

講 座

デ ー タ 解 析 入 門

デ ー タ 解 析 入 門 5

戸 田 準

5. 回帰分析法

ある特性 y が別の特性 x の影響を受けるとき、あるいは x が原因で y が結果であるようなとき、 y を x の関数で表現しておけば便利である。もっとも簡単な場合は、次のように y を x の一次式で表す。

$$y = \beta_0 + \beta_1 x + \epsilon$$

このとき、 y を目的変数または従属変数、 x を説明変数または独立変数という。 ϵ は誤差である。実験では、 x が $x_1, x_2, \dots, x_i, \dots, x_n$ のときの y の観測値 $y_1, y_2, \dots, y_i, \dots, y_n$ を得る。データの形式は表 5.1(1) のようになる。

上の式は、 y の x に対する単回帰モデルといわれ、化学では分析の検量線としてなじみのものである。ある化学物質の濃度を x とし、そのときの反応の度合（たとえば吸光度など）を y とする。既知濃度の x について y を測定して、 β_0, β_1 を決めておけば、濃度未知の試料について y を測定して、逆に x を推定することができる。

y に影響を与える説明変数が 1 つではなく、2 つ以上考えなくてはならないこともある。説明変数の数が p のときのモデルは次のようになる。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

これを重回帰モデルといい、 β_0 を定数項、 $\beta_1, \beta_2, \dots$ を回帰係数という。データの形式は表 5.1(2) のようになる。重回帰モデルで、 $p=1$ の場合が単回帰モデルとなる。

表 5.1 の形式のデータから、回帰モデルの回帰係数の推定、検定を行うのが回帰分析である。

表 5.1. 回帰分析データ

(1) 単回帰分析

	1	2	...	i	...	n
説明変数 x	x_1	x_2	...	x_i	...	x_n
目的変数 y	y_1	y_2	...	y_i	...	y_n

$$\text{回帰モデル: } y = \beta_0 + \beta_1 x + \epsilon$$

(2) 重回帰分析

	1	2	...	i	...	n
説明変数 x_1	x_{11}	x_{12}	...	x_{1i}	...	x_{1n}
x_2	x_{21}	x_{22}	...	x_{2i}	...	x_{2n}
...	...	...	...	...	...	...
x_j	x_{j1}	x_{j2}	...	x_{ji}	...	x_{jn}
...	...	...	...	...	...	...
x_p	x_{p1}	x_{p2}	...	x_{pi}	...	x_{pn}
目的変数 y	y_1	y_2	...	y_i	...	y_n

$$\text{回帰モデル: } y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$$

5.1 単回帰分析

単回帰分析において推定しなくてはならない量は β_0, β_1 である。これらの推定は最小 2 乗法によるが、途中の議論は省いて必要なことだけを示す。必要な計算の詳細は表 5.2 にまとめた。表 5.2 の内容はとくに説明の必要はないであろう。これを参照しながら例を提示することにする。

[例 5.1] 寒天の濃度とゲルの破断強度の間には、一方が大きくなれば他方も大きくなるという単調な関係があることは大方の認めるところであろう。実際に、どのような関数で表現できるであろうか。寒天濃度を 5 水準とり、それぞれのゲルの破断強度を測定して表 [例 5.1] (1) のような結果を得た。寒天濃度を説明変数 x 、破断強度を目的変数 y とする。

まず、簡単なモデル 1 を想定してみる（表 [例 5.1] A）。

$$\text{モデル 1 } y = \beta_0 + \beta_1 x + \epsilon$$

回帰係数 β_1 の推定に必要な統計量は、説明変数 x の平方和 S_{xx} と、説明変数 x と目的変数 y の積和 S_{xy}

TODA Jun

著者紹介〔略歴〕1959 年 3 月東京大学農学部農芸化学科卒業。同年 4 月武田薬品工業(株)入社。1997 年 1 月同社定年退職。〔専門分野〕食品科学。〔趣味〕絵画(鑑賞)。〔連絡先〕〒666-0143 兵庫県川西市清和台西 3-3-17 (自宅)。

