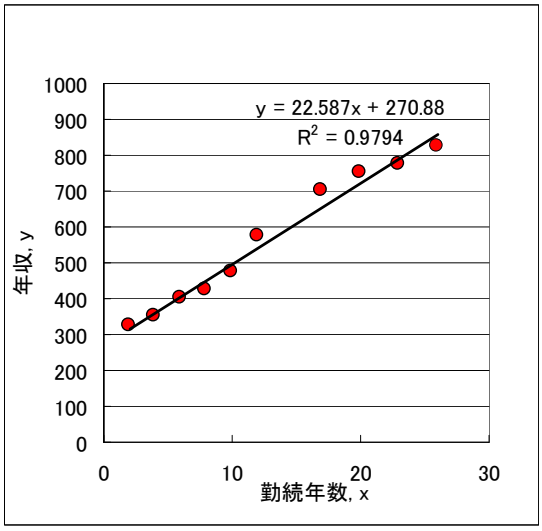
単回帰分析

１つの変数ｘから，１つの変数ｙを推定する．

例） 勤続年数と年収の関係を分析する．
直線で関係式を表現する．

$y = ax + b$　ｘ：説明変数
ｙ：目的変数

勤続年数, x	年収, y
2	325
4	350
6	400
8	425
10	475
12	575
17	700
20	750
23	775
26	825



単回帰分析 －最小２乗法による－

推定の誤差の２乗和を考え，これを最小にするようにa,bを決定する．

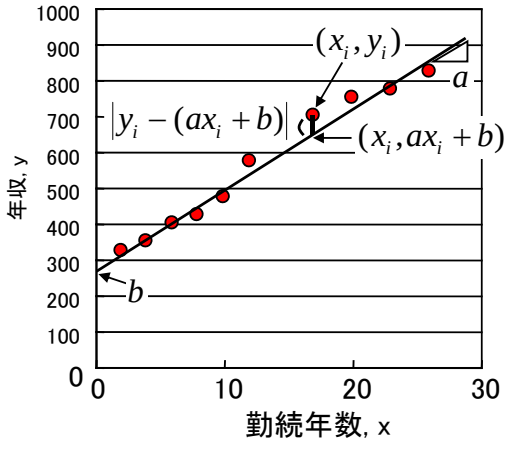
$$S(a,b) = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 \rightarrow \min.$$

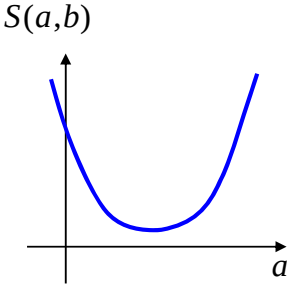
観測値　推定値

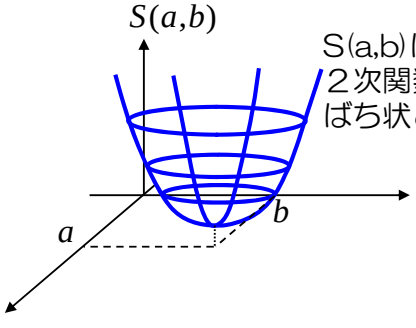
a, b:回帰パラメータ

条件：

S(a,b)をa, bでそれぞれ微分したものが０でなければならない．







S(a,b)はa, bに関する２次関数であり，すりばち状となる．

回帰関係の計算手順

推定の誤差の２乗和を考え，これを最小にするようにa,bを決定する.

$$S(a,b)=\sum_{i=1}^{n}\left[y_i-(ax_i+b)\right]^2 \quad \cdots(1)$$

S(a,b)をa, bでそれぞれ微分したものが０でなければならない.

$$\frac{\partial S(a,b)}{\partial a}=\sum_i 2\left[y_i-(ax_i+b)\right](-x_i)=0 \quad \cdots(2)$$

$$\frac{\partial S(a,b)}{\partial b}=\sum_i 2\left[y_i-(ax_i+b)\right](-1)=0 \quad \cdots(3)$$

（２），（３）は以下のように書き直せる.

$$\sum_i\left[y_i-(ax_i+b)\right]x_i=0 \quad \cdots(4)$$

$$\sum_i\left[y_i-(ax_i+b)\right]=0 \quad \cdots(5)$$

（４），（５）を整理すると，

$$a\sum_i x_i^2+b\sum_i x_i=\sum_i x_i y_i \quad \cdots(6)$$

$$a\sum_i x_i+nb=\sum_i y_i \quad \cdots(7)$$

これは，a,bに関する線形方程式になっている.
これを正規方程式(normal equations)という.
マトリクスで書けば，

$$\begin{bmatrix}\sum_i x_i^2 & \sum_i x_i\\ \sum_i x_i & n\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}\sum_i x_i y_i\\ \sum_i y_i\end{bmatrix} \quad \cdots(8)$$

a,bは以下の式により計算できる.

$$\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}\sum_i x_i^2 & \sum_i x_i\\ \sum_i x_i & n\end{bmatrix}^{-1}\begin{bmatrix}\sum_i x_i y_i\\ \sum_i y_i\end{bmatrix} \quad \cdots(9)$$

重回帰分析

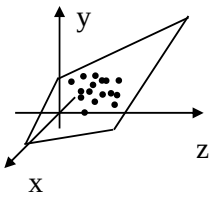
２変数ｘ，ｚから，１つの変数ｙを推定する.

$$\bar{y}_{xz}=ax+bz+c \quad \cdots(1)$$

１変数の場合と同様，推定の誤差の２乗を評価して，これを最小とするように，回帰係数 a, b, cを決定する.

$$S(a,b,c)=?$$

関数S(a,b,c)を極小とするa,b,cを求めるために，各変数による偏微分をとり，０とおいて解く.



