統計ライブラリー

回帰分析

佐和隆光 著

朝倉書店

回帰分析

佐和隆光

朝倉書店

著者略歴

昭和17年　和歌山県に生まれる
昭和40年　東京大学経済学部卒業
昭和44年　京都大学助教授(経済研究所)となり，
現在教授．　経済学博士

統計ライブラリー

回帰分析

昭和54年4月20日　初版第1刷
昭和58年10月25日　第4刷

著者　佐和隆光

発行者　朝倉邦造
東京都新宿区新小川町6-29

発行所
株式会社 朝倉書店
東京都新宿区新小川町6-29
郵便番号　162
電話 東京(260)0141(代)
振替口座 東京6-8673番
自然科学書協会会員

教弘印刷・渡辺製本

はしがき

数多くある統計手法のうち，もっとも実用に供されることの多いのが，回帰分析(regression analysis)である．その応用領域は，自然諸科学，工学から人文・社会諸科学に至るまで，はなはだ広範囲に及んでいる．また，その歴史も古く，19世紀初頭のガウスによる最小2乗法の発見にまで遡れる．実際，回帰分析の理論の大まかな骨子は，今から150年ほど前に，天文学や測地学への応用を旨として，ガウスによって形づくられた．その後，生物学，農学，経済学等へと次第に応用領域を広げるとともに，理論面での彫琢と精緻化がなされてきた．純理論的には，線形代数と十分統計量の一般理論にもとづいて，線形回帰分析の理論は，きれいに体系化されている．だがしかし，回帰分析の応用を目指す者にとっては，理論体系の美しさを賞玩してすますわけにはいかない．

統計学のこの分野における，過去十数年間の動向を振り返ってみると，大別して二つの流れが認められる．ひとつは，線形回帰の理論を抽象代数的に再構成して，より一層の精緻化を目指す立場である．いまひとつは，いわゆる「データ解析学派」の主導のもとに，推測統計的枠組みにしばられることなく，残差分析や非線形問題などを，プラグマティックに取り扱おうという立場である．様々な新しい手法が提案され，計算プログラムが開発され，広く実用に供されている．

本書では，こうした最近の研究動向をふまえた上で，理論面での自己完結性(self-sufficiency)をねらうとともに，データ解析学派の研究成果の積極的導入をはかった．つまり前半部において，線形代数と数理統計の一般理論をベースに，線形回帰分析の理論を，ほぼ完全な証明つきで展開する．そして後半部において，回帰モデルの妥当性を診断するための方法，モデル改善のためのてだて，等々のプラクティカルな手法について述べる．前半部と後半部が，木に竹を接いだかのようにならないよう，私なりに工夫をこらしたつもりである．も

ちろん，回帰分析に関連する，ありとあらゆる手法を，200ページ足らずの本書に網羅することは望むべくもない．紙幅の都合で割愛せざるをえなかったものも少なくない．どこまで成功しているかは別として，回帰分析の理論と応用にかんする中級程度のテキストブックに，「何を書き何を書くべきでないか」について，私なりに細心の気配りをしたつもりである．また，読者の理解をたすけるために，数値例や応用例をなるたけ多く盛りこむように努めた．

本書の構成については，第1章の終りに書いたので，ここでの再述はさけたい．

本書を執筆する過程で，多くの方々から，ひとかたならぬお世話をこうむった．原稿と校正刷りを精読され，いくたの的確なコメントと議論を提供してくださった片岡佑作(京都産業大学)，加納悟(横浜国立大学)の両氏には，この場をかりて，あつく御礼を申しあげたい．また，筆者が統計学を学び始めて以来の恩師である竹内啓先生には，本書の執筆を思いたった頃から，折にふれて数数の有益な助言を賜った．あつく謝意を表したい．最後に，すみずみまで心を配って本書の刊行にあたられた，朝倉書店の編集部の方々に，心からの感謝の気持ちを表する次第である．

1979年3月　京都紫野にて

佐　和　隆　光

北大助教授 齋藤 堯幸著

統計ライブラリー **多次元尺度構成法**

A 5 判 256頁 価3200円

心理学，社会学，政治学をはじめとした社会科学から，医学，生物学を主とした自然科学に至るまで広い範囲にわたってその応用範囲を拡げている多次元尺度構成法について応用例を中心にしてきわめて実際的にかつ懇切丁寧にまとめられている

東工大教授 工学博士 日野 幹雄著

統計ライブラリー **スペクトル解析**

A 5 判 312頁 価3900円

〔内容〕ランダム変動の表現・自己相関関数とスペクトル，相互相関とクロス・スペクトル，定常性，情報エントロピーとスペクトル，線型システムの理論，スペクトル計算の誤差理論，データ処理の手法，さらにすすんだスペクトルの概念。

