

データサイエンスのための統計学

2次元のデータ

by 尚晋 (名古屋大学経済学研究科助教)

on 2025 年 4 月 30 日

目次

- * 2次元のデータ
- * 散布図と分割表
- * 相関係数
- * 直線のあてはめ

目次

- * 2次元のデータ
- * 散布図と分割表
- * 相関係数
- * 直線のあてはめ

» 2次元のデータに関して

- * 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。 p 個の変数を取り扱う場合、 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、 $p = 2$ とし、2変数 x, y の関係を考えてみる。 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。

相関に関して注意ポイント

- * 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。

目次

- * 2次元のデータ
- * 散布図と分割表
- * 相関係数
- * 直線のあてはめ

» 散布図

散布図とは x と y 両方が量的データである場合、この2次元データを平面にプロットした図は散布図という。

- * 2次元のデータを n 個の点, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ として図示すると、 x と y の関係ははっきりする。例：図 3.1 と図 3.2
- * 散布図上で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分かる。逆に各点が _____ x と y は関係がない。

» 散布図

散布図とは x と y 両方が量的データである場合、この2次元データを平面にプロットした図は散布図という。

- * 2次元のデータを n 個の点, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ として図示すると、 x と y の関係ははっきりする。例：図 3.1 と図 3.2
- * 散布図上で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分かる。逆に各点がバラバラに散らばれば x と y は関係がない。

》 散布図の例

関係あるか？

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

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)

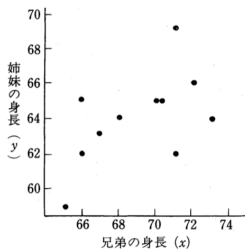


図 3.1 例(1)をグラフにした図 ($n=11$)

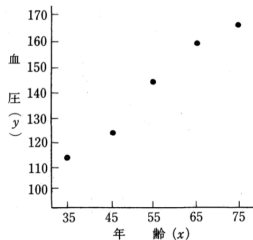


図 3.2 例(2)をグラフにした図 ($n=5$)

図 3.2 では、 y と x の間に明らかな関係があるが、図 3.1 にはそれはない。

» 散布図の例

正(負)の相関関係

- * 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「_____がある」ということが多い.
- * 「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.

» 散布図の例

正(負)の相関関係

- * 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.
- * 「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.

》 散布図の例

正 (負) の相関関係

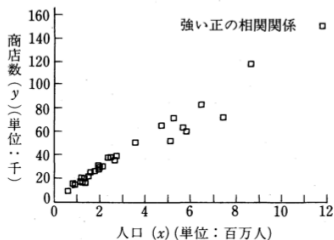


図 3.3 人口と小売商店数の散布図

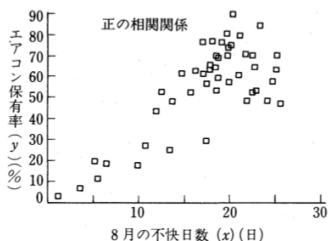


図 3.4 8月の平均不快日数^{*}とルーム・エアコンの保有率の散布図

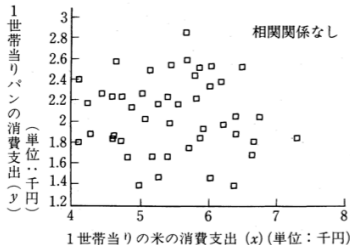


図 3.5 世帯・月あたり米の消費支出とパンの消費支出の散布図

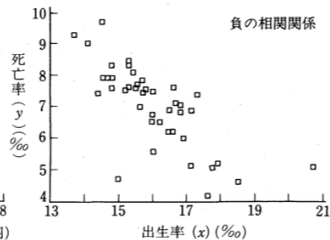


図 3.6 出生率と死亡率の散布図

» 分割表

分割表とは 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。例:表 3.4 の 2x2 分割表, 表 3.5 の 5x2 分割表
- * 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), 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。

» 分割表

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:%)

目次

- * 2次元のデータ
- * 散布図と分割表
- * 相関係数
- * 直線のあてはめ

» 積率相関係数

定義

積率相関係数とは 相関の程度を示す指標のことであり、よく用いられる定義は、**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 _____ である場合に用いられる。データが $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 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。つねに $-1 \leq r_{xy} \leq 1$ の範囲にある。