表 5.2. 単回帰分析

(1) データ

	1	2	...	i	...	n	合計	平均値
説明変数 x	x_1	x_2	...	x_i	...	x_n	X	$\bar{x}$
目的変数 観測値 y	y_1	y_2	...	y_i	...	y_n	Y	$\bar{y}$
推定値 $\hat{y}$	$\hat{y}_1$	$\hat{y}_2$	...	$\hat{y}_i$	...	$\hat{y}_n$	Y	

$$\begin{aligned} \text{合計: } X &= x_1 + x_2 + \cdots + x_i + \cdots + x_n & \text{平均値: } \bar{x} &= X/n \\ Y &= y_1 + y_2 + \cdots + y_i + \cdots + y_n & \bar{y} &= Y/n \end{aligned}$$

(2) 回帰モデル $y = \beta_0 + \beta_1 x + \varepsilon$

(3) 統計量

$$\begin{aligned} \text{平方和} \quad S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - X^2/n \\ S_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - Y^2/n \\ \text{積和} \quad S_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - XY/n \end{aligned}$$

(4) 回帰係数の推定

$$\begin{aligned} \beta_1 \text{ の推定値: } b_1 &= S_{xy}/S_{xx} \\ \beta_0 \text{ の推定値: } b_0 &= \bar{y} - b_1 \bar{x} \end{aligned}$$

(5) 検定 (分散分析)

$$\begin{aligned} \text{帰無仮説 } H_0 &: \beta_1 = 0 \\ \text{対立仮説 } H_1 &: \beta_1 \neq 0 \end{aligned}$$

全平方和	全自由度	
$S_T = S_{yy}$	$\phi_T = n - 1$	
回帰による平方和	回帰による自由度	回帰による平均平方
$S_R = b_1 S_{xy} = S_{xy}^2 / S_{xx}$	$\phi_R = 1$	$V_R = S_R / \phi_R$
残差平方和	残差自由度	残差平均平方
$S_e = S_T - S_R$	$\phi_e = \phi_T - \phi_R = n - 2$	$V_e = S_e / \phi_e$

分散分析表

	平方和	自由度	平均平方	分散比
回帰 R	S_R	ϕ_R	V_R	$F_0 = V_R / V_e$
残差 e	S_e	ϕ_e	V_e	
計 T	S_T	ϕ_T		

有意水準: α , $F_0 \geq F(1, n-2, \alpha)$ のとき帰無仮説を棄却する

(6) 寄与率 $r^2 = S_R / S_T$

$$\text{相関係数 } r = \sqrt{S_R / S_T}$$

(7) 回帰式による推定 $\hat{y} = b_0 + b_1 x$

- (注) 1. 目的変数の観測値 y の平均と推定値 $\hat{y}$ の平均値は等しい
 2. 回帰による平方和 S_R は, 推定値 $\hat{y}$ の平方和に等しい
 3. 相関係数 r は, 目的変数の観測値 y と推定値 $\hat{y}$ の相関係数である

である。回帰係数 β_1 の推定値を b_1 とする。 b_1 と説明変数および目的変数の平均値から、定数項 β_0 の推定値 b_0 が得られる (表 A(4))。

回帰分析の検定は、分散分析による。このときの帰無仮説は $\beta_1 = 0$ である。 $\beta_1 = 0$ ということは、 $y = \beta_0 + \varepsilon$ 、すなわち誤差範囲で y は x の値に関係なく一定で

ある、ということを意味する。帰無仮説が棄却できれば、たしかに y は x によって変わる、 y には x の影響があるということになる。

検定統計量は分散分析表の最右列にある分散比 F_0 である (表 A(5))。この例では F 表を参照するまでもなく、帰無仮説を棄却することができる。寒天ゲル

データ解析入門

表 [例 5.1] 単回帰分析・寒天ゲルの濃度と破断強度

A.

(1) データ: $n=5$

		1	2	3	4	5
説明変数: 寒天濃度 (%)	x	0.50	0.75	1.00	1.25	1.50
目的変数: 破断強度 (dyn/cm ²)	y	0.48	1.51	2.89	4.53	6.28
推定値	$\hat{y}$	0.21	1.68	3.14	4.60	6.06

[未発表]

(2) モデル 1: $y = \beta_0 + \beta_1 x + \epsilon$

(3) 統計量

	合 計	平均値	平方和	分 散
x	$T_x = 5.00$	$\bar{x} = 1.000$	$S_{xx} = 0.6250$	$V_{xx} = 0.1563$
y	$T_y = 15.69$	$\bar{y} = 3.138$	$S_{yy} = 21.5867$	$V_{yy} = 5.3967$
$x \times y$	積和	$S_{xy} = 3.6550$	共分散	$V_{xy} = 0.9138$

(4) 推 定 β_1 の推定値 $b_1 = S_{xy}/S_{xx} = 5.848$
 β_0 の推定値 $b_0 = \bar{y} - b_1 \bar{x} = -2.710$

(5) 検 定

帰無仮説 $H_0: \beta_1 = 0$ 対立仮説 $H_1: \beta_1 \neq 0$

全平方和, 回帰による平方和, 残差平方和, これらに対応する自由度, 平均平方を表 5.2 の公式にしたがって計算する.

