回帰分析における
偏回帰係数推定値と相関係数の符号逆転現象

猪原正守 大阪電気通信大学

概 要

回帰分析は，目的変数 y と複数の説明変数 $x_1, x_2, \ldots, x_p$ の間の因果関係を説明する統計モデルとして，工学，自然科学，医学，社会科学などの広い分野で活用される．しかし，回帰モデルの説明変数 $x_i (i = 1, 2, \ldots, p)$ に対する偏回帰係数に対する推定値 $\hat{\beta}_i$ が，目的変数 y と説明変数 x_i の相関係数と異なる符号になる現象が発生することがあるため，統計学の初心者や多くの技術者に混乱を与えている．この論文では，そうした符号逆転現象が，いくつかの層別母集団における回帰分析における回帰係数の真値が異なることによって引き起こされる可能性を数理的に明らかにし，数値例を与える．

keywords:回帰分析，偏相関係数，条件付き分布，層別解析

1　はじめに

統計的品質管理 (Statistical Quality Control：SQC) に関する手法を学び始めると，目的変数 y と説明変数 $x_1, x_2, \ldots, x_p$ の間の因果関係を解析する手法として，早い段階で回帰分析を学習する (猪原，飯塚，岩崎 (2014)，永田，棟近 (2014))．

事例 某社の化学工場では，製品 A の収率 $y(\%)$ を高めるため原料に副原料 $x_1(g/l)$ を添加している．しかし，目標とする収率を達成できないときがあるため，その原因分析を行ったところ，「主原料の酸度 $x_2(pH)$ が影響しているのではないか」との意見があり，最近 30 日間の製造記録から無作為にデータを収集して，表 1 のデータ表を得た[1]．

表 1　データ

No.	副原料 x_1	酸度 x_2	収率 y	No.	副原料 x_1	酸度 x_2	収率 y
1	37	51	73.5	16	35	47	70.4
2	41	53	72.8	17	40	52	71.5
3	38	55	73.7	18	39	46	69.3
4	40	50	71.3	19	38	47	69.6
5	35	44	70.1	20	38	49	69.4
6	37	47	70.8	21	42	57	73.5
7	42	56	74.7	22	39	52	70.9
8	36	48	71.0	23	35	48	70.2
9	42	61	76.6	24	40	51	72.7
10	36	49	71.6	25	35	48	71.4
11	34	47	69.9	26	39	47	68.5
12	38	50	71.4	27	38	49	69.8
13	37	48	71.8	28	35	48	71.4
14	39	53	73.5	29	40	55	73.4
15	37	47	70.3	30	39	54	71.5

[1]この事例は，猪原，飯塚，岩崎 (2014) を参考に作成したものである．

2015 年 5 月 11 日受理

表１のデータから基本統計量，分散・共分散および相関係数を求めると，表２と表３のようになる．

表２ 基本統計量

変数	最大値	最小値	平均値	標準偏差	変動係数	ひずみ	とがり
x_1	42	34	38.0	2.282	0.060	0.068	-0.815
x_2	61	44	50.3	3.834	0.076	0.922	0.636
y	76.6	68.5	71.55	1.806	0.025	0.785	0.673

表３ 分散共分散と相関係数

変数	x_1	x_2	y
x_1	5.206	0.760	0.551
x_2	6.645	14.700	0.867
y	2.271	6.002	94.635

ただし．表３における対角線から下側は共分散であり，上側は相関係数である．

表１のデータから，副原料 x_1 と収率 y および酸度 x_2 と収率 y の散布図を作成すると，図１と図２が得られる．

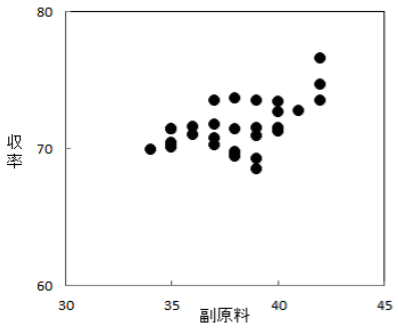

図１ 副原料と収率

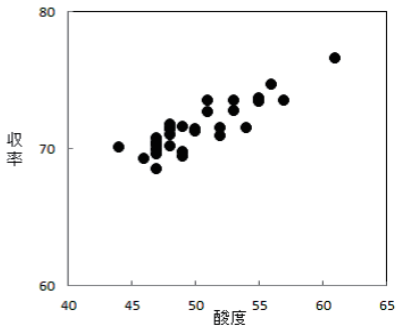

図２ 酸度と収率

また，これまでの考え方の基本となる収率 y の副原料 x_1 の回帰式は

$$\hat{y} = 71.55 + 0.63(x_1 - 38.0) \tag{1.1}$$

と推定され，副原料 x_1 を１単位だけ増加すると収率 y は 0.63(%) だけ増加することがわかる．

一方，収率 y と副原料 x_1 および酸度 x_2 の回帰式は

$$\hat{y} = 71.55 - 0.20(x_1 - 30.8) + 0.50(x_2 - 50.3) \tag{1.2}$$

と推定される．

収率 y の変動に対する回帰による寄与率は $R^2 = 0.511$ から $R^2 = 0.778$ へと増加するが，(1.2) における x_1 の回帰係数に対する推定値が，(1.1) のプラスから (1.2) のマイナスへと変化している．この符号の反転現象に遭遇することで，回帰分析に対する初学者や多くの技術者は (1.2) の結果を技術的に解釈することに困惑することとなる．

