



臨床検査データを合成・縮約するための主成分分析

第2回 回帰と相関

山西 八郎

Synthesizing and Reducing the Coefficient of Variation by Principal Component Analysis

Part 2: Regression and Correlation analysis

Hachiro Yamanishi

Summary Regression analysis aims to define the relationship between two variables with a regression equation. There are four different methods of regression analysis, which need to be used differently depending on the measurement error assumptions. Regression analysis by the method of least squares is generic, but only the measurement error of the variable placed on the Y-axis is reflected in the regression equation. In this point, Deming regression analysis yields a regression line that best reflects measurement error, but standardized major axis regression analysis is the most practical and reasonable method of analysis. Correlation analysis, on the other hand, aims to evaluate the degree of linearity of the bivariate with a correlation coefficient. However, even if the correlation coefficient is statistically significant, it is necessary to examine whether the correlation is substantively meaningful.

I. はじめに

回帰分析と相関分析は、測定法評価や観測データ間の関係を解析するうえにおいて極めて重要な統計的手法である。しかし、回帰分析における回帰式の決定方法の選択や、相関係数の意味について誤解されている場合が少なくない。また、回帰分析に属する主成分回帰は、本シリーズの主テーマである主成分分析の根本となる分析法である。そこで今回は、これらの分析法の数理に軸足を置いて解説する。なお、以下の解説では文献^{1) 2) 3) 4)}を参考とした。

II. 回帰分析

回帰分析の主目的は、目的変数: y と説明変数: x を、回帰式: $y=ax+b$ として関係付けることにある。ここで、説明変数に係る係数 a は、説明変数が1単位変化したときの目的変数の変化量を意味しており、 b は説明変数が0のときの目的変数の値: 初期値を表している。回帰式の決定方法には以下に示す4つの方法がある。

1. 最小二乗法による方法

これは、各プロットから回帰直線までの y 軸

天理大学 医療学部
〒632-0018 奈良県天理市別所町80-1 天理大学別所キャンパス

Tenri University Department of Clinical Laboratory Science,
80-1 Bessho-cho, Tenri, Nara 632-0018, Japan

連絡先: 山西 八郎
Tel: +81-743-63-7811
E-mail: yamaha@sta.tenri-u.ac.jp

方向の距離 d の2乗和が最小となるように回帰式の傾き a を決定する方法である (Fig. 1A)。この場合、距離 d は y 軸に配置した測定値の誤差を意味しており、この誤差が最小となるように回帰直線が決定される。言い換えると、 x に置いた測定値の誤差は仮定していないので、正確には「 x を基準とした最小二乗法」と表現できる。この方法は、標準物質や標準的測定法での値を x 軸に、日常測定法での測定値を y 軸に配置した回帰分析や、検量線の作成には適用できるが、測定誤差を無視できない (日常) 測定法間の回帰分析に用いることは妥当ではない。

一方、Fig. 1Bは、各プロットから回帰直線までの x 軸方向の距離 d の2乗和が最小となるように回帰式を決定する「 y を基準とした最小二乗法」であるが、通常回帰分析に単独で適用されることはない。

Table 1に「 x を基準とした最小二乗法」によ

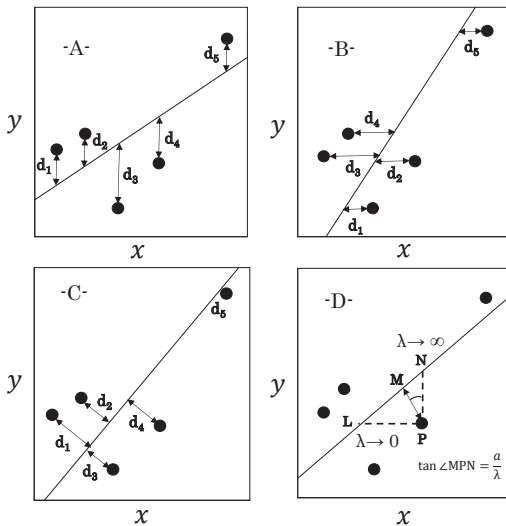


Fig. 1 Method of determining the regression line

る回帰直線の決定プロセスを示す。まず、 x の測定値から平均を引いた値を x の偏差とし、その2乗和を S_{xx} とする。同様に y の偏差を求め、 x と y の偏差の積和を S_{xy} とすると、回帰式の傾きは次式から算出される。

