第3回：2次元のデータ

尚　晋
大学院経済学研究科　助教

2025年4月30日

今日のポイント

1. 2次元のデータとは
2. 散布図と分割表
3. 相関係数
4. 直線および平面のあてはめ

1　2次元のデータについて

- 単一の変数xでなく，2変数x, y, あるいは3変数x, y, zなどを観測して，n個(n組)のデータを得る場合，そのデータを**多次元データ**という。
- 一般に、p個の変数をとり扱う場合，p次元データという。
- 多次元データの統計学は，多くの変数を一括してその間の**関係**を扱うもの.
- 簡単のため、$p = 2$、2変数x, yの関係を考えてみると，xとyの間に区別をもうけず対等に見る見方や方法を**相関**といい，xからy(あるいはyからx)を見るとき，**回帰**という.
- 相関関係として見るのが良い：例えば、身長と体重、どちらがどちらを決めるとも言えない。
- 単に相関関係があるだけでなく，ある一方が他方を左右する(決定する)という一方向の関係にある場合，分析には回帰分析の方法がふさわしい：年齢と血圧、所得と貯蓄。

2　散布図と分割表

2.1　散布図

- 2次元のデータをn個の点, $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$ として図示すると、 xとyの関係ははっきりする。
- xとy両方が**量的データ**である場合、横軸にx、縦軸にyをとって、各観測対象を平面にプロットした図は**散布図**という。
- 散布図上で各点がバラバラに散らばればxとyは関係がなく，逆に点の分布が何らかの傾向を示せば xとyとは関係がありそうとわかる.

表3.1: 2次元データの例(1):11家族内での兄弟と姉妹の身長の組

x :	男(兄弟)	71	68	66	67	70	71	70	73	72	65	66	(インチ)
y :	女(兄弟)	69	64	65	63	65	62	65	64	66	59	62	(インチ)

表3.2: 2次元データの例(2):年齢階級(中点)と血圧の平均

x :	年齢階級	35	45	55	65	75	(歳)
y :	血圧の平均	114	124	143	158	166	(mmHg)

図3.1と図3.2は表3.1の11家族内での兄弟と姉妹の身長の組と表3.2の年齢階級(中点)と血圧の平均を基づいてプロットした散布図である。図3.1の散布図の例では、兄弟の身長と姉妹の身長の関係はあいまいで見出し難い.図3.2の散布図の例では、年齢が高進すると血圧も上昇する傾向がはっきり表れている.

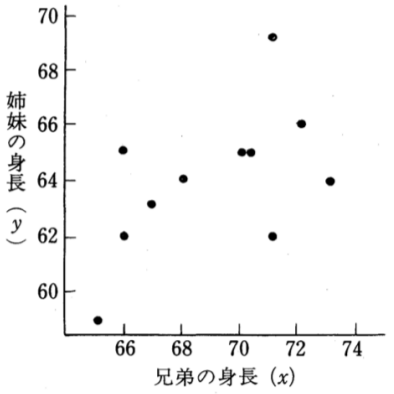

図3.1　例(1)をグラフにした図（n=11）

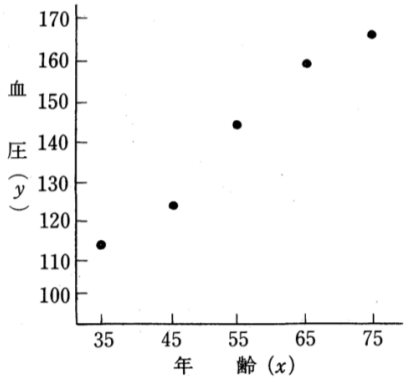

図3.2　例(2)をグラフにした図（n=5）

図3.3から図3.6は，いずれも散布図の例であり，日本の47都道府県について作成したものである.一方が増加すれば他方は増加するか減少するか，あるいはそのような傾向自体が強いか弱いか，散布図を見れば一目瞭然である.後に述べる相関係数rの値よりも，現象の質的側面の観察ではすぐれている.統計学とは「数字の計算」であるという考え方が不十分である一つの例証である.関係の分析はまず散布図から入るべきであろう.