分散分析表

	平方和	自由度	平均平方	分散比
回帰 R	$S_R = 21.3744$	$\phi_R = 1$	$V_R = 21.3744$	$F_0 = 302$
残差 e	$S_e = 0.2122$	$\phi_e = 3$	$V_e = 0.0707$	
計 T	$S_T = 21.5867$	$\phi_T = 4$		

(6) 寄与率 $r^2 = S_R/S_T = 0.9902$
 相関係数 $r = \sqrt{S_R/S_T} = 0.9951$

(7) 回帰式による推定値 $\hat{y} = b_0 + b_1 x = -2.710 + 5.848x$ ((1) 最下行)

の破断強度は寒天濃度で説明できる.

回帰による平方和 S_R を全平方和 S_T で割った値を寄与率 (または決定係数) という. 寄与率は, 全体の情報のうち回帰式によって説明できる情報の割合を表すと考えてよい. 寄与率が 0.99 ということは, 回帰式でほとんど誤差なく説明できることを意味している. 寄与率を r^2 と書くのは, その平方根が相関係数 r となるからである.

普通なら, このモデルで十分に満足できる. しかし, ゲルの弾性率や破断強度のような物理的な値 S とゲル化剤の濃度 C との間には, 次のような n 乗則が成立するといわれている.

$$S = aC^n$$

そこで, この形式のモデルを適用してみることにする.

$$\text{モデル 2 } y = ax^v$$

このモデルは線形ではないので, 両辺の対数をとる.

$$\log y = \log a + v \log x$$

ここで, $X = \log x$, $Y = \log y$, $A = \log a$ とおくと, 次のように線形となる. ここに誤差を考えることにする.

$$Y = A + vX + \epsilon$$

こうすると, モデル 1 と同じ形になる. そこで X を説明変数, Y を目的変数としてあらためて回帰分析を行う. モデル 1 と同じ手順で解析を行う. 対数変換後のデータ, 統計量は表 B に記してある.

モデル 1 がもともとよくあてはまっているので, さらによい結果は期待できない. 検定の結果や寄与率が

表 [例 5.1] の続き

B.

(1) データ

	1	2	3	4	5
説明変数: $X = \log x$	-0.3010	-0.1249	0	0.0969	0.1761
目的変数: $Y = \log y$	-0.3188	0.1790	0.4609	0.6561	0.7980
推定値 $\hat{Y}$	-0.2775	0.1344	0.4266	0.6532	0.8384

(2) モデル 2: $y = ax^v$

$$\log y = \log a + v \log x$$

$$X = \log x, Y = \log y, A = \log a$$

$$Y = A + vX + \epsilon$$

(3) 統計量

	合 計	平均値	平方和	分 散
X	$T_X = -0.1530$	$\bar{X} = -0.0306$	$S_{XX} = 0.1419$	$V_{XX} = 0.0355$
Y	$T_Y = 1.7752$	$\bar{Y} = 0.3550$	$S_{YY} = 0.7830$	$V_{YY} = 0.1958$
$X \times Y$	積和	$S_{XY} = 0.3320$	共分散	$V_{XY} = 0.0830$

(4) 推 定 v の推定値: $n = S_{XY}/S_{XX} = 2.3389$

$$A \text{ の推定値: } a = \bar{Y} - n \bar{X} = 0.4266$$

(5) 検 定

帰無仮説 $H_0: v = 0$ 対立仮説 $H_1: v \neq 0$

全平方和, 回帰による平方和, 残差平方和, これらに対応する自由度, 平均平方を表 5.2 の公式にしたがって計算する.

