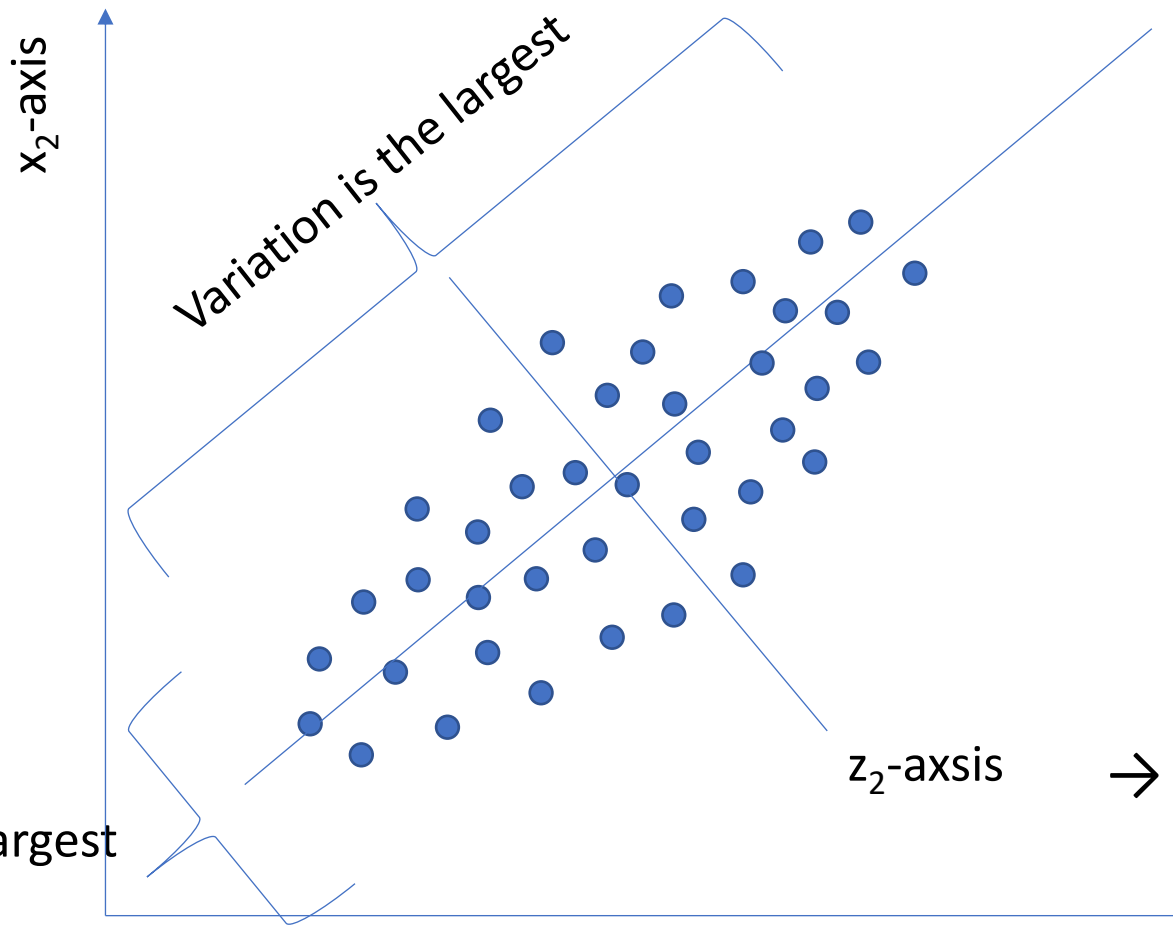
Summary of Principal Component Analysis

Principal Component Analysis (PCA)

What is principal component analysis (PCA) ?

- One of multivariate analysis
- A method that summarizes multiple quantitative explanatory variables into a smaller number of indicators called “principal component”
- The axes are set in order of orientation of dispersion, from largest to smallest: first principal component, second principal component, ...
- Reduced to the same number of principal components as the number of data series

Appendix

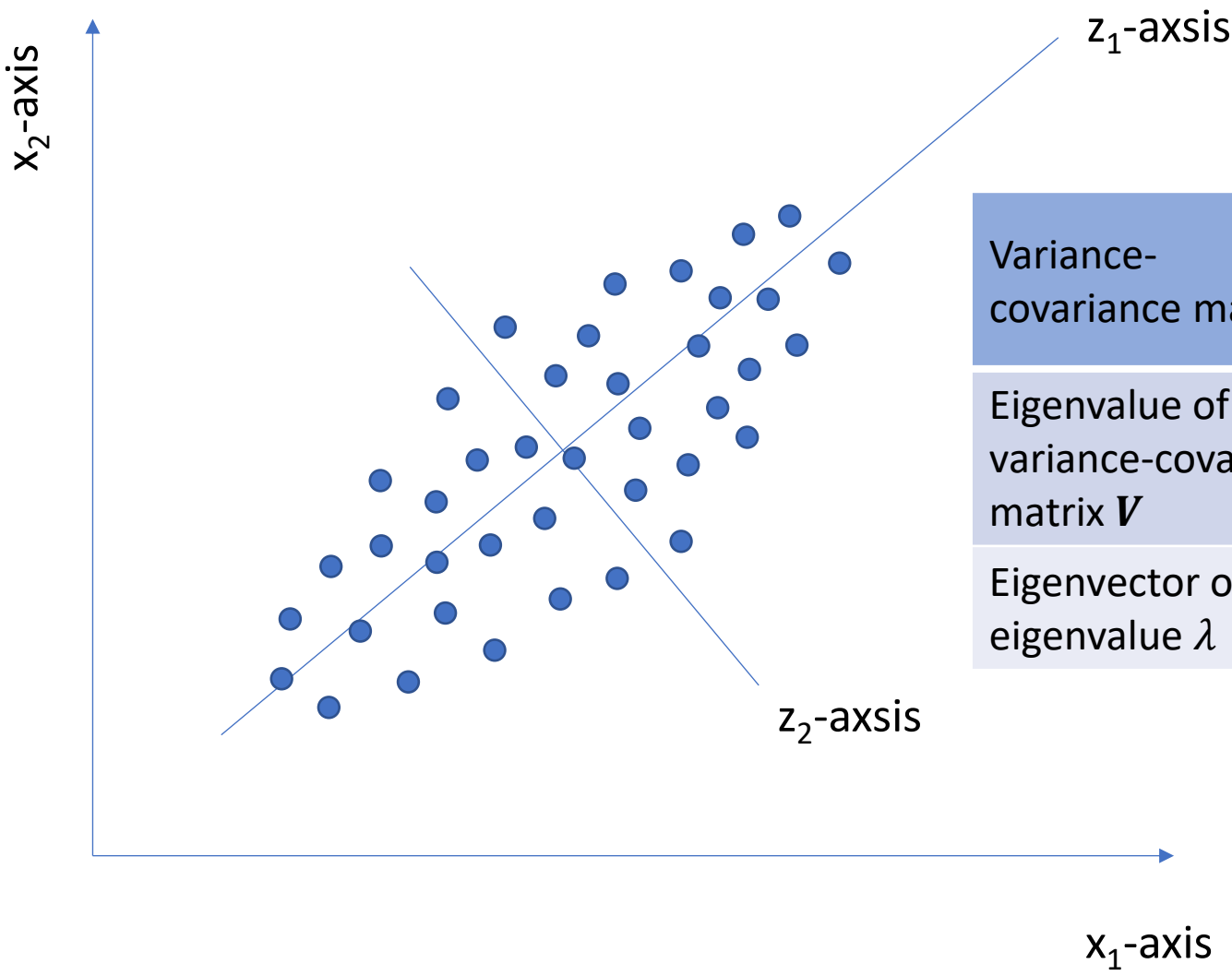


→ Axis with the largest variation

Each axis is orthogonal to each other.