- 二つの変数間の関係のことを，一般に**相関関係**と呼ぶが，特に統計学では二つの変数の間に**直線関係**に近い傾向が見られるときに「相関関係がある」ということが多い.

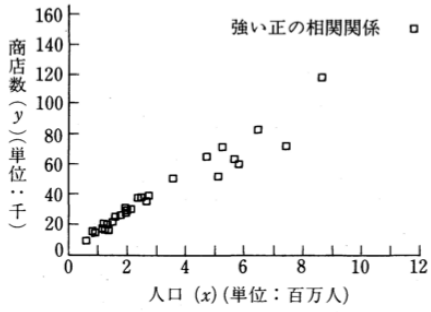

図3.3 人口と小売商店数の散布図

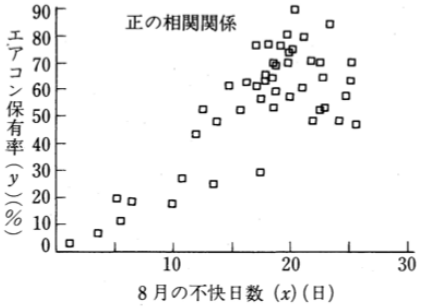

図3.4 8月の平均不快日数*)とルーム・エアコンの保有率の散布図

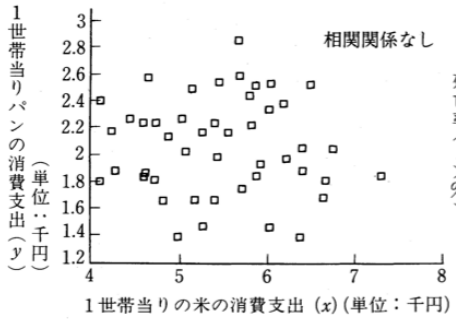

図3.5 世帯・月あたり米の消費支出とパンの消費支出の散布図

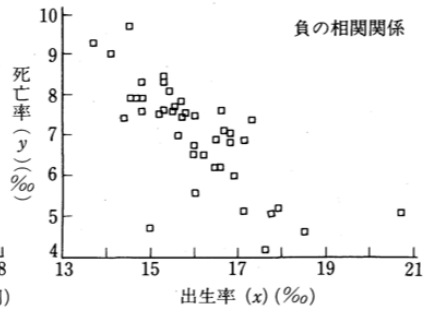

図3.6 出生率と死亡率の散布図

- 「正の相関関係がある」,「負の相関関係がある」,直線的な傾向の程度は「強い」「弱い」と表現する.

2.2 分割表

- 両方とも質的データの場合には，2変量データの度数分布表を**分割表**という.クロス表ともいう。

- 縦方向にある変数を表側，横方向にある変数を表頭と呼ぶ.表側の項目と表頭の項目が交差(クロス)する細目のます目に対応する度数を書き込んだものが分割表である。

表3.4は1989年の東京大学大学院における修士課程・博士課程の別と日本人・留学生の別に学生数を分割表にしたものである.表3.5の5x2分割表は全学について同様のものを作成した分割表である。

分割表で二つのデータの関係を見るには，相対度数を用いる.表3.5の相対度数を計算した表が表3.6である.相対度数は3種類できる.横方向の相対度数(各ますの上段の数字)，縦方向の相対度数(各ますの中段の数字)，右下コーナーのますに表れるデータ全体の大きさを分母とした相対度数(各ますの下段の数字)の三つである.

両方のデータとも量的データであっても，適当な階級に分ければ分割表ができる，

3 相関係数

3.1 積率相関係数

- 相関係数とは相関の程度を示す指標のことであり，多くの定義があるが，そのうちもっともよく用いられるものは，ピアソンの積率相関係数であり，単に相関係数というときには通常これをさしている.積率相関係数はデータがともに量的変数である場合に用いられる.