- * 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 _____ と呼ぶ。

» 積率相関係数

定義

積率相関係数とは 相関の程度を示す指標のことであり、よく用いられる定義は、**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。データが $(x_i, y_i), (x_2, y_2), \dots, (x_n, y_n)$ 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。つねに $-1 \leq r_{xy} \leq 1$ の範囲にある。

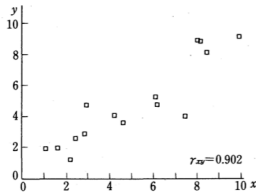
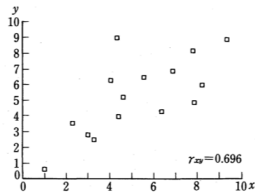
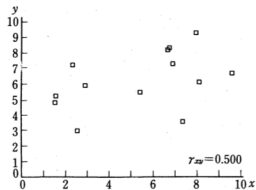
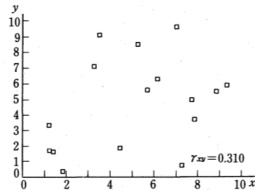
- * 上式の分母で、 $\sqrt{\Sigma(x_i - \bar{x})^2/n}$ 、 $\sqrt{\Sigma(y_i - \bar{y})^2/n}$ は、変数 x, y のそれぞれの標準偏差 S_x, S_y である。
- * 分子は $C_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y})/n$ は x の偏差 $x_i - \bar{x}$ と y の偏差 $y_i - \bar{y}$ を同時に考えたとき (偏差の積は偏差積と呼ぶ) の全データについての平均で、これを**共分散**と呼ぶ。

» 積率相関係数

例

$r_{xy} = 1$ の場合、正の完全相関という。逆に $r_{xy} = -1$ の場合、負の完全相関という。

* 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。

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

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

» 相関関係と因果関係

重要!

- * 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_____は異なる。
- * 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、人口と商店数の例では、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_____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だけでなく、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、或いは相関係数の値が低いものもある。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。

» 相関関係と因果関係

重要!

- * 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。
- * 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、人口と商店数の例では、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だけでなく、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、或いは相関係数の値が低いものもある。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。

» みかけ上の相関と偏相関係数

導入問題

- * 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_____。

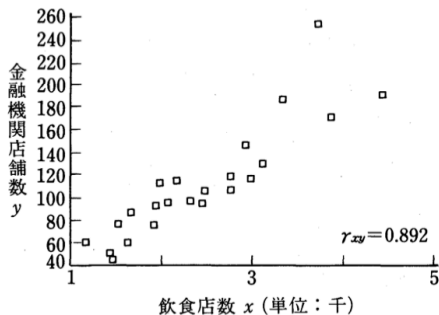


図 3.12 飲食店数と金融機関店舗数

» みかけ上の相関と偏相関係数

導入問題

- * 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。

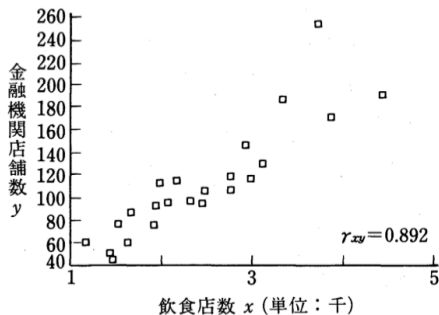


図 3.12 飲食店数と金融機関店舗数

» みかけ上の相関と偏相関係数

第3の変数

* 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者の関係は図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

と呼ばれる。容易に人の判断を誤らせることがある。

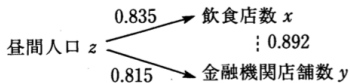


図3.15 見かけ上の相関

共通原因をもつ2変数は、強く相関することがある。

» みかけ上の相関と偏相関係数

第 3 の変数

* 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 者の関係は図 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。容易に人の判断を誤らせることがある。

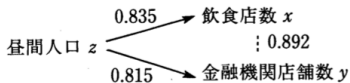


図 3.15 見かけ上の相関

共通原因をもつ 2 変数は、強く相関することがある。

» みかけ上の相関と偏相関係数

偏相関係数

- * 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 $r_{12.3}$ と書き、

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

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、 $r_{12.3} = \underline{\hspace{2cm}}$ 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。

» みかけ上の相関と偏相関係数

偏相関係数

- * 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 $r_{12.3}$ と書き、

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

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, $r_{12.3} = 0.665$ 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。

» 順位相関係数

定義

順位相関係数とは 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 $\text{rank} R_i, R'_i$ 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よく使われる。