→ Axis with the second largest variation

Appendix



Variance-covariance matrix	$\boldsymbol{V}$	$\begin{pmatrix} V[X_1] & Cov[X_1, X_2] \\ Cov[X_2, X_1] & V[X_2] \end{pmatrix}$
Eigenvalue of variance-covariance matrix $\boldsymbol{V}$	λ	
Eigenvector of eigenvalue λ	$\boldsymbol{v}$	

Appendix

$$\boldsymbol{V}\boldsymbol{v} = \lambda\boldsymbol{v}$$

$$\begin{pmatrix} V[X_1] & Cov[X_1, X_2] \\ Cov[X_2, X_1] & V[X_2] \end{pmatrix} \boldsymbol{v} = \lambda\boldsymbol{v}$$

From this, the following eigenvectors and eigenvalues are obtained

$$(\boldsymbol{v_1}, \boldsymbol{v_2}) \quad (\lambda_1, \lambda_2)$$

where, λ_1 : eigenvalue for eigenvector v_1

λ_2 : eigenvalue for eigenvector v_2

and, $\lambda_1 > \lambda_2$

Appendix

Here,

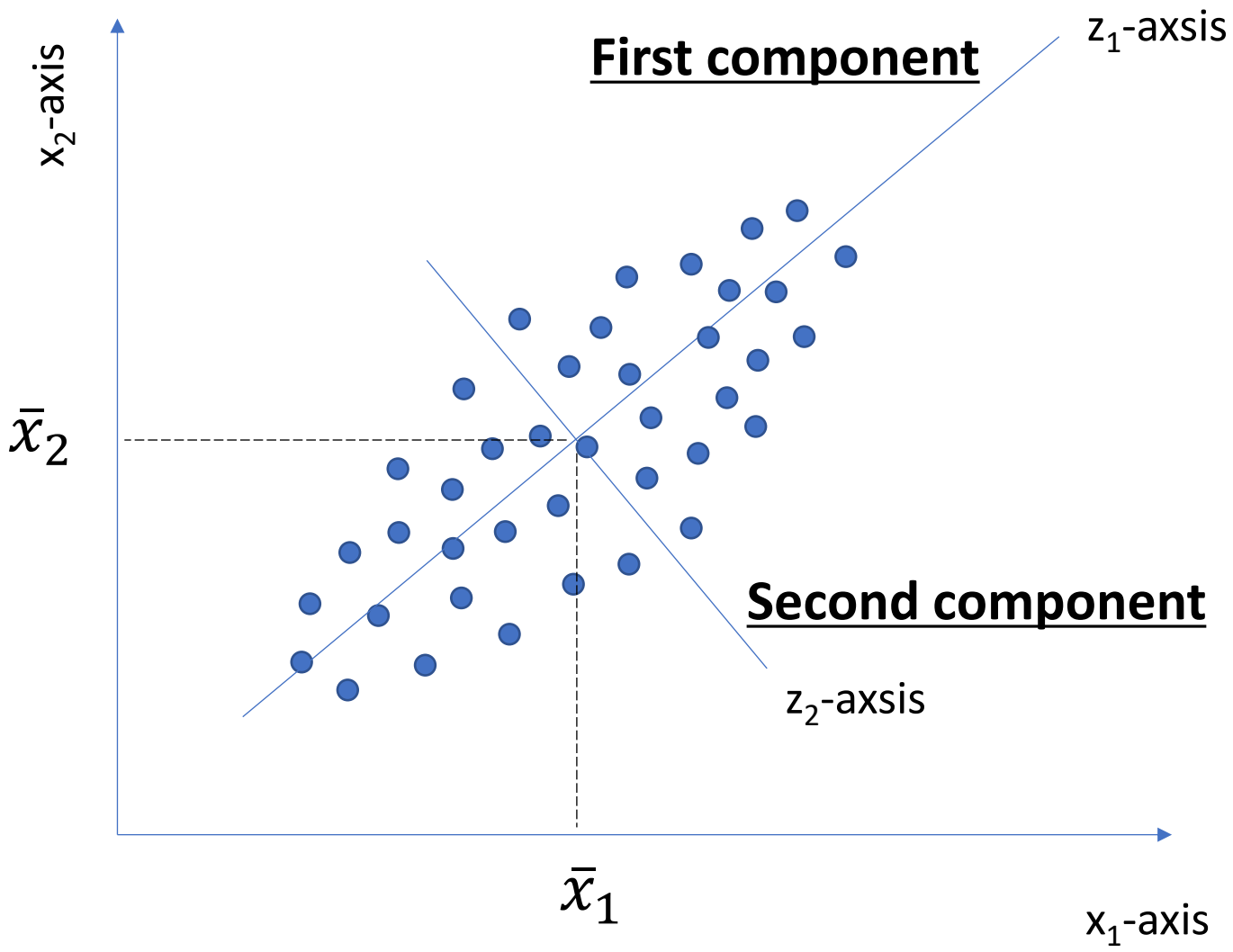
Eigenvector v_1 expresses an orientational vector of the new axis z_1.

Eigenvector v_2 expresses an orientational vector of the new axis z_2.

Eigenvalue λ_1 expresses a variation size for an orientation of the new axis z_1.

Eigenvalue λ_2 expresses a variation size for an orientation of the new axis z_2.

◆ Principal component score (PCS)

