データ解析

http://coconut.sys.eng.shizuoka.ac.jp/data/06/

静岡大学工学部
安藤和敏

2006.10.12

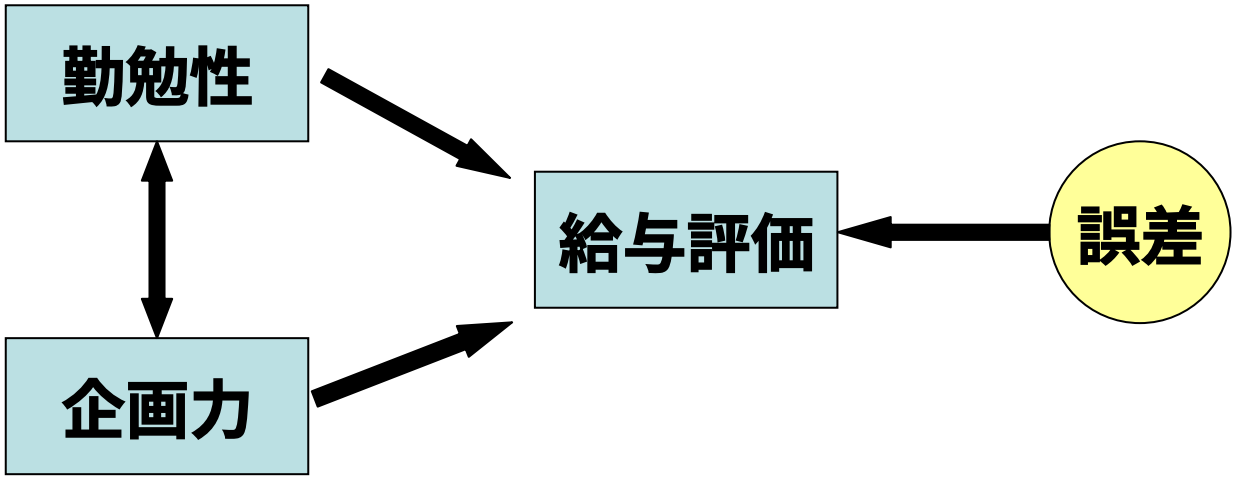
1-4 関係，原因･結果をイメージにするパス図

多変量解析におけるモデルを直感的に理解するための図

ある会社の社員のデータ

社員No	社交性	勤勉性	企画力	判断力	給与評価
1	7	6	7	8	10
2	4	5	5	4	4
3	6	8	4	4	8
4	5	5	5	5	8
5	6	6	4	5	6
6	6	5	6	6	7
7	4	4	6	6	8