$$\text{傾き} : a = \frac{S_{xy}}{S_{xx}} = \frac{82}{88} = 0.932$$

回帰式の初期値は、後述するいずれの回帰直線も x と y の平均 (重心) を通過することより、次式により算出される。

$$\text{初期値} : b = \bar{y} - a \times \bar{x} = 9.4 - 0.932 \times 6 = 3.808$$

したがって、Table 1 の回帰式は、 $y = 0.932x + 3.808$ となる。

2. 標準主軸回帰

y だけでなく x 軸に配置した測定値にも同程度の誤差があると仮定して適用される回帰方法である。その傾きは、「 x を基準とした最小二乗法」による傾き a_1 と、「 y を基準とした最小二乗法」とした傾き a_2 の幾何平均として算出される。

$$a = \sqrt{a_1 \times a_2}$$

Table 1における y を基準とした回帰直線の傾きを2.135とすると、標準主軸回帰式の傾きは、

$$a = \sqrt{0.932 \times 2.135} = 1.411$$

初期値は、

$$b = \bar{y} - a \times \bar{x} = 9.4 - 1.411 \times 6 = 0.934$$

と決定される。

標準主軸回帰は、(日常) 測定法評価に対し最も実用的かつ妥当な分析法である。また、最小二乗法による回帰分析はエクセルのワークシート関数 (SLOPE関数) で対応できるが、標準主軸回帰もエクセルで実行可能である。Fig. 2にその方法を示す。

Table 1 The process of determining a regression line by the method of least squares.

				x の偏差	x の偏差の二乗	y の偏差	x と y の偏差の積
	x	y		-5	25	-1.4	7
	1	8		-3	9	0.6	-1.8
	3	10		-1	1	-7.4	7.4
	5	2		2	4	-2.4	-4.8
	8	7		7	49	10.6	74.2
	13	20					
平均	6	9.4	総和	0	$S_{xx}=88$	0	$S_{xy}=82$

	A	B	C	D	E	F
1		Y	X			①
2		120	180	Xを基準		1.211946
3		100	119			
4		70	130			
5		80	170			②
6		121	148	Yを基準		1.474121
7		150	150			
8		101	151			
9		60	140			
10		300	300			③
11		81	131	傾き		1.33662
12	Mean	118.3	161.9	切片		-98.099
13						④

$$\textcircled{1} = \text{SLOPE}(\text{B2:B11}, \text{C2:C11})$$

Yの値 Xの値

$$\textcircled{2} = 1 / \text{SLOPE}(\text{C2:C11}, \text{B2:B11})$$

逆数 Xの値 Yの値

$$\textcircled{3} = (\text{F2} * \text{F6}) ^ 0.5$$

①と②の幾何平均

$$\textcircled{4} = \text{B12} - \text{F11} * \text{C12}$$

Yの平均 - ③ × Xの平均

Fig. 2 Running a standard principal axis regression in Excel

まず、 x と y の測定値をシートに入力し、① x を基準とした回帰式の傾きを次式で求める。

$$\textcircled{1} = \text{SLOPE}(\text{B2:B11}, \text{C2:C11})$$

次に、② y を基準とした回帰式の傾きを次式で求める。

$$\textcircled{2} = 1/\text{SLOPE}(\text{C2:C11}, \text{B2:B11})$$

①と②の幾何平均を③により、初期値を④により算出すると、回帰式は、

$$y = 1.34x - 98.1$$

と決定される。

3. 主成分回帰 (Fig. 1 C)