後藤昌司・畠中駿逸・田崎武信訳

統計ライブラリー

コックス **二値データの解析**

—医学・生物学への応用—

A 5 判 236頁 価3400円

医学・薬学・生物学・農学系の研究者，技術者のための適用例が数多く収められ，さらに日本版にはプログラムも加えられ，実際に役立つようまとめられている。〔内容〕線形ロジスティックモデル，単一パラメータの解析，複雑な解析，他。

東工大教授 Ph.D 池田 央　立教大助教授 工学博士 岡太彬訓訳

統計ライブラリー

アプトン **調査分類データの解析法**

A 5 判 176頁 価2500円

統計調査に必然的に生ずる2元分類および多元分類表の統一的分析法を初学者にもわかるように平易に解説。クロス分類表の古典的連関分析から始まって近年急速に発展した対数線形モデルを詳解調査の実際家や統計調査に係る研究者に好適の書

脇本和昌・後藤昌司・松原義弘著

多変量グラフ解析法

A 5 判 208頁 価2600円

多次元データのもつ意味を総合的に伝達また解析する手法として開発された顔形グラフ，体形グラフ，星座グラフ，木形グラフ，パイプロット法，非線形変換グラフ，樹形図等の多変量グラフ解析法を平易に解説。巻末にプログラムを付した。

クラスカル・ウィッシュ著　高根芳雄訳

人間科学の統計学 1 **多次元尺度法**

A5変型 120頁 価1500円

統計学や数学についてあまり専門的な知識を必要としないで多次元尺度法について簡潔にわかりやすくかつ実際的にまとめられている。〔目次〕多次元尺度法の基礎概念，布置の解釈，次元数の決め方，個人差をとらえる多次元尺度法，他。

アッシャー著　広瀬弘忠訳

人間科学の統計学 2 **因果分析法**

A5変型 108頁 価1500円

因果分析に関心をもつ学生，研究者，実務家のためのテキストとしては現在のところ最良の書といわれている。簡潔でわかりやすく，かつ実際的にまとめられている。〔目次〕序論，Simon-Blalock法，逐次・非逐次パス解析，結語，他。

A．J．リヒトマン他著　長谷川政美訳

人間科学の統計学 3 **生態学的推論**

A5変型 96頁 価1500円

集団データ解析に関連した一般的問題を実例を中心にわかりやすく簡潔にまとめられた入門書〔目次〕集積偏倚と標準化されない係数—定式化の問題，集積偏倚と標準化された測度，集積偏倚の問題に対する解，結論—集積・計算および理論，他

定価は 1983 年 10 月現在のものです.

目　　次

1。回帰分析への誘い

人間社会や自然現象における，多少とも複雑な事象を統計解析しようとすれば，個々の変量を個別にとりあげて分析するだけでは不十分である．そのためには，複数個の変量(多変量)間の関係のあり方を総括的に分析する必要にせまられることが多い．この本のテーマである回帰分析(regression analysis)とは，多変量間の関係を解析するための，もっとも基本的な統計手法のひとつにほかならない．

多変量間の関係を解析するための統計手法一般のことを，通常，多変量解析法(multivariate analysis)という．近年，電子計算機の発達にたすけられて，多変量解析法の応用は，すこぶる盛んである．なかでも回帰分析が，もっとも広く実用されている．自然科学であれ人文・社会科学であれ，多変量のあいだの因果関係や相互依存関係を解析する必要に迫られることが，多いからであろう．また回帰分析は，“予測”という営みとも大いに関わっている．私たちは常日頃，意識するとしないにかかわらず，回帰分析のお世話になっている．病気の診断，天気予報，選挙予測，景気予報，等々の背後には，陰に陽に，回帰分析が控えているのである．

回帰分析の理論と応用について詳しく説くのが，言うまでもなく，この本のねらいである．回帰分析のすべてを習得するには，最後のページまで読みきって頂くほかないけれども，まずはじめにこの章で，回帰分析というものの鳥瞰的スケッチを試み，初学者のための誘いとしたい．すでに初歩的な回帰分析にお馴じみの読者は，この章をとばして，第2章から始められても一向にさしつかえない．