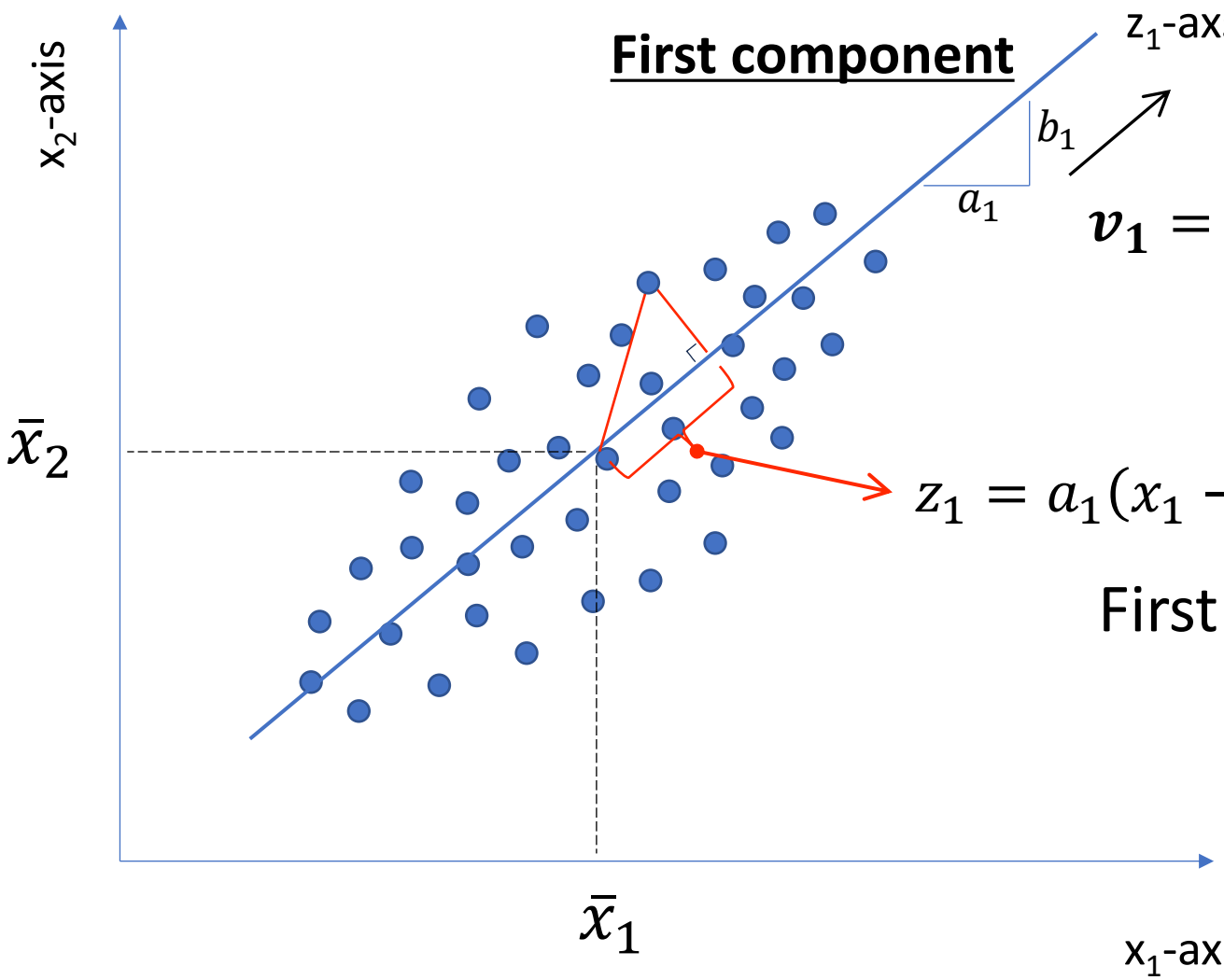


$\bar{x}_1$: Mean of x_1

$\bar{x}_2$: Mean of x_2

Principal component score (Cont.)

