

第12回：回帰分析

尚 晋
大学院経済学研究科 助教

2025年6月24日

今日のポイント

1. 回帰分析と回帰係数の推定
2. 偏回帰係数の統計的推測
3. 重回帰分析

1 回帰分析	1	1.2.2 回帰方程式の当てはまりと決定係数 η^2	5
1.1 回帰分析	2	1.3 偏回帰係数の統計的推測	7
1.2 回帰係数の推定	3	1.3.1 偏回帰係数の標本分布	7
1.2.1 最小二乗法による回帰係数の推定	3	1.3.2 偏回帰係数の検定	8
		1.4 重回帰分析	10

1 回帰分析

統計的方法の中で最も広く応用されているのは回帰分析である。その有用性、理論的根拠、使い易さからして、これをおいて統計学は語れない。推定も統計的検定もここに使われている。現在は、コンピューター・パッケージの普及も目ざましく、計算結果を読むためにも、回帰分析の基本知識は欠かせない。

- 回帰分析(regression analysis)は、2変数 X, Y のデータがあるとき、その関係を定量的に表現する回帰方程式(regression equation)を導出することを主な目的とする統計手法である。
2次元データの節で記述したと違って、「回帰」を「母集団」と「標本」という考え方の中に置くことにより、回帰にも統計的推測の方法を導入することが目的である。
- 説明される変数を Y で表し、これを従属変数、被説明変数、内生変数などと呼ぶ。また、説明する変数を X で表し、独立変数、説明変数、外生変数などと呼ぶ。
- 回帰分析の目的は、 X と Y との定量的な関係の構造(モデル model ということがある)を求めることがあるが、注意すべき点は、 Y を X によって説明(explain)しようすることに主眼を置いている。この点において、変数間の関係の有無を調べることに焦点を当てる相関分析とは本質的に異なる。

例：表13.1のように、東京の気圧と、前日の福岡における日平均海面気圧のデータが集計できたとする。天気は西から東へ変化するから、東京の気圧を予想する上で、前日の福岡の気圧は重要なデータであると考えられる。この間の関係を分析することは、東京の天気予報を行う上で役に立つと考えられる。

表13.1：東京および前日の福岡の日平均海面気圧(mb)

番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
月日	5/7	5/8	5/9	5/10	5/11	5/12	5/13	5/14	5/15	5/16	5/17	5/18	5/19	5/20	5/21	5/22	5/23	5/24	5/25	5/26
東京の 気圧	1019.4	1005.7	1002.0	1006.7	1005.1	1010.1	1016.7	1011.0	999.5	1006.9	1001.9	1007.5	1014.4	1014.3	1014.6	1009.0	1006.7	1009.4	1011.8	1009.4
前日の 福岡の気圧	1018.4	1007.6	1006.2	1009.9	1010.8	1013.2	1016.2	1009.1	1008.1	1012.5	1006.4	1006.3	1012.2	1015.0	1017.4	1016.5	1012.1	1008.7	1009.2	1009.2

(出典：中村繁・北村幸房『気象データマニュアル』理科年表読本，丸善，1987)

1.1 回帰分析

集計したデータにおいて、福岡の前日の気圧を説明変数 X 、東京の気圧を被説明変数 Y とすると、図13.1(福岡の前日の気圧と東京の気圧の散布図(相関係数:0.803))のような散布図が得られる。この図から、東京の気圧について次の2点がわかる：

1. 福岡の前日の気圧が増加するに従って、東京の気圧も増加する傾向がある。
2. 福岡の前日の気圧が同一であっても、ばらつきが存在する。

したがって、東京の気圧を、福岡の前日の気圧によって変化する部分と、それ以外のばらつきによる部分に分けて表すことができる。福岡の気圧によって変化する部分を y とし、 x の線形関数として、

$$y = \beta_1 + \beta_2 x \quad (1)$$

とする。

この式は y の x 上への回帰方程式(regression equation)あるいは回帰関数(regression function)と呼ばれ、 y が x の線形関数である場合を線形回帰(linear regression)，それ以外を非線形回帰(non-linear regression)と呼ぶ。ここでは線形回帰のみを扱うが、関数変換等によって非線形モデルであっても線形モデルに変換可能な場合や、線形モデルで近似できるものも多い。