この回帰法は、各プロットから回帰直線に下した垂線の長さの2乗和が最小となるように傾きが決定される。前述のごとく、主成分回帰が変数を合成・縮約するための主成分分析に応用されている。また、近年では新しい精度管理法としても利用されている⁹⁾。しかし、主成分回帰ではバラツキの大きな変数の軸方向に回帰直線が傾く傾向にあり、また x と y の単位が異なると意味のない回帰式が得られる場合があるので、測定法評価のために適用されることはない。

4. Deming回帰

x と y の測定誤差を最も忠実に反映した回帰式が得られる。Deming回帰では、 x と y の測定試料を二重測定し分散を求め、それぞれの分散の平均を誤差分散： e_x 、 e_y として誤差分散比： λ を算出する。

$$\lambda = \frac{e_y}{e_x}$$

なお、二重測定の値を k_1 、 k_2 とすると、その分

散は次式により算出される。

$$\frac{(k_1 - k_2)^2}{2}$$

そして、 y の偏差平方和を S_{yy} とすると、Deming回帰式の傾き a は次式より算出される。

$$a = \frac{S_{yy} - \lambda S_{xx} + \sqrt{(S_{xy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}}$$

λ の視覚的な意味はFig. 1Dに示すように、

$$\tan \angle \text{MPN} = \frac{a}{\lambda}$$

の関係が成立しており、 λ によって決まる角度 ($\angle \text{MPN}$) での各プロットと回帰直線までの距離PMの2乗和が最小となるように傾きを決定することにある。ここで、 λ を $+\infty$ に設定すると、Deming回帰式は x を基準とした最小二乗法に、 λ を0に近づけると y を基準とした最小二乗法による回帰式に近似する。また、 $\lambda=1$ のとき、Deming回帰式は主成分回帰式に近似する。

一方、 λ が不明の場合は、 x と y の誤差分散をそれぞれの偏差平方和で置き換えると、Deming回帰式は標準主軸回帰式に近似する。Table 1のデータにおいて、 $\lambda=2$ とした場合のDeming回帰式を以下に示す。ちなみに、Deming回帰の実行には専用の統計ソフトが必要となる。

$$y = 1.41x + 0.944$$

5. 回帰係数 (傾き) の有意性

回帰式の傾き a の標準誤差： s_b と、回帰直線からの標準偏差： s は次式で定義され、

$$s_b = \frac{s}{\sqrt{S_{xx}}}$$

$$s = \sqrt{\frac{S_{yy} - a \times S_{xy}}{n - 2}}$$

次式で得られる統計量 t が自由度 $n-2$ の t 分布にしたがうことを利用してその有意性を検定する。

$$t = \frac{a}{s_b}$$

II-1で求めた x を基準とした最小二乗法による傾きの s_b は0.611であることより、 t 値は1.523となり、自由度3の t 分布における5%有意点3.182以下であることから、有意な回帰係数ではないと判断できる。

Ⅲ. 相関分析

x と y の直線関係を相関係数： r で提示することを主目的とする。ただし、相関係数の絶対値が0に近似しているので、2変量間には関係がないと誤解される場合がある。散布図上でプロットが円を描くとき、相関係数は0に近づくが、 x と y の間には a, b, r を定数とした場合、 $(x-a)^2 + (y-b)^2 = r^2$ に近似した関係が成立しているからである。

1. 相関係数算出のプロセス

相関係数は x と y の偏差（中心化データともよばれる）を、それぞれのSDで割った値（標準化データ）同士の積の平均として算出される（Fig. 3）。ただし、平均を求める際の分母は自由度の関係から「標本数-1」とする。また、 x と y の偏差の積の平均が共分散に相当する。