Appendix



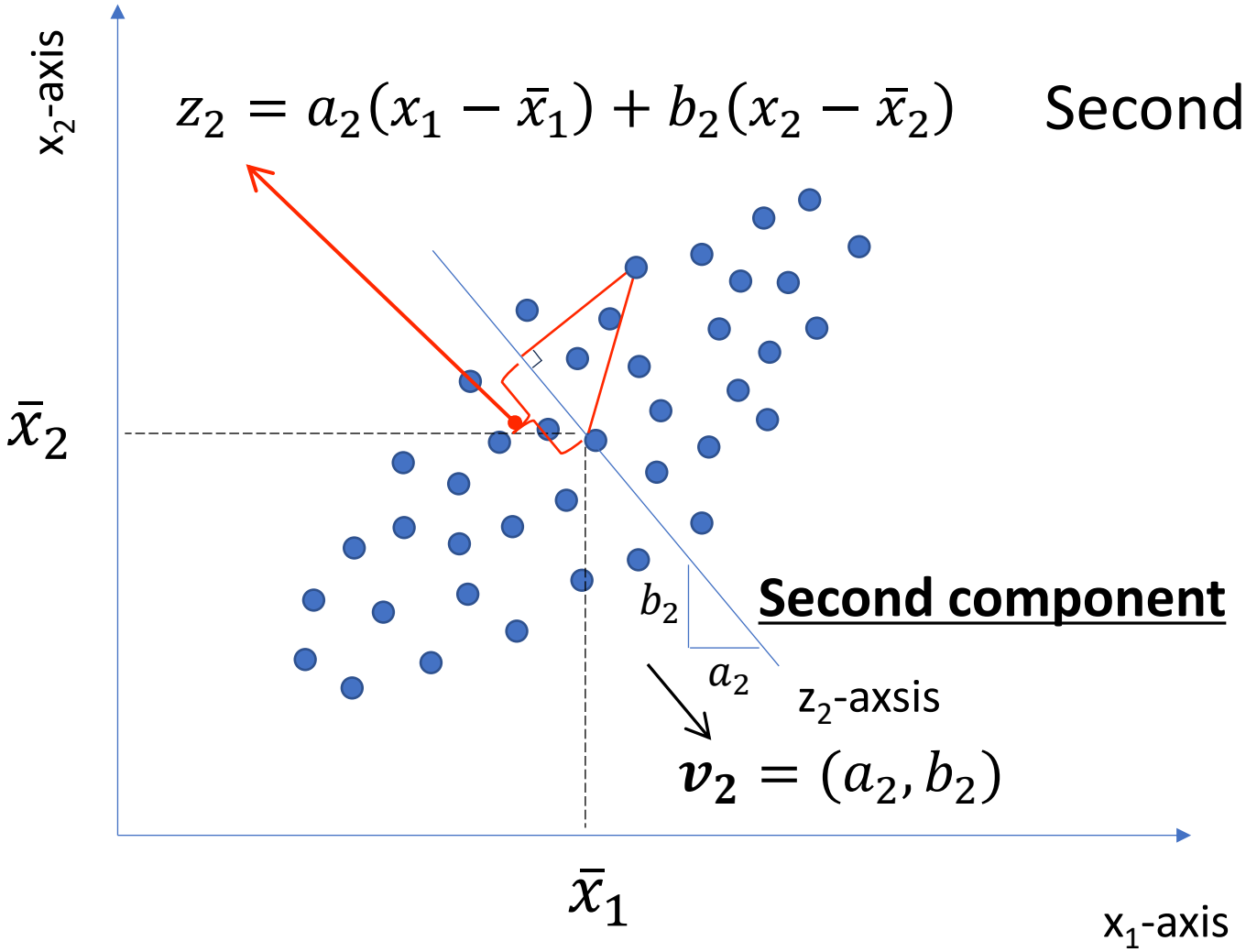
$$\boldsymbol{v_1} = (a_1, b_1)$$

$$z_1 = a_1(x_1 - \bar{x}_1) + b_1(x_2 - \bar{x}_2)$$

First principal component score

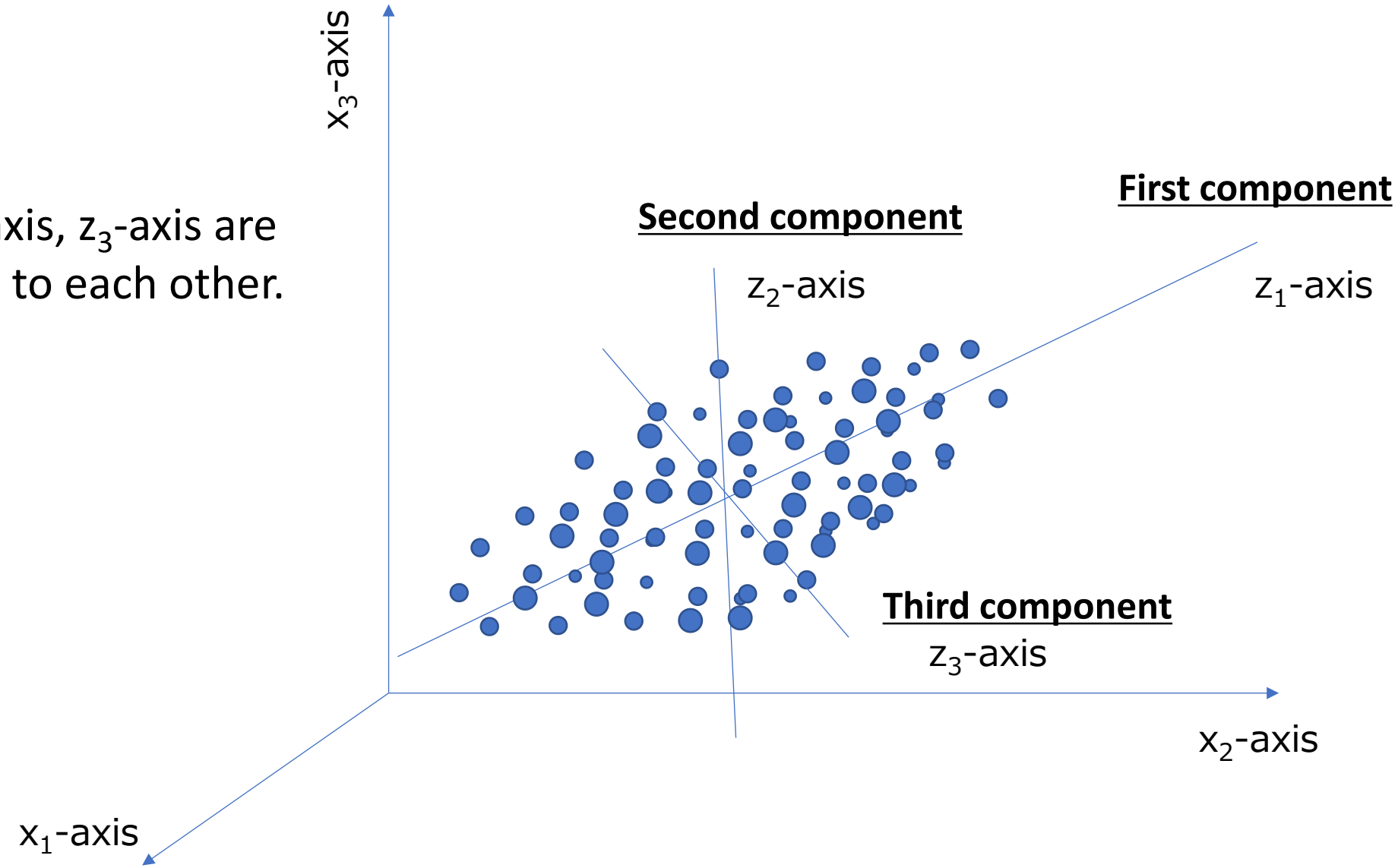
Principal component score (Cont.)

Appendix



◆ Principal Component Analysis with 3 variables

z_1-axis, z_2-axis, z_3-axis are orthogonal to each other.



◆ Contribution ratio

Appendix

An eigenvalue of each principal component divided by its sum is called the contribution ratio.

$$\text{Contribution ratio of first component} = \frac{\text{Eigenvalue of first component}}{\text{Sum of all eigenvalues}} = \frac{\lambda_1}{\lambda_1+\lambda_2}$$

$$\text{Contribution ratio of second component} = \frac{\text{Eigenvalue of second component}}{\text{Sum of all eigenvalues}} = \frac{\lambda_2}{\lambda_1+\lambda_2}$$

- It indicates how much of the total variation is accounted for by each principal component, with larger values indicating principal components with higher relative explanatory capability.

For example, when $\lambda_1 = 57.94$, $\lambda_2 = 26.95$,

Contribution ratio of first component is **57.94 / (57.94 + 26.95) = 0.68**

then, this means that the first component has 68% of the information.

For 4 or more variables, similarly, transformed into components around n mutually orthogonal axes,

$$\boldsymbol{V}\boldsymbol{v} = \lambda\boldsymbol{v}$$

$$\begin{pmatrix} V[X_1] & Cov[X_1,X_2] & \cdots & Cov[X_1,X_{n-1}] & Cov[X_1,X_n] \\ Cov[X_2,X_1] & V[X_2] & & Cov[X_2,X_{n-1}] & Cov[X_2,X_n] \\ \vdots & & \ddots & \vdots & \\ Cov[X_{n-1},X_1] & Cov[X_{n-1},X_2] & & V[X_{n-1}] & Cov[X_{n-1},X_n] \\ Cov[X_n,X_1] & Cov[X_n,X_2] & \cdots & Cov[X_n,X_{n-1}] & V[X_n] \end{pmatrix} \boldsymbol{v} = \lambda\boldsymbol{v}$$

From the above equation, the following eigenvectors and eigenvalues are obtained,

$$(\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_n}) \quad (\lambda_1, \lambda_2, \cdots, \lambda_n)$$

and, $\lambda_1 > \lambda_2 > \cdots > \lambda_n$

Appendix

First PCS $z_1 = a_{1(1)}(x_1 - \bar{x}_1) + a_{2(1)}(x_2 - \bar{x}_2) + a_{3(1)}(x_3 - \bar{x}_3) + \cdots + a_{n(1)}(x_n - \bar{x}_n)$

$\boldsymbol{v_1} = (a_{1(1)}, a_{2(1)}, a_{3(1)}, , , , , a_{n(1)})$: eigenvector of the first component

Second PCS $z_2 = a_{1(2)}(x_1 - \bar{x}_1) + a_{2(2)}(x_2 - \bar{x}_2) + a_{3(2)}(x_3 - \bar{x}_3) + \cdots + a_{n(2)}(x_n - \bar{x}_n)$

$\boldsymbol{v_2} = (a_{1(2)}, a_{2(2)}, a_{3(2)}, , , , , a_{n(2)})$: eigenvector of the second component

Third PCS $z_3 = a_{1(3)}(x_1 - \bar{x}_1) + a_{2(3)}(x_2 - \bar{x}_2) + a_{3(3)}(x_3 - \bar{x}_3) + \cdots + a_{n(3)}(x_n - \bar{x}_n)$

$\boldsymbol{v_3} = (a_{1(3)}, a_{2(3)}, a_{3(3)}, , , , , a_{n(3)})$: eigenvector of the third component

.
.
.

n-th PCS $z_n = a_{1(n)}(x_1 - \bar{x}_1) + a_{2(n)}(x_2 - \bar{x}_2) + a_{3(n)}(x_3 - \bar{x}_3) + \cdots + a_{n(n)}(x_n - \bar{x}_n)$

$\boldsymbol{v_n} = (a_{1(3)}, a_{2(3)}, a_{3(3)}, , , , , a_{n(n)})$: eigenvector of the n-th component

PCS: Principal component score

$\bar{x}_1$: mean of x_1 $\bar{x}_2$: mean of x_2 $\bar{x}_3$: mean of x_3 $\bar{x}_n$: mean of x_n

Appendix

Up to what number of principal components are used in the evaluation?

$$\text{Contribution ratio of first component} = \frac{\text{Eigenvalue of first component}}{\text{Sum of all eigenvalues}} = \frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\cdots+\lambda_n}$$

$$\text{Contribution ratio of second component} = \frac{\text{Eigenvalue of second component}}{\text{Sum of all eigenvalues}} = \frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3+\cdots+\lambda_n}$$

$$\vdots$$

$$\text{Contribution ratio of n-th component} = \frac{\text{Eigenvalue of n-th component}}{\text{Sum of all eigenvalues}} = \frac{\lambda_n}{\lambda_1+\lambda_2+\lambda_3+\cdots+\lambda_n}$$

The cumulative contribution ratio is one of indicators.