例：弹性モデル 関係式が $z = aw^b$ であるとする。このモデルは弹性モデルと呼ばれ、価格の変化に対する消費量や生産量の変化を分析するのにしばしば用いられる。両辺の対数をとると、

$$\log z = \log a + b \log w$$

と表せるので、 $y = \log z$ 、 $x = \log w$ 、 $\beta_1 = \log a$ 、 $\beta_2 = b$ とすれば、式(1)と同じ形の線形回帰方程式が得られる。

例：指数回帰 関係式が $z = ab^x$ であるとする。このモデルは指数回帰と呼ばれ、経済成長や細菌の増殖など、指数関数的に増加する現象の分析に用いられる。両辺の対数をとると、

$$\log z = \log a + x \log b$$

であるから、 $y = \log z$ 、 x はそのまま、 $\beta_1 = \log a$ 、 $\beta_2 = \log b$ とすれば、やはり線形の回帰方程式が得られる。

母回帰方程式 i 番目の東京の気圧を Y_i 、前日の福岡の気圧を x_i 、ばらつきの部分を ε_i とすると、母集団において次のように表せる：

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (2)$$

このモデルは母回帰方程式(population regression equation)と呼ばれ、 β_1, β_2 を母(偏)回帰係数(population (partial) regression coefficient)(正確には、「母集団偏

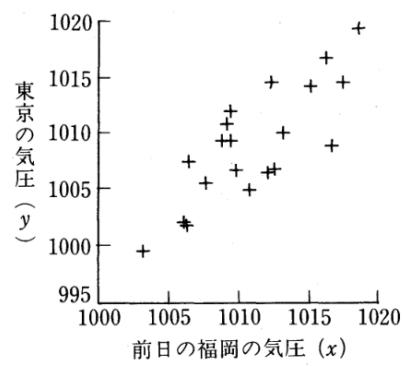


図13.1(福岡の前日の気圧と東京の気圧の散布図(相関係数:0.803))

「回帰係数」という。「偏」は略してもよい)と呼ぶ。これは母集団の値であり、一般には未知である。これについて推定・検定を行うのが回帰分析である。

ここで、 x_i は確率変数ではなく、すでに観測された確定した値である。また、 ε_i は誤差項(error term)または擾乱項(disturbance term)と呼ばれ、以下の3条件を満たす確率変数である：

- (a) 期待値は 0: $E(\varepsilon_i) = 0 \quad (i = 1, 2, \dots, n)$
- (b) 分散は一定で $\sigma^2: V(\varepsilon_i) = \sigma^2 \quad (i = 1, 2, \dots, n)$
- (c) 異なった誤差項は無相関: $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \quad (i \neq j)$

このことから、式(2)のモデルは次のような形で期待値を持つ：

$$E(Y_i) = \beta_1 + \beta_2 X_i \quad (i = 1, 2, \dots, n) \quad (3)$$

すなわち、定まった X_i に対応して、変数 Y_i は誤差項(ε_i)を含んだ確率変数であり、その確率変数の取りうる値の期待値は $\beta_1 + \beta_2 X_i$ となる。

1.2 回帰係数の推定

1.2.1 最小二乗法による回帰係数の推定

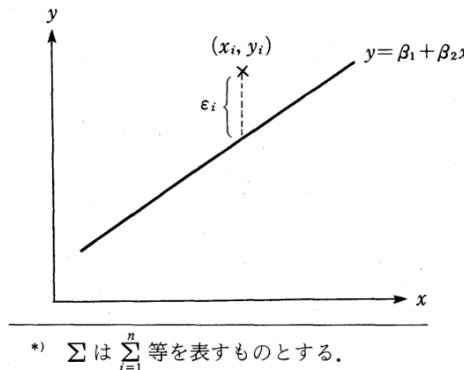


図 13.2 最小二乗法の考え方

なるべく誤差 ε_i が小さくなるように、 β_1, β_2 を決めて直線をあてはめたいが、そのためには $\sum \varepsilon_i^2$ を最小にする基準をとる。一例としては、図 13.4 を見よ。

式(2)の回帰式において、母回帰係数 β_1, β_2 の推定を考える。

いま図13.2のような回帰式を考えた場合、 X_i によって説明できない誤差項は、

$$\varepsilon_i = Y_i - (\beta_1 + \beta_2 X_i), \quad i = 1, 2, \dots, n$$

である。 ε_i の符号の影響を除くために、次のように平方和を定義する：

$$S = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - (\beta_1 + \beta_2 X_i))^2$$