パス図

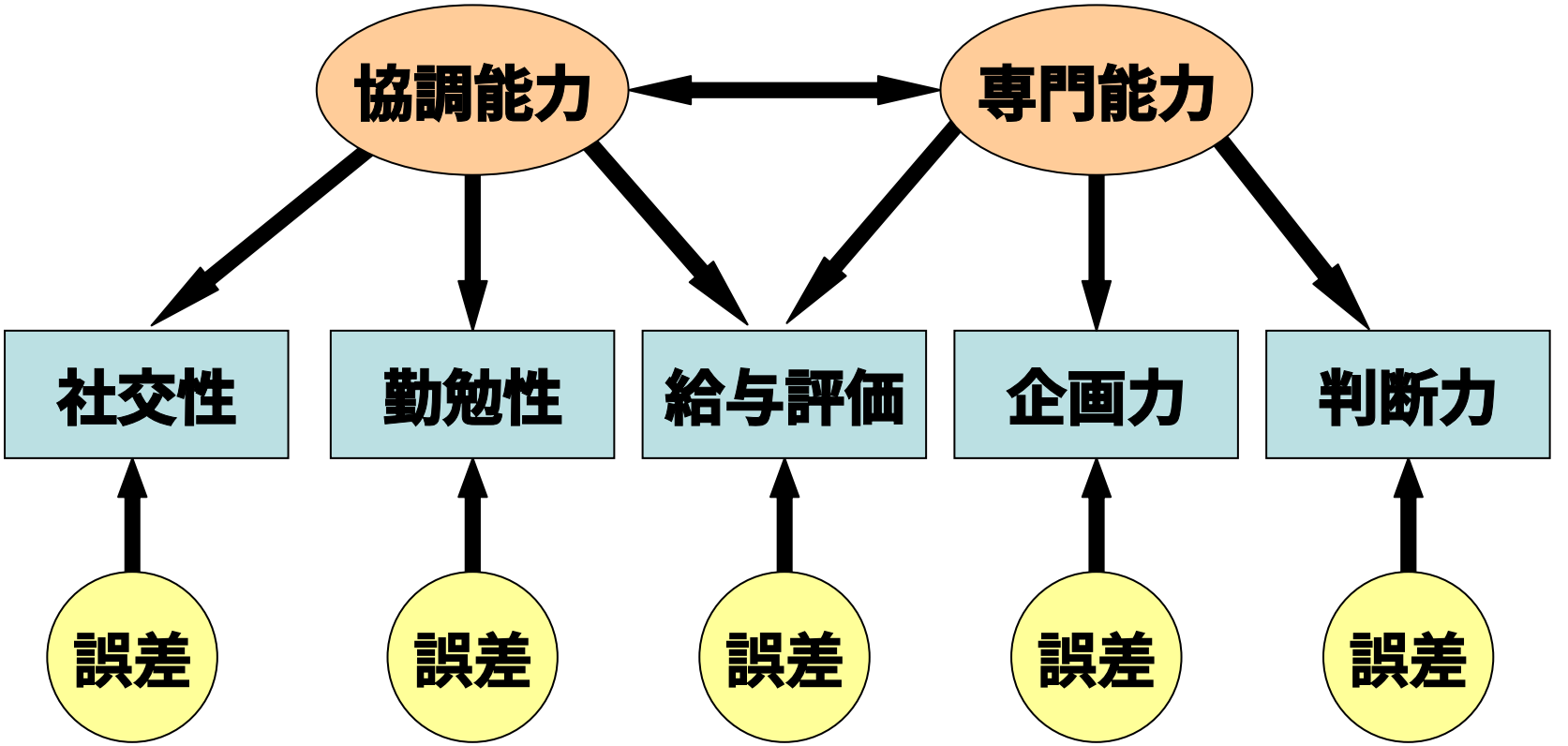


データに現れる変数（観測変数）を四角で囲む．

変数間の因果関係を矢線で示す．

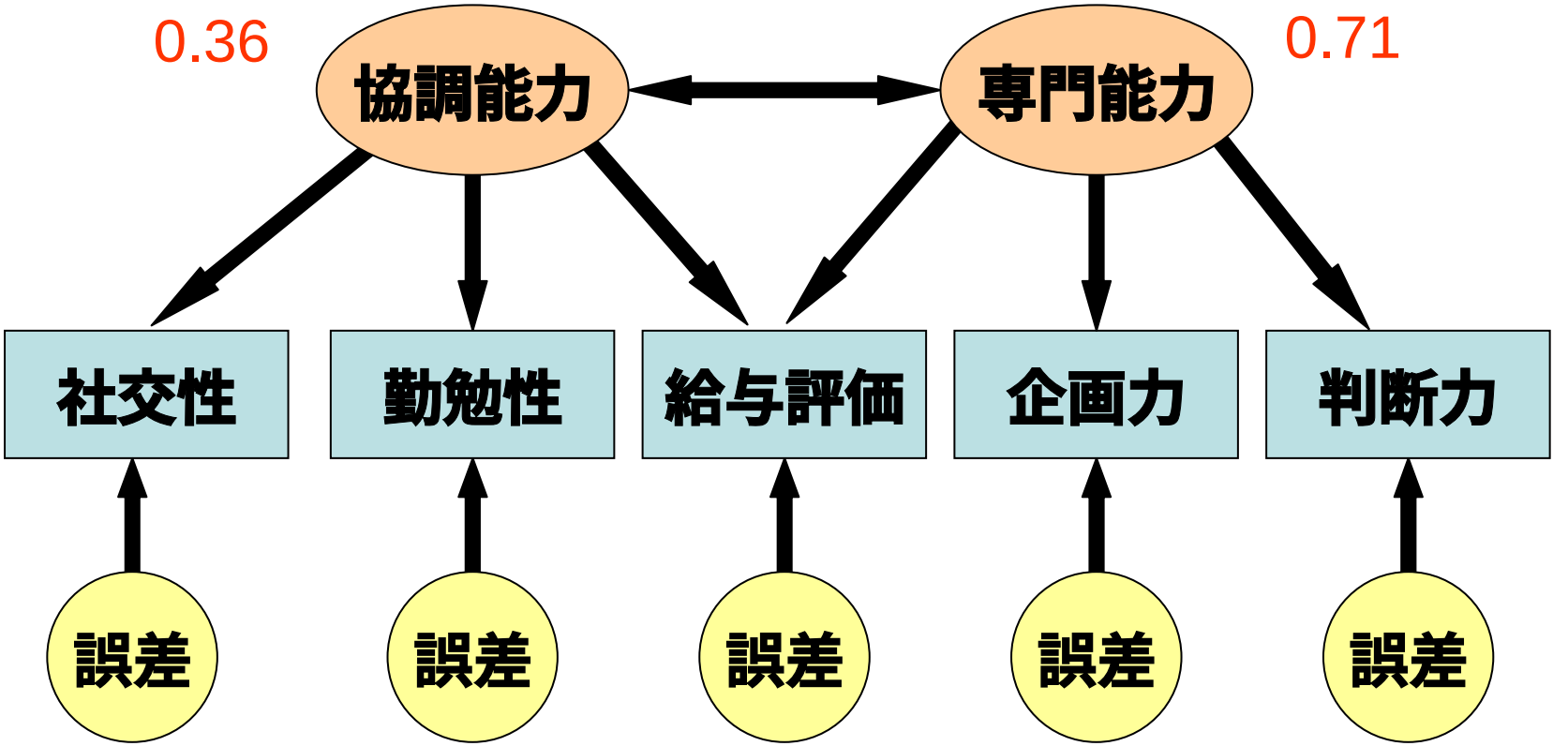