表 3.4　東京大学大学院の学生構成

	日本人	留学生	合　計
修 士 課 程	2,415	274	2,689
博 士 課 程	2,002	620	2,622
合　　計	4,417	894	5,311

（単位：人）

表 3.5　東京大学学部・大学院の学生構成

	日本人	留学生	合　計
学　　部	14,871	96	14,967
学部研究生	252	17	269
修 士 課 程	2,415	274	2,689
博 士 課 程	2,002	620	2,622
大学院研究生	143	454	597
合　　計	19,683	1,461	21,144

（単位：人）

分割表の例と相対度数のいろいろ

表 3.4 は最も簡単な 2×2 分割表，表 3.5 は，5×2 分割表．表 3.6 は表 3.5 を相対度数で示したもの．通常は表側（ひょうそく）がより根元的なので，横方向の相対度数が表示されることが多い．

（出典：The University of Tokyo 1989〜1990）

表 3.6　東京大学学部・大学院の学生構成
横比（上段），縦比（中段），および全度数に対応する相対度数（下段）．

	日本人	留学生	合　計
学　　部	99.4	0.6	100.0
	75.6	6.6	70.8
	70.3	0.5	70.8
学部研究生	93.7	6.3	100.0
	1.3	1.2	1.3
	1.2	0.1	1.3
修 士 課 程	89.8	10.2	100.0
	12.3	18.8	12.7
	11.4	1.3	12.7
博 士 課 程	76.4	23.6	100.0
	10.2	42.4	12.4
	9.5	2.9	12.4
大学院研究生	24.0	76.0	100.0
	0.7	31.1	2.8
	0.7	2.1	2.8
合　　計	93.1	6.9	100.0
	100.0	100.0	100.0
	93.1	6.9	100.0

（単位：%）

データが $(x_i, y_i), (x_2, y_2), ..., (x_n, y_n)$で与えられた場合，変数$x$と$y$の間の相関係数は:

$$r_{xy} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})/n}{\sqrt{\Sigma(x_i-\bar{x})^2/n}\sqrt{\Sigma(y_i-\bar{y})^2/n}} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}}$$

で定義される.

- 上式の分母で，$\sqrt{\Sigma(x_i-\bar{x})^2/n}$、$\sqrt{\Sigma(y_i-\bar{y})^2/n}$は，変数$x, y$のそれぞれの標準偏差$S_x, S_y$である.
- 分子は$C_{xy} = \Sigma(x_i-\bar{x})(y_i-\bar{y})/n$は$x$の偏差$x_i-\bar{x}$と$y$の偏差$y_i-\bar{y}$を同時に考えたとき(偏差の積は偏差積と呼ぶ)の全データについての平均で，これを**共分散**と呼ぶ.
- 相関係数のつねに$-1 \leq r_{xy} \leq 1$の範囲にある.
- $r_{xy} = 1$の場合、これを正の完全相関という.逆に負の完全相関は$r_{xy} = -1$のときである，
- 図3.7から図3.10は概ね$r_{xy} = 0.9, 0.7, 0.5, 0.3$の場合のデータの散らばり方を示している.
- データの大きさが$n = 15$ぐらいであれば，相関係数は$r_{xy} = 0.7$程度でも実際にはかなり弱い相関であり$r_{xy} = 0.5$ならば事実上関係があるようには見受けられないことに注意してほしい.

3.2　相関関係と因果関係

- 相関係数が高いことは，一般的には強い相関関係があるということであるが，このことは必ずしもその二つのデータの間に因果関係があるということではない.**相関関係**と**因果関係**は異なる.