この S は、 Y_i が X_i で説明できない部分の総和を表すから、できるだけ小さい方が望ましいとされる。 S を最小にする β_1, β_2 の推定量を求める方法を最小二乗法(method of least squares)と呼び、それによって得られる推定量を最小二乗推定量(least squares estimator)という。

S を最小にするには、 β_1, β_2 に関して一次の偏微分してゼロとおく：

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n (Y_i - \beta_1 - \beta_2 X_i) = 0 \quad (4a)$$

$$\frac{\partial S}{\partial \beta_2} = -2 \sum_{i=1}^n (Y_i - \beta_1 - \beta_2 X_i) X_i = 0 \quad (4b)$$

これを整理すると,

$$\begin{cases} n\beta_1 + \beta_2 \sum X_i = \sum Y_i \\ \beta_1 \sum X_i + \beta_2 \sum X_i^2 = \sum X_i Y_i \end{cases}$$

を得る。これを正規方程式(normal equations)と呼ぶ。

この連立方程式を解くと、 \bar{X}, \bar{Y} をそれぞれ X と Y の標本平均として、

$$\begin{cases} \hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \\ \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \end{cases} \quad (5)$$

この推定量 $\hat{\beta}_1, \hat{\beta}_2$ を標本(偏)回帰係数と呼ばれる。次の回帰式

$$Y = \hat{\beta}_1 + \hat{\beta}_2 X \quad (6)$$

を標本回帰方程式(sample regression equation)または標本回帰直線(sample regression line)と呼ぶ。 $\hat{\beta}_1, \hat{\beta}_2$ は、それぞれ、その傾き、 y 切片という意味をもっている。なお、

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

は $E(Y_i)$ の推定量となるが、これを回帰値 regressed value ということもある。

例えば、福岡と東京の気圧データにおいて :

$$\begin{aligned} \sum(X_i - \bar{X})^2 &= 335.6 \\ \sum(X_i - \bar{X})(Y_i - \bar{Y}) &= 329.7 \end{aligned}$$

よって、

$$\begin{aligned} \hat{\beta}_2 &= \frac{329.7}{335.6} = 0.9822 \\ \hat{\beta}_1 &= 1009.1 - 0.9822 \times 1011.0 = 16.09 \end{aligned}$$

したがって、標本回帰方程式は $Y = 16.09 + 0.9822X$

となる(図13.3)。このことから、福岡の前日の気圧の lmb の変化は、翌日の東京の気圧に 2% 程度減殺された変動をもたらすことがわかる。

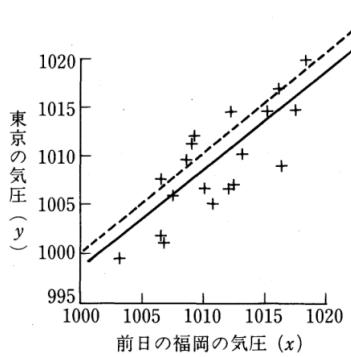


図 13.3 東京の気圧を前日の福岡の気圧から予測する標本回帰方程式 $y = 16.09 + 0.9822 x$ 。点線は、両者は等しいと考えた場合の予測式 $y = x$ で、傾き = 1 である。回帰方程式の方がごくわずか傾きがゆるやかである。なお、 x のこの範囲では $y = x$ の方が上側にあることに注意する。

ここで、実測値 Y_i の、回帰方程式に定められた回帰値 \hat{Y}_i からのずれは、

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$$

は X で説明されずに残った分であり、 \hat{e}_i は回帰残差(residual)と呼ばれる。

\hat{e}_i は誤差項 ε_i の推定量であるが、 $\hat{\beta}_1, \hat{\beta}_2$ は式(4) から求められるので：

$$\sum \hat{e}_i = 0, \quad \sum \hat{e}_i X_i = 0 \quad (7)$$

となる(回帰残差の和と、回帰残差と説明変数の積和は 0 になる)。 \hat{e}_i の平均値は 0 であり、また、 \hat{e}_i と X_i とはベクトルとして直交していることを意味している。これは、最小二乗法それ