誤差を丸で囲む．

潜在変数を含むパス図



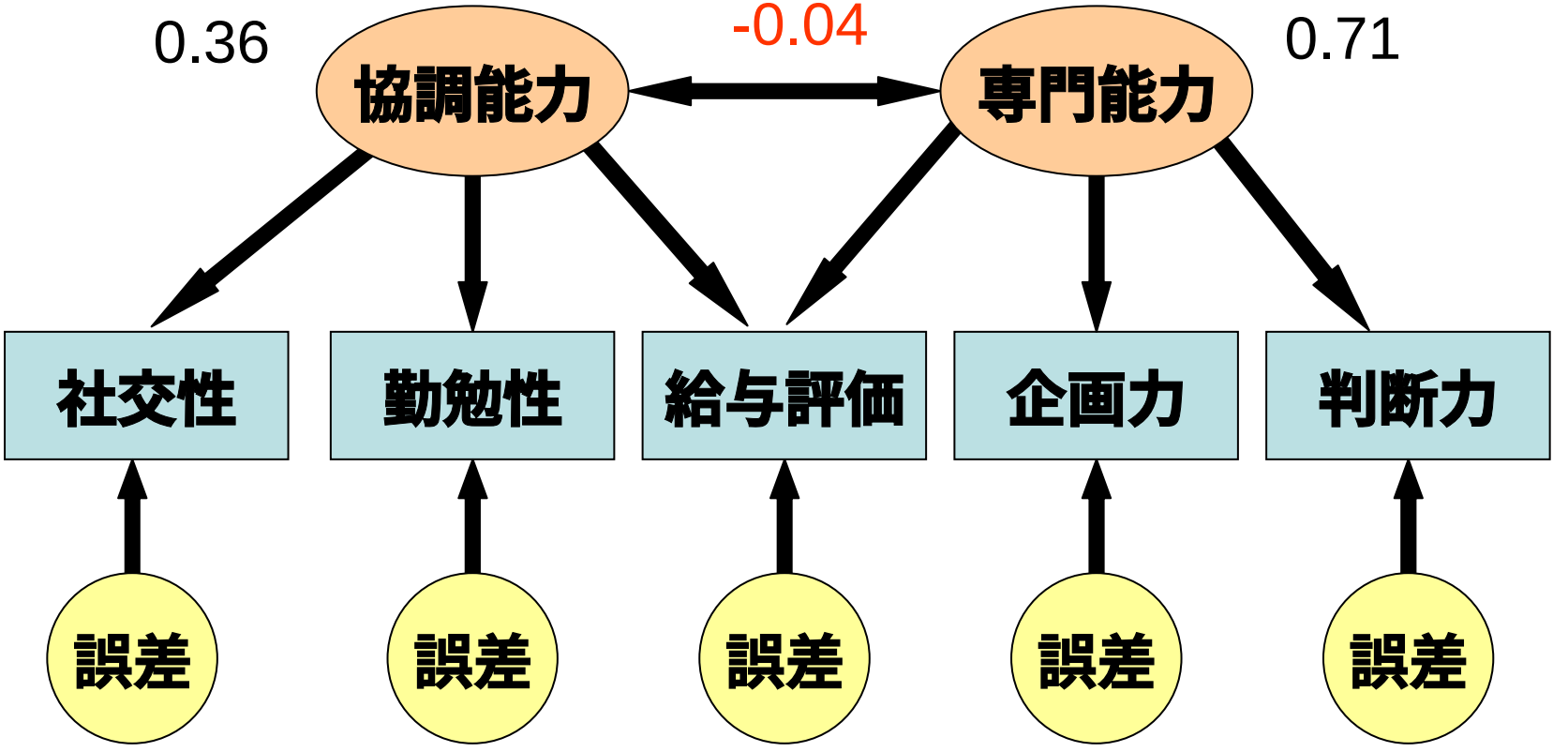
データに現れない変数をモデルに組み込むことがある．
そのような変数を**潜在変数**と呼んで，楕円で囲む．