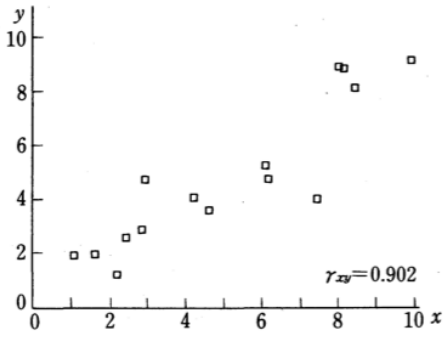

図 3.7　相関係数 $r_{xy}=0.902$ の散布図

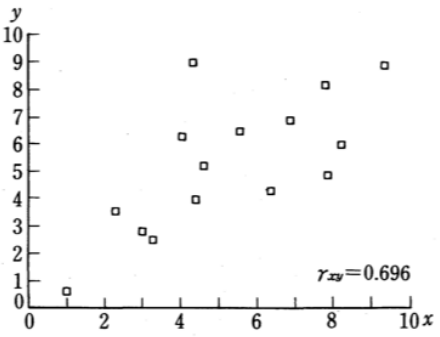

図 3.8　相関係数 $r_{xy}=0.696$ の散布図

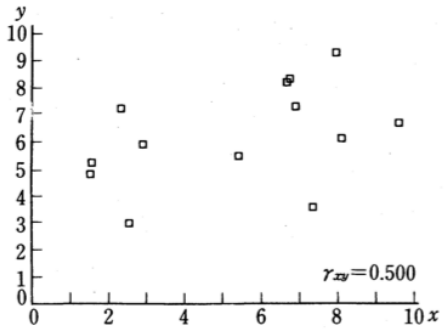

図 3.9　相関係数 $r_{xy}=0.500$ の散布図

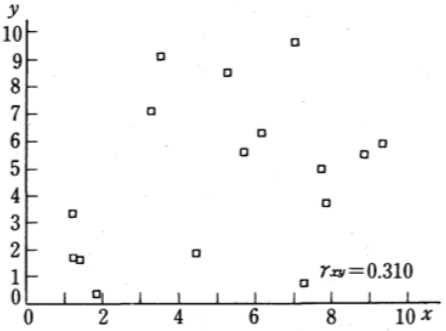

図 3.10　相関係数 $r_{xy}=0.310$ の散布図

- たとえば身長と体重の間には相関関係があるが，どちらがどちらを決めるともいえないので因果関係とはいえない.
- これに対して，人口と商店数の例などは，人口が商店数を決めていると考えられるので相関関係があると同時に因果関係がある.
- 相関関係とは二つのデータ間の直線的な関係のことであるが，因果関係には直線というような単純な関係でなく，はるかに複雑な関係も含まれる.因果関係であっても相関関係にはなく，相関係数の値も低くなるものもある.例えば：$y = (x-8)^2$,データの大きさは15とし、相関係数は0となる。

3.3　みかけ上の相関と偏相関係数

図3.12は東京都23区について，xに飲食店の数をとり，yに金融機関の店舗数をとって，各区をプロットした散布図である.図からは，飲食店の多いところには金融機関も多いということになる.相関係数も $r_{xy}=0.892$と非常に高い.きわめて強い正の相関関係が認められる.しかし，これは常識的に考えると，両者の間には直接的な関係はない。

実はこの二つの変数は，人口，とくに居住者の人口である昼間人口という第3の変数を間にはさんで，強い正の相関関係が観察されるのである.この3者の関係を図示すれば図3.15のようになる.このとき，昼間人口をはさんで，飲食店数と金融機関店舗数の間には相関関係が生じるが，このような相関関係はみかけ上の相関と呼ばれる.みかけ上の相関は容易に人の判断を誤らせることがある.