本論文では，この回帰係数推定値の相関係数や単回帰分析における回帰係数推定値との符号逆転現象に対する統計理論の立場に基づく説明法をレビューした後，層別回帰の立場からの説明を与える．統計理論からの説明に比べて，層別回帰による説明は理論性には欠けるが分かりやすい説明になっている．また，こうした符号逆転現象に遭遇したときは，層別解析を検討する必要性を示唆する．このことが技術者に与える知見の意義についても，数値例を通して説明する．

2　回帰分析の基本

2.1　回帰モデル

目的変数 y と p 個の説明変数 $x_1, x_2, \ldots, x_p$ との因果関係を解析するため，目的変数 y と説明変数 $x_1, x_2, \ldots, x_p$ の間に，回帰モデル

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e \tag{2.1}$$

を仮定し，表 1 のような変数 $(y, x_1, x_2, \ldots, x_p)$ に対する n 組の観測値から，未知パラメータ $\beta_0, \beta_1 \cdots, \beta_p$ および誤差分散 $V(e) = \sigma^2$ を最小 2 乗法などの推定法によって求める方法を回帰分析という．また，説明変数が $p = 1$ 個の場合を単回帰分析，$p > 2$ の場合を重回帰分析と区別し，重回帰分析における回帰係数を偏回帰係数と呼ぶことがある．

2.2　回帰推定値

観測された n 組のデータ $(y_1, x_{11}, \ldots, x_{1p}), \ldots, (y_n, x_{n1}, \ldots, x_{np})$ から計算される基本統計量を

$$\begin{cases} \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \ \ \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \ \ S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \ \ S_{jk} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \\ S_{jy} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(y_i - \bar{y}) \ \ (j, k = 1, 2, \ldots, p) \end{cases} \tag{2.2}$$

とするとき，未知パラメータは

$$\begin{cases} \hat{\beta}_j = \sum_{k=1}^{p} S^{jk} S_{ky} = S^{j1} S_{1y} + \cdots + S^{jp} S_{py}, \ \ (j = 1, 2, \ldots, p) \\ \hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_p \bar{x}_p) \end{cases} \tag{2.3}$$

によって推定される．ただし，$S^{-1} = (S^{jk})$ は，説明変数 $x_1, \ldots, x_p$ の偏差積和平方和行列 $S = (S_{jk})$ の逆行列である．

回帰による変数 y の観測値 y_i に対する予測値 $\hat{y}_i$ は

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \tag{2.4}$$

によって与えられ，予測値の変動 S_R は

$$S_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{j=1}^{p} \hat{\beta}_j S_{jy} \tag{2.5}$$

によって与えられる．特に，単回帰分析 $(p = 1)$ の場合には

$$\hat{\beta}_1 = \frac{S_{1y}}{S_{11}}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1, \ S_R = \hat{\beta}_1 S_{1y} \tag{2.6}$$

である，ここで，$R^2 = S_R / S_{yy}$ を回帰による寄与率という．

3　偏回帰係数の符号逆転現象の説明

本章では，第 1 章で述べた単回帰分析における回帰係数と重回帰分析における回帰件数の推定値の符号が逆転する現象について，統計学の立場と層別回帰の立場から説明を行う．

3.1　第３変数の影響による説明

表１のデータが得られた製造工程では，オペレーターが，原料の酸度 x_2 の値に応じて投入する副原料の量 x_1 を調整している可能性があるため，副原料の投入量 x_1 の値が，酸度 x_2 の値によって影響を受けているかもしれない．

そこで，変数 x_1 から変数 x_2 による影響を除去するため，変数 x_1 の変数 x_2 による回帰予測値 $\hat{x}_1$ を求め，変数 $x_{1\cdot 2} = x_1 - \hat{x}_1$ によって収率 y を推定すると，

$$\hat{y} = 15.37 - 0.20 x_{1\cdot 2} \tag{3.1}$$

となって，回帰係数の推定値は偏回帰係数の推定値 $\hat{\beta}_1$ と一致する．

もし，副原料の投入量 x_1 を原料の酸度 x_2 による予測値 $\hat{x}_1$ に設定することが正しいものとすると，変数 $x_{1\cdot 2}$ は，オペレーターによる調整誤差を意味することになるため，(3.1) は，調整誤差 $x_{1\cdot 2}$ と収率 y の因果関係を表していると解釈することができる．すなわち，調整誤差がプラス側のときは収率が低下し，マイナス側のときは収率が増加していると解釈できる．

3.2　条件付き分布による説明

(1) 偏相関係数による説明

回帰分析において，説明変数 x_1, x_2 は制御された変数や外的に与えられた変数であると考えることが一般的であるが，それらを含む変数 $\mathbf{x}^t = (x_1, x_2, x_3) = (x_1, x_2, y)$ が３次元正規分布に従う確率変数であると仮定する．そして，それらの平均ベクトルと分散行列を

$$\begin{cases} \mu = E(\mathbf{x}) = (\,\mu_1, \quad \mu_2, \quad \mu_3\,)^t \\ \Sigma = V(\mathbf{x}) = (\sigma_{jk}) \end{cases} \tag{3.2}$$

であるとする．ただし，記号 t は，ベクトルと行列の転置を表すものとする．

このとき，変数 x_2 を与えたときの変数 x_1, x_3 の条件付き平均は

$$\begin{aligned} \mu_{1\cdot 2} &= E(x_2|x_1) = \mu_1 - \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2) \\ \mu_{3\cdot 2} &= E(x_3|x_1) = \mu_3 - \frac{\sigma_{32}}{\sigma_{22}}(x_2 - \mu_2) \end{aligned}$$

