1 Multiple regression analysis

重回帰分析

1.1 Objectives

- Estimate one continuous value from a linear combination of multiple (more than one) types of variables.
- Eg. Baseball

$$R_{un} = a_1 B_{atting} + a_2 H_{omerun} + a_0 \qquad (1.1)$$

R_{un}: Run (score)
B_{atting}: Batting average
H_{omerun}: Number of home runs

- Runs that a team earns is likely to be predicted based on the team's batting average and number of home runs hit by a team in a year.

1.1 Objectives

Club	Run	Batting ave.	Home runs
Tigers	597	.262	145
Giants	531	.255	82
…	…	…	…
Eagles	534	.256	105

This data is available in baseball201x.mat.

1.2 Theory
1.2.1 Model

$$y_j = a_1(x_{1j} - \overline{x_1}) + \cdots a_p(x_{pj} - \overline{x_p}) + a_0 + \epsilon_j \qquad (1.2)$$

y_j: Objective variable (目的変数)/Dependent variable (従属変数). Values to be estimated.
x_j: Explanatory variable (説明変数)/Independent variable (独立変数). Values to explain the objective variable.
j: Suffix of samples. $j = 1, 2, \ldots, n$. Specify the baseball club.
n: Number of samples. Twelve clubs: $n = 12$.
a_i: (Partial) Regression coefficient (偏回帰係数). Weights of explanatory variables.

1.2.1 Model

p: Number of explanatory variables

ϵ_j: Error (誤差) or residual (残差). Difference between the observed (観測値) and predicted (推定値) values. ϵ_j is a random variable with the mean being 0. $\epsilon_j \perp \epsilon_k$. Errors randomly vary around zero.

$\overline{x_j}$: Mean of x_j.

$$\overline{x_i} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$$

1.2.1 Model

- Determine partial regression coefficients with the least squares sum of errors
- For all n samples, (1.2) holds and can be written by using matrices and vectors as follows:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11}-\overline{x_1} & \dots & x_{i1}-\overline{x_i} & \dots & x_{p1}-\overline{x_p} & 1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ x_{1j}-\overline{x_1} & \dots & x_{ij}-\overline{x_i} & \dots & x_{pj}-\overline{x_p} & \vdots \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ x_{1n}-\overline{x_1} & \dots & x_{in}-\overline{x_i} & \dots & x_{pn}-\overline{x_p} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_p \\ a_0 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_j \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (1.3)$$

$(n \times 1)$ $(n \times (p+1))$ $((p+1) \times 1)$ $(n \times 1)$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\epsilon} \qquad (1.4)$$

1.2.1 Model

- Scalar variable … Italic font
- Vector … Italic and bold font
- Matrix … Capital letter in Italic and bold font

1.2.2 Mathematical principles

- Least squares estimation of $\boldsymbol{a}$ is determined when the sum of squared errors is minimized. The sum is given by

$$\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_n^2 = [\epsilon_1 \quad \dots \quad \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \to \min. \quad (1.5)$$

- From (1.4), the error vectors are

$$\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{a} \qquad (1.6)$$

$$\boldsymbol{\epsilon}^{\mathrm{T}} =$$

1.2.2 Mathematical principles

- Least squares estimation of $\boldsymbol{a}$ is determined when the sum of squared errors is minimized. The sum is given by

$$\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_n^2 = [\epsilon_1 \quad \cdots \quad \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \to \min. \quad (1.5)$$

- From (1.4), the error vectors are

$$\begin{aligned} \boldsymbol{\epsilon} &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{a} \\ \boldsymbol{\epsilon}^\mathrm{T} &= \boldsymbol{y}^\mathrm{T} - \boldsymbol{a}^\mathrm{T}\boldsymbol{X}^\mathrm{T} \end{aligned} \quad (1.6)$$

1.2.2 Mathematical principles

- Using (1.6), the sum of squared errors is

$$\boldsymbol{\epsilon}^\mathrm{T}\boldsymbol{\epsilon} = \quad (1.7)$$

1.2.2 Mathematical principles

- Using (1.5) and (1.6), the sum of squared errors is

$$\begin{aligned} \boldsymbol{\epsilon}^\mathrm{T}\boldsymbol{\epsilon} &= (\boldsymbol{y}^\mathrm{T} - \boldsymbol{a}^\mathrm{T}\boldsymbol{X}^\mathrm{T})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{a}) \quad (1.7) \\ &= \boldsymbol{y}^\mathrm{T}\boldsymbol{y} - \boldsymbol{a}^\mathrm{T}\boldsymbol{X}^\mathrm{T}\boldsymbol{y} - \boldsymbol{y}^\mathrm{T}\boldsymbol{X}\boldsymbol{a} + \boldsymbol{a}^\mathrm{T}\boldsymbol{X}^\mathrm{T}\boldsymbol{X}\boldsymbol{a} \\ &= \boldsymbol{y}^\mathrm{T}\boldsymbol{y} - 2\boldsymbol{a}^\mathrm{T}\boldsymbol{X}^\mathrm{T}\boldsymbol{y} + \boldsymbol{a}^\mathrm{T}\boldsymbol{X}^\mathrm{T}\boldsymbol{X}\boldsymbol{a} \end{aligned}$$