分散分析表

	平方和	自由度	平均平方	分散比
回帰 R	$S_R = 0.7765$	$\phi_R = 1$	$V_R = 0.7765$	$F_0 = 357$
残差 e	$S_e = 0.00652$	$\phi_e = 3$	$V_e = 0.00217$	
計 T	$S_T = 0.7830$	$\phi_T = 4$		

(6) 寄与率 $r^2 = S_R/S_T = 0.9917$

$$\text{相関係数 } r = \sqrt{S_R/S_T} = 0.9958$$

(7) 回帰式による推定値 $\hat{Y} = a + nX = 0.4266 + 2.3389X$ ((1) の最下行)

らみて, あてはまりはモデル 1 と同等であるといえるが, ここではモデル 2 を採用したい.

回帰分析では, このようにつくつかのモデルが同程度にあてはまることがよくある. そのとき, どのモデルを採用するかということは研究者の裁量にまかされる. 統計的にあてはまりがよいということとともに, 固有分野の知見や了解事項などを考慮してきめるべきである. それぞれの分野の知識や知見と整合する回帰モデルを考えなくてはならない.

この例でいえば, この分野の研究者ならば初めからモデル 2 を仮定するであろう. 参考までにモデル 1 を考えてみたものと理解していただきたい.

5.2 重回帰分析

説明変数の数 p が 2 以上の場合である. 重回帰分析を行うには, p 元の連立 1 次方程式 (正規方程式) を解くために, $p \times p$ 行列の逆行列を計算しなくてはならない. p が大きくなるとこの計算には大変な労力が必要だが, コンピュータが計算してくれる. しかし, 回帰分析の意味を理解するために, 小さいデータの解析を手計算で試みることをおすすめしたい.

以下, 表 5.3 に沿って説明していく. 計算の方法, 手順はこの表を参照されたい.

データの形式は表 5.1 (2) と同じものである. 回帰モデルを設定したら (表 (2)), p 個の説明変数間の平

データ解析入門

方積和を計算する (表(3)). S_{jj} は変数 x_j の平方和, S_{jk} は変数 x_j と変数 x_k の積和である. 平方和は自分自身との積和である. p 個の変数を 2 個ずつ対にして計算するから, 全部で $p(p+1)/2$ 個の値が得られる. これを表(3) のように $p \times p$ の平方和積和行列にまとめる. この行列は対称行列で, $S_{jk} = S_{kj}$ の関係がある. 同じように, 説明変数 x_j と目的変数 y との積和 S_{jy} を計算する. これは p 元のベクトルである.

次に, 平方和積和行列の逆行列を計算する. 逆行列の計算法にもいろいろあるが, これはコンピュータがやってくれる. 最近では, ポケットコンピュータにも逆行列計算機能を持ったものがある. 逆行列も $p \times p$ の対称行列である. その要素を $S^{\#}$ とする. この逆行列と説明変数と目的変数の積和ベクトルから, 回帰係数の推定値 $b_1, \dots, b_p$ を計算することができる. そのあとで, 定数項 b_0 の推定値を計算する (表(4)).

推定値が計算できたら, 分散分析による検定を行う. ここでも帰無仮説は「すべての回帰係数は 0」ということである. すなわち, 目的変数 y の値は説明変数 $x_1, \dots, x_p$ の値の影響を受けない, というのである. 以下, 表(5) に示した統計量を計算して, 分散分析表を作る. 帰無仮説が棄却できたら, 寄与率, 重相関係数などを計算し (表(6)), 必要があれば回帰式により推定値を計算する (表(7)).

[例 5.2] フルクトースとグルコースの混合溶液の「甘味度」を, これらの濃度の関数として表現するために行った実験の一部である. フルクトースの濃度 F とグルコースの濃度 G を変えた 12 個の糖溶液を調製した. 20 人の判定者が, 10% スクロース溶液の「甘味の強さ」を基準 (100) として, フルクトースとグルコースの混合溶液の「甘味の強さ」をマグニチュード推定法で評価した. 20 人の評価の幾何平均を甘味度 (Y) とする (表 [例 5.2] (1)).

フルクトースの濃度 F とグルコースの濃度 G が説明変数, 甘味度 Y が目的変数となるべきものであるが, 感覚量をマグニチュード推定法で測定した場合は, 感覚量は物理量のべき関数で表現できるとされているため, ここでは初めから (自然) 対数に変換して回帰分析を行った. すなわち, フルクトースの濃度の対数 $\ln F$ を x_1 , グルコースの濃度の対数 $\ln G$ を x_2 とし, 「甘味度」の対数 $\ln Y$ を y として $p=2$ の重回帰分析を行う (自然対数としても常用対数としても, 本質的な違いはない).