条件付き分散共分散は

$$\begin{aligned} \sigma_{11\cdot 2} &= V(x_1|x_2) = \sigma_{11} - \frac{\sigma_{12}\sigma_{21}}{\sigma_{22}} \\ \sigma_{33\cdot 2} &= V(x_3|x_2) = \sigma_{33} - \frac{\sigma_{32}\sigma_{23}}{\sigma_{22}} \\ \sigma_{13\cdot 2} &= C(x_1, x_3) = \sigma_{13} - \frac{\sigma_{12}\sigma_{23}}{\sigma_{22}} \end{aligned}$$

で与えられる．特に，変数 x_1 を与えたときの x_2 と x_3 の条件付き相関係数（「偏相関係数」という）は

$$\rho_{13\cdot 2} = \frac{\sigma_{13\cdot 2}}{\sqrt{\sigma_{11\cdot 2}\sigma_{33\cdot 2}}} \tag{3.3}$$

となる．

ここで，(3.2) の平均ベクトルと分散共分散行列が，表２と表３によって推定されるとすれば，x_1 と x_3 の偏相関係数は

$$r_{13\cdot 2} = \frac{-12.826}{\sqrt{63.861 \times 23.574}} = -0.331$$

となる．すなわち，変数 x_2 を与えたときの変数 x_1 と変数 $x_3(=y)$ の相関係数 $r_{1y\cdot 2}$ は，変数 x_2 を考慮しない場合の相関係数 $r_{1y}=0.551$ とは違ってマイナスの値になっている．また，変数 x_2 を与えたときの変数 x_1 と変数 $x_3(=y)$ との回帰関係は

$$\begin{aligned}\hat{x}_{3\cdot 2} &= \mu_{3\cdot 2}+\frac{\sigma_{13\cdot 2}}{\sigma_{11\cdot 2}}(x_{1\cdot 2}-\mu_{1\cdot 2})\\ &= 51.06+\frac{-12.826}{23.574}(x_{1\cdot 2}-15.26)\\ &= 51.06-0.20(x_{1\cdot 2}-15.26)\end{aligned} \tag{3.4}$$

となる．

(3.4) は，酸度 x_2 の値を与える, すなわち，酸度の値を事前情報として与えたとき，オペレータの副原料投入調整量 $x_{1\cdot 2}$ と収率 y の間に成立している因果関係を与えたことになる．

(2) 変数間の相関関係による説明

(3.4) における回帰係数は

$$\begin{aligned}\frac{\sigma_{13\cdot 2}}{\sigma_{11\cdot 2}} &= \frac{\sigma_{13}-\frac{\sigma_{12}\sigma_{23}}{\sigma_{22}}}{\sigma_{11}-\frac{\sigma_{12}\sigma_{21}}{\sigma_{22}}}\\ &= \frac{\sigma_{13}\sigma_{22}-\sigma_{12}\sigma_{23}}{\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21}}\left(=\frac{\sigma_{1y}\sigma_{22}-\sigma_{12}\sigma_{2y}}{\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21}}\right)\end{aligned} \tag{3.5}$$

であるから，σ_{1y} と σ_{2y} がプラスであっても，$\sigma_{1y}\sigma_{22}-\sigma_{12}\sigma_{2y}<0$ ならば，偏回帰係数の符号はマイナスになる．このことは，$\sigma_{13\cdot 2}$ が (3.3) における変数 x_1 と変数 x_3 の偏相関係数の分子であることから，偏相関係数 $\rho_{13\cdot 2}$ の符号と (3.5) の符号が一致することを意味している．

ここで，

$$\begin{pmatrix}\sigma^{11} & \sigma^{12}\\ \sigma^{21} & \sigma^{22}\end{pmatrix}=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}^{-1}$$

とおくと，(3.5) は

$$\frac{\sigma_{13\cdot 2}}{\sigma_{11\cdot 2}}=\sigma^{11}\sigma_{1y}+\sigma^{12}\sigma_{2y}$$

である．したがって，(2.3) に注意すると，x_2 を与えたときの y と x_1 の回帰関係における回帰係数 (3.5)，x_1 と y の偏相関係数 $\rho_{1y\cdot 2}$，および，x_1 と x_2 による y の重回帰式における偏回帰係数 β_1 は，すべて同じ符号になっていることがわかる．

ここで，(3.5) を少し変形すると，

$$\frac{\sigma_{1y}}{\sqrt{\sigma_{11}\sigma_{yy}}}-\frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}\frac{\sigma_{2y}}{\sqrt{\sigma_{22}\sigma_{yy}}}=\rho_{1y}-\rho_{12}\rho_{2y}$$

であることから，これらの符号は $\rho_{1y}-\rho_{2y}\rho_{12}$ の符号とも一致している．

したがって，x_2 と y の相関係数 ρ_{2y} がプラスであるとき，これらの符号は，$\frac{\rho_{1y}}{\rho_{2y}}-\rho_{12}$ によって決定することになる．一般に,「偏相関係数 β_1 の符号が y の x_1 への単回帰分析における回帰係数の符号と逆転するのは，x_1 と x_2 の相関関係の強さに起因している」と巧みな説明がなされるのは，このことに由来していることがわかる．

4　層別データの回帰分析による説明

3.1 節から 3.3 節までの説明を SQC の初心者が理解することはやさしいことではない．特に，条件付き分布による説明を物理的な意味を含めて理解することは容易でない．そこで，本節では，層別された部分母集団における y と変数 x_1, x_2 の層別回帰分析の考え方を用いて説明する．