このような場合は**偏相関係数**を用いた方がよい.偏相関係数とは，変数1から変数3まで三つの変数があるとき，変数3の影響を除いたあとの変数1と変数2の間の相関係数のことで，一般に$r_{12\cdot3}$と書き，

$$r_{12\cdot3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

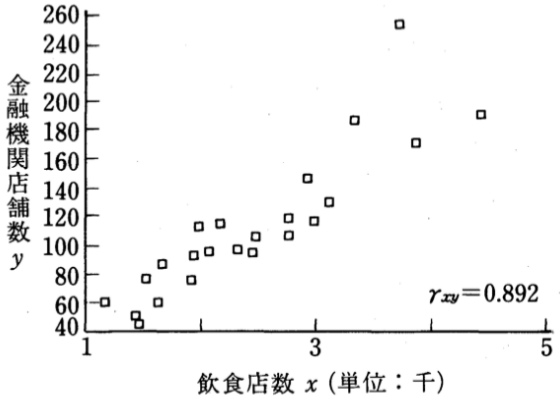

図 3. 12　飲食店数と金融機関店舗数

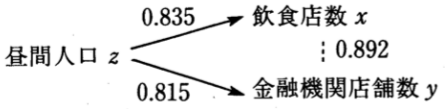

図 3. 15　見かけ上の相関
共通原因をもつ 2 変数は，強く相関することがある．

と定義する．

飲食店数を変数1,金融機関店舗数を変数2,昼間人口を変数3とすると，$r_{12} = 0.892$,$r_{13} = 0.835$,$r_{23} = 0.815$より，$r_{12\cdot 3} = 0.665$となり，昼間人口の影響を除いたあとの，飲食店数と金融機関店舗数の関係はさほど強くないことがわかる．

3.4　順位相関係数

順位相関係数とは，二つの質的基準(量的変数を大小関係でこれに変換してもよい)がある場合に，観測対象iの，二つの基準による順位rankR_i, R_i'の間の相関を示す指標である.スピアマンの定義によるものと，ケンドールの定義によるものがしばしば用いられる．

スピアマンの順位相関係数r_sは

$$r_s = 1 - \frac{6}{n^3-n}\Sigma_{i=1}^{n}(R_i - R_i')^2$$

と定義する．

ケンドールの順位相関係数r_kは，観測対象の対$(i,j)(i,j = l,2,...,n)$を考え，下記の式で．

$$r_k = \frac{\Sigma_{i=2}^{n}\Sigma_{j=1}^{i-1}([(R_i-R_j)(R_i'-R_j')>0]-[(R_i-R_j)(R_i'-R_j')<0])}{{}_nC_2}$$

と定義する．ただし，[.]は命題が真の場合1を，偽の場合0を与える指示関数のこと．意味は，例えば，2つの観測値$(R_i, R_j), (R_i', R_j')$を取り出したとき，もし正順：$R_i > R_j$且つ$R_i' > R_j'$あるいは$R_i < R_j$且つ$R_i' < R_j'$の場合，$[(R_i - R_j)(R_i' - R_j') > 0]$が+1の値を与える．正順の対の数と逆順の対の数の差を対の全数${}_nC_2$の中の割合で定義されること．

例：NHK放送世論調査所が1978年に行った全国県民意識調査によると，好きな花の順番は下記の表でまとめた．

男:　桜　菊　バラ　梅　ゆり　チューリップ　カーネーション　椿
女:　菊　バラ　桜　ゆり　梅　カーネーション　チューリップ　椿

順位は表3. 9の通りとなる.男女間での順位相関係数は，スピアマンの順位相関係数は$r_s = 0.81$,ケンドールの順位相関係数は$r_k = 0.714$となる.男女の嗜好は比較的似ていることがわかる.

表3.9: 好きな花の順番:順位相関の例

	桜	菊	バラ	梅	ゆり	チューリップ	カーネーション	椿
男:	1	2	3	4	5	6	7	8
女:	3	1	2	5	4	7	6	8

3.5　時系列と自己相関

補足 3.1 時系列と自己相関

データ$x_1, x_2, ..., x_n$が時間的に観測されたものであるとき，これらを一般に時系列という.

時系列では$1, 2, ..., n$の番号は時間を表現するものであって，$i < j$ならばjはiより時間的に後である.