スピアマンの順位相関係数とは スピアマンの順位相関係数 r_s は

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

と定義する。順位は全く同じなら、順位相関係数は 1。

ケンドールの順位相関係数とは ケンドールの順位相関係数 r_k は、観測対象の対 $(i, j) (i, j = 1, 2, \dots, n)$ を考え、下記の式で。

$$r_k = \frac{\sum_{i=2}^n \sum_{j=1}^{i-1} ([(R_i - R_j)(R'_i - R'_j) > 0] - [(R_i - R_j)(R'_i - R'_j) < 0])}{n C_2}$$

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 $n C_2$ の中の割合で定義されること。

» 順位相関係数

例

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

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

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 $r_s = 0.88$, ケンドールの順位相関係数は $r_k = 0.71$ となる. 男女の嗜好は比較的似ていることがわかる.

» ケンドールの順位相関係数の計算

$i = 2$

$$\begin{aligned}
 & \Sigma_{j=1}^{2-1} ([(R_i - R_j)(R'_i - R'_j) > 0] - [(R_i - R_j)(R'_i - R'_j) < 0]) \\
 &= [(R_2 - R_1)(R'_2 - R'_1) > 0] - [(R_2 - R_1)(R'_2 - R'_1) < 0] \\
 &= [(2 - 1)(1 - 3) > 0] - [(2 - 1)(1 - 3) < 0] \\
 &= [-2 > 0] - [-2 < 0] \\
 &= 0 - 1 = -1
 \end{aligned}$$

$i = 3$

$$\begin{aligned}
 & \Sigma_{j=1}^{3-1} ([(R_i - R_j)(R'_i - R'_j) > 0] - [(R_i - R_j)(R'_i - R'_j) < 0]) \\
 &= ([(R_3 - R_1)(R'_3 - R'_1) > 0] - [(R_3 - R_1)(R'_3 - R'_1) < 0]) \\
 & \quad + ([(R_3 - R_2)(R'_3 - R'_2) > 0] - [(R_3 - R_2)(R'_3 - R'_2) < 0]) \\
 &= ([(3 - 1)(2 - 3) > 0] - [(3 - 1)(2 - 3) < 0]) \\
 & \quad + ([(3 - 2)(2 - 1) > 0] - [(3 - 2)(2 - 1) < 0]) \\
 &= ([-2 > 0] - [-2 < 0]) + ([1 > 0] - [1 < 0]) \\
 &= -1 + 1 = 0
 \end{aligned}$$

目次

- * 2次元のデータ
- * 散布図と分割表
- * 相関係数
- * 直線のあてはめ

» 最小二乗法

回帰とは

- * 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, またその結果はどう解釈するか, その大まかな方法を考える.
- * 表 3.2 のデータ例 (2) と図 3.2 を見れば, おおよそ 1 次式となる:

$$y = bx + a$$

ここで, 係数の値 b, a が定まれば, この直線が定まり.

» 最小二乗法

係数の決め方

- * 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**

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

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**

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

となる. これを**正規方程式**ということがある.

» 最小二乗法

回帰直線

- * これを a, b の二元連立一次方程式として解くと,

$$\begin{cases} b = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \\ a = \bar{y} - b \bar{x} \end{cases} \quad (2)$$

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.

» 回帰分析の例

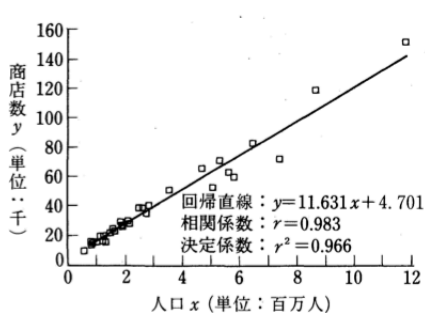


図 3.22 回帰分析の結果(1)

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。

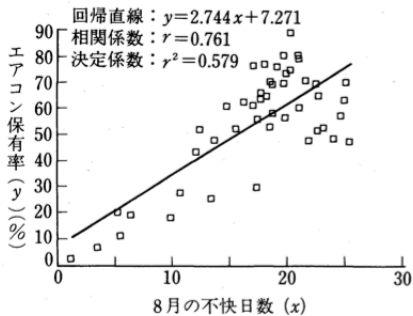


図 3.23 回帰分析の結果(2)

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。

» 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。(100 r^2 として, % ということもある).

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

- * 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.