4.1　単回帰の場合

前節までは，重回帰分析における偏回帰係数と単回帰分析における回帰係数の符号逆転現象を説明することを考えてきた．本章では，母集団が s 個の層に分割されていて，それぞれの層における変数 y と変数 x_1 の回帰関係が，次式で与えられるものとする．

$$y_j^{(r)} = \beta_0 + \beta_1^{(r)} x_{j1}^{(r)} + \epsilon_j^{(r)},\ (r = 1, \ldots, s; j = 1, 2, \ldots, n_r) \tag{4.1}$$

4.1.1　層別された回帰モデル

各層における回帰係数 $\beta_1^{(r)}\ (r = 1, 2, \ldots, s)$ の最小２乗推定値は

$$\hat{\beta}_1^{(r)} = \frac{\sum_{j=1}^{n_r} \left(x_{j1}^{(r)} - \bar{x}_1^{(r)}\right)\left(y_j^{(r)} - \bar{y}^{(r)}\right)}{\sum_{j=1}^{n_r} \left(x_{j1}^{(r)} - \bar{x}_1^{(r)}\right)^2} \tag{4.2}$$

であるから，誤差に関する正規性の仮定の下で，

$$\hat{\beta}_1^{(r)} \sim N\left(\beta_1^{(r)}, \frac{\sigma^2}{S_{11}^{(r)}}\right) \tag{4.3}$$

である．したがって，仮説

$$\begin{cases} H_0 : \beta_1^{(r)} = \beta_1^{(t)} \\ H_1 : \beta_1^{(r)} \neq \beta_1^{(t)} \end{cases} \tag{4.4}$$

に対する有意水準 $100\alpha\%$ の検定は，統計量

$$t_0 = \frac{\hat{\beta}_1^{(r)} - \hat{\beta}_1^{(t)}}{\sqrt{\left(\frac{1}{S_{11}^{(r)}} + \frac{1}{S_{11}^{t)}}\right)\hat{\sigma}^2}} \tag{4.5}$$

と，棄却域

$$R : |t_0| \geq t(n - 2s, \alpha) \tag{4.6}$$

によって行われる．ただし，

$$S_e^{(r)} = \sum_{j=1}^{n_r} \left(y_j^{(r)} - \bar{y}_j^{(r)}\right)^2 \tag{4.7}$$

に対して，

$$\hat{\sigma}^2 = \frac{S_e^{(1)} + S_e^{(2)} + \cdots + S_e^{(s)}}{n - 2s} \tag{4.8}$$

とする．

4.2　誤った単回帰モデル

モデル (4.1) が成立しているとき，誤った単回帰モデル

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \ (i = 1, 2, \ldots, n) \tag{4.9}$$

を当てはめたとき，回帰推定値は

$$\hat{\beta}_1 = \frac{S_{1y}}{S_{11}} \tag{4.10}$$

によって与えられるから，モデル (4.1) の下で，

$$\begin{aligned} S_{1y} &= \sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y}) \\ &= \sum_{r,j}\left(x_{j1}^{(r)}-\bar{x}_1\right)\left(y_j^{(r)}-\bar{y}\right) \\ &= \sum_{r,j}\left(x_{j1}^{(r)}-\bar{x}_1\right)\left\{\left(\beta_1^{(r)}x_{j1}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(t)}\right)+\left(\epsilon_j^{(r)}-\bar{\epsilon}\right)\right\} \end{aligned}$$

に注意すると，誤差に対する正規性の下で，$\hat{\beta}_1$ は，次の期待値と分散をもつ正規分布に従っている．

$$\begin{cases} E\left(\hat{\beta}_1\right) = \frac{1}{S_{11}}\sum_{r,j}\left(x_{j1}^{(r)}-\bar{x}_1\right)\left(\beta_1^{(r)}x_{j1}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(r)}\right) \\ V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{11}} \end{cases} \tag{4.11}$$

したがって，(4.11) において，$\beta_1^{(1)}=\beta_1^{(2)}=\cdots=\beta_1^{(s)}=\beta_1$ のとき，$E\left(\hat{\beta}_1\right)=\beta_1$ であることに注意すると，次の命題が成立する．

命題 1　変数 x_1 と y に関するモデル (4.9) の回帰推定値 $\hat{\beta}_1$ が与えられたとき，モデル (4.1) に対する仮説

$$\begin{cases} H_0 : \beta_1^{(1)}=\beta_1^{(2)}=\cdots=\beta_1^{(s)}\left(=\beta_1^{(0)}\right) \\ H_1 : \text{少なくとも一つは等号が成立しない} \end{cases} \tag{4.12}$$

に対する有意水準 $100\alpha\%$ の検定は，統計量

$$t_0 = \frac{\hat{\beta}_1-\beta_1^{(0)}}{\sqrt{\frac{\hat{\sigma}^2}{S_{11}}}} \tag{4.13}$$

と棄却域

$$R : |t_0| \geq t(n-2,\alpha) \tag{4.14}$$

によって与えられる．ただし，$\hat{\sigma}^2$ は，(4.9) 式による残差分散

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2} \tag{4.15}$$

である．

証明　帰無仮説は，モデル (4.9) を意味していることから自明である．

4.2.1　誤った重回帰モデル

モデル (4.1) が成立しているとき，誤った重回帰モデル

$$y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\epsilon_i,\ \ (i=1,2,\ldots,n) \tag{4.16}$$