Fig. 4Aは測定データの、またFig. 4Bは標準化データに変換した値での散布図であるが、変換後もプロットの位置関係に変化はないことが確認できる。ここで、 x と y の標準化データ同志の積は、そのプロットと原点を対角とする四角形の面積に相当する。つまり、Fig. 3における x と y の標準化データ同志の積の平均は、Fig. 4Bにおける5つの四角形の面積の平均を意味している。したがって、第1象限と第3象限にあるプロットからつくられる四角形の面積は（+）の符号となるので、その面積から算出される相関係数は（+）の符号、すなわち正の相関関係を表すことになる。また、第2、4象限での x と y の標準化データは逆の符号となるため、相関係数は（-）の符号、すなわち負の相関性を意味することになる。一方、プロットが第1~4象限に散在していると、その平均を求める際に正と負の符号をもつ面積が互いにキャンセルし合うために、相関係数は低く算出されることになる。

相関係数算出のプロセスを式で展開すると、まず x と y の標準化データの積は次式で表される。

測定データから平均を引く			中心化データをSDで割る		
測定データ			中心化データ		
x	y		x	y	$x \times y$
1	8		-5	-1.4	7
3	10		-3	0.6	-1.8
5	2		-1	-7.4	7.4
8	7		2	-2.4	-4.8
13	20		7	10.6	74.2
平均	6.0	9.4	平均	0	0
SD	4.7	6.6	SD	4.7	6.6
			標準化データ		
			x	y	$x \times y$
			-1.06	-0.21	0.22
			-0.64	0.09	0.06
			-0.21	-1.12	0.24
			0.43	-0.36	-0.15
			1.48	1.60	2.37
			平均	0	0
			SD	1	1

積和 ÷ (n - 1) = 共分散 積和 ÷ (n - 1) = 相関係数

Fig. 3 Process of calculating correlation coefficients

$$\frac{(x_i - \bar{x})}{SD_x} \times \frac{(y_i - \bar{y})}{SD_y}$$

次に相関係数に相当する積の平均は、標本数を n とすると、

$$\left\{ \sum_{i=1}^n \frac{(x_i - \bar{x})}{SD_x} \times \frac{(y_i - \bar{y})}{SD_y} \right\} \times \frac{1}{n}$$

となるので、整理すると、

$$\frac{\sum_{i=1}^n (x_i - \bar{x}) \times (y_i - \bar{y})}{n} \times \frac{1}{SD_x \times SD_y}$$

上式の第1項は x と y の共分散: $COV(x, y)$ であるので、相関係数は、

$$\text{相関係数} = \frac{COV(x, y)}{SD_x \times SD_y}$$

で定義できる。また、測定データが標準化されていると、 x と y のSDは1となるので、相関係数は共分散に等しくなる。

2. $-1 \leq \text{相関係数} \leq 1$ であることの証明

次式はCauchy-Schwarzの式とよばれるもので、 n 個の x_1 から x_n のそれぞれの二乗値の総和と、 n 個の y_1 から y_n のそれぞれの2乗値の総和の積は、 n 個の x_i と y_i の積の総和の二乗以上になることを示している。

$$\left(\sum_{i=1}^n x_i^2 \right) \left(\sum_{i=1}^n y_i^2 \right) \geq \left(\sum_{i=1}^n x_i y_i \right)^2$$

そこで、 n 個の x と y の平均を $\bar{x}$ 、 $\bar{y}$ とし、不等式の x と y をそれぞれの偏差に置き換えて両辺を n^2 で割ると、左辺は x と y の分散の積、右辺は x と y の共分散: $COV(x, y)$ の2乗に書き換えることができる。

$$\begin{aligned} \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right) \left(\sum_{i=1}^n (y_i - \bar{y})^2 \right) &\geq \left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right)^2 \\ \frac{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)}{n} \times \frac{\left(\sum_{i=1}^n (y_i - \bar{y})^2 \right)}{n} &\geq \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} \right)^2 \end{aligned}$$