1.1 2変数回帰

1.1.1 相関と関係

まずはじめに，2個の変量間の関係について考えてみよう．たとえば，身長と体重，血圧と年齢，所得と知能指数，夫の年齢と妻の年齢，(体積一定の気体の)温度と圧力，施肥量と収穫高，入試の成績と入学後の成績，等々．これらの2変量間には，なんらかの“関係”が存在するであろうことを，誰もが知っている．しかし，“関係”の中味は多種多様である．たとえば(体積一定の)気体の温度と圧力は，測定誤差を別にすれば，「2変量の積が一定値に等しい」というエクザクトな関係で結ばれる．身長と体重のあいだに，そうした厳密な関係が存在するとは誰も思わない．しかし，「身長の高い人は総じて体重も重かろう」という“傾向”としての相関関係が存在することはまちがいない．一般に，2個以上の変量が「かなりの程度の規則正しさをもって，増減をともにする関係」のことを**相関**(correlation)関係という．所得とIQの間にも，やはり一方が高ければ他方も高かろう，という共変的傾向が認められるであろう．だからといって「IQの差が所得格差を決める」と結論するのは，いささか速断である．2変量が相関しているからといって，ただちに一方の変量が他方の変量を決定する，という一方向の因果関係を帰納するのは誤りである．もちろん，相関関係の存在が，因果関係の存在を確証するための有力な拠りどころとなることは確かである．

見せかけの相関(spurious correlation)ということも，しばしば起こりがちである．たとえば，「血圧と所得の間に正の相関がある」という命題は，おそらく統計的に真であろう．しかし，これらの2変量の間には，体重と身長のような(同じ個体のサイズを2種類の尺度で測るという)直接的相関関係は認められないし，いわんや一方が原因で他方が結果という因果的関係もありそうにない．日本のような年功序列賃金の国では，所得は年齢と正の相関をもち，また一般に，血圧と年齢との間にも正の相関が存在すると考えられる．その結果として，個人の所得と血圧の間に正の相関が認められることになるのであろう．

以上に述べたように，2個の変量が相関しているということは，両者の間に何らかの“関係”が存在していることを，たんに示唆するにすぎない．それが

因果関係なのか，あるいは第三の変量を介しての見せかけの相関なのかは，つまるところ，(所与のデータ以外の)先験情報にもとづいて判断すべき問題である．また，統計学でいうところの相関関係は，「線形な共変関係」という限定的な意味につかわれることが多い．したがって，「関係はあっても相関はない」というケースも，間々ありうることに注意しておこう．

1.1.2 相関係数

2変量 X と Y にかんする n 個の観測値 (x_i, y_i) $(i=1, 2, \cdots, n)$ が与えられたとしよう．2変量間の“関係”を記述するための，もっとも基本的な統計量は**相関係数**(correlation coefficient)

$$r=\frac{s_{xy}}{s_x s_y} \tag{1.1}$$

である．ただし

$$\begin{aligned} &s_{xy}=n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}), \\ &s_x^2=n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \qquad s_y^2=n^{-1}\sum_{i=1}^{n}(y_i-\bar{y})^2, \\ &\bar{x}=n^{-1}\sum_{i=1}^{n}x_i, \qquad \bar{y}=n^{-1}\sum_{i=1}^{n}y_i \end{aligned} \tag{1.2}$$

である．r の絶対値が1を超えないことを，次のようにして示すことができる．任意の実数にたいして

$$t^2 s_x^2-2ts_{xy}+s_y^2=n^{-1}\sum\{(y_i-\bar{y})-t(x_i-\bar{x})\}^2\geq 0 \tag{1.3}$$

が成りたつ．このことは，左辺の2次関数の判別式が非正であることを意味する，すなわち

$$s_{xy}^2-s_x^2 s_y^2\leq 0. \tag{1.4}$$

これは $r^2\leq 1$，すなわち $|r|\leq 1$ を意味する．$|r|=1$ となるのは判別式がゼロのとき，すなわち (1.3) の左辺をゼロにするような実数 t が存在する場合である．すなわち

$$y_i-\bar{y}=t_0(x_i-\bar{x}), \qquad i=1, 2, \cdots, n \tag{1.5}$$

となる実数 $t_0(\neq 0)$ が存在する場合にかぎり，$|r|=1$ となる．別の言葉でいいかえれば，n 個の観測値が一直線上に並ぶときにかぎり，$|r|=1$ となる．$t_0>0$ のとき $r=+1$ となり，$t_0<0$ のとき $r=-1$ となる．