を当てはめたとすると，偏回帰係数 β_1 の推定値は

$$\hat{\beta}_1 = S^{11}S_{1y}+S^{12}S_{2y} \tag{4.17}$$

で与えられる．

ここで，

$$\begin{aligned} S_{1y} &= \sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y}) \\ &= \sum_{r=1}^{s}\sum_{j=1}^{n_r}\left(x_{j1}^{(r)}-\bar{x}_1\right)\left\{\left(\beta_1^{(r)}x_{j1}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(t)}\right)+\left(\epsilon_j^{(r)}-\bar{\epsilon}\right)\right\} \\ S_{2y} &= \sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y}) \\ &= \sum_{r=1}^{s}\sum_{j=1}^{n_r}\left(x_{j2}^{(r)}-\bar{x}_2\right)\left\{\left(\beta_1^{(r)}x_{j1}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(t)}\right)+\left(\epsilon_j^{(r)}-\bar{\epsilon}\right)\right\} \end{aligned}$$

であるから，モデル (4.1) の下で，誤差の正規性が成立するならば，$\hat{\beta}_1$ は，次の期待値と分散をもつ正規分布に従う．

$$\begin{aligned} E(\hat{\beta}_1) &= E(S^{11}S_{1y}+S_{12}S_{2y}) \\ &= S^{11}\sum_{r,j}\left(x_{j1}^{(r)}-\bar{x}_1\right)\left(\beta_1^{(r)}x_{j1}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(t)}\right) \\ &+ S^{12}\sum_{r,j}(x_{j2}-\bar{x}_2)\left(\beta_1^{(r)}x_{j1}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(t)}\right) \qquad (4.18) \\ V(\hat{\beta}_1) &= S^{11}\sigma^2 \qquad (4.19) \end{aligned}$$

この議論を，$p(>2)$ に一般化することは容易であるから，次の命題が成立する．

命題 2 偏回帰係数の推定値 $\hat{\beta}_1$ が，回帰モデル

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\epsilon_i,\quad (i=1,2,\ldots,n) \qquad (4.20)$$

の下で得られたとする．このとき，モデル (4.1) に対する仮説

$$\begin{cases} H_0:\beta_1^{(1)}=\beta_1^{(2)}=\cdots=\beta_1^{(s)}\left(=\beta_1^{(0)}\right) \\ H_1:\text{少なくとも一つは等号が成立しない} \end{cases} \qquad (4.21)$$

の有意水準 $100\alpha\%$ の検定は，統計量

$$t_0=\frac{\hat{\beta}_1-\beta_1^{(0)}}{\sqrt{S^{11}\hat{\sigma}^2}} \qquad (4.22)$$

と棄却域

$$R:|t_0|\geq t(n-p-1,\alpha) \qquad (4.23)$$

によって与えられる．ただし，$\hat{\sigma}^2$ は，(4.20) による残差分散

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p-1} \qquad (4.24)$$

である．

証明 帰無仮説の下で，$\beta_1^{(1)} = \cdots = \beta_1^{(s)} (= \beta_1)$ が成立する．このとき，(4.18) から，$E(\hat{\beta}_1) = \beta_1$ である．

ここで，(4.22)〜(4.24) で与える検定は，重回帰モデル (4.20) 式に対する仮説

$$\begin{cases} H_0 : \beta_1 = \beta_1^{(0)} \\ H_1 : \beta_1 \neq \beta_1^{(0)} \end{cases} \tag{4.25}$$

に対する有意水準 $100\alpha\%$ の検定と同一であることに注意せよ．

4.3 重回帰の場合

いま，話を簡単にするため $p = 2$ として，変数 x_1, x_2 と変数 y の間に，回帰関係

$$y_j^{(r)} = \beta_0 + \beta_1^{(r)} x_{j1}^{(r)} + \beta_2^{(r)} x_{j2}^{(r)} + \epsilon_j^{(r)}, \ (r = 1, \ldots, s; j = 1, 2, \ldots, n_r) \tag{4.26}$$

が成立しているとする．ここで，回帰切片 β_0 に対して層別の違いを仮定していないのは制約条件のように思われるが，すべての変数を，平均 0，分散 1 となるように標準化した場合，回帰切片は 0 になるため，制約にはならないことに注意せよ．

4.3.1 層別重回帰モデル

モデル (4.26) の偏回帰係数に対する推定値は

$$\sum_{r=1}^{s} \sum_{j=1}^{n_r} \left(y_j^{(r)} - \beta_0 - \beta_1^{(r)} x_{j1} - \beta_2^{(r)} x_{j2}^{(r)} \right)^2 \tag{4.27}$$

の $(\beta_1^{(r)}, \beta_2^{(r)})$ $(r = 1, 2, \ldots, s)$ に関する最小化の解として

$$\begin{cases} \hat{\beta}_1^{(r)} = S^{11(r)} S_{1y}^{(r)} + S^{12(r)} S_{2y}^{(r)} \\ \hat{\beta}_2^{(r)} = S^{21(r)} S_{1y}^{(r)} + S^{22(r)} S_{2y} \end{cases} \tag{4.28}$$

である．ただし，$j, k = 1, 2$ に対して

$$\begin{aligned} S_{jk}^{(r)} &= \sum_{l=1}^{n_r} \left(x_{lj}^{(r)} - \bar{x}_j^{(r)} \right) \left(x_{lk}^{(r)} - \bar{x}_k^{(r)} \right) \\ S_{jy}^{(r)} &= \sum_{l=1}^{n_r} \left(x_{lj}^{(r)} - \bar{x}_j^{(r)} \right) \left(y_l^{(r)} - \bar{y}^{(r)} \right) \end{aligned}$$

であって，$S^{(r)-1} = \left(S^{jk(r)} \right)$ は，行列 $S^{(r)} = \left(S_{jk}^{(r)} \right)$ の逆行列である．