分散はSDの二乗値であるので

$$SD_x^2 \times SD_y^2 \geq COV(x, y)^2$$

より、両辺を SD_x^2 と SD_y^2 の積で割ると、右辺は x と y の相関係数の二乗に等しくなる。

$$1 \geq \left(\frac{COV(x, y)}{SD_x \times SD_y} \right)^2$$

$$r_{xy}^2 \leq 1$$

この不等式の解より、 $-1 \leq \text{相関係数} \leq 1$ が導かれる。

3. 相関係数の有意性⁴⁾

標本数を n としたとき、次式により算出される t 値が自由度 $n-2$ の t 分布にしたがうことを利用して検定される。

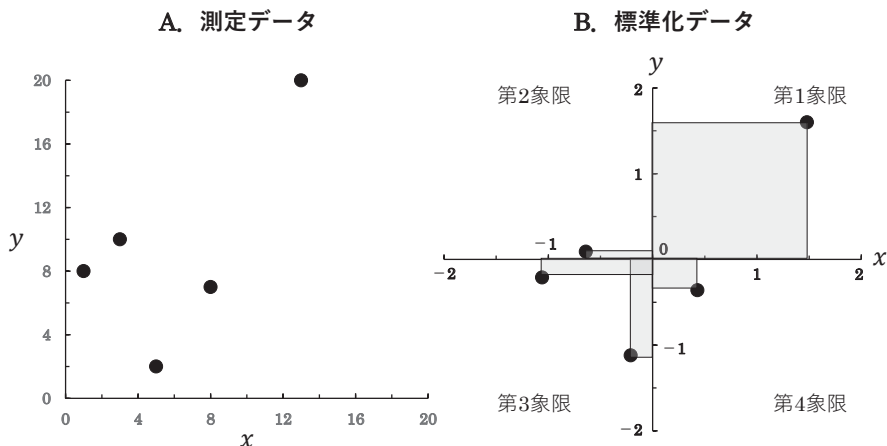


Fig. 4 Meaning of the product of standardized data

$$t = r \times \sqrt{\frac{n-2}{1-r^2}}$$

Fig. 3のデータより得られるt値は、

$$t = r \times \sqrt{\frac{n-2}{1-r^2}} = 0.672 \times \sqrt{\frac{3}{1-0.672^2}} = 1.573$$

自由度3のt分布における5%有意点は3.182で、 $3.182 > 1.573$ より相関係数は有意でないと判断できる。

ここで、相関係数の有意性はサンプル数に大きく依存する。サンプル数が50以上であれば、相関係数が0.3程度でも有意となる。しかし、その散布図からは2変量間に直線関係を見て取することはほとんどできない。統計学的に有意であっても、実質的に意味のある相関関係であるかを考察する必要がある。

4. 順位相関係数

x と y のそれぞれの測定値について、大きさの順番で順位を割り当て、その順位をデータとして計算されるノンパラメトリックな相関分析法である。極端に大きく飛び離れたデータがあるとき、そのデータにより大きな面積を有する四角形ができるため、相関係数が大きく見積もれる場合に有用となる。

Ⅳ. 終わりに

回帰分析と相関分析について解説した。特に、回帰分析における回帰式決定法の選択が重要である。これは日常測定法の比較において、 x を基準とした最小二乗法による評価報告が決して少ないからである。相関分析において、予想していたよりも相関性が低い場合、重心付近の測定データをいくつ追加しても相関係数には大きく反映されないことを付け加えておく。

本論文内容に関連する著者らの利益相反：なし

文献

- 1) 市原 清志：バイオサイエンスの統計学，206-240，南江堂，東京（1990）
- 2) P.G.ホーエル：初等統計学，182-206，培風館，東京（1979）
- 3) 山西八郎：正規分布と基本統計量，寺子屋統計教室，48-53，情報機構，東京（2022）
- 4) 山西八郎：臨床検査試薬性能評価に必要な統計手法 回帰と相関，検査と技術，51：114-121，2023.
- 5) 上野民生，宮子 弘，犀川哲典：PCA管理システムによる検査過誤検出，日本臨床検査自動化学会会誌，36：337-343，2011.