サンプル
(x_i, z_i, y_i)

整理すると

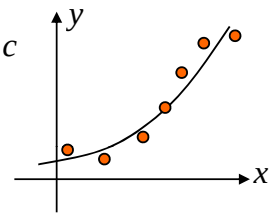
マトリクスで書けば

これを解いて回帰係数を得る.

重回帰分析　-高次項を用いた回帰-

課題：１変数の高次式から，１つの変数ｙを推定する．

$y = ax^2 + bx + c$



回帰係数を算出する計算式を導きなさい．

―決定係数と相関係数―

得られた標本について，回帰式（モデル式）によるあてはまりの程度を数値化する．

目的変数yの分散は以下の式で表される．

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \cdots(10)$$

一方，モデル式によって推定されたｙの値の，実測値からのばらつきを，以下の式によって評価する．

$$S_{y\cdot x}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - (ax_i + b)]^2 \quad \cdots(11)$$

（もし，モデル式による予測が完璧ならば，このばらつきは０になる．）

いま，

$$S_r^2 \equiv S_y^2 - S_{y\cdot x}^2 \quad \cdots(12)$$

という測度を考え，（１０）式の分散との比をとる．

$$r^2 = \frac{S_r^2}{S_y^2} = \frac{S_y^2 - S_{y\cdot x}^2}{S_y^2} = 1 - \frac{S_{y\cdot x}^2}{S_y^2} \quad \cdots(13)$$

モデル式による予測が確からしいほど，S_r^2はS_y^2に近づく．すなわち，r^2は１に近づく．

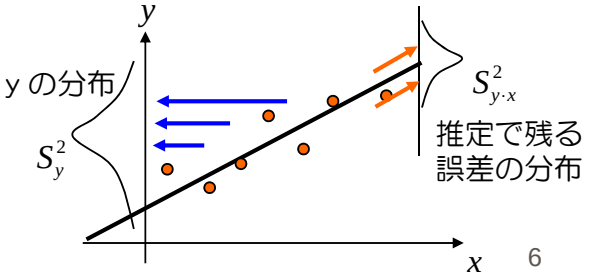
ｒ２は決定係数(coefficient of determination)と呼ばれ，以下の範囲をとる．

$$0 \le r^2 \le 1$$

また

$$r = \pm\sqrt{r^2} \qquad -1 \le r \le 1$$

を相関係数と呼ぶ．ｒの符号は回帰係数aの符号に合わせる．



回帰式の相関係数と２変数の相関係数との関係（補足資料）

回帰の誤差は，以下のように書き直せる．

$$
\begin{aligned}
S_{y\cdot x}^2 &= \frac{1}{n}\sum_i \left[y_i - (ax_i + b) \right]^2 \\
&= \frac{1}{n}\sum_i \left[(y_i - \bar{y}) + \bar{y} - (ax_i + b) \right]^2 \\
&= \frac{1}{n}\sum_i \left[(y_i - \bar{y}) + a\bar{x} + b - (ax_i + b) \right]^2 \\
&= \frac{1}{n}\sum_i \left[(y_i - \bar{y}) - a(x_i - \bar{x}) \right]^2 \\
&= S_y^2 - 2aS_{xy} + a^2 S_x^2 \\
&= S_y^2 - 2\frac{S_{xy}^2}{S_x^2} + \frac{S_{xy}^2}{S_x^4} S_x^2 \\
&= S_y^2 - \frac{S_{xy}^2}{S_x^2} = S_y^2 \left(1 - \frac{S_{xy}^2}{S_x^2 S_y^2} \right)
\end{aligned}
$$

最適化されたa,bに対して以下の関係が成り立つ

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

$\Leftarrow S_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$

$\Leftarrow a = \frac{S_{xy}}{S_x^2}$

したがって

$$\frac{S_{y\cdot x}^2}{S_y^2} = 1 - \frac{S_{xy}^2}{S_x^2 S_y^2} = 1 - r^2$$

すなわち，

$$r^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

これは，以前に定義した２変数間の相関係数，

$$r = \frac{1}{n}\sum_i \frac{(x_i - \bar{x})}{S_x}\frac{(y_i - \bar{y})}{S_y} = \frac{S_{xy}}{S_x S_y}$$

と同等である．

重回帰分析 –重決定係数，重相関係数–

２変数の場合も，単回帰と同様に，推定値の，実測値からの分散を考えることができる．

予測式，

$$\bar{y}_{xz} = ax + bz + c \quad \cdots(1)$$

を用いてｙを推定したときの，誤差の分散は

$$S_{y\cdot xz}^2 = \frac{1}{n}\sum_i [y_i - (ax_i + bz_i + c)]^2$$

いま，以下に示す２つの分散の差を考える．

$$S_r^2 = S_y^2 - S_{y\cdot xz}^2$$

ただし，

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \cdots(10)$$

である．

以下のように，２つの分散の比を考える．

$$R^2 = \frac{S_r^2}{S_y^2} = \frac{S_y^2 - S_{y\cdot xz}^2}{S_y^2} = 1 - \frac{S_{y\cdot xz}^2}{S_y^2}$$

R^2は，ｙの分散のうちｘとｚで説明される部分の割合を示している．

R^2　：重決定係数

$R = \sqrt{R^2}$　：重相関係数　$0 \le R \le 1$

２変数以上の説明変数に対して，目的変数との方向の一致，不一致を言えないため．