ここで，

$$\begin{aligned} S_{1y}^{(r)} &= \sum_{l=1}^{n_r} \left(x_{l1}^{(r)} - \bar{x}_1^{(r)} \right) \left(y_l^{(r)} - \bar{y}^{(r)} \right) \\ &= \sum_{l=1}^{n_r} \left(x_{l1}^{(r)} - \bar{x}_1^{(r)} \right) \left\{ \beta_1^{(r)} \left(x_{l1}^{(r)} - \bar{x}_1^{(r)} \right) + \beta_2^{(r)} \left(x_{l2}^{(r)} - \bar{x}_2^{(r)} \right) + \left(\epsilon_l^{(r)} - \bar{\epsilon}^{(r)} \right) \right\} \\ &= \beta_1^{(r)} S_{11}^{(r)} + \beta_2^{(r)} S_{12}^{(r)} + \sum_{l=1}^{n_r} \left(x_{l1}^{(r)} - \bar{x}_1^{(r)} \right) \left(\epsilon_l^{(r)} - \bar{\epsilon}^{(r)} \right) \end{aligned}$$

$$
\begin{aligned}
S_{2y}^{(r)} &= \sum_{l=1}^{n_r}\left(x_{l2}^{(r)}-\bar{x}_2^{(r)}\right)\left(y_l^{(r)}-\bar{y}^{(r)}\right)\\
&= \sum_{l=1}^{n_r}\left(x_{l2}^{(r)}-\bar{x}_2^{(r)}\right)\left\{\beta_1^{(r)}\left(x_{l1}^{(r)}-\bar{x}_1^{(r)}\right)+\beta_2^{(r)}\left(x_{l2}^{(r)}-\bar{x}_2^{(r)}\right)+\left(\epsilon_l^{(r)}-\bar{\epsilon}^{(r)}\right)\right\}\\
&= \beta_1^{(r)}S_{21}^{(r)}+\beta_2^{(r)}S_{22}^{(r)}+\sum_{l=1}^{n_r}\left(x_{l2}^{(r)}-\bar{x}_2^{(r)}\right)\left(\epsilon_l^{(r)}-\bar{\epsilon}^{(r)}\right)
\end{aligned}
$$

などに注意すると，誤差に関する正規性の下で，$(\hat{\beta}_1^{(r)},\hat{\beta}_2^{(r)})$ は，次の期待値と分散・共分散をもつ 2 次元正規分布に従う．

$$
\begin{cases}
E\left(\hat{\beta}_1^{(r)}\right)=\beta_1^{(r)}\\
E\left(\hat{\beta}_2^{(r)}\right)=\beta_2^{(r)}\\
V\left(\hat{\beta}_1^{(r)}\right)=S^{11(r)}\sigma^2\\
V\left(\hat{\beta}_2^{(r)}\right)=S^{22(r)}\sigma^2\\
Cov\left(\hat{\beta}_1^{(r)},\hat{\beta}_2^{(r)}\right)=S^{12(r)}\sigma^2
\end{cases}
\tag{4.29}
$$

4.3.2 重回帰モデル

モデル (4.26) が成立しているときに，誤ったモデル

$$
y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\epsilon_i,\ \ (i=1,2,\ldots,n) \tag{4.30}
$$

を当てはめると，偏回帰係数 β_1 の最小 2 乗推定値は (4.17) で与えられる．

このとき，

$$
\begin{aligned}
S_{1y} &= \sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})\\
&= \sum_{r,j}\left(x_{j1}^{(r)}-\bar{x}_1\right)\left\{\left(\beta_1^{(r)}x_{j1}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(t)}\right)+\left(\beta_2^{(r)}x_{j2}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_2^{(t)}x_{k2}^{(t)}\right)+\left(\epsilon_j^{(r)}-\bar{\epsilon}\right)\right\}\\
S_{2y} &= \sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})\\
&= \sum_{r,j}\left(x_{j2}^{(r)}-\bar{x}_2\right)\left\{\left(\beta_1^{(r)}x_{j1}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(t)}\right)+\left(\beta_2^{(r)}x_{j2}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_2^{(t)}x_{k2}^{(t)}\right)+\left(\epsilon_j^{(r)}-\bar{\epsilon}\right)\right\}
\end{aligned}
$$

であるから，誤差に関する正規性の下で，$\hat{\beta}_1$ は，次の期待値と分散をもつ正規分布に従う．

$$
E(\hat{\beta}_1)=E(S^{11}S_{1y}+S^{12}S_{2y}) \tag{4.31}
$$

$$
\begin{aligned}
&= S^{11}\sum_{r,j}(x_{j1}^{(r)}-\bar{x}_1)(\beta_1^{(r)}x_{j1}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(t)})+S^{11}\sum_{r,j}(x_{j1}^{(r)}-\bar{x}_1)(\beta_2^{(r)}x_{j2}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_2^{(t)}x_{k2}^{(t)})\\
&+ S^{12}\sum_{r,j}(x_{j2}^{(r)}-\bar{x}_2)(\beta_1^{(r)}x_{j1}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_1^{(t)}x_{k1}^{(t)})+S^{12}\sum_{r,j}(x_{j2}^{(r)}-\bar{x}_2)(\beta_2^{(r)}x_{j2}^{(r)}-\frac{1}{n}\sum_{t,k}\beta_2^{(t)}x_{k1}^{(t)})
\end{aligned}
$$

$$
V(\hat{\beta}_1) = S^{11}\sigma^2 \tag{4.32}
$$

ここで，(4.31) において，$\beta_1^{(r)}=\beta_1,\beta_2^{(r)}=\beta_2\ \ (r=1,2,\ldots,s)$ が成立するとき，$E(\hat{\beta}_1)=\beta_1$ であることに注意すると，次の自明な命題が得られる．

