10　相関係数と最小二乗法

実験データの解析では，2 種類以上のデータの間にどのような関係があるのかを調べる必要があることが多い。ここでは，そのような場合に重要になる「相関係数」と「最小二乗法」について学ぶ。

10.1　相関係数 correlation coefficient

n 組のデータ $(x_1, y_1), \cdots, (x_n, y_n)$ があるとして，偏差二乗和 S_{xx}, S_{yy} と偏差積和 S_{xy} を次のように定義する。

$$S_{xx} = \sum_{i=1}^{n} (x_i - \langle x \rangle)^2 \tag{10.1}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \langle y \rangle)^2 \tag{10.2}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \langle x \rangle)(y_i - \langle y \rangle) \tag{10.3}$$

ただし $\langle x \rangle, \langle y \rangle$ は，それぞれ x_i, y_i の平均である。

$$\langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{10.4}$$

$$\langle y \rangle = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{10.5}$$

このとき，相関係数 r は次のように定義される。

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \tag{10.6}$$

相関係数は $-1 \leq r \leq 1$ の値をとる。

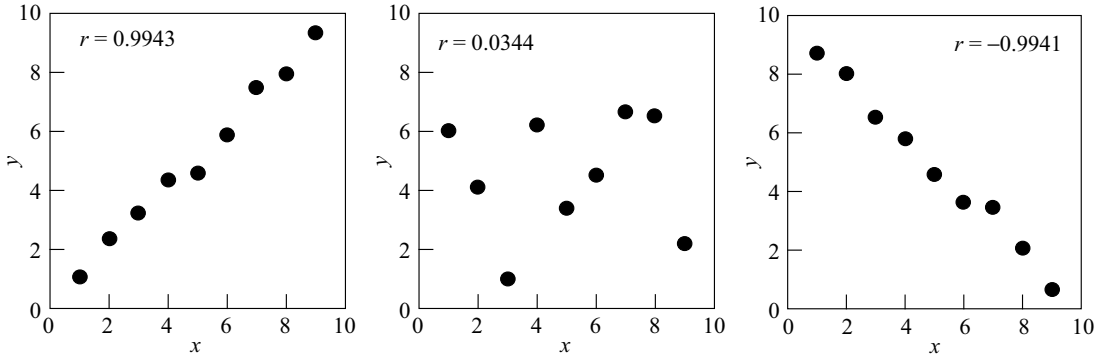
次の量 V_{xy} を共分散 covariance と呼ぶ。

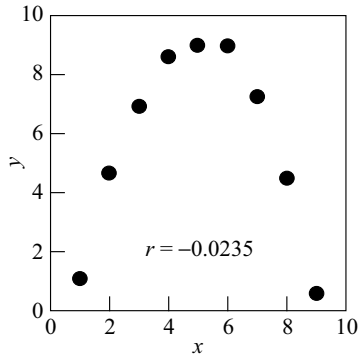
$$V_{xy} = \frac{S_{xy}}{n} \tag{10.7}$$

目的によって，共分散の定義式の分母は n ではなく $(n-1)$ とする場合もある。

10.2　相関図

横軸に x_i, 縦軸に y_i をとった図を「相関図」という。





r が 1 に近いときには相関図は右上がりの直線になる。r が 0 に近いときにはデータはバラバラである。r が −1 に近いときには右下がりの直線になる。

ただし，左下の場合を見てわかるように，相関係数が 0 に近いからといって，x_i と y_i の間に一定の関係が存在しないとはいえない。

10.3 Excel による相関係数の計算

Excel 2007

1. Web ページから相関係数説明用ファイルをダウンロードし，Excel で開く。

	A	B	C	D	E	F	G	H	I	J	K
1	データ	x	ya	yb	yc	yd	偏差の積	ya	yb	yc	yd
2		1	1.053776	6.032997	8.71773	1.11241					
3		2	2.345283	4.128298	8.033126	4.687426					
4		3	3.236039	1.007231	6.545628	6.925556					
5		4	4.351326	6.214122	5.819709	8.599049					
6		5	4.583612	3.412494	4.572589	8.975689					
7		6	5.866946	4.51902	3.609331	8.953419					
8		7	7.488722	6.650281	3.452077	7.241908					
9		8	7.953277	6.546639	2.05829	4.487617					
10		9	9.340896	2.20379	0.676785	0.60003					
11											
12	平均										
13	偏差二乗和										
14	積和										
15	共分散										
16	共分散	excel									
17	相関係数										
18	相関係数	excel									