自体の特徴であり、母集団にかかわらず常に成り立つ重要な性質である。

誤差項 ε_i の分散 σ^2 は回帰方程式のあてはまりの良さを表すが、その推定量は

$$s^2 = \frac{\sum \hat{e}_i^2}{n - 2} \quad (8)$$

となる。回帰残差の平方和を $n - 2$ で割るのは、2つのパラメータ $\hat{\beta}_1, \hat{\beta}_2$ を推定したため自由度が2減るからである(もう一つの解釈： e_i は式(7)の二つの条件を満たすべきため制限が加わり、自由度が2失われているためである)。

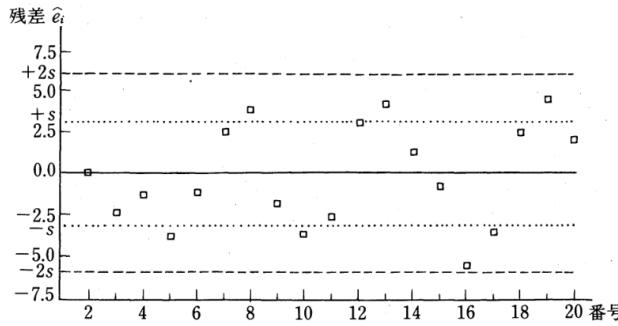


図 13.4 福岡の前日の気圧から東京の気圧を予測する回帰方程式の回帰残差 \hat{e}_i のプロット。
残差 $\hat{e}_i = Y_i - \hat{Y}_i$ のプロットは各データの個々の傾向を点検するときに用いられ、また残差の全般的な傾向も一目で観察される。ここでも、 \hat{e}_i にはある周期性が表れている。また、すべての e_i が $\pm 2s (s=3.148)$ の中に入っているので、予測の誤差は小さいといえるであろう(本図はTSPの出力による)。

気圧の例で、回帰残差は、

$$\hat{e}_i = Y_i - 16.09 - 0.9822X_i$$

で計算されるから、これをプロットすると、図13.4のようになる。また、残差平方和 $\sum \hat{e}_i^2 = 178.40$ 、自由度18なので、

$$s^2 = \frac{178.40}{18} = 9.911, \quad s = \sqrt{9.911} \approx 3.148$$

よって、回帰値のあてはまりの良し悪しは、この s の値を基礎として判断される。この s を推定値の標準誤差(standard error of estimates)といい、 $s.e$ で表す。この値が小さいほど回帰式はよく適合していると考えられる。

\hat{e}_i の多くは $\pm 2s$ の範囲に入り、誤差が正規分布に従う場合には約95%の観測値がこの範囲に入る。 $|\hat{e}_i/s|$ が非常に大きいものは外れ値の可能性が高い。従って、外れ値を含むようなデータの分析には、あらかじめその可能性を除くなどの注意が必要である。

なお、例えば東京の気圧はそのまま mb(ミリバール)で、福岡の気圧を水銀柱 mmHg の単位で表したとする時、単位を変更しただけであるから、2変数間の関係は変わらないが、 $\hat{\beta}_2$ の推定値は大きく異なる。

このように、説明変数 X_i を別の単位に変更した場合、回帰係数の値は変わるが、2変数間の関係性は変わらない。この影響を除くため、変数を標準化して回帰を行うことが多い。このときの偏回帰係数を標準化(偏)回帰係数 standardized (partial) regression coefficientという。この場合、切片の推定量はつねに0にな、傾きの推定量は相関係数に一致する。

1.2.2 回帰方程式の当てはまりと決定係数 η^2

回帰方程式(2)がどの程度よく当てはまっているか、すなわち、 X がどの程度よく Y を説明しているかは、モデルの妥当性・有効性を考える上で重要である。

X が Y のばらつきをほとんど説明できないのであれば、この回帰式はあまり価値がないといえる。逆に、 X が Y のばらつきの多くを説明するのであれば、この回帰式の価値は高いといえる。

モデルの当てはまりの良さを測る基準として一般に使われるが、決定係数 η^2 である。 Y_i のばらつきの総和変動(全変動)は

$$\sum(Y_i - \bar{Y})^2$$

であり、これは回帰式で説明できる変動と説明できない変動の二つに分けられる。

回帰式(2)によって説明できる変動は

$$\sum(\hat{Y}_i - \bar{Y})^2$$

残差二乗和(残差変動)は回帰によって説明できない変動

$$\sum \hat{e}_i^2 = \sum(Y_i - \hat{Y}_i)^2$$

である。従って次が成り立つ：

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

決定係数 η^2 は、 Y_i の変動のうち、回帰式によって説明できる割合を示すものであり、次式で定義される：

$$\eta^2 = 1 - \frac{\sum \hat{e}_i^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \quad (9)$$