命題 3 偏回帰係数の推定値 $\hat{\beta}_1$ が，回帰モデル

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \ \ (i = 1, 2, \ldots, n) \tag{4.33}$$

から得られたとする．このとき，モデル

$$y_j^{(r)} = \beta_0 + \beta_1^{(r)} x_{i1}^{(r)} + \beta_2 x_{i2}^{(r)} + \cdots + \beta_p x_{ip}^{(r)} + \epsilon_i^{(r)}, \ \ (r = 1, \ldots, s; j = 1, \ldots, n_r)$$

に対する仮説

$$\begin{cases} H_0 : \beta_1^{(1)} = \cdots = \beta_1^{(s)} = \beta_1^{(0)}, \ \ (r = 1, 2, \ldots, s) \\ H_1 : \text{少なくとも一つは等号が成立しない} \end{cases} \tag{4.34}$$

の有意水準 $100\alpha\%$ の検定は，統計量

$$t_0 = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{\sqrt{S^{11}\hat{\sigma}^2}} \tag{4.35}$$

と棄却域

$$R : |t_0| \geq t(n - p - 1, \alpha) \tag{4.36}$$

によって与えられる．ただし，$\hat{\sigma}^2$ は，(4.33) による残差分散

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p - 1} \tag{4.37}$$

である．

ここで，(4.35)〜(4.37) は，モデル (4.33) に対する仮説

$$\begin{cases} H_0 : \beta_1 = \beta_1^{(0)} \\ H_1 : \beta_1 \neq \beta_1^{(0)} \end{cases} \tag{4.38}$$

に対する有意水準 $100\alpha\%$ の検定を与えていることに注意せよ．

5　2つの数値例

例 1 第 1 章の事例は，原料を 2 つの生産地から調達していることが判明したため，表 4 のように再整理することができた.

表 4 データ

調達先 P_1				調達先 P_2			
No.	副原料 x_1	酸度 x_2	収率 y	No.	副原料 x_1	酸度 x_2	収率 y
1	37	51	73.5	16	42	56	74.7
2	35	48	71.4	17	40	52	71.5
3	38	55	73.7	18	39	46	69.3
4	35	48	70.2	19	38	47	69.6
5	35	44	70.1	20	38	49	69.4
6	37	47	70.8	21	42	57	73.5
7	35	47	70.4	22	39	52	70.9
8	36	48	71	23	40	50	71.3
9	35	48	71.4	24	40	51	72.7
10	36	49	71.6	25	41	53	72.8
11	34	47	69.9	26	39	47	68.5
12	38	50	71.4	27	38	49	69.8
13	37	48	71.8	28	42	61	76.6
14	39	53	73.5	29	40	55	73.4
15	37	47	70.3	30	39	54	71.5

このとき，調達先 P_1 と P_2 における $n_1 = n_2 = 15$ 個のデータから，収率 y を副原料の投入量 x_1 と酸度 x_2 によって回帰すると

$$\begin{cases} P_1 : & \hat{y} = 48.35 + 0.147x_1 + 0.364x_2 \\ P_2 : & \hat{y} = 31.34 + 0.552x_1 + 0.354x_2 \end{cases}$$

であって，それぞれの回帰による寄与率は $R^2_{P_1} = 0.809$, $R^2_{P_2} = 0.921$ である.

ただし，この場合の誤差分散は $\hat{\sigma}^2 = 0.381$ と推定されるため，仮説

$$\begin{cases} H_0 : \beta_1^{(1)} = \beta_1^{(2)} \\ H_1 : \beta_1^{(1)} \neq \beta_1^{(2)} \end{cases} \tag{5.1}$$

は，検定統計量の値 $|t_0| = 1.578$ と有意水準 5% の棄却限界 $t(26, 0.05) = 2.056$ を比較して，有意水準 5% で有意ではないが，棄却限界 $t(26, 0.25) = 1.177$ と比較すると有意水準 25% では有意である．すなわち，$\beta_1^{(1)} \neq \beta_1^{(2)}$ でないとはいえない.

例 2 表 5 は，某社における 3 台の設備 A_1, A_2, A_3 を用いて生産している自動車用部品の製造条件 x_1, x_2 と製品の圧縮強度 $y(kg \cdot f/mm^2)$ のデータを調べたものである.

表 5 数値例

設備 A_1				設備 A_2				設備 A_3			
No.	x_1	x_2	y	No.	x_1	x_2	y	No.	x_1	x_2	y
1	59	57	134	11	55	43	171	21	47	51	168
2	51	56	141	12	52	50	178	22	54	57	169
3	59	54	140	13	61	54	182	23	52	46	156
4	64	61	142	14	62	62	185	24	60	52	166
5	58	70	148	15	56	59	193	25	48	44	163
6	36	36	132	16	43	47	178	26	45	47	166
7	56	59	149	17	50	52	177	27	45	47	164
8	69	69	148	18	57	54	176	28	57	54	169
9	55	47	129	19	41	38	156	29	44	37	140
10	47	38	125	20	58	61	196	30	68	67	182

表 5 対する分散共分散と相関係数は

表 5 分散共分散と相関係数

変数	x_1	x_2	y
x_1	202.107	0.828	0.202
x_2	193.700	270.7000	0.293
y	101.270	169.789	1241.041