x と y_a, x と y_b, x と y_c, x と y_d という 4 つの組み合わせについて相関係数を求める。前節の図に対応している。

2. ワークシート名を「soukan」から「相関係数」に変更する。
3. 課題提出用のファイル名で，Excel 形式で保存する。
4. 平均と偏差二乗和については Excel の関数で計算できる。
   (i) B12 に「=AVERAGE(B2:B10)」，B13 に「=DEVSQ(B2:B10)」と入力。
   (ii) B12, B13 をコピーして，C12:F13 の四角い範囲に貼り付ける。
5. 偏差積和の計算準備として，平均からの偏差の積を H2:K10 の四角い範囲に計算する。
   (i) H2 に「=($B2−$B$12)∗(C2−C$12)」と入力。
   (ii) H2 をコピーして，H2:K10 の四角い範囲に貼り付ける。
   - 「$B2」という参照は，列は B のまま固定（絶対参照）で，行は現在の行から見た相対位置。
   - 「C$12」という参照は，列は現在の列からみた相対位置，行は 12 のまま固定（絶対参照）。
   - 「$B$12」という参照は，列は B，行は 12 で，両方とも固定（絶対参照）。
   - 今回の例では x が共通に用いられているので，このような書き方をするとコピー & ペーストが楽。
6. 偏差積和と共分散の計算
   (i) C14 に「=SUM(H2:H10)」，C15 に「=C14/COUNT(C2:C10)」と入力。
   (ii) C14, C15 をコピーして，D14:F15 の四角い範囲に貼り付ける。
   (iii) Excel には, 共分散を計算する関数が用意されているので, C16 に「=COVAR($B2:$B10,C2:C10)」と入力。
   (iv) C16 をコピーして D16:F16 の範囲に貼り付ける。
7. 相関係数の計算

(i) C17 に「=C14/SQRT(C13*$B13)」と入力。
(ii) C17 をコピーして D17:F17 の範囲に貼り付ける。
(iii) Excel には, 相関係数を計算する関数が用意されているので, C18 に「=CORREL($B2:$B10,C2:C10)」と入力。C18 をコピーして D18:F18 の範囲に貼り付ける。

10.4 Excel による相関図の作成

例として x と y_d の組み合わせについて

Excel 2007

1. Ctrl をうまく用いて，B1:B10 と F1:F10 の 2 つの範囲が選択されている状態にする（このとき B1:F10 のひとつの四角い範囲にならないように注意する）。
2.「挿入」タブ →「グラフ」グループ →「散布図」クリックで現れるメニューの最初のもの（プロットのみで線なし）を選択。

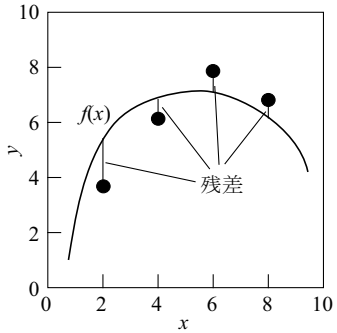
3. 目盛り等の細かい指定はグラフを描いたあとで，必要に応じて変更する。

Excel 2003

1. メニュー [挿入] → [グラフ] →「散布図」で，形式は一番上のプロットのみ（線なし）を選択 → [完了]。

10.5 最小二乗法によるフィッティングの原理

n 組のデータ $(x_1,y_1),\cdots,(x_n,y_n)$ があり，x と y の関係は独立なパラメータ (定数や係数) を $(m+1)$ 個 $(a_0,\cdots,a_m)$ 含む関数 $y=f(x)$ で表すことができると期待されているとする。このとき，残差二乗和 S を次のように定義する。



$$S=\sum_{i=1}^{n}[y_i-f(x_i)]^2 \tag{10.8}$$

S が最小になるように $(m+1)$ 個のパラメータ $a_0,\cdots,a_m$ を決める方法を「最小二乗法」least mean-square method という。