潜在変数を含むパス図



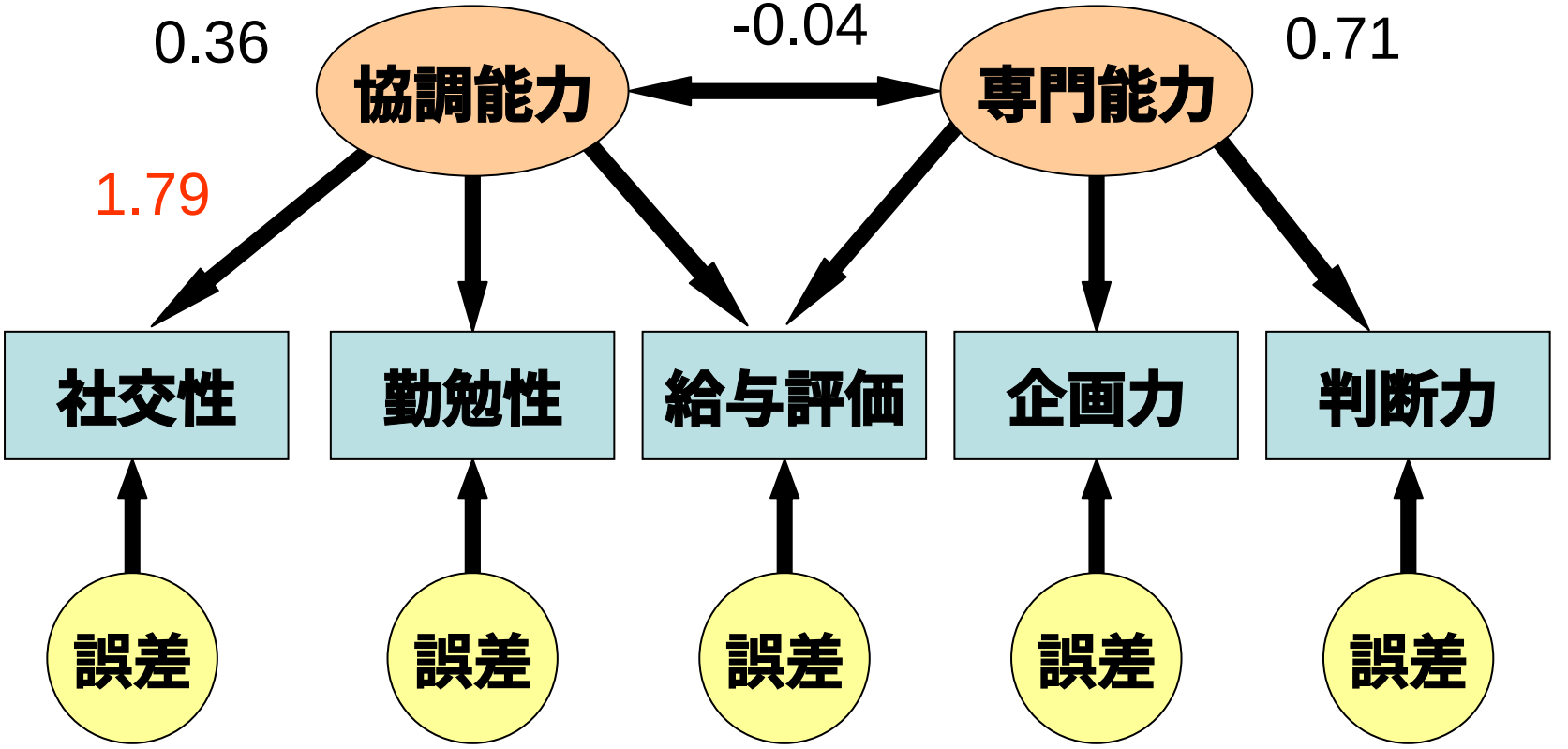
楕円の横にその変数の分散書き込むことがある．(分散はその変数の持つ情報量をあらわすということを先週述べた．)

