Lesson 11

Statistics IV

Todays topics

- **linear regression (線形回帰)**
- 単回帰
- 重回帰
- 自己回帰
- モデル選択 **AIC**

来嶋 秀治 (Shuji Kijima)
Dept. Informatics,
Graduate School of ISEE

Linear regression

Ex. Advertisement

year	1	2	3	4	5	6	7	8
x: ad. cost	8	11	13	10	15	19	17	20
y: sale amount	115	124	138	120	151	186	169	193

Question

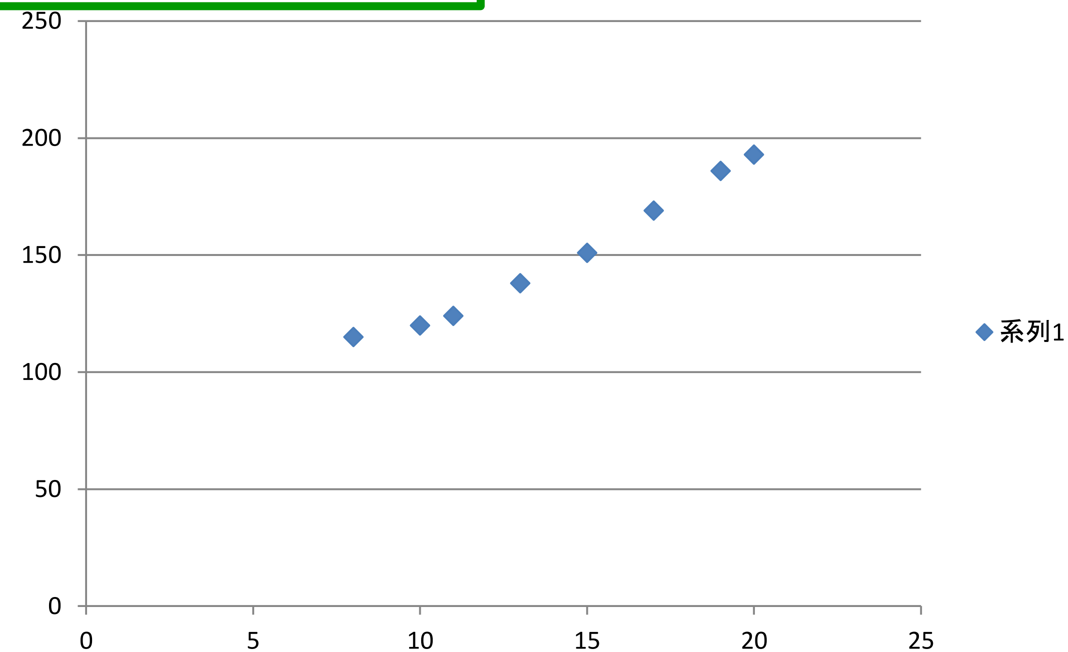
How does y increase, as x increasing?

Ex. Advertisement

year	1	2	3	4	5	6	7	8
x: ad. cost	8	11	13	10	15	19	17	20
y: sale amount	115	124	138	120	151	186	169	193

Question

How does y increase, as x increasing?



Least Square Estimator

year	1	2	3	4	5	6	7	8
x: ad. cost	8	11	13	10	15	19	17	20
y: sale amount	115	124	138	120	151	186	169	193

Question

How does y increase, as x increasing?

Linear regression (線形回帰)

Suppose $y_i = \alpha + \beta x_i + e_i$ where $e_i \sim \mathrm{N}(0, \sigma^2)$.

Estimate α and β such that

$$\min \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2$$

Least Square Estimator

Linear regression (線形回帰)

Suppose $y_i = \alpha + \beta x_i + e_i$ where $e_i \sim \mathrm{N}(0, \sigma^2)$.

Estimate α and β such that $\min \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2$

$$\frac{\partial}{\partial \alpha} g(\alpha, \beta) = \sum_{i=1}^{n} -2(y_i - (\alpha + \beta x_i))$$

$$\frac{\partial}{\partial \beta} g(\alpha, \beta) = \sum_{i=1}^{n} (-2x_i)(y_i - (\alpha + \beta x_i))$$

$\hat{\beta} =$

$\hat{\alpha} =$

Least Square Estimator

Linear regression (線形回帰)

Suppose $y_i = \alpha + \beta x_i + e_i$ where $e_i \sim \mathrm{N}(0, \sigma^2)$.