Appendix

For example, $\lambda_1 = 57.94$, $\lambda_2 = 26.95$, $\lambda_3 = 12.34$, $\lambda_4 = 8.65$, $\lambda_5 = 5.27$,

Contribution ratio of first component is 57.94 / (57.94 + 26.95 + 12.34 + 8.65 + 5.27) = 0.521

Contribution ratio of second component is 26.95 / (57.94 + 26.95 + 12.34 + 8.65 + 5.27) = 0.242

Contribution ratio of third component is 12.34 / (57.94 + 26.95 + 12.34 + 8.65 + 5.27) = 0.111

Contribution ratio of fourth component is 8.65 / (57.94 + 26.95 + 12.34 + 8.65 + 5.27) = 0.078

Contribution ratio of fifth component is 5.27 / (57.94 + 26.95 + 12.34 + 8.65 + 5.27) = 0.048

- The cumulative contribution ratio are

Up to the first : 0.521

Up to the second : 0.521 + 0.242 = 0.763

Up to the third : 0.521 + 0.242 + 0.111 = 0.874

Up to the fourth : 0.521 + 0.242 + 0.111 + 0.078 = 0.952

Up to the fifth : 0.521 + 0.242 + 0.111 + 0.078 + 0.048 = 1.000

For example, “exceeding 80%” is selected as an indicator,

- By up to the third component are accumulated, the cumulative contribution ratio exceeds 80%.
- Then, up to the third component is used for evaluation of this analysis.

◆ How to calculate eigenvalue and eigenvector

$$\boldsymbol{V}\boldsymbol{v} = \lambda\boldsymbol{v} \qquad \boldsymbol{v} \neq \boldsymbol{0}$$

$$(\lambda\boldsymbol{I} - \boldsymbol{V})\boldsymbol{v} = \boldsymbol{0}$$

$$\boldsymbol{v} \neq \boldsymbol{0}$$

$$\boldsymbol{I} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$ Unit matrix

to do this,

$$det(\lambda\boldsymbol{I} - \boldsymbol{V}) = 0$$

is necessary.

Example)

To calculate eigenvalues and eigenvectors of the following matrix,

$$\boldsymbol{V} = \begin{pmatrix} 4 & 1 \\ -2 & 1 \end{pmatrix}$$

$$det\left(\lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 4 & 1 \\ -2 & 1 \end{pmatrix}\right) = 0$$

$$det\begin{pmatrix} \lambda - 4 & -1 \\ 2 & \lambda - 1 \end{pmatrix} = 0$$

$$\Leftrightarrow \lambda^2 - 5\lambda + 6 = 0$$

$$\Leftrightarrow \lambda = 2, 3$$

Appendix

Case, $\lambda = 2$

$$(\lambda \boldsymbol{I} - \boldsymbol{V})\boldsymbol{v} = \boldsymbol{0}$$

Substitute $\lambda = 2$ and $\boldsymbol{V} = \begin{pmatrix} 4 & 1 \\ -2 & 1 \end{pmatrix}$ into the above equation,

$$\left(2\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 4 & 1 \\ -2 & 1 \end{pmatrix}\right)\boldsymbol{v_1} = \boldsymbol{0}$$

$$\begin{pmatrix} -2 & -1 \\ 2 & 1 \end{pmatrix}\boldsymbol{v_1} = \boldsymbol{0}$$

and, determined as $\boldsymbol{v_1} = (a_1, b_1)$

Appendix

$$\begin{pmatrix} -2a_1 & -b_1 \\ 2a & b_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} -2a_1 - b_1 = 0 \\ 2a_1 + b_1 = 0 \end{cases}$$

$$\begin{cases} a_1 = -\frac{1}{2}k \\ b_1 = k \end{cases}$$ (k is real number)

Eigenvector, which belongs to eigenvalue $\lambda = 2$ is $\begin{pmatrix} -\frac{1}{2} \\ 1 \end{pmatrix}$.

Appendix

Case, $\lambda = 3$

$$(\lambda \boldsymbol{I} - \boldsymbol{V})\boldsymbol{v} = \boldsymbol{0}$$

Substitute $\lambda = 3$ and $\boldsymbol{V} = \begin{pmatrix} 4 & 1 \\ -2 & 1 \end{pmatrix}$ into the above equation,

$$\left(3\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 4 & 1 \\ -2 & 1 \end{pmatrix}\right)\boldsymbol{v_2} = \boldsymbol{0}$$

$$\begin{pmatrix} -1 & -1 \\ 2 & 2 \end{pmatrix}\boldsymbol{v_2} = \boldsymbol{0}$$

and, determined as $\boldsymbol{v_2} = (a_2, b_2)$

Appendix

$$\begin{pmatrix} -a_2 & -b_2 \\ 2a_2 & 2b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} -a_2 - b_2 = 0 \\ 2a_2 + 2b_2 = 0 \end{cases}$$

$$\begin{cases} a_2 = -k \\ b_2 = k \end{cases}$$ (k is real number)

Eigenvector, which belongs to eigenvalue $\lambda = 3$ is $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

◆ Points for principal component analysis

Appendix

- Each principal component calculated by principal component analysis is an objective quantity.

  It is necessary for the analyst to interpret the principal component analysis results.

  Such as, What does each principal component mean?

  It is not always possible to interpret the meaning of each principal component.

- When the units of each variable, etc. are different, principal component analysis is performed using a correlation matrix instead of a variance-covariance matrix.

$$\text{variance-covariance matrix} = \begin{pmatrix} \sigma 1 & C12 & C13 & C1n \\ C21 & \sigma 2 & C23 & C2n \\ C31 & C32 & \sigma 3 & C3n \\ Cn1 & Cn2 & Cn3 & \sigma n \end{pmatrix}$$

$$\text{correlation matrix} = \begin{pmatrix} 1 & r12 & r13 & r1n \\ r21 & 1 & r23 & r2n \\ r31 & r32 & 1 & r3n \\ rn1 & rn2 & rn3 & 1 \end{pmatrix}$$

σ: variance
C: Covariance

r: correlation coefficient

Appendix

◆ Utilization of Principal Component Analysis

- Each principal component obtained by principal component analysis can be subjected to multiple regression analysis.

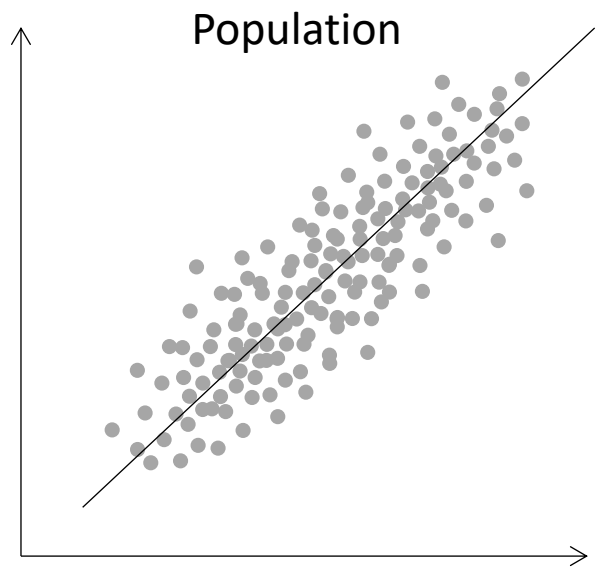
  -> Principal component regression (PCR)

  ➢ Since each principal component is independent of each other, there is no risk of multicolinearity.

