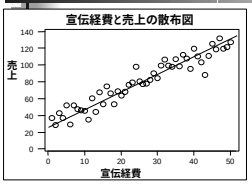
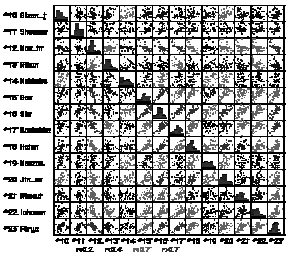
情報意味論 (5) 回帰 復習

櫻井彰人
慶應義塾大学理工学部

散布図



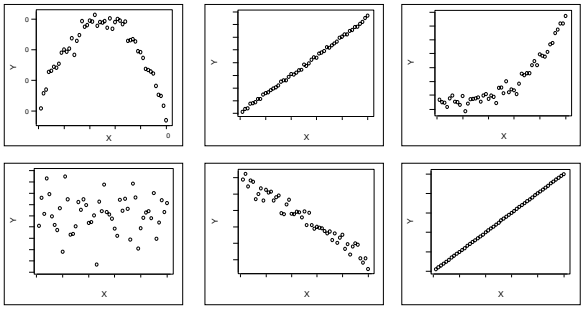
散布図：二変数を縦軸・横軸にとり、観察されたデータをプロットしたもの



多変数の場合、2変数ずつ組み合わせて散布図を描くことができる

http://aoki2.si.gunma-u.ac.jp/lecture/Dosuu/sanpuzu-2.png

様々な二変数間の関係



相関関係

- 分布図（散布図）を作成したとき、X軸成分とY軸成分に何らかの関係が見出せる場合がある。

- 正の相関：
  第一の変量の値が大きくなれば、他方も大きくなる。
- 負の相関：
  第一の変量の値が大きくなれば、他方は小さくなる。

- 相関関係を定量化する方法はいくつかある。
  - ピアソンの積率相関係数
  - スピアマンの順位相関係数、ケンドールの順位相関係数
  - 属性相関係数（クラメール係数、ファイ係数、コンティンジェンシー係数）

（積率）相関係数

二変数 X と Y の間の（積率）相関係数 は、その変数間の関係の「線型性」の程度を表す。

母集団の相関係数（ρ で表すことが多い）は -1 から 1 までの値をとる
サンプルの相関係数は r で表すことが多い

ρ=−1	完全な負の線型関係
-1< ρ <0	負の線型関係
ρ=0	線型関係がない
0< ρ <1	正の線型関係
ρ=1	完全な正の線型関係

線型以外の関係があっても、ρ に反映されないことがある

共分散と相関係数

二変数XとYの共分散:

$$Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]$$

但し μ_X と μ_Y は、変数Xと Yの母集団の平均値

母集団の相関係数は:

$$\rho=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

標本の相関係数は:

$$r=\frac{SS_{XY}}{\sqrt{SS_X SS_Y}}$$

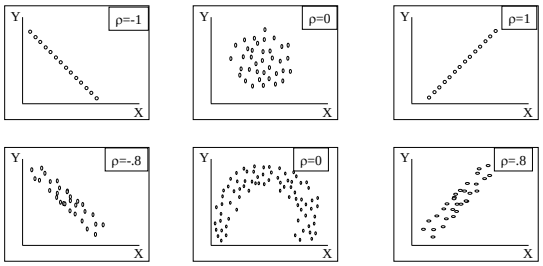
偏差平方和と偏差積和

$$SS_x=\sum(x-\bar{x})^2=\sum x^2-\frac{(\sum x)^2}{n}$$

$$SS_y=\sum(y-\bar{y})^2=\sum y^2-\frac{(\sum y)^2}{n}$$

$$SS_{xy}=\sum(x-\bar{x})(y-\bar{y})=\sum xy-\frac{(\sum x)(\sum y)}{n}$$

相関係数との関係



Excel による相関分析

回帰統計									
重相関 R	0.926424								
重決定 R2	0.858261								
補正 R2	0.826764								
標準誤差	43245.26								
観測数	12								
分散分析表									
	自由度	変動	分散	測された分散	有意 F				
回帰	2	1.02E+11	5.1E+10	27.24859	0.000152				
残差	9	1.68E+10	1.87E+09						
合計	11	1.19E+11							
	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%	
切片	-109435	68205.53	-1.60449	0.14307	-263727	44856.79	-263727	44856.79	
X 値 1	922.7078	159.3518	5.790383	0.000263	562.2287	1283.187	562.2287	1283.187	
X 値 2	7219.01	3257.642	2.216023	0.053913	-150.294	14588.31	-150.294	14588.31	

残差出力		
観測値	予測値 : Y	残差
1	477634.1	31904.87
2	396814.4	-17260.4
3	209255.7	-36240.7
4	173356.4	22015.57
5	356711.2	8596.774
6	333023	70762.97
7	254717.4	-14160.4
8	207023.7	-29585.7
9	420665.2	-38392.2
10	241392.8	47598.18
11	346131.7	14780.29
12	366192.3	-60019.3

意味のある相関・ない相関

・相関があっても、必ずしも変数Xと変数Yの間に論理的な因果関係があるわけではない。相関がない場合も同様。
・変数Xと変数Yとの関係において、隠れた変数Zの存在が影響していることがある。(擬似相関)