表 1.1　数値例 1

首回り (X)	腕の長さ (Y)
38	81
40	82
34	78
41	81
34	75
38	79
42	83
36	79
35	77
39	80

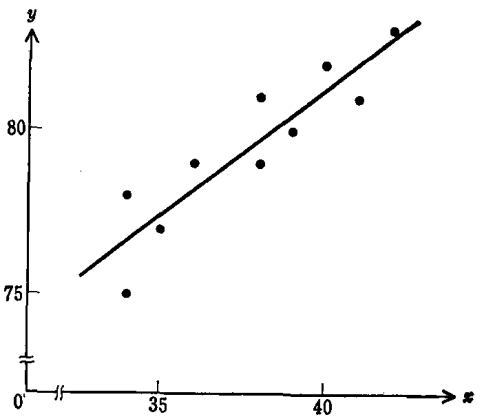

図 1.1　数値例 1 の散布図 ($r=0.9059$)

表 1.2　数値例 2

施肥量 (X)	収穫高 (Y)
10	9.5
20	12.0
30	14.5
40	16.0
50	16.0
60	17.5
70	18.0
80	19.0
90	19.0
100	19.5

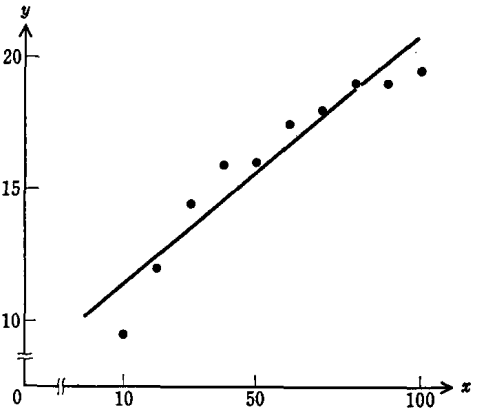

図 1.2　数値例 2 の散布図 ($r=0.9451$)

以上の演算から明らかなように，相関係数によって測れるのは，線形な関係からの乖離の程度であって，より広義の(非線形をも含めた)関係の尺度とはなりえない．散布図が図 1.1 のようになるとき，X と Y の関係は線形に近く，相関係数によって関係の強弱を測るのは妥当である．他方，散布図が図 1.2 のようであれば，X と Y の関係が非線形である可能性が高く，この場合，相関係数は適切な関係の尺度とはなりえない．

「関係の尺度」としての相関係数には，上に述べたような限界があるけれども，適当に工夫をこらすことによって，その適用範囲を拡げることができる．たとえば，表 1.2 の変量 X と Y を対数変換(各変量の対数をとる)して，あら

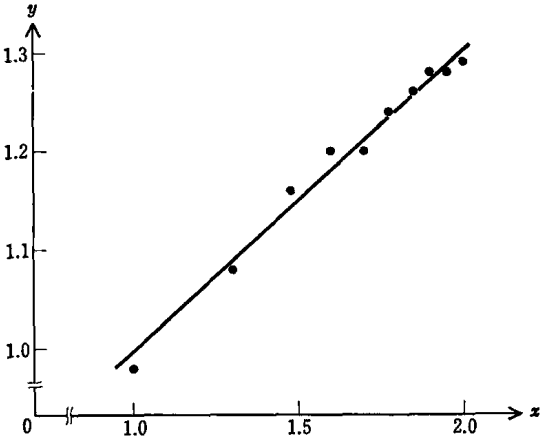

図 1.3 数値例2のデータを対数変換した散布図
($r=0.9923$)

ためて観測値をプロットしなおしてみると，図1.3のようになる．変換された観測値($\log x_i$, $\log y_i$)は直線のまわりにばらついている．すなわち，$\log X$と$\log Y$の関係は，直線関係に近いものとなり，相関係数によってそれらの関係の程度を測ることが妥当とみなされる．実際，対数変換された変量間の相関係数は0.9923となり，変換前のそれ0.9451よりも大きい．

1.1.3 線形回帰モデル

2個の変量間に有意な相関関係が認められたとしよう．このような関係の背後にある“構造”を解析してみたい，あるいは，そうした関係を“予測”に役だてたいと考えるのが，自然な発想の展開であろう．

たとえば，施肥量(X)と収穫高(Y)との間には，図1.2にみるように，「Xを増やせばYも増える」という関係が明確に認められる．こうした関係を，**因果関係**(causal relationship)とみなすことに，誰しも異論はあるまい．いうまでもなく，Xが原因(またはインプット)であり，Yが結果(またはアウトプット)である．両者の関係を$Y=f(X)$と書いてみる．ところでしかし，収穫高に影響するのは施肥量だけではない．地味，気温，降雨量等も，収穫高を決める重要な因子であろう．図1.2にプロットされた観測値が，施肥量以外の因子を等しく制御した実験によって得られたものであるとしても，すべての(x_i, y_i)が，なんらかの簡単な数式を厳密に満足するとは，とうてい考えられない．地味そ