Appendix 4

Uncertainty Evaluation using Single Regression Analysis

Appendix



For example, case of single regression analysis,

Results of regression analysis of data (x_i, y_i) is given by the following equation,

$$y = \beta_1 x + \beta_0 \qquad \text{Equation 5-1}$$

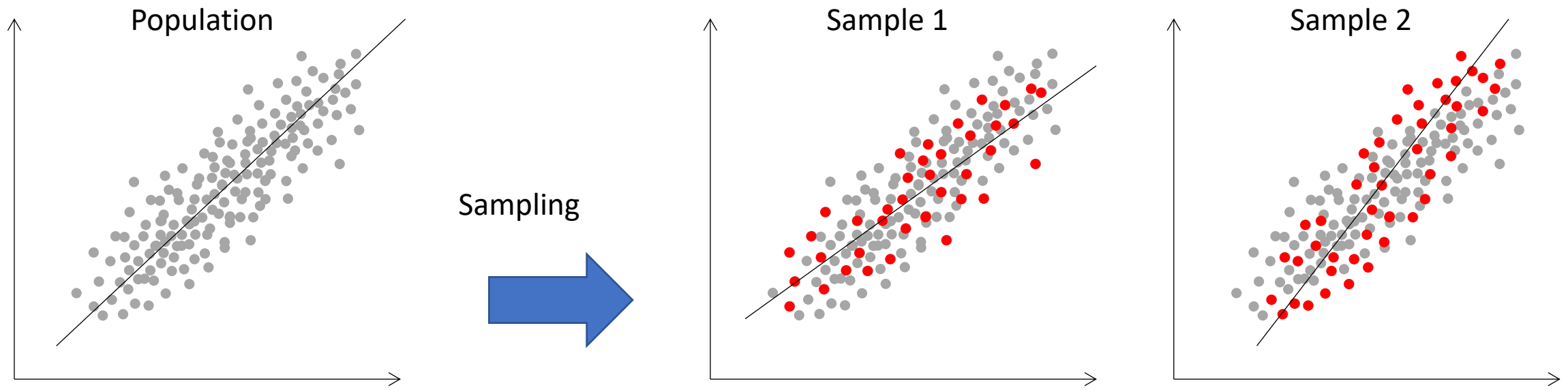
However, the above equation and values of β_1 and β_0 give true value,

Regression equation given by least square method is estimate, expressed by the following,

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0 \qquad \text{Equation 5-2}$$

$\hat{\beta}_1$, $\hat{\beta}_0$ are respectively, estimated value of true value of β_1, β_0.

Appendix



The regression equation given the sampled data (sample) will vary from sample to sample and will not yield exactly the same results.

Distribution $\hat{\beta}_1$ follows $N\left(\beta_1, \quad \frac{\sigma^2}{S_{xx}}\right)$ Equation 5-3

$$S_{xx} = \sum_i (x_i - \bar{x})^2$$ Equation 5-5

Distribution $\hat{\beta}_0$ follows $N\left(\beta_0, \quad \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$ Equation 5-4

Since $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ is calculated from $\hat{\beta}_0, \hat{\beta}_1, \hat{y}$ is also estimated value.

Appendix

Distribution $\hat{y}$ follows $N\left(\beta_1 x + \beta_0 , \quad \left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right\}\sigma^2\right)$ $S_{xx} = \sum_i (x_i - \bar{x})^2$

Equation 5-6

Since x is included in $\hat{y}$, It can be seen that as the value of x changes, the variance of $\hat{y}$ also changes.

Derivation of variance of $\hat{y}$ (1)

Derivation of variance of $\hat{\beta}_1$, variance of $\hat{\beta}_0$, variance of $\hat{y}$

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$$ Equation 5-2

$$V(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} = ({\sigma_{\beta 1}}^2)$$ Equation 5-7

$$V(\hat{\beta}_0) = \frac{\sigma^2 \sum {x_i}^2}{n\sum(x_i - \bar{x})^2} = ({\sigma_{\beta 0}}^2)$$ Equation 5-8

$$\sigma_{\beta 0} = \sigma\sqrt{\frac{\sum {x_i}^2}{n\sum(x_i - \bar{x})^2}} = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$ Equation 5-9

Appendix

$$\sigma_y{}^2 = \left(\frac{\partial y}{\partial \hat{\beta}_0}\right)^2 \sigma_{\beta 0}{}^2 + \left(\frac{\partial y}{\partial \hat{\beta}_1}\right)^2 \sigma_{\beta 1}{}^2 + \left(\frac{\partial y}{\partial x}\right)^2 \sigma_x{}^2$$
$$+2\left(\frac{\partial y}{\partial \hat{\beta}_0}\right)\left(\frac{\partial y}{\partial \hat{\beta}_1}\right) C_{\beta 0 \cdot \beta 1} + 2\left(\frac{\partial y}{\partial \hat{\beta}_0}\right)\left(\frac{\partial y}{\partial x}\right) C_{\beta 0 \cdot x} + +2\left(\frac{\partial y}{\partial \hat{\beta}_1}\right)\left(\frac{\partial y}{\partial x}\right) C_{\beta 1 \cdot x}$$

Equation 5-10

$$\frac{\partial y}{\partial \hat{\beta}_0} = 1 \qquad \frac{\partial y}{\partial \hat{\beta}_1} = x$$

$\sigma_x{}^2 = 0$ $\hat{\beta}_0$ and x are independent of each other.

$\hat{\beta}_1$ and x are independent of each other.

From these,

$$\sigma_y{}^2 = \sigma_{\beta 0}{}^2 + x^2 \sigma_{\beta 1}{}^2 + 2x\, C_{\beta 0 \cdot \beta 1}$$

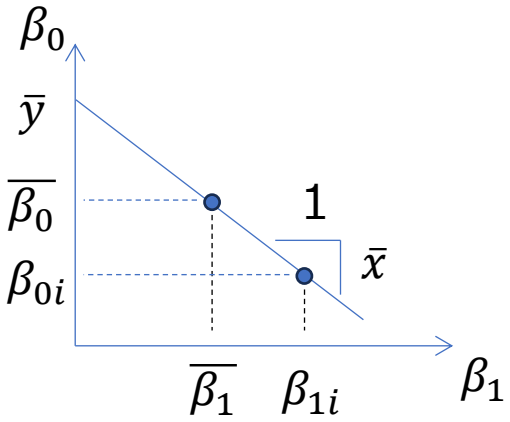
Equation 5-11

Appendix

Relationship between $\hat{\beta}_0$ and $\hat{\beta}_1$ is

$$\hat{\beta}_0 = -\hat{\beta}_1 \bar{x} + \bar{y}$$

Equation 5-12



$$C_{\beta 0 \cdot \beta 1} = \frac{\sum\{(\beta_{0i} - \overline{\beta_0})(\beta_{1i} - \overline{\beta_1})\}}{n}$$

$$= \frac{-\sum\left\{(\beta_{1i} - \overline{\beta_1})^2 \bar{x}\right\}}{n}$$

$$= -\frac{\bar{x}\sum(\beta_{1i} - \overline{\beta_1})^2}{n}$$

$$= -\bar{x} \cdot {\sigma_{\beta 1}}^2 \qquad = -\frac{\bar{x}\sigma^2}{\sum(x_i - \bar{x})^2}$$

Equation 5-13

$$\beta_{0i} - \overline{\beta_0} : \beta_{1i} - \overline{\beta_1} = -\bar{x} : 1$$

$$-\bar{x}(\beta_{1i} - \overline{\beta_1}) = \beta_{0i} - \overline{\beta_0}$$

Then,

$${\sigma_y}^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} + \frac{x^2}{\sum(x_i - \bar{x})^2} - \frac{2x\bar{x}}{\sum(x_i - \bar{x})^2}\right) = \sigma^2\left\{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right\}$$

Equation 5-14

Derivation of variance of $\hat{y}$ (2)