である[2]．したがって，変数 x_1 と変数 y との関係は正比例の関係であると期待されるが，変数 y を変数 x_1 と変数 x_2 の上に回帰すると，

$$\hat{y} = 188.41 - 0.318x_1 + 0.855x_2$$

となる．すなわち，変数 x_1 と変数 y との関係が，相関係数で期待したものと逆転している．

この現象に対して，x_2 を与えたときの x_1 と y の条件付き分散共分散

表 6 条件付き分散共分散

変数	x_1	y
x_1	63.505	-5.225
y	-5.225	1134.545

から計算した，変数 x_1 と変数 y の偏相関係数が

$$r_{1y\cdot 2} = \frac{-5.225}{\sqrt{63.505 \times 1134.545}} = -0.0195$$

となる．また，変数間の相関係数には

$$r_{1y} - r_{12}r_{2y} = 0.202 - 0.828 \times 0.293 = -0.041$$

の関係がある．これらの事実から偏相関係数の推定値 $\hat{\beta}_1$ と単回帰による x_1 の回帰係数の符号が逆転するという説明を与えることができ，これが従来の方法である．

[2]対角線の上側が相関係数である．

これに対して，各設備における変数 y の変数 x_2 への単回帰分析を行うと，

$$\begin{cases} \hat{y}^{(1)} = 138.80 + 0.633(x_1 - 54.7) \\ \hat{y}^{(2)} = 179.20 + 1.308(x_1 - 52.0) \\ \hat{y}^{(3)} = 250.60 + 4.308(x_1 - 50.2)) \end{cases}$$

となることがわかる．

ここで，多重比較になることを考慮しないで，それぞれの回帰係数が異なるかどうかを検定すると

$$\begin{cases} H_0 : \beta_1^{(1)} = \beta_1^{(2)} \\ H_1 : \beta_1^{(1)} \neq \beta_1^{(2)} \end{cases}$$

に対する検定統計量の値は $|t_0^{(12)}| = 2.838 > t(18, 0.05) = 2.445$ となって，有意水準 5% で回帰係数の異なることが示される．同様に，設備 A_2 と設備 A_3，設備 A_3 と設備 A_1 に対する検定統計量を求めると $|t_0^{(23)}| = 2.722, |t_0^{(31)}| = 2.736$ となって，それぞれの回帰係数は有意水準 5% で異なっていると判定できる．

一方，$\hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)}$ の平均を β_1 の帰無仮説の値とした仮説

$$\begin{cases} H_0 : \beta^{(1)} = \beta_1^{(2)} = \beta_1^{(3)} (= 2.038) \\ H_1 : \text{少なくとも一つは等号が成立しない} \end{cases} \tag{5.2}$$

は，検定統計量の値

$$t_0 = \frac{-0.318 - (2.038)}{\sqrt{0.00175 \times 376.035}} = -2.954$$

と 5% 棄却限界 $t(27, 0.05) = 2.052$ から，有意水準 5% で有意であると判定できる．

なお，層別モデルに基づく回帰係数の推定値 $\hat{\beta}_1^{(r)}, (r = 1, 2, 3)$ に対して，

$$F_0 = \frac{\sum_r \left(\hat{\beta}_1^{(r)} - \beta_1^{(0)}\right)^2 S_{11}^{(r)}}{S_e} \cdot \frac{n-6}{3} \tag{5.3}$$

は，(5.2) の帰無仮説の下で，自由度 $(3, n-6)$ の F 分布に従う．このことを利用して，検定統計量を計算すると，

$$F_0 = \frac{(0.633 - 2.038)^2 \times 758.40 + (1.308 - 2.038)^2 \times 450.50 + (4.308 - 2.038)^2 \times 552.00}{1524.387} \times \frac{30-6}{3} = 27.354$$

であるから，$F(3, 24; 0.05) = 3.01$ と比較して，(5.2) の帰無仮説は有意水準 5% で棄却される．

ここでの議論は，「偏回帰係数の符号が単回帰係数の符号と異なっていれば，層別解析における回帰係数（因果関係）が層ごとに異なっている」と命題化できるものではない．しかし，上記のような仮説検定を行うことで，層別解析を行う必要性があることを明確にできるという意味で，技術者に対して有意な情報を与えていることがわかる．

6　まとめ

回帰分析において目的変数 y と説明変数 $x_j (j = 1, 2, \ldots, p)$ の回帰係数の推定値と相関係数の符号が逆転することに対して，第３変数による影響，条件付き分布と偏相関係数および層別解析の考え方を用いて検討してきた．はじめの３つは，多くの多変量統計解析において偏回帰係数の符号逆転現象を説明する方法として与えられているものであるが，初学者が理解することは容易で

はない．一方，当該変数（ここでは，変数 x_2）と変数 y の層別された単回帰分析における回帰係数の値が異なることを理解することは難しいことではない．

むしろ，変数間の相関関係または単回帰分析における回帰係数がプラス（又はマイナス）であると考えられるときに，重回帰分析から得られる偏回帰係数の符号がマイナス（又はプラス）の値をとるとき，「層別した回帰分析における回帰係数が一様でないかどうか」を検定することの意義は大きいと考えられる．もし実務的に層別解析の必要性が明らかになれば，それに対応した実務的な処理を行うことで重要な情報を得ることが期待できる．

謝辞 本原稿の審査過程において，文章の精査をいただき，貴重なコメントをいただいた審査員に対して感謝申し上げます．

参考文献

[1] 猪原正守，飯塚悦功，岩崎日出男: 『回帰分析』，（一般財団法人）日本科学技術連盟，品質管理ベーシックコーステキスト，第 10 章，2014 年．

[2] 永田靖，棟近雅彦：『多変量解析入門』，サイエンス社，2014 年．

[3] 荒木孝治：『「R」と R コマンダーではじまる多変量解析』，日科技連出版社，2007 年．