潜在変数を含むパス図



両矢印の上には共分散を記入することもある.

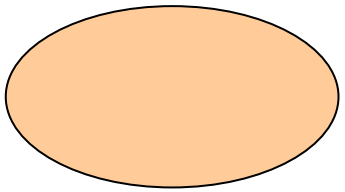
潜在変数を含むパス図



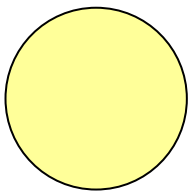
影響の強さは，矢線の上につけた数値で表現する．この数値を**パス係数**と呼ぶ．

パス図のまとめ

…… 観測変数



…… 潜在変数



…… 誤差

…… 因果関係

…… 関連

第2章 Excelで学ぶ重回帰分析

- 単回帰分析
- 重回帰分析

重回帰分析

浜松駅周辺の中古マンションのデータ

マンションNo	広さ（平米）	築年数（年）	価格（千万円）
1	51	16	3.0
2	38	4	3.2
3	57	16	3.3
4	51	11	3.9
5	53	4	4.4
6	77	22	4.5
7	63	5	4.5
8	69	5	5.4
9	72	2	5.4
10	73	1	6.0

重回帰分析によってわかること

1. 価格は，広さと築年数によってどのように予測できるか.
2. 予測できるとすれば，その精度はどれくらいか.
3. 同じ地区で広さ70m^2，築年数10年，価格5.8千万円のマンションを提示された．この価格は妥当か.

1. 価格と広さと築年数は以下の関係にあると推定される．

$$\text{価格} = 1.02 + 0.0668 \times \text{広さ} - 0.0808 \times \text{築年数}$$

2. 寄与率は 0.933 で上式の精度は十分高い.
3. 広さ=70，築年数=10を代入すると，価格=4.89となるので，5.8千万円は相場より高い.