同じxでも系列の異時点間の相関関係を表すのが，自己相関係数, 系列相関係数である.たとえば，yとしてデータを一時点だけずらして$(x_1, x_2), (x_2, x_3), ..., (x_{n-1}, x_n)$から相関係数を作ってみよう.この場合，$x$と$y$の相関係数を

$$r_1 = \frac{\Sigma_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})/(n-1)}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2/n}$$

で計算する．これを遅れlag1の自己相関係数という.

一般に，遅れhの自己相関係数は

$$r_h = \frac{\Sigma_{i=1}^{n-h}(x_i - \bar{x})(x_{i+h} - \bar{x})/(n-h)}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2/n}$$

自己相関係数は図3.21(コレログラムといわれる)のように表すことができる．遅れ$h = 6$でr_hのピークが見られ，周期性がたしかめられる.

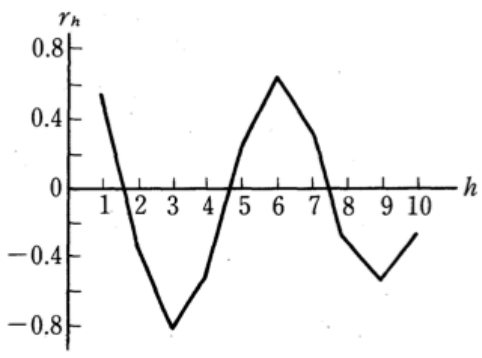

図 3. 21　自己相関係数のコレログラム

4　直線のあてはめ

4.1　最小二乗法

2変数x,yを考えよう.たとえば，表3.2,図3.2のデータ例(2)でみたようにx=年齢，y=血圧としよう.xはyをある程度決定するが，このように2変数xとyの間に，一方xが他方yを左右ないしは決定する関係があるとき，xを**独立変数**(あるいは，**説明変数**),yを**従属変数**という(あるいは，**被説明変数**).統計学では，このxとyの関係を**回帰**という見方や方法で扱う.

ここでは，2次元データをどのように要約整理して，そこからデータの意味するところを知るか，その大まかな方法を考える.

データ例(2)を見れば，おおよそ1次式

$$y = bx + a$$

係数の値b, aが定まれば，この直線が定まり，xからyが決定される様子やしくみが明らかになる.

b, aを決める方法は最小二乗法によるあてはめを考えよう.すなわち，x_iから予想されるyの値$bx_i + a$と現実の値y_iが，最も小さいへだたりをもつのが，最適な直線$y = bx + a$の引き方である.したがって，二乗和

$$L = \Sigma_{i=1}^{n}\{y_i - (bx_i + a)\}^2$$

を最小にするa, bの値を求める.

Lは，a, bの二変数関数の2次式だから，最小を求めるためにa, bでそれぞれ偏微分して0とおくと，結果として

$$\begin{cases} na + (\Sigma x_i)b = \Sigma y_i \\ (\Sigma x_i)a + (\Sigma x_i^2)b = \Sigma x_i y_i \end{cases} \tag{1}$$

となる.これを**正規方程式**ということがある.これをa, bの二元連立一次方程式として解くと，

$$\begin{cases} b = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}^2} \\ a = \bar{y} - b\bar{x} \end{cases} \tag{2}$$

のように，a, bの値が得られる．得られたa, bによる1次式をyのx上への回帰方程式，あるいは回帰直線，bはその傾きで**偏回帰係数**と呼ばれる．aは**回帰直線**のy**切片**である．このようにして，もっとも良くあてはまる直線が得られる.

表3.2の例(2)の年齢と血圧のデータに適用してみると，最も良くあてはまる直線(回帰直線)は$y = 1.38x + 65.1$となる.この回帰直線を用いて，各x_iの値に対するyのあてはめ値(対応するyの値)を求めることができる.これを$\hat{y_i}$と書こう.y_iと$\hat{y_i}$の差(外れ)は$d_i = y_i - \hat{y_i}$が全体として小さいほどあてはまりはよい.