Estimate α and β such that $\min \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2$

$$\frac{\partial}{\partial \alpha} g(\alpha, \beta) = \sum_{i=1}^{n} -2(y_i - (\alpha + \beta x_i)) \quad \xrightarrow{\frac{\partial}{\partial \alpha} g(\alpha,\beta) = 0} \quad \alpha + \beta \bar{x} = \bar{y}$$

$$\frac{\partial}{\partial \beta} g(\alpha, \beta) = \sum_{i=1}^{n} (-2x_i)(y_i - (\alpha + \beta x_i)) \quad \xrightarrow[\frac{\partial}{\partial \beta} g(\alpha,\beta) = 0]{} \quad \alpha \bar{x} + \beta \overline{x^2} = \overline{xy}$$

$$\hat{\beta} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Ex. Advertisement

year	1	2	3	4	5	6	7	8
x: ad. cost	8	11	13	10	15	19	17	20
y: sale amount	115	124	138	120	151	186	169	193

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{113}{8} = 14.125$$

$$\overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1196}{8} = 149.5$$

$$\overline{x^2} := \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1729}{8} = 216.25$$

$$\overline{xy} := \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{17810}{8} = 2226.25$$

$$\hat{\beta} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{2226.25 - 14.125 \times 149.5}{216.125 - 14.125^2} = 6.9$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 149.5 - 6.9 \times 14.125 = 52.1$$

Q: Unbiased estimator?

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{\mathrm{E}\left[s_{xy}\right]}{s_x^2} = ?$$

$$\mathrm{E}[\hat{\alpha}] = \mathrm{E}\left[\overline{y} - \hat{\beta}\overline{x^2}\right] = \mathrm{E}[\overline{y}] - \mathrm{E}\left[\hat{\beta}\right]\overline{x} = ?$$

which are unbiased estimators**?**

Q: Unbiased estimator?

$$\mathrm{E}[\bar{y}] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[y_i] = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) = \alpha + \beta\bar{x}$$

$$\mathrm{E}[\overline{xy}] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[x_i y_i] = \frac{1}{n}\sum_{i=1}^{n}x_i(\alpha + \beta x_i) = \alpha\bar{x} + \beta\overline{x^2}$$

$$\mathrm{E}\left[s_{xy}\right] = \mathrm{E}[\overline{xy} - \bar{x}\cdot\bar{y}] = \alpha\bar{x} + \beta\overline{x^2} - \bar{x}(\alpha + \beta\bar{x}) = \beta s_x^2$$

Hence,

$$\mathrm{E}\left[\hat{\beta}\right] = \frac{\mathrm{E}\left[s_{xy}\right]}{s_x^2} = \beta$$

$$\mathrm{E}[\hat{\alpha}] = \mathrm{E}\left[\bar{y} - \hat{\beta}\overline{x^2}\right] = \mathrm{E}[\bar{y}] - \mathrm{E}\left[\hat{\beta}\right]\bar{x} = \alpha$$

which are unbiased estimators.

Variance (Gauss-Markov theorem)

$$\mathrm{Var}[\hat{\alpha}] := \mathrm{E}[(\hat{\alpha}-\alpha)^2] = a(x)\sigma^2 \quad \text{where } a(x) := \frac{\overline{x^2}}{n\cdot s_x^2}$$

$$\mathrm{Var}[\hat{\beta}] := \mathrm{E}\left[\left(\hat{\beta}-\beta\right)^2\right] = b(x)\sigma^2 \quad \text{where } b(x) := \frac{1}{n\cdot s_x^2}$$

Thus

$$\hat{\alpha} \sim \mathrm{N}(\alpha, a(x)\sigma^2)$$
$$\hat{\beta} \sim \mathrm{N}(\beta, b(x)\sigma^2)$$

Remark

$\mathrm{Var}[\hat{\alpha}]$ and $\mathrm{Var}[\hat{\beta}]$ decrease as $s_x^2 = \frac{\sum_{i=1}^{n}(x_i^2-(\overline{x})^2)}{n}$ increases.

$\Rightarrow$ Observe x in a wide range, then we obtain a good estimator.

Thm.