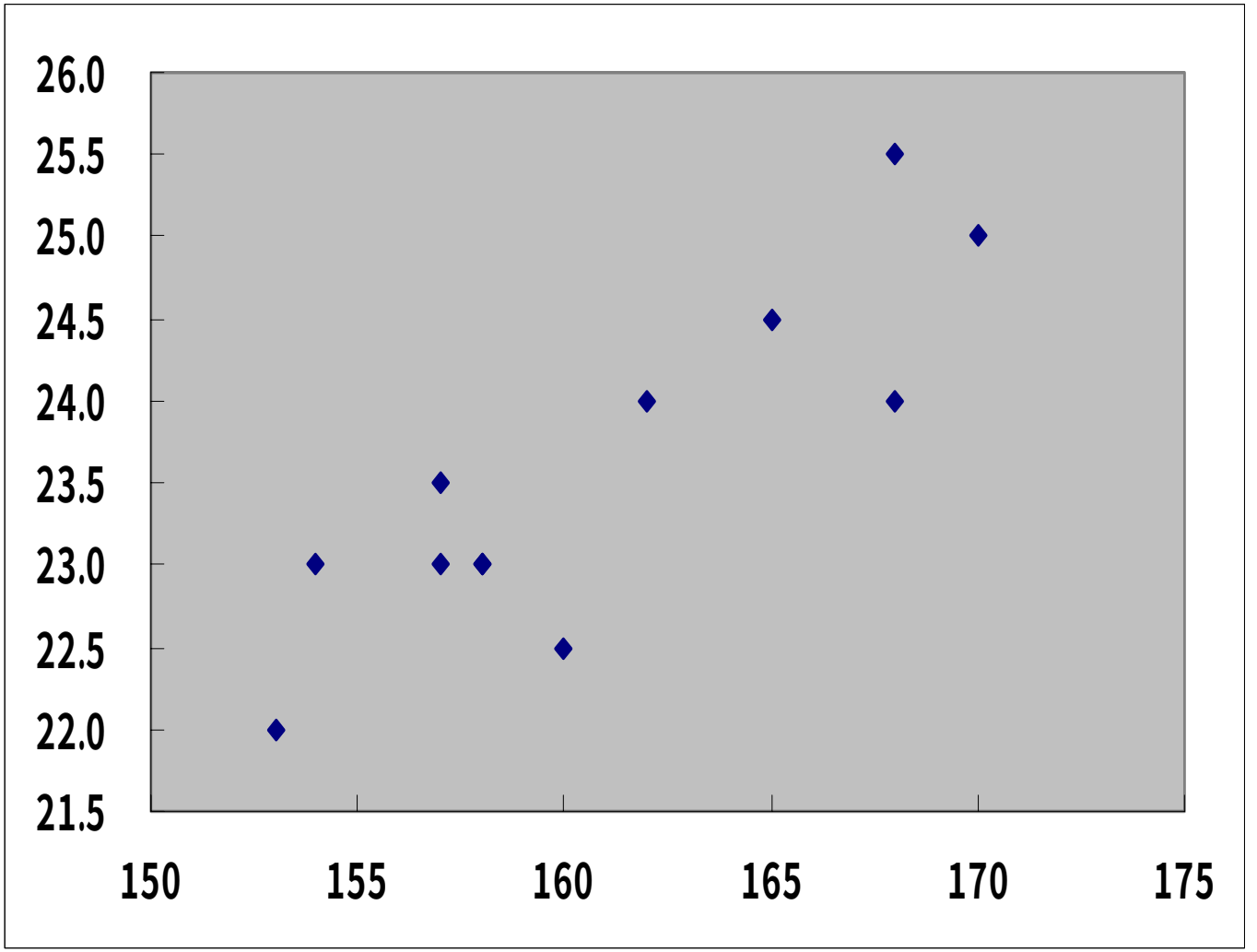
2-1 1変数を1変数から予測する単回帰分析

単回帰分析は重回帰分析の最も単純な特別な場合．
重回帰分析の理解のための基礎となる．

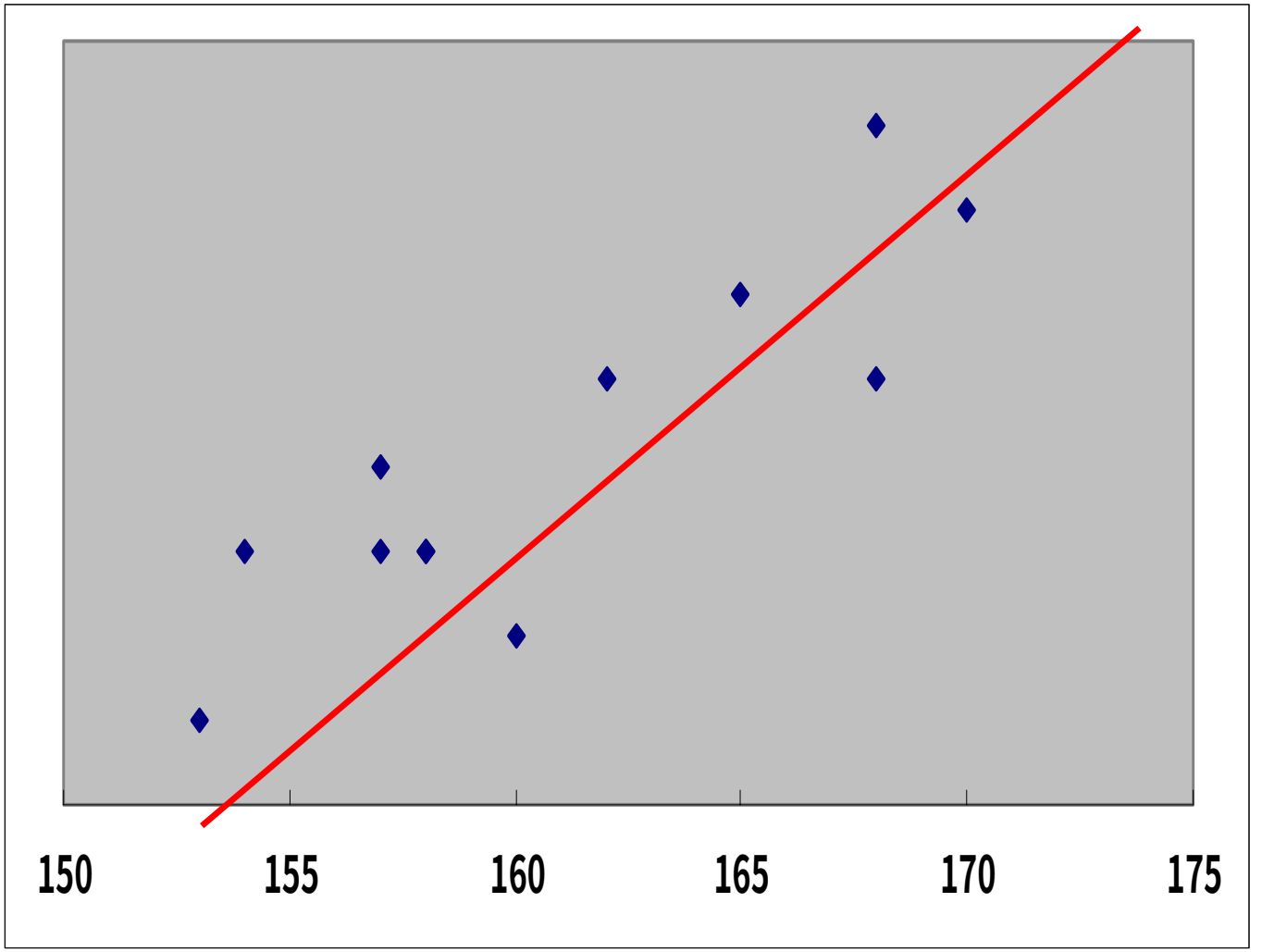
単回帰分析のデータ

番号	身長(x)	靴サイズ(y)
1	162	24.0
2	165	24.5
3	168	25.5
4	160	22.5
5	158	23.0
6	153	22.0
7	158	23.0
8	168	24.0
9	157	23.0