Derivation of variance of $\hat{\beta}_1$, variance of $\hat{\beta}_0$, variance of $\hat{y}$

$$y = \hat{\beta}_1(x - \bar{x}) + \bar{y}$$ Equation 5-15

$$\sigma_{\beta 1}{}^2 = \sum_{i=1}^{n} \left(\frac{\partial \beta_1}{\partial y_i}\right)^2 \sigma^2$$ Equation 5-16

$$\frac{\partial \beta_1}{\partial y_i} = \frac{1}{\Delta}\left(nx_i - \sum x_i\right)$$ Equation 5-17

$$\Delta = n\sum x_n{}^2 - \left(\sum x_i\right)^2 = n\left\{\sum (x_i - \bar{x})^2\right\}$$ Equation 5-18

$$\sigma_{\beta 1}{}^2 = \frac{n\sigma^2}{n\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$ Equation 5-19

Appendix

$$\sigma_y{}^2 = \left(\frac{\partial y}{\partial \hat{\beta}_1}\right)^2 \sigma_{\beta 1}{}^2 + \left(\frac{\partial y}{\partial \bar{x}}\right)^2 \sigma_{\bar{x}}{}^2 + \left(\frac{\partial y}{\partial \bar{y}}\right)^2 \sigma_{\bar{y}}{}^2 + \left(\frac{\partial y}{\partial x}\right)^2 \sigma_x{}^2$$

Equation 5-20

$$\frac{\partial y}{\partial \hat{\beta}_1} = x - \bar{x} \qquad \frac{\partial y}{\partial \bar{x}} = -\hat{\beta}_1 \qquad \frac{\partial y}{\partial \bar{y}} = 1 \qquad \frac{\partial y}{\partial x} = 1$$

$$\sigma_x{}^2 = 0 \qquad \sigma_{\bar{x}}{}^2 = 0 \qquad \sigma_{\bar{y}}{}^2 = \frac{\sigma^2}{n}$$

From these,

$$\sigma_y{}^2 = (x - \bar{x})^2 \sigma_{\beta 1}{}^2 + \frac{\sigma^2}{n}$$

$$= (x - \bar{x})^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} + \frac{\sigma^2}{n}$$

$$= \sigma^2 \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}$$

Equation 5-21

◆ Unbiased variance

In general, variance σ^2 of measurement value is unknown.

Due to this, unbiased variance is obtained from measurement values.

When defined unbiased variance as u^2, $u^2 \fallingdotseq \sigma^2$

$$u^2 = \frac{1}{n-2}\sum_{i=1}^{n}\{y_i - y(x_i)\}^2$$

Equation 5-22

y_i is measurement value, $y(x_i)$ is estimated value of y_i obtained from regression equation.

Appendix

n-2 in denominator of equation 5-22 is “degree of freedom of residual”.

And “degree of freedom of residual” is obtained as (**n-1-p**) by subtracting “degree of freedom of regression **p** ” from “Degree of freedom of total variation (**n-1**) “

n: total number of data

$$(n-1-p) = (n-1) - p$$

Equation 5-23

Degree of freedom of total variation

degree of freedom of regression

In case, since “**p**” of single regression is (**p=1**),

Degree of freedom of residual on single regression analysis is “**n-2**”.

Interval estimation of single regression analysis

- Interval estimation of single regression coefficient, intercept

- Interval estimation of estimates by single regression analysis
  - Evaluate estimated value y for x and confidential interval by single regression analysis.

◆ Confidence interval of y(x) by regression analysis

When x by regression analysis is changed, we want to know confidence interval of $\hat{y}$.

Population variance is known → Normal distribution is used.

Population variance is unknown → t-distribution is used.

Population variance of a distribution which $\hat{y}$ follows is usually **UNKNOWN**.

Distribution which $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ follows: $$N\left(\beta_1 x + \beta_0 , \quad \left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right\}\sigma^2\right)$$ Equation 5-6

$$Y_i = bx_i + ax_i + \varepsilon_i \qquad \varepsilon_i \text{ follows } N(0, \sigma^2).$$ Equation 5-7

Since σ^2 is unknown value, variance is estimated from sample

Appendix

→ σ^2 is estimated from variance of residual.

Variance of residual: Sum square of residual divided by the degree of freedom of the residual

Degree of residual: Value obtained by subtracting regression degree of freedom from total degree of freedom → (n-1) – 1 = n-2

Degree of freedom of regression analysis : the number of explanatory variables

i.e., **Degree of freedom for regression** in the case of (multiple) regression analysis with p explanatory variables : p

- Single regression analysis → the number of explanatory variable p=1
  → Degree of freedom of regression : 1

→ **Degree of freedom of residual** in the case of (multiple) regression analysis with p explanatory variables : n-1-p

Then, estimate of σ^2 is calculated from variance of residual $V_E = \frac{S_E}{n-2}$ Equation 5-8

Since population variance of distribution which follows predicted value $\hat{y}$ is unknown, and V_E is relaced with estimated value.

Appendix

Then, t-distribution is used for interval estimation.

Variation factor	Sum square	Degree of freedom	Variance	Variance ratio
Regression: R	S_R	$\phi_R = 1$	$V_R = \frac{S_R}{\phi_R}$	$\frac{V_R}{V_E}$
Residual: E	S_E	$\phi_E = n - 2$ $(= \phi_T - \phi_R)$	$V_E = \frac{S_E}{\phi_E}$	
Total: T	S_T	$\phi_T = n - 1$		

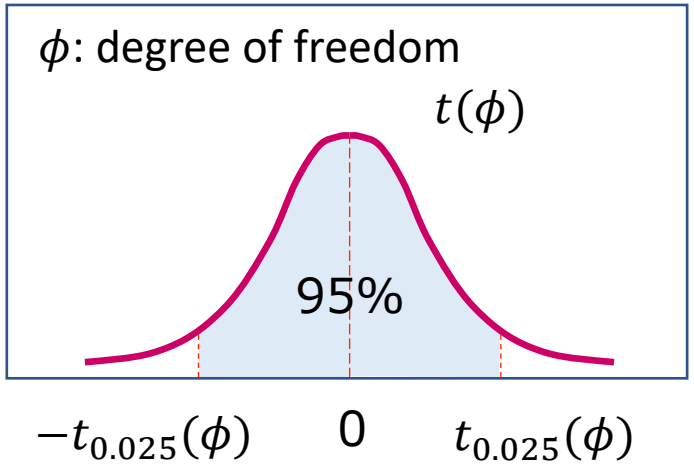
t-statistic $$t = \frac{\hat{\beta}_1 x + \hat{\beta}_0 - (\beta_1 x + \beta_0)}{\sqrt{\left\{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right\} V_E}}$$ Equation 5-9

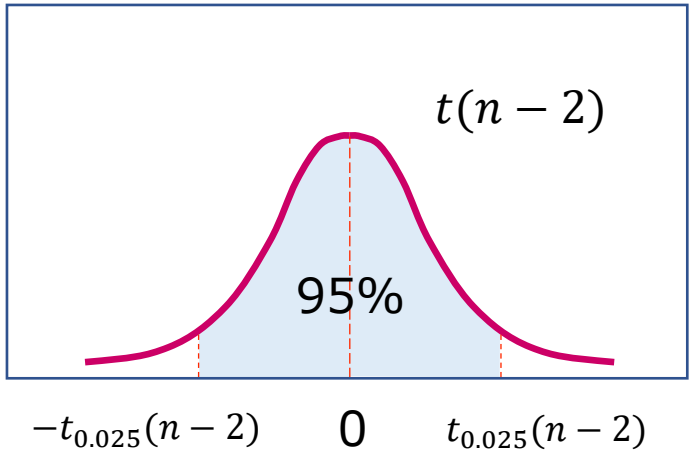
$$S_{xx} = \sum_i (x_i - \bar{x})^2$$ Equation 5-5

- $\hat{\beta}_1 x + \hat{\beta}_0$: calculated result of $\hat{y}$ from sample
- $\beta_1 x + \beta_0$: Population mean of distribution which $\hat{y}$ follows
- $\left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right\} V_E$: Estimator of population variance of distribution which $\hat{y}$ follows

follows t-distribution of degree of freedom n-2.