- For derivation, you may use the following equation about scalars.

$$\boldsymbol{a}^\mathrm{T}\boldsymbol{X}^\mathrm{T}\boldsymbol{y} = \boldsymbol{y}^\mathrm{T}\boldsymbol{X}\boldsymbol{a} \quad (1.8)$$

1.2.2 Mathematical principles

- The coefficients $\boldsymbol{a}$ that minimizes $\boldsymbol{\epsilon}^\mathrm{T}\boldsymbol{\epsilon}$ is given by solving the following about $\boldsymbol{a}$.

$$\frac{\partial \boldsymbol{\epsilon}^\mathrm{T}\boldsymbol{\epsilon}}{\partial \boldsymbol{a}} = \quad (1.9)$$

$$= \boldsymbol{0}$$

- Then, the least square estimate of $\boldsymbol{a}$ is given by

$$\boldsymbol{a} = \quad (1.10)$$

Mathematical review I:

Derivative of a scalar with respect to a vector

s: Scalar
$\boldsymbol{v} \in \mathbb{R}^{p\times 1}$
$\boldsymbol{b} \in \mathbb{R}^{p\times 1}$
$\boldsymbol{W} \in \mathbb{R}^{p\times p}$

$$\boldsymbol{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix} \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix} \qquad \frac{\partial s}{\partial \boldsymbol{v}} = \begin{bmatrix} \frac{\partial s}{\partial v_1} \\ \vdots \\ \frac{\partial s}{\partial v_p} \end{bmatrix}$$

Eg.

$$\frac{\partial \boldsymbol{b}^{\mathrm{T}}\boldsymbol{v}}{\partial \boldsymbol{v}} = \boldsymbol{b} \qquad \frac{\partial \boldsymbol{b}^{\mathrm{T}}\boldsymbol{v}}{\partial \boldsymbol{b}} = \boldsymbol{v} \qquad \frac{\partial \boldsymbol{b}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{b}}{\partial \boldsymbol{b}} = 2\boldsymbol{W}\boldsymbol{b}$$

1.2.2 Mathematical principles

- The coefficients $\boldsymbol{a}$ that minimizes $\boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}$ is given by solving the following about $\boldsymbol{a}$.

$$\frac{\partial \boldsymbol{\epsilon}^{\mathrm{T}}\boldsymbol{\epsilon}}{\partial \boldsymbol{a}} = -2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{a} \qquad (1.9)$$

$$= \boldsymbol{0}$$

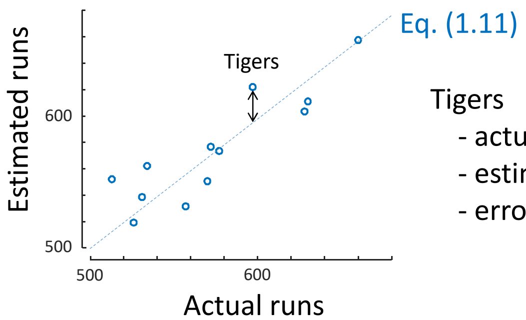
- Then, the least square estimate of $\boldsymbol{a}$ is given by

$$\boldsymbol{a} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \qquad (1.10)$$

1.3 Example of estimation

- By using (1.10), the partial regression coefficients of (1.1) are solved, and the estimation equation is

$$R_{un} = a_1 B_{atting} + a_2 H_{omerun} + a_0$$
$$= 4.38 \times 10^3 B_{atting} + 0.838 \times H_{omerun} + 575 \quad (1.11)$$



Tigers
- actual runs: 597
- estimate: 621
- error: 24

1.3 Example of estimation

- We may say that the runs of Tigers are unexpectedly small considering its batting average and number of home runs.
  - Estimated runs is 622.
  - Actual runs is 597.

1.3 Example of estimation

- Runs that baseball clubs earn in a year are estimated by the number of single hits and home runs.

$$R_{un} = a_1 H_{omerun} + a_2 S_{ingle} + a_3 T_{wobase}$$
$$+ a_4 T_{hreebase} + a_0$$
$$= 1.82 \times H_{omerun} + 0.75 \times S_{ingle} + 1.13 \times T_{wobase}$$
$$+ 2.24 \times T_{hreebase} - 577$$

- We expect that the team earns
  - 1.82 runs from a home run
  - 1.13 runs from a two-base hit
  - 0.75 run from a single hit

1.3 Example of estimation

- Homework (optional, fro your own study)
  - Compute the ***a*** values for the following model.

$$W_{inning-rate} = a_1 \times Run + a_2 \times ERA + a_0$$

- Next week
  - How can we improve the estimation and reduce the error?