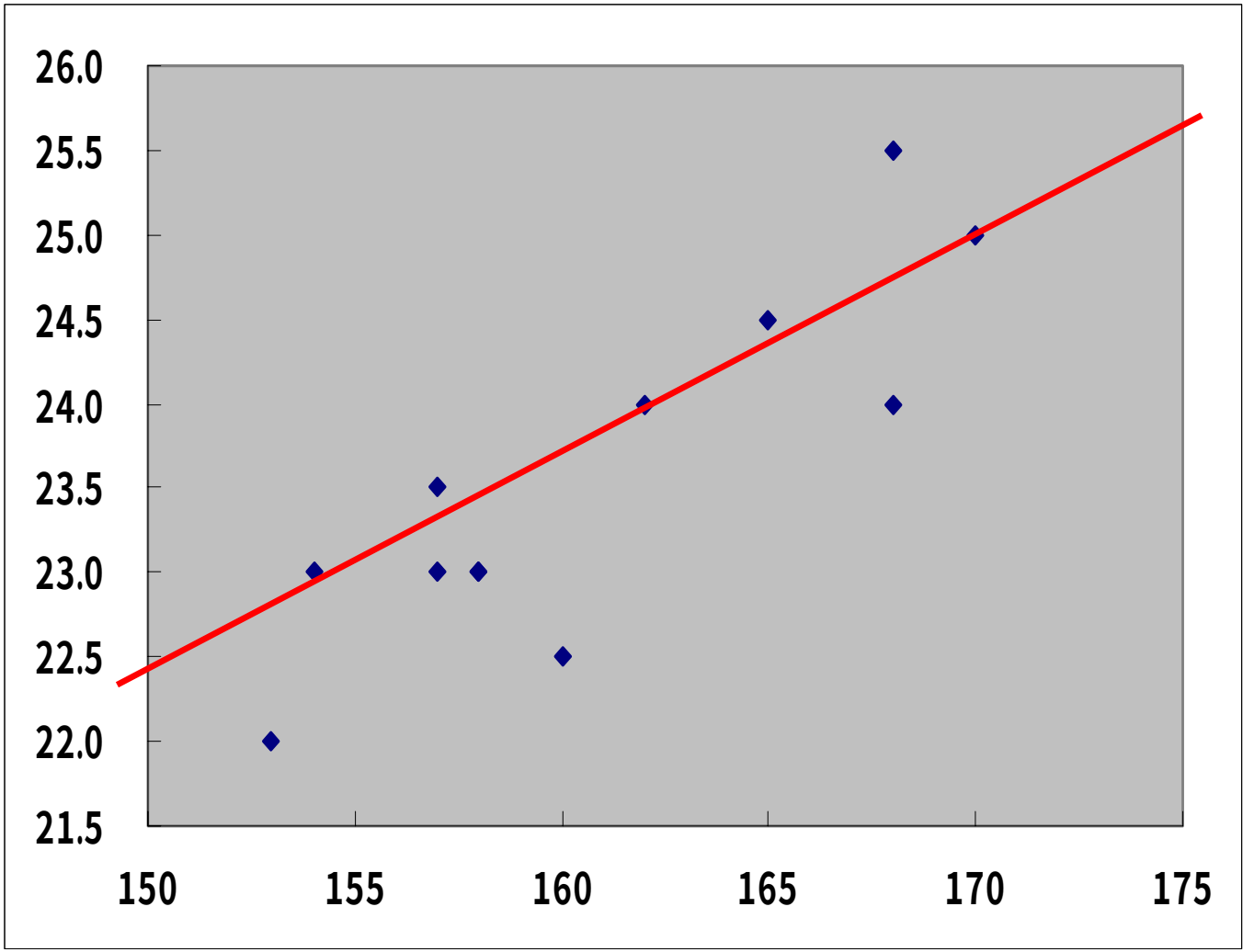
散布図



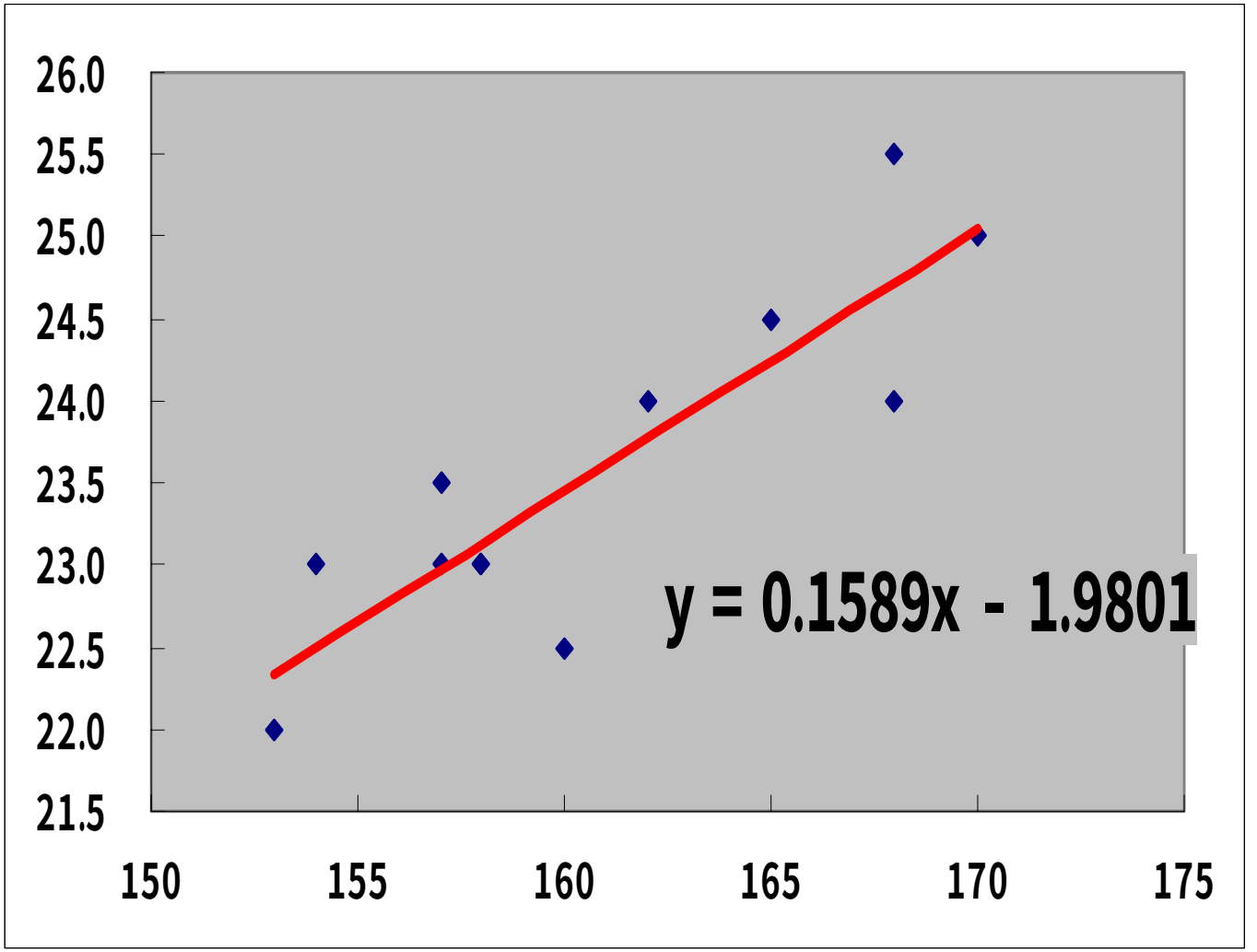
直線のあてはめ（1）



直線のあてはめ (2)



直線のあてはめ（3）


26.0
25.5
25.0
24.5
24.0
23.5
23.0
22.5
22.0
21.5
150
155
160
165
170
175
y = 0.1589x - 1.9801