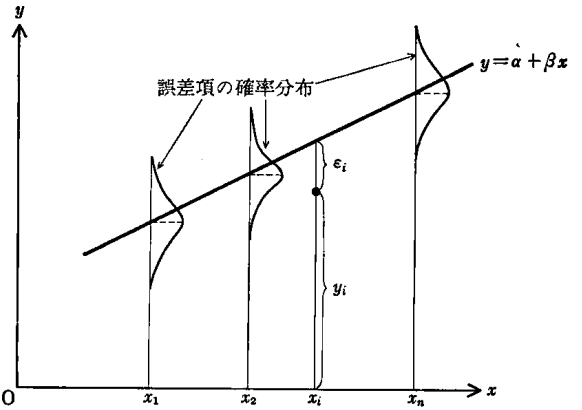

図 1.4　線形回帰モデル

の他の因子を完璧に等しく制御することは，実際問題として不可能であろうし，また自然界の因果関係が，簡単な数学的関数によって正確に表現される保障はない．すでに§1.1.2でみたように，施肥量(X)と収穫高(Y)との関係は，たんなる線形式よりも対数線形式によって，より良く近似できそうである．そこで

(1.6)　　$$\log Y=\alpha+\beta\log X+\varepsilon$$

という関係式を想定してみる．α と β は未知の母数(パラメータ)であり，ε は**確率誤差項**である．すなわち，α と β で決まる対数線形式(図 1.3 の直線)が，X と Y の基本的な関係式としてまずある．実際の観測値 (x_i, y_i) のすべてが，こうした単純な関係式を厳密に満たすわけではない．そこで，対数線形式と観測値とのズレを，確率的な誤差項 ε によるものと解釈しようというわけである．すなわち，各 (x_i, y_i) にたいしては，

(1.7)　　$$\log y_i=\alpha+\beta\log x_i+\varepsilon_i,\qquad (i=1,2,\cdots,10)$$

となる(図 1.4 参照)．

誤差項 ε は，平均が 0 で分散が一定の確率変数であり，異なる観測値に対応する誤差項はおたがいに無相関である，と仮定される．こうして定式化されたモデルのことを，(対数)**線形回帰モデル**という．さらに「ε が正規分布にしたがう」という仮定を追加したモデルのことを，**線形正規回帰モデル**という．

(1.6) の両辺を微分すれば，$(dY/Y)/(dX/X)=\beta$ という関係が導かれる．す

なわち β は，施肥量の増加率にたいする収穫高の増加率の比(弾力性係数)であり，この値を推定することじたい有意味であろう．また，α と β の値を推定すれば，施肥量の変化に対応する収穫量の変化を“予測”することができる．

回帰モデルの右辺にある変数を**独立変数**または**説明変数**とよび，左辺にある変数のことを**従属変数**または**被説明変数**とよぶ．未知母数 β のことを**回帰係数**とよぶ．

1.1.4 2次元正規分布の回帰関数

次に図 1.1 にプロットされたデータについて考えてみよう．首回り(X)と腕の長さ(Y)とのあいだには，線形な関係が存在することが，図から読みとれる．しかし，この関係を因果関係とみなすのは，どうみても不適切である．X と Y は，ひとりの人間の身体の異なる部位の測定値であって，それらの間の関係は，因果の序列をともなわない，純粋な相関関係である．そこで (X, Y) を，2次元正規分布にしたがう確率変数とみなし，(x_i, y_i) $(i=1, 2, \cdots, n)$ をそうした母集団からのランダム標本とみなすことにしよう．母集団分布の密度関数を

$$(1.8)\quad f_{XY}(x,y)=\frac{1}{(2\pi)\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

と書くことにする．

さて，レディメードのワイシャツメーカーにとって，成人男子の首回りと腕の長さの関係のあり方は，大きな関心事であろう．首回りが x($X=x$)の人の腕の長さ(Y)は，平均的に，どのくらいと予想されるか．さらに一歩進んで，首回りが x の人の腕の長さ(Y)の90%信頼区間を求めたい．すなわち，条件つき期待値 $\mu_{Y\cdot x}=E(Y|X=x)$ の点推定と区間推定に，ワイシャツメーカーは関心をもつ．2次元正規分布の仮定のもとで(詳しくは§3.1.2参照)，$\mu_{Y\cdot x}$ は

$$(1.9)\quad \mu_{Y\cdot x}=E(Y|X=x)=\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

という x の線形式になる．$Y-\mu_{Y\cdot x}=\varepsilon$ とすれば，$E(\varepsilon|X=x)=0$，その分散は

$$(1.10)\quad V(\varepsilon|x)=\sigma_y^2(1-\rho^2)$$