表 [例 5.2] には, 表 5.3 に対応させて具体的な推

定, 検定に必要な統計量を示した. この例でも, 寄与率はきわめて高い. この種の実験では, 統計的な検定が有意になることは自明であり, 回帰分析の目的は回帰係数の値を推定することである. この例では, フルクトースの係数が 1.09, グルコースの係数が 0.74 で, 一般にいわれているフルクトースとグルコースの甘味度と一致する係数が得られたといえる.

(注) 例 5.2 は 2 成分系の呈味相互作用を一般的に表現する関数を求めるために行った研究の一部である. 実際には, ここに記載したモデルではなく, 非線形のモデルを採用した. 本例は数値例として理解していただきたい.

この例は $p=2$ で, 説明変数の数が少ない. p の数が多くなると, 帰無仮説が棄却されても p 個の説明変数の中に意味のあるものがあることを指摘するだけで, 説明変数のすべてが目的変数に影響を与えているということにはならない. 意味のある説明変数を見つけるために種々の手法がある. たとえば, 回帰係数の各々について有意性の検定を行う方法とか, はじめから p 個の説明変数を含むモデルを考えるのではなく, 目的変数に対する影響の大きい説明変数からモデルに取り入れていく変数選択法などがある. これらについて論ずることは本稿の目的を超えるので割愛するが, 回帰分析法を有効に活用するには実質的な面と統計的な面からの多面的な検討が必要であることを指摘しておきたい.

5.3 補 足

(1) 線形モデル

一般的な回帰モデルとして, 次のモデルを示した.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

このようなモデルは線形モデルといわれる. 「線形」という意味を説明しておく. 上のモデルで $y; x_1, x_2, \dots, x_p$ は観測値として得られているが, $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ の値は未知である. 上のモデルは, これらの未知の値の 1 次式である. 数字では 1 次式のことを線形式ともいう. 未知の値について線形であるという意味で, 線形モデルとよぶ.

プディングのようなゲル状食品のテクスチャーの嗜好を考えてみよう. テクスチャーの嗜好に対応する物理的特性としては破断強度のようなものが考えられる. そして, かたすぎてもやわらかすぎても嗜好は低下し, どこかに嗜好が最大になる点があると考えても無理はないであろう. そこで, 嗜好度 y を目的変数とし, 破

表5.3. 重回帰分析

(1) データ

	1	2	...	i	...	n	合 計	平均値
説明変数 x_1	x_{11}	x_{12}	...	x_{1i}	...	x_{1n}	X_1	$\bar{x}_1$
...	...	...	...	...	...	...	...	...
x_j	x_{j1}	x_{j2}	...	x_{ji}	...	x_{jn}	X_j	$\bar{x}_j$
...	...	...	...	...	...	...	...	...
x_p	x_{p1}	x_{p2}	...	x_{pi}	...	x_{pn}	X_p	$\bar{x}_p$
目的変数 観測値 y	y_1	y_2	...	y_i	...	y_n	Y	$\bar{y}$
推定値 $\hat{y}$	$\hat{y}_1$	$\hat{y}_2$	...	$\hat{y}_i$	...	$\hat{y}_n$	Y	$\bar{y}$

合計: $X_j = x_{j1} + x_{j2} + \dots + x_{ji} + \dots + x_{jn}$, 平均値: $\bar{x}_j = X_j/n$ ($j=1, \dots, p$)

$Y = y_1 + y_2 + \dots + y_i + \dots + y_n$, $\bar{y} = Y/n$

(2) 回帰モデル $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_j x_j + \dots + \beta_p x_p + \epsilon$

(3) 統計量

説明変数 x_j と x_k の積和: $S_{jk} = \sum_{i=1}^n (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k) = \sum_{i=1}^n x_{ji} x_{ki} - X_j X_k / n$ ($j, k=1, \dots, p$)

説明変数 x_j と目的変数 y の積和: $S_{jy} = \sum_{i=1}^n (x_{ji} - \bar{x}_j)(y_i - \bar{y}) = \sum_{i=1}^n x_{ji} y_i - X_j Y / n$ ($j=1, \dots, p$)