Let $\hat{\sigma}^2 := \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (\hat{\alpha}+\hat{\beta}x_i)\right)^2$,

then $E[\hat{\sigma}^2] = \sigma^2$ and $\frac{(n-2)\,\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$ hold.

omit the proof (not easy)

Hypothesis testing

Hypothesis testing for β

The central limit theorem suggests that

$$\frac{\hat{\beta} - \beta}{\sqrt{\mathrm{Var}[\hat{\beta}]}} \sim \mathrm{N}(0,1)$$

Then, its Studentization is

$$T_n := \frac{\hat{\beta} - \beta}{\frac{\widehat{\sigma^2}}{n \cdot s_x^2}} \quad \left(\simeq \frac{\hat{\beta} - \beta}{\sqrt{\mathrm{Var}[\hat{\beta}]}} \right)$$

Thm.

$T_n \sim t_{n-2}$

Ex. Advertisement

year	1	2	3	4	5	6	7	8
x: ad. cost	8	11	13	10	15	19	17	20
y: sale amount	115	124	138	120	151	186	169	193

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{113}{8} = 14.125$$

$$\overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1196}{8} = 149.5$$

$$\overline{x^2} := \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1729}{8} = 216.25$$

$$\overline{xy} := \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{17810}{8} = 2226.25$$

$$\overline{x}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = 14.125^2 = 199.5$$

$$\overline{x} \cdot \overline{y} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = 14.125 \times 149.5 = 2111.6875$$

Ex. Advertisement

year	1	2	3	4	5	6	7	8
x: ad. cost	8	11	13	10	15	19	17	20
y: sale amount	115	124	138	120	151	186	169	193

$$\hat{\beta} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = 6.9$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 52.1$$

$$\widehat{\sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \left(\hat{\alpha} + \hat{\beta} x_i \right) \right)^2 = 21.418$$

$$T_n = \frac{\hat{\beta} - \beta}{\sqrt{\frac{\widehat{\sigma^2}}{n \cdot s_x^2}}} = \frac{6.9 - 0}{\sqrt{\frac{4.628^2}{132.875}}} = 17.19$$

$t_{6-2}^* = 2.477$ ⇨ null hypothesis $\beta = 0$ is rejected.

Ex.

Least Square Estimator

	1	2	3	… …	37	38
x: applied dose (投与量)	2.32	2.39	2.61		7.78	8.28
y: observed value (観測数値)	2.88	3.21	3.01		5.88	6.67

Question

How does y increase, as x increasing?

Linear regression (線形回帰)

Suppose $y_i = \alpha + \beta x_i + e_i$ where $e_i \sim \mathrm{N}(0, \sigma^2)$.

Estimate α and β such that

$$\min \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2$$

Hypothesis testing for β

The central limit theorem suggests that

$$\frac{\hat{\beta} - \beta}{\sqrt{\mathrm{Var}[\hat{\beta}]}} \sim \mathrm{N}(0,1)$$

Then, its Studentization is

$$T_n \coloneqq \frac{\hat{\beta} - \beta}{\sqrt{\frac{\widehat{\sigma^2}}{n \cdot s_x^2}}} \quad \left(\simeq \frac{\hat{\beta} - \beta}{\sqrt{\mathrm{Var}[\hat{\beta}]}} \right)$$

Thm.

$T_n \sim t_{n-2}$

ex

$$\hat{\alpha} = 2.07 \ , \hat{\beta} = 0.49 \ , \hat{\sigma}^2 = 0.47^2, s_x^2 = \cdots$$

Q: β = 0?

- administration does not work? (投与効果はない？)

$$T_n := \frac{\hat{\beta} - \beta}{\sqrt{\frac{\widehat{\sigma^2}}{n \cdot s_x^2}}} = \frac{0.49 - 0}{\sqrt{\frac{0.47^2}{38 \times ???}}} = \cdots$$

If $|T_n| > |z^*_{36}|$ then

null hypothesis of β = 0 is rejected

Multiple Linear Regression

Least Square Estimator

Multiple linear regression (多重線形回帰)

Suppose that $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i$ and $e_i \sim \mathrm{N}(0, \sigma^2)$.

Estimate $\boldsymbol{\beta}$ such that

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta})$$

where $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ and $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}$

Proposition

The optimum solution is $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$.

Furthermore, $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator.