S が最小になるためには，$0\le j\le m$ のすべての j に関して，次の条件が同時に満たされなければならない。

$$\frac{\partial S}{\partial a_j}=-\sum_{i=1}^{n}2\frac{\partial f(x_i)}{\partial a_j}[y_i-f(x_i)]=0 \tag{10.9}$$

このようにしてパラメータを決定して得られた曲線 は，「回帰曲線」regression curve（直線の場合は「回帰直線」）と呼ばれる。

回帰曲線が実測値をどの程度うまく再現しているかを評価するためには，いくつかの目安がある。

1. 標準誤差 σ standard error（小さいほどよい）

$$\sigma = \sqrt{\frac{S}{n-(m+1)}} \tag{10.10}$$

2. 実測値 y_i と予測値 $f(x_i)$ の相関係数 r （ 1 に近いほどよい）

$$r = \frac{\sum [y_i - \langle y \rangle][f(x_i) - \langle f(x) \rangle]}{\sqrt{\sum [y_i - \langle y \rangle]^2 \sum [f(x_i) - \langle f(x) \rangle]^2}} \tag{10.11}$$

ただし $\langle f(x) \rangle$ は次のように定義する。

$$\langle f(x) \rangle = \frac{1}{n} \sum_{i=1}^{n} f(x_i) \tag{10.12}$$

3. 決定係数 R^2 （ 1 に近いほどよい）

$$R^2 = 1 - \frac{S}{\sum [y_i - \langle y \rangle]^2} = r^2 \tag{10.13}$$

10.6 直線フィッティング

$$f(x) = a_0 + a_1 x \tag{10.14}$$

このとき，残差二乗和 は次のように書ける。

$$S = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i)]^2 \tag{10.15}$$

S が最小になるには，次の 2 つの条件を同時に満たす必要がある。

$$\frac{\partial S}{\partial a_0} = -\sum_{i=1}^{n} 2(y_i - a_0 - a_1 x_i) = 0 \tag{10.16}$$

$$\frac{\partial S}{\partial a_1} = -\sum_{i=1}^{n} 2x_i(y_i - a_0 - a_1 x_i) = 0 \tag{10.17}$$

この条件を整理すると，a_0 と a_1 に関する次のような方程式になる。

$$\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \tag{10.18}$$

この方程式は「正規方程式」と呼ばれる。これを次のような記号で書くことにする。

$$\mathbf{Aa} = \mathbf{b} \tag{10.19}$$

前回の授業を思い出せば，この方程式はすぐに解くことができる。

$$\mathbf{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \mathbf{A}^{-1}\mathbf{b} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} \tag{10.20}$$

10.7 Excel による直線フィッティング

直線フィッティングの場合，Excel には，(x_1, y_i) のデータセットから傾きや切片を求める関数が用意されている。

Excel 2007

1. Web から直線回帰用ファイルをダウンロードし，Excel で開く。

	A	B	C	D	E
1	x	y	xy	f(x)	残差
2	1	5.328618			
3	2	6.564359			
4	3	8.781574			
5	4	11.05264			
6	5	13.15991			
7	6	14.73167			
8	7	16.96247			
9	8	18.63649			
10	9	21.20744			
11	A			b	
12					
13					
14	A(inv)		a	x=A(inv)b	excel
15			a0		
16			a1		
17		sigma	r	r^2	R^2
18					
19	excel				