目的変数 y の平方和: $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - Y^2 / n$

平方和積和行列

	x_1	x_2	...	x_k	...	x_p
x_1	S_{11}	S_{12}	...	S_{1k}	...	S_{1p}
x_2	S_{21}	S_{22}	...	S_{2k}	...	S_{2p}
...	...	...	...	...	...	...
x_j	S_{j1}	S_{j2}	...	S_{jk}	...	S_{jp}
...	...	...	...	...	...	...
x_p	S_{p1}	S_{p2}	...	S_{pk}	...	S_{pp}

説明変数と目的変数の積和

y
S_{1y}
S_{2y}
...
S_{jy}
...
S_{py}

平方和積和行列の逆行列

	x_1	x_2	...	x_k	...	x_p
x_1	S^{11}	S^{12}	...	S^{1k}	...	S^{1p}
x_2	S^{21}	S^{22}	...	S^{2k}	...	S^{2p}
...	...	...	...	...	...	...
x_j	S^{j1}	S^{j2}	...	S^{jk}	...	S^{jp}
...	...	...	...	...	...	...
x_p	S^{p1}	S^{p2}	...	S^{pk}	...	S^{pp}

(4) 回帰係数の推定

β_j の推定値: $b_j = \sum_{k=1}^p S^{*k} S_{ky}$ ($j=1, \dots, p$)

β_0 の推定値: $b_0 = \bar{y} - \sum_{k=1}^p b_k \bar{x}_k$

(5) 検 定 (分散分析)

帰無仮説 $H_0: \beta_1 = \beta_2 = \dots = \beta_j = \dots = \beta_p = 0$

対立仮説 $H_1: \text{少なくとも1つの}\beta_j\text{が}0\text{ではない}$

全平方和

$S_T = S_{yy}$

回帰による平方和

$S_R = \sum_{k=1}^p b_k S_{ky}$

残差平方和

$S_e = S_T - S_R$

全自由度

$\phi_T = n - 1$

回帰による自由度

$\phi_R = p$

残差自由度

$\phi_e = \phi_T - \phi_R = n - p - 1$

回帰による平均平方

$V_R = S_R / \phi_R$

残差平均平方

$V_e = S_e / \phi_e$

データ解析入門

分散分析表

	平方和	自由度	平均平方	分散比
回帰 R	S_R	ϕ_R	V_R	$F_0 = V_R/V_e$
残差 e	S_e	ϕ_e	V_e	
計 T	S_T	ϕ_T		

有意水準： α , $F_0 \geq F(p, n-p-1, \alpha)$ のとき帰無仮説を棄却する

$$(6) \text{ 寄与率 } R^2 = S_R/S_T$$

$$\text{重相関係数 } R = \sqrt{S_R/S_T}$$

(7) 回帰式による推定

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_j x_j + \cdots + b_p x_p = b_0 + \sum_{k=1}^p b_k x_k$$

(注) 1. 目的変数の観測値 y の平均と推定値 $\hat{y}$ の平均値は等しい

2. 回帰による平方和 S_R は, 推定値 $\hat{y}$ の平方和に等しい

3. 重相関係数 R は, 目的変数の観測値 y と推定値 $\hat{y}$ の相関係数である

$$4. \text{ 逆行列の性質: } \left\{ \begin{array}{l} \sum_{j=1}^p S_{ji} S^{jk} = \delta_{ik} \\ \sum_{i=1}^p S^{ji} S_{ek} = \delta_{jk} \end{array} \right\}, \quad \delta_{jk} = \begin{cases} 1 & (j=k) \\ 0 & (j \neq k) \end{cases}$$

断強度 x を説明変数として最大値を持つようなモデルを考える。最大値を持つモデルとして x の 2 次式を仮定すると, モデルは次のようにかける。

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

x^2 の項があるが, これも線形モデルである。なぜなら, 未知の値 $\beta_0, \beta_1, \beta_2$ については 1 次式であるからである。あらためて $x_1 = x, x_2 = x^2$ とおけばモデルは次のようになり線形であることは明らかである。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$y = ax^n$ 形は, a と n が推定すべき未知の値であれば, このままでは線形ではない。対数変換により $Y = \log y, X = \log x, A = \log a$ とすれば, 次のように線形モデルに変換できることは例で示した通りである。

$$Y = A + nX$$

(2) モデルの重要性

実験範囲では線形モデルがうまくあてはまっても, 本質的な関係も線形であると考えすることはできない。変数の変域が狭いときには, 本質的な関係がどんなものであっても, 説明変数と目的変数の関係は 1 次式でよく説明できることが多い。実用的な立場で行う実験では, もともと最適値付近のごく狭い範囲しか変数を動かさないで, 線形モデルで十分な場合が多いが, このようなときは, 常に変数の変動範囲を付記して回帰式を提示すべきである。しかし, 基礎的な研究では, 本質的な関係を見出すことに挑戦すべきであり, まず理論的あるいは経験的に事前にモデルを設定して, そこに含まれる未知数を推定することが, 広い意味の回