このようにして，散布図(図3.3,図3.4)に対しても，回帰直線をあてはめてみよう(図3.22,図3.23).とくに，傾きbに意味がある．血圧データでは，年齢が1歳進むと血圧は1.38mmHg上昇する．また，人口(x)と商店数(y)の例では，人口が千人増加すれば商店は12店増加し，不快日数(x)とエアコン保有率(y)の例では，不快日数1日の増加はエアコン保有率を2.7%増加させる.

4.2　決定係数

決定係数とは，独立変数(説明変数)xが従属変数(被説明変数)yを決定する強弱の度合を表すものという.r^2で表す．($100r^2$として，%でいうこともある).

$$r^2 = \frac{\Sigma(\hat{y_i} - \bar{y_i})}{\Sigma(y_i - \bar{y})} = 1 - \frac{\Sigma(y_i - \hat{y_i})}{\Sigma(y_i - \bar{y})}$$

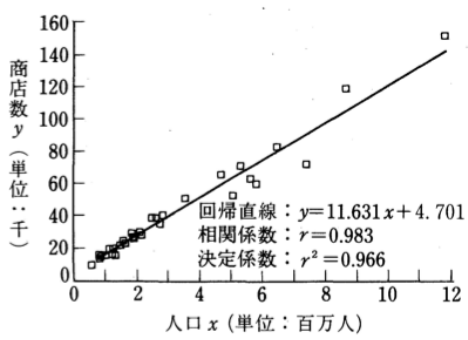

図 3.22 回帰分析の結果(1)
図 3.3 の散布図に回帰直線を引いたもの．決定係数 r^2 が高く，回帰直線は信頼できる．

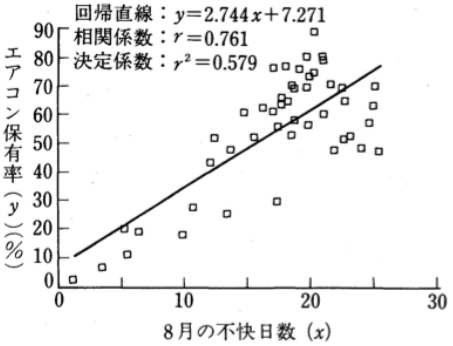

図 3.23 回帰分析の結果(2)
図 3.4 の散布図に回帰直線を引いたもの．決定係数 $r^2=0.579$ は良い値とはいえない．

例として，人口と商店数の場合(図3.22)は，$r=0.983$から決定係数は$r^2=0.96(96.6\%)$で，不快日数とエアコン保有率の場合(図3.23)は$r=0.761$から，決定係数は$r^2=0.579(57.9\%)$となる.後者の方が決定の度合が低いことがわかる.図からもそれは理解されるであろう.**r^2が大きいほど回帰の効果も大きいこと**となる.決定係数を考える仕組みは図3.24である.

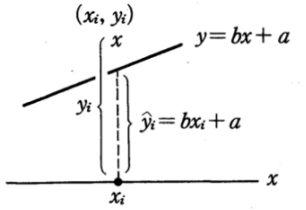

図 3.24 決定係数を考えるしくみ

補足 4.1 回帰と相関の考え方には，つながりがある.あてはめられた回帰直線の傾き，つまり回帰係数bは別の形で表されると，相関係数rと比較しよう.

$$\begin{cases} b=\dfrac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} \\ r=\dfrac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}} \end{cases} \tag{3}$$

比べると，bとrの間には，

$$b=r\frac{S_y}{S_x}$$

という関係が成立することがわかる.すなわち，相関係数r_{xy}の表すxとyの関係の密接度は，実はxとyの間の直線関係のあてはまりの良さという意味をもつ.

今日のキーワード

散布図，分割表，共分散，(積率)相関係数，順位相関係数(スピアマン，ケンドール)，因果関係，見かけ上の相関，最小二乗法，決定係数