【例】 X:子どもの手の大きさ Y:筆記能力→正の相関
隠れた変数Z:年齢
【例】 X:年収 Y:100m競争のタイム →負の相関
隠れた変数Z:年齢

相関係数の検定

H_0: ρ=0 **(線型関係なし)**
H_1: ρ≠0 **(線型関係なしとはせず)**
(H_0**は帰無仮説、** H_1**は対立仮説)**

検定統計量: $t_{(n-2)} = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$

$t_{(n-2)}$ **は, 自由度が** n − 2 **の** t **分布に従う。**
有意確率を $P = \Pr\{|t| \geq t_{(n-2)}\}$ **とすればよい。**

数値例:

$$t_{(n-2)} = \frac{r}{\sqrt{\dfrac{1-r^2}{n-2}}} = \frac{0.9824}{\sqrt{\dfrac{1-0.9651}{25-2}}} = \frac{0.9824}{0.0389} = 25.25$$

$t_{0.005} = 2.807 < 25.25$
H_0 は有意水準1%で棄却される

参考: http://aoki2.si.gunma-u.ac.jp/lecture/Corr/corr.html

順位相関係数(順序変量の場合に用いられる相関係数)

スピアマンの順位相関係数

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

ただし $d_i = x_i - y_i, \quad i = 1,2,\cdots n$

・相関係数 r の性質はそのまま当てはまる。
−1≦r≦1
絶対値が大きいほど、相関が強い
マイナスの時、負の相関、プラスのとき正の相関

【例】 Jリーグ16チームの年間順位と観客動員数順位

線型単回帰モデル

母集団の線型単回帰モデル:

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{非ランダム または 構造的な 要素}} + \underbrace{\varepsilon}_{\text{ランダムな 要素}}$$

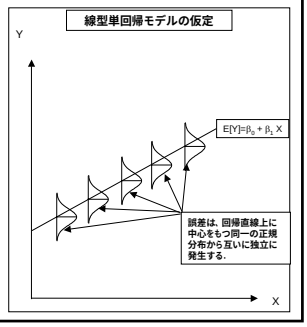
ここで Y **は 被説明(従属) 変数, すなわち、説明ないし予測したい変数;** X **は 説明(独立) 変数; そして** ε **は 誤差項, このモデルの唯一のランダムな要素, 従って,** Y **に含まれるランダム性の唯一の源.**

β_0 は非ランダムな要素の切片
β_1 は非ランダムな要素の傾き

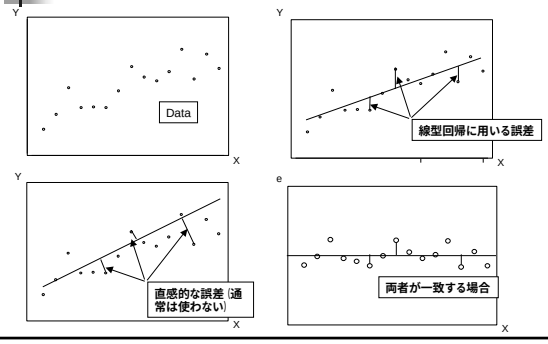
Y **の条件付期待値:** $E[Y|X] = \beta_0 + \beta_1 X$

線型単回帰モデルの仮定

- X と Y の関係は線型である
- 説明変数 X はランダムではない; Y に含まれる唯一のランダム性は誤差項 ε_i から生ずる.
- 誤差 ε_i は平均 0 分散 σ^2 の正規分布に従う. 誤差は互いに独立である. すなわち; $\varepsilon \sim N(0,\sigma^2)$



回帰直線と回帰誤差の考え方



正規方程式

$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ を最小化することを考える

$f(b_0, b_1) \equiv \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$ とおくと

$$\frac{\partial f}{\partial b_0} = -2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial f}{\partial b_1} = -2\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i)$$

となる。 これらを0とする停留点を求めることにする。
その方程式は、 次のようになる

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

正規方程式の解

偏差 平方和と偏差積和：

$$SS_x = \sum (x - \bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$SS_y = \sum (y - \bar{y})^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

回帰係数の推定量：

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{SS_{XY}}{SS_X}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

残差分散と回帰係数の推定

回帰における自由度:
$df = (n-2)$ (nから、推定したパラメータ(b_0とb_1)のそれぞれにつき1自由度を減じたもの)

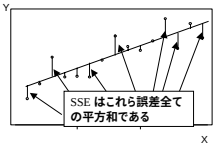
残差平方和：

$$SSE = \sum (Y - \hat{Y})^2 = SS_Y - \frac{(SS_{XY})^2}{SS_X}$$

$$= SS_Y - b_1 SS_{XY}$$

残差の不偏分散：
s^2はσ^2の不偏推定量:

$$s^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{(n-2)}$$



数値例:

$SSE = SS_Y - b_1 SS_{XY}$
$= 66855898 - (1.255333776)(51402852.4)$
$= 2328161.2$

$MSE = \frac{SSE}{n-2} = \frac{2328161.2}{23}$
$= 101224.4$

$s = \sqrt{MSE} = \sqrt{101224.4} = 318.158$