帰分析の活用法であろう。

たとえば, 説明変数 x が 0 のとき目的変数 y も 0 になるはずである, あるいはなるべきであるという現象も多い。化学分析における検量線はこのような性質のものであろう。このような場合には, はじめから定数項を除いたモデルで考えてもよい。

閾値の測定を行うときのように, 物理量に対する反応率を表現することを考えてみる。反応率は 0 と 1 の間におさまるものであるから, 実験範囲で反応率が物理量の 1 次式で表されたとしても普遍性はない。プロビット分析のモデルなど, 反応率が 0 と 1 の範囲におさまるモデルをあてはめるべきである。

(3) 計画された実験のデータと“なりゆき”のデータ

原因と結果の関係にあると考えられる特性群があるとき, 原因と考えられる特性を説明変数とし, 結果と考えられる特性を目的変数として回帰分析を行うことがある。また, 制御したい特性を直接制御することはむずかしいが, それに影響を与える特性で制御が容易なものがあるとき, 制御可能な特性を説明変数, 制御がむずかしい特性を目的変数として回帰分析を行うこともある。ところが, このような関係が予想できない場合でも, 表 5.1 に示したようなデータがあれば形式的に回帰分析を適用することはできる。そのとき, どれを説明変数とし, どれを目的変数にするか自明ではないこともある。このようなときは, 回帰分析よりも相関分析の方が適当であろう。

表 [例 5.2] 重回帰分析：フルクトース・グルコース混合溶液の甘味度

(1) データ： $p=2$, $n=12$

	Fructose F	Glucose G	甘味度 Y	$\ln F$ x_1	$\ln G$ x_2	$\ln Y$	
						y	$\hat{y}$
1	2.25	0.75	14.4	0.8109	-0.2877	2.6672	2.5680
2	4.50	1.50	51.4	1.5041	0.4055	3.9396	3.8372
3	9.00	3.00	204.1	2.1972	1.0986	5.3186	5.1064
4	18.00	6.00	485.6	2.8904	1.7918	6.1854	6.3756
5	1.50	1.50	11.6	0.4055	0.4055	2.4510	2.6358
6	3.00	3.00	41.9	1.0986	1.0986	3.7353	3.9050
7	6.00	6.00	196.3	1.7918	1.7918	5.2796	5.1741
8	12.00	12.00	515.5	2.4849	2.4849	6.2451	6.4433
9	0.75	2.25	9.7	-0.2877	0.8109	2.2721	2.1767
10	1.50	4.50	22.7	0.4055	1.5041	3.1224	3.4459
11	3.00	9.00	157.1	1.0986	2.1972	5.0569	4.7151
12	6.00	18.00	443.3	1.7918	2.8904	6.0942	5.9843
平均値				1.3493	1.3493	4.3640	

[和田武夫・河村貫二・戸田 準：農化, 59, 487 (1985)]

(2) モデル： $y=\beta_0+\beta_1x_1+\beta_2x_2+\epsilon$

(3) 統計量：

平方和積和行列

説明変数と目的変数の積和

	x_1	x_2	y
x_1	$S_{11}=9.6759$	$S_{12}=4.8481$	$S_{1y}=14.1568$
x_2	$S_{21}=4.8481$	$S_{22}=9.6759$	$S_{2y}=12.4372$

平方和積和行列の逆行列

	x_1	x_2
x_1	$S^{11}=0.1380$	$S^{12}=-0.0691$
x_2	$S^{21}=-0.0691$	$S^{22}=0.1380$

(4) 推 定

 β_1 の推定値： $b_1=S^{11}S_{1y}+S^{12}S_{2y}=1.0936$ β_2 の推定値： $b_2=S^{21}S_{1y}+S^{22}S_{2y}=0.7374$ β_0 の推定値： $b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2=1.8933$