Appendix





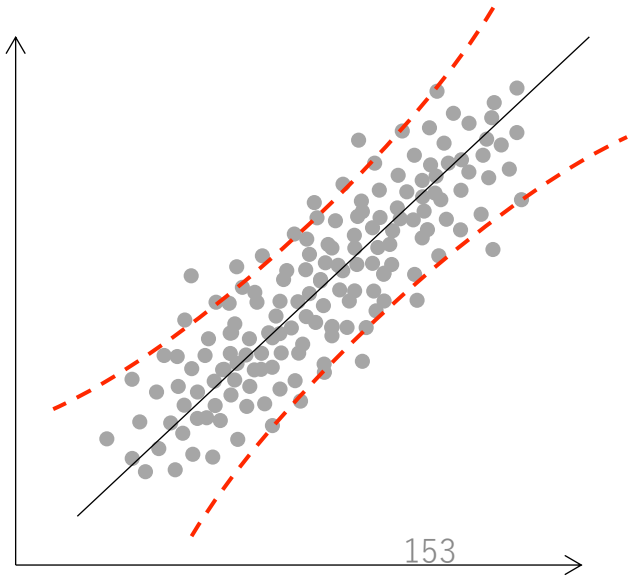
$\phi = \mathrm{n} - 2$

$$-t_{0.025}(n-2) \leq \frac{\hat{\beta}_1 x + \hat{\beta}_0 - (\beta_1 x + \beta_0)}{\sqrt{\left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right\} V_E}} \leq t_{0.025}\,(n-2)$$

Equation 5-10

$$\hat{\beta}_1 x + \hat{\beta}_0 - t_{0.025}(n-2)\sqrt{\left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right\} V_E} \leq \beta_1 x + \beta_0 \leq \hat{\beta}_1 x + \hat{\beta}_0 + t_{0.025}(n-2)\sqrt{\left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right\} V_E}$$

Equation 5-11



Example)

x_i	y_i	$\hat{y}_i$	residual
1	7.7	3.04	-1.34
2	7.5	5.48	2.02
3	6.4	7.92	-1.52
4	11.3	10.36	0.94
5	13.6	12.8	0.80
6	14.0	15.24	-1.24
7	18.2	17.68	0.52
8	21.5	20.12	1.38
9	20.1	22.56	-2.46
10	25.9	25.00	0.90
$\bar{x}$ =5.5	$\bar{y}$ =14.02		

$S_{xx} = \sum(x_i - \bar{x})^2$=82.5

$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$=201.3

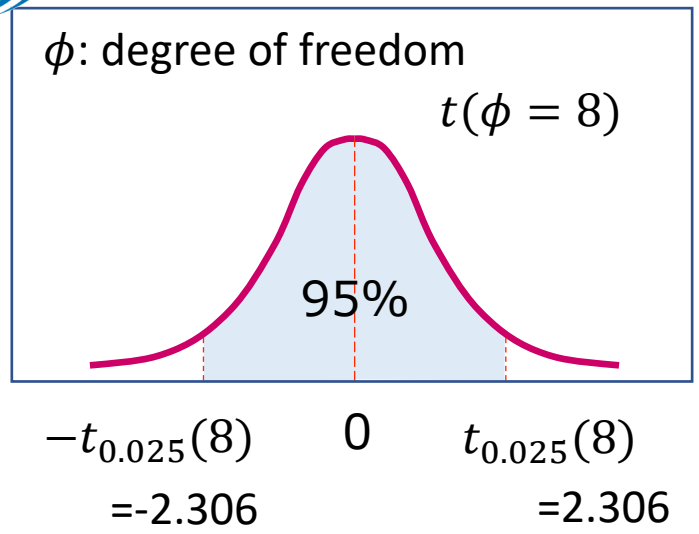
$\hat{\beta}_1$ =2.44

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ =0.6

V_E =2.5355

$$\hat{y} = 2.44x + 0.6$$

Appendix



Measurement value

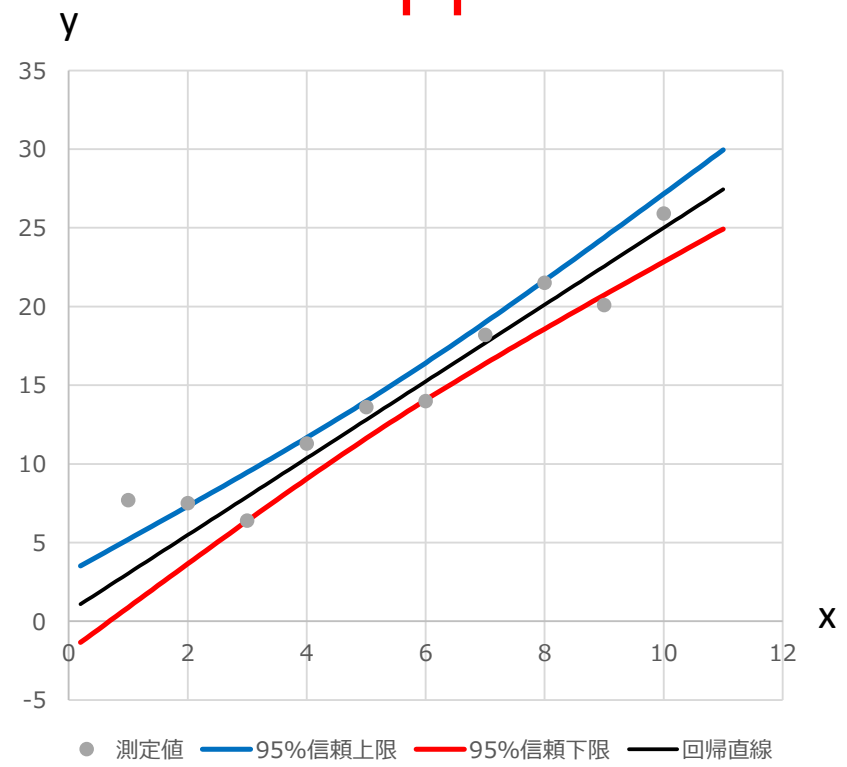
Single regression line

95% UCB

95% LCB

UCB: upper confidence bound

LCB: lower confidence bound



$$\hat{\beta}_1 x + \hat{\beta}_0 - t_{0.025}(n-2)\sqrt{\left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right\}V_E} \leq \beta_1 x + \beta_0 \leq \hat{\beta}_1 x + \hat{\beta}_0 + t_{0.025}(n-2)\sqrt{\left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right\}V_E}$$

$$2.44x + 0.6 - 2.306\sqrt{\left\{\frac{1}{10} + \frac{(x-5.5)^2}{82.5}\right\}2.5355} \leq \beta_1 x + \beta_0 \leq 2.44x + 0.6 + 2.306\sqrt{\left\{\frac{1}{10} + \frac{(x-5.5)^2}{82.5}\right\}2.5355}$$