2. 「ホーム」→「セル」→「書式」→「ワークシートの移動またはコピー」で，「相関係数」と同じファイルにワークシートを移動する。
3. ワークシート名を「kaiki」から「直線回帰」に変更する。
4. まず C2:C10 に $x_i y_i$ を求めておく。
   (i) C2 に「=A2∗B2」と入力。
   (ii) C2 をコピーして C3:C10 の範囲に貼付け。
5. 行列 $\mathbf{A}$ の各要素を A12:B13 の四角い範囲に計算し，逆行列 $\mathbf{A}^{-1}$ を A15:B16 の四角い範囲に計算する。
   (i) A12 はデータ数 n なので，「=COUNT(A2:A10)」。
   (ii) A13 と B12 はともに x_i の和なので，「=SUM(A2:A10)」。
   (iii) B13 は x_i^2 の和なので，「=SUMSQ(A2:A10)」。
   (iv) 逆行列は A15:B16 を選択した状態で「=MINVERSE(A12:B13)」として Ctrl+Shift+Enter。
6. ベクトル $\mathbf{b}$ の各要素を D12:D13 に計算する。
   (i) D12 は y_i の和なので，「=SUM(B2:B10)」。
   (ii) D13 は $x_i y_i$ の和なので，「=SUM(C2:C10)」。
7. $\mathbf{a} = \mathbf{A}^{-1}\mathbf{b}$ を D15:D16 に計算すれば，D15 が回帰直線の切片，D16 が傾きになる。
   (i) D15:D16 を選択した状態で「=MMULT(A15:B16,D12:D13)」としてから Ctrl+Shift+Enter。
8. 実は回帰直線の切片と傾きを求める Excel 関数が用意されている。
   (i) E15 に「=INTERCEPT(B2:B10,A2:A10)」と入力。
   (ii) E16 に「=SLOPE(B2:B10,A2:A10)」と入力。
   [Excel 関数の説明] 「INTERCEPT」は直線回帰した場合の切片を，「SLOPE」は傾きを求める。これらの引数は 2 つあるが，はじめの引数で y の範囲を指定してから次の引数で x の範囲を指定するので注意。
9. 各 x_i について D2:D10 に $f(x_i)$ を計算し，E2:E10 に残差 $(y_i - f(x_i))$ を計算する。
   (i) D2 に「=$D$15+$D$16∗A2」，E2 に「=B2−D2」と入力。
   (ii) D2:E2 をコピーして，D3:E10 の四角い範囲に貼り付け。
10. 標準誤差 σ，y_i と $f(x_i)$ の相関係数 r，相関係数の二乗 r^2，決定係数 R^2 を求める。

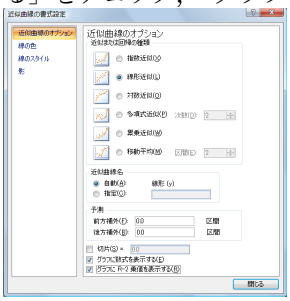
(i) B18 に「=SQRT(SUMSQ(E2:E10)/ (COUNT(A2:A10)−2))」。
(ii) C18 に「=CORREL(B2:B10,D2:D10)」。
(iii) D18 には「=C18^2」。
(iv) E18 には「=1−SUMSQ(E2:E10)/ DEVSQ(B2:B10)」。
(v) σ を求めるには Excel 関数がある。B19 に「=STEYX(B2:B10,A2:A10)」。
(vi) R^2 を求めるには Excel 関数がある。E19 に「=RSQ(B2:B10,A2:A10)」。