単回帰分析の目的（の一つ）

与えられたデータに「最もよくあてはまる」直線

回帰方程式

$$y = a + bx \quad \cdots\cdots (1)$$

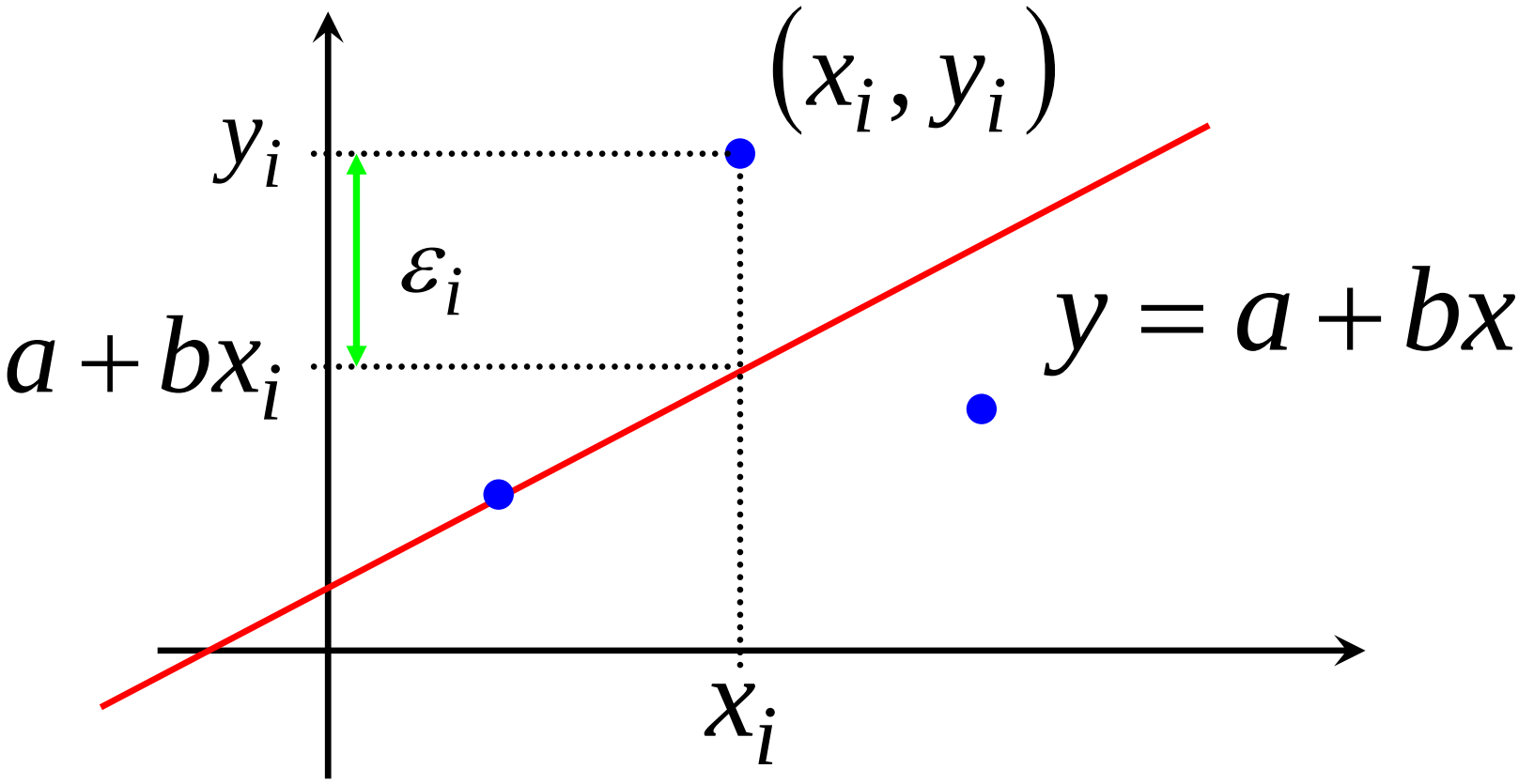
を求めること．

「最もよくあてはまる直線」ってどういうこと？

単回帰分析のデータ

個体番号	変数 x	変数 y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
n	x_n	y_n

残差 $\varepsilon_i = y_i - (a + bx_i)$



残差平方和 Q

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \{y_i - (a + bx_i)\}^2$$

Qを a と b を変数にもつ2変数関数として見て，$Q(a,b)$を最小にする a と b が，データに「最もよくあてはまる」直線を与えると考える．

このようにしてa と b を求める方法を**最小2乗法**と呼ぶ．

どのようにして$Q(a,b)$を最小にする a と bをもとめるのかを見ていく．

一般に多変数関数の極値（最大値，最小値）を求めるには，各変数で偏微分して0と置いた方程式系を解けばよい

$$\begin{cases} \dfrac{\partial Q}{\partial a} = \displaystyle\sum_{i=1}^{n} -2\{y_i - (a + bx_i)\} = 0, \\ \dfrac{\partial Q}{\partial b} = \displaystyle\sum_{i=1}^{n} -2x_i\{y_i - (a + bx_i)\} = 0 \end{cases}$$

連立方程式を解く（1）

$$\begin{cases} \sum_{i=1}^{n} \{y_i - (a + bx_i)\} = 0, \\ \sum_{i=1}^{n} x_i \{y_i - (a + bx_i)\} = 0 \end{cases}$$

連立方程式を解く（2）

$$\sum_{i=1}^{n}\{y_i-(a+bx_i)\}=0$$

$$\bar{y}=a+b\bar{x}$$

$$\sum_{i=1}^{n}x_i\{y_i-(a+bx_i)\}=0$$

連立方程式を解く（3）

$$\sum_{i=1}^{n} x_i \left\{ y_i - (a + bx_i) \right\}$$

$$= \sum_{i=1}^{n} x_i \left\{ y_i - (\bar{y} - b\bar{x} + bx_i) \right\}$$

$$= \sum_{i=1}^{n} x_i \left\{ (y_i - \bar{y}) - b(x_i - \bar{x}) \right\}$$

$$+ \sum_{i=1}^{n} \bar{x} \left\{ (y_i - \bar{y}) - b(x_i - \bar{x}) \right\}$$

$$= \sum_{i=1}^{n} (x_i - \bar{x}) \left\{ (y_i - \bar{y}) - b(x_i - \bar{x}) \right\}$$

$$= ns_{xy} - bns_x^2$$

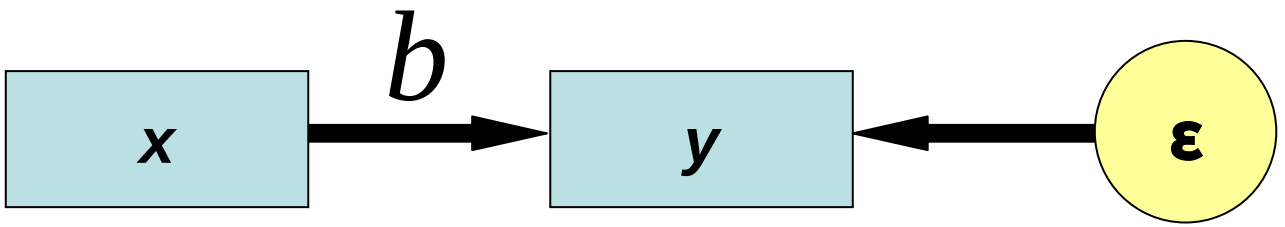
連立方程式の解

$$b = \frac{s_{xy}}{s_x^2},$$

$$a = \bar{y} - b\bar{x} = \bar{y} - \frac{s_{xy}}{s_x^2}\bar{x}$$

単回帰分析のパス図

$$y = a + bx$$



本日のまとめ

- パス図の読み方，書き方を理解した．
- 回帰分析に関わる用語：回帰方程式，説明変数，目的変数，などを理解した．
- 最小2乗法の考え方，及び，回帰方程式の求め方を理解した．
- Excelを用いて単回帰分析を行う方法を理解した．