Multi linear regression

Proposition

The optimum solution is $\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$.

$$\nabla (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}) = -2X^\top (\boldsymbol{y} - X\boldsymbol{\beta})$$

Rem 10-1. (see Apex.)

$$\nabla\big((A\boldsymbol{x})^\top (A\boldsymbol{x})\big) = 2A^\top A\boldsymbol{x}$$

$$X^\top \boldsymbol{y} - X^\top X\boldsymbol{\beta} = 0$$

$$X^\top X\boldsymbol{\beta} = X^\top \boldsymbol{y}$$

Apex. Proof of Rem. 10-1

Rem 10-1. (see Apex.)

$$\nabla\left((A\boldsymbol{x})^\top(A\boldsymbol{x})\right) = 2A^\top A\boldsymbol{x}$$

$$A\boldsymbol{x} = \begin{pmatrix} \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_d^\top \end{pmatrix}\boldsymbol{x} = \begin{pmatrix} a_{11}x_1 + \cdots + a_{1d}x_d \\ \vdots \\ a_{d1}x_1 + \cdots + a_{dd}x_d \end{pmatrix}$$

linearity of ∇

$$\nabla\left((A\boldsymbol{x})^\top(A\boldsymbol{x})\right) = \nabla\sum_{i=1}^{d}(a_{i1}x_1 + \cdots + a_{id}x_d)^2 = \sum_{i=1}^{d}\nabla(a_{i1}x_1 + \cdots + a_{id}x_d)^2$$

Since

$$\frac{\partial(a_{i1}x_1 + \cdots + a_{id}x_d)^2}{\partial x_j} = 2(a_{i1}x_1 + \cdots + a_{id}x_d)a_{ij}$$

we have

$$\nabla(a_{i1}x_1 + \cdots + a_{id}x_d)^2 = 2(a_{i1}x_1, \dots, a_{id}x_d)\begin{pmatrix} a_{i1} \\ \vdots \\ a_{id} \end{pmatrix} = 2\left(\boldsymbol{a}_i^\top\boldsymbol{x}\right)\boldsymbol{a}_i$$

Claim (next slide)

$$\sum_{i=1}^{d} c_i\boldsymbol{y}_i = (\boldsymbol{y}_1 \dots \boldsymbol{y}_d)\boldsymbol{c}$$

Then,

!

$$\sum_{i=1}^{d}\nabla(a_{i1}, \dots, a_{id})^2 = 2\sum_{i=1}^{d}(\boldsymbol{a}_i^\top\boldsymbol{x})\boldsymbol{a}_i = 2(\boldsymbol{a}_1 \dots \boldsymbol{a}_d)\begin{pmatrix} \boldsymbol{a}_1^\top\boldsymbol{x} \\ \vdots \\ \boldsymbol{a}_d^\top\boldsymbol{x} \end{pmatrix} = 2A^\top A\boldsymbol{x}$$

□

Apex. Proof of Rem. 10-1 (contd.)

Claim

$$\sum_{i=1}^{d} c_i \boldsymbol{y_i} = (\boldsymbol{y_1} \dots \boldsymbol{y_d})\boldsymbol{c}$$

$$(\mathrm{r.h.s.}) = \begin{pmatrix} y_{11} & \cdots & y_{d1} \\ \vdots & \ddots & \vdots \\ y_{1d} & \cdots & y_{dd} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_d \end{pmatrix}$$

$$= \begin{pmatrix} y_{11}c_1 + \cdots + y_{1d}c_d \\ \vdots \\ y_{d1}c_1 + \cdots + y_{dd}c_d \end{pmatrix}$$

$$= c_1 \begin{pmatrix} y_{11} \\ \vdots \\ y_{1d} \end{pmatrix} + \cdots + c_d \begin{pmatrix} y_{d1} \\ \vdots \\ y_{dd} \end{pmatrix}$$

$$= \sum_{i=1}^{d} c_i \boldsymbol{y_i} = (\mathrm{l.h.s.})$$

□

Nonlinear Regression

Least Square Estimator

Nonlinear regression (非線形回帰)

Suppose

$y_i = h(x_i) + e_i$ and $e_i \sim \mathrm{N}(0, \sigma^2)$ where

$h(x) = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3 + \cdots.$

Estimate h such that

$$\min_{\beta} \sum_{i=1}^{n} (y_i - h(x_i))^2$$

AIC (Akaike Information criteria: 赤池情報量)

$$\mathrm{AIC} = -2 \sum_{i=1}^{n} \log f\left(X_i; \hat{\theta}_n^{\mathrm{ML}}\right) + 2\dim(\theta)$$