10.8 Excel のグラフの機能を利用

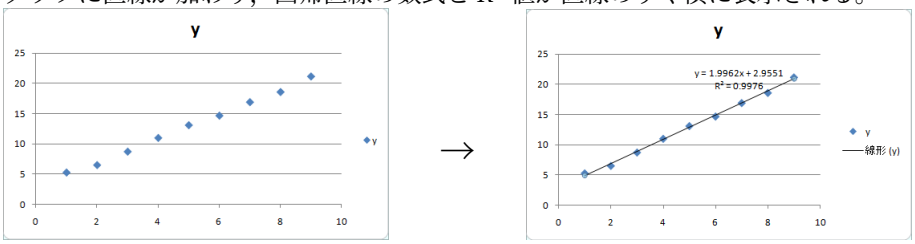
Excel では，グラフを描くともっと簡単に回帰直線を得ることができる。

Excel 2007

1. x_i と y_i の相関図を描くために A1:B10 の四角い範囲を選択する。
2.「挿入」タブ →「グラフ」グループ →「散布図」クリックで現れるメニューの最初のもの（プロットのみで線なし）を選択。
3. 目盛り等の細かい設定は必要に応じて指定すること。
4. プロットのシンボルのうちどれかにマウスポインタを合わせ右クリック →「近似曲線の追加」。
5.「近似曲線のオプション」で，「近似または回帰の種類」を「線形近似」に，「グラフに数式を表示する」をチェック，「グラフに R-2 乗値を表示する」をチェックして「閉じる」。



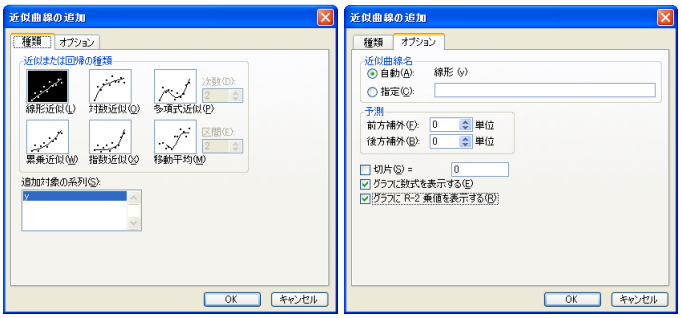
6. グラフに直線が加わり，回帰直線の数式と R^2 値が直線のすぐ横に表示される。



7. 一次関数（直線）以外の関数へのフィッティングも用意されている。

Excel 2003

1. メニュー [挿入] → [グラフ] →「散布図」で，形式は一番上のプロットのみ（線なし）を選択 → [完了]。
2. プロットのシンボルのうちどれかにマウスを合わせ右クリック → [近似曲線の追加]。
3.「種類」は左上の「線形近似」（一次関数のこと）。



4.「オプション」で「グラフに数式を表示する」,「グラフに R-2 乗値を表示する」を ON にして [OK]。

10.9　多項式フィッティングの原理

より一般的な場合として，y が x に関する m 次の多項式として表すことができると期待されているとする。

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m \tag{10.21}$$

このとき，残差の二乗和 S は次のように定義される。

$$S = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i + \cdots + a_m x_i^m)]^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m)^2 \tag{10.22}$$

S が最小になるには，次の $(m+1)$ 個の条件を同時に満たす必要がある。

$$\frac{\partial S}{\partial a_0} = -\sum_{i=1}^{n} 2\,(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m) = 0 \tag{10.23}$$

$$\frac{\partial S}{\partial a_1} = -\sum_{i=1}^{n} 2x_i\,(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m) = 0 \tag{10.24}$$

$\cdots$

$\cdots$

$$\frac{\partial S}{\partial a_m} = -\sum_{i=1}^{n} 2x_i^m\,(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m) = 0 \tag{10.25}$$

そして，正規方程式は次のように書ける。

$$\begin{pmatrix} n & \sum x_i & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_i^m & \sum x_i^{m+1} & \cdots & \sum x_i^{2m} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^m y_i \end{pmatrix} \tag{10.26}$$

この方程式の解は，次のように書くことができる。

$$\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} n & \sum x_i & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_i^m & \sum x_i^{m+1} & \cdots & \sum x_i^{2m} \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^m y_i \end{pmatrix} \tag{10.27}$$

特に $m = 2$ の場合について書くと次のようになる。

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{pmatrix} \tag{10.28}$$