(5) 検 定

帰無仮説 $H_0:\beta_1=\beta_2=0$ 対立仮説 $H_1:\beta_1, \beta_2$ のうち少なくともいずれか1つが0ではない全平方和，回帰による平方和，残差平方和及びそれぞれに対応する自由度と平均平方，分散比 F_0 を表 5.3 の公式によって計算する。

分散分析表

	平方和	自由度	平均平方	分散比
回帰 R	$S_R=24.6536$	$\phi_R=2$	$V_R=12.3268$	$F_0=242$
残差 e	$S_e=0.4576$	$\phi_e=9$	$V_e=0.0508$	
計 T	$S_T=25.1112$	$\phi_T=11$		

(6) 寄与率 $R^2=S_R/S_T=0.9818$ 重相関係数 $R=\sqrt{S_R/S_T}=0.9908$ (7) 回帰式による推定値 $\hat{y}=b_0+b_1x_1+b_2x_2$ ：(1) の最右列

データ解析入門

この議論に関連して、計画された実験のデータと“なりゆき”のデータを考えてみたい。計画された実験とは、温度や時間、原材料の配合などの条件を実験者が一定の水準に設定して操作し、目的の特性を測定する実験である。実験者が設定した条件が説明変数となり、特性の観測値が目的変数となる。これに対して、“なりゆき”のデータとは出来上がったものを実験材料として入手して、いくつかの特性を測定したものである。そのなかの1つを目的変数とし、他の特性を説明変数として回帰分析を適用する例もみられるが、説明変数を特定の値に設定することはできない。実験というより観察というほうが適当である。このような“なりゆき”のデータの解析は予測というより説明であり、本格的な実験に先立つ探索的なものと位置付けるのがよいだろう。

回帰分析を予測という側面からみれば、説明変数には人為的に制御できる特性をとりあげたい。予測のためには、統計的に有意であるというだけでは使いものにならないことは、分析化学における検量線を考えてみれば明らかである。

おわりに

問題を設定し、それを解決するために調査や実験を計画・実施して結果をまとめるまでにはいくつか過程がある。さらに、このような調査・実験をいくつか積み重ねて、一つの研究が完結する。本稿で述べてきたことは、これらの一連の過程のほんの一部にかかわるものにすぎない。すなわち、観測値や測定値が得られたあとのデータの解析法を解説しただけである。データを統計的に解析したからといって、データの質が向上するわけではない。調査や実験で得られた情報のすべては観測値に含まれており、観測値にしかない。データ解析にできることは、その情報を効率よく取り出し、問題に則して整理することである。そのような意味で、本稿では言及しなかったが、データを得るまでの過程が妥当であったかどうかということが重要であ

る。

いまでは、測定機器の多くが自動化され、統計計算のパソコンソフトも容易に入手できるようになって、実験やデータの集計・解析は格段に省力化された。このような状況が科学や技術の展開に寄与するという見解に異論はない。ただ、実験や解析が手間と時間のかかるものであれば、実際に手を付ける前に慎重に検討した上で実験にとりかかるであろうが、簡単にできるとなると事前の検討が十分になされないのではないかという危惧はある。

専門書を揃えた書店に行くと、統計学関係の単行書がずらっと並んでいる。大学の教科書として書かれたものが正統的なものであろうが、統計学的手法をあくまで手段と考えておられる方々からは、使いにくいという声も聞く。数式を見ただけで理解できる人もいるが、多くの人はそうではない。しかし、統計学的な論理や手法が家政学や生活科学で有効なものであれば、その分野に似合うように仕立て直して自分の中に取り込むことも必要であろう。統計学の素人である筆者があえて本稿を執筆したのも、そのような試みの一つのもりである。手始めに、小さいデータを手計算で解析してみることをおすすめしたい。理解の助けになるはずである。本稿がそのような動機付けになれば望外の幸いである。本稿には、単純な計算間違いだけでなく、重要な誤りもあるかもしれないし、厳密性を欠いた部分もある。ご批判ご意見を賜れば歓迎するところである。

終りにあたって、データの引用をご快諾いただいた諸先生、とくに未発表のデータをご提供いただいた先生方にお礼を申し上げる。また、原稿の段階で査読をお願いし、貴重なご意見をいただいた関西福祉大学教授和田武夫博士に感謝する。さらに、本稿の執筆をおすすめいただいた本誌前編集委員長 お茶の水女子大学教授畑江敬子先生に深謝する。