η^2 は $0 \leq \eta^2 \leq 1$ の範囲にあり、 $\eta^2 = 1$ はすべての点が回帰直線上にあり完全に説明できている場合、 $\eta^2 = 0$ は全く説明できていない場合である(図13.5参照)。

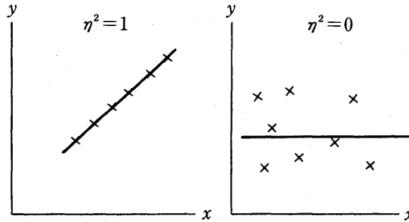


図 13.5 決定係数の両極端

左の $\eta^2=1$ の場合は、 x は完全に y を説明する。このような場合は一見して関係があることが明らかである。右の $\eta^2=0$ の場合は、標本回帰方程式は説明の意味をもたず、回帰分析は全く効果がない。現実は、両者の中間にある。また、 $\eta^2=(\text{相関係数})^2$ となることが証明できる。3.4 節参照。

線形回帰式 (2) において、決定係数は標本相関係数 r を用いて以下のようにも表される：

$$\eta^2 = r^2$$

偏回帰係数は標本から最小二乗法によって求められるが、式(2)のような直線だけでなく他の曲線も回帰として考えられる。しかし、なぜ最小二乗法による直線が採用されるのか、その優位性を見てみよう。

まず、 $\hat{\beta}_1, \hat{\beta}_2$ は以下の意味で優れている：

- 不偏性：

$$E(\hat{\beta}_1) = \beta_1, \quad E(\hat{\beta}_2) = \beta_2 \quad (10)$$

- 分散の最小性：

$$\text{Var}(\hat{\beta}_2) = E(\hat{\beta}_2 - \beta_2)^2 = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \quad (11a)$$

$$\text{Var}(\hat{\beta}_1) = E(\hat{\beta}_1 - \beta_1)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2} \right) = \frac{\sigma^2 \sum X_i^2}{n \sum(X_i - \bar{X})^2} \quad (11b)$$

- 最良線形不偏推定量(BLUE)：他の線形で不偏な推定量 $\hat{\beta}'_1 = \sum c_{1i} Y_i$, $\hat{\beta}'_2 = \sum c_{2i} Y_i$ に対して、この時必ず

$$\text{Var}(\hat{\beta}_1) \leq \text{Var}(\hat{\beta}'_1), \quad \text{Var}(\hat{\beta}_2) \leq \text{Var}(\hat{\beta}'_2)$$

を満たす。最小二乗推定量 $\hat{\beta}_1, \hat{\beta}_2$ が最良線形不偏推定量であることをガウス・マルコフの定理という。

したがって、 $\hat{\beta}_1, \hat{\beta}_2$ は線形で不偏な推定量の中で分散が最小となる最良推定量であることがわかる。

このような推定量は、**最良線型不偏推定量 (BLUE: Best Linear Unbiased Estimator)** と呼ばれる。

1.3 偏回帰係数の統計的推測

回帰分析の目的は、標本偏回帰係数 $\hat{\beta}_1, \hat{\beta}_2$ を推定するだけではない。この値をもとに、母偏回帰係数 β_1, β_2 についての検定を行うことも目的の一つである。

たとえば、東京と福岡の気圧データにおいて、両者は定数しか違わないという仮説(すなわち $Y = X + c$)を考えられる。このとき $\beta_1 = 1$ という仮説が問題となりうる。実際、標本からは $\hat{\beta}_1 = 0.9822$ が得られており、この仮説検定を考えることは意味がある。

また、体重 Y と身長 X のデータから「標準体重」の式 $Y = 0.9(X - 100)$ のあてはまりの良否を確認することもできる。

このように、 $\hat{\beta}_1, \hat{\beta}_2$ の値をもとに、母集団の母偏回帰係数についてのさまざまな仮説を検定する方法を以下に述べる。そのためには、まず $\hat{\beta}_1, \hat{\beta}_2$ の標本分布について理解する必要がある。

1.3.1 偏回帰係数の標本分布

これらの標本分布を導出するために、今までの式(2)に対する仮定(a),(b),(c)に加えて、誤差項 $\varepsilon_1, \dots, \varepsilon_n$ は独立で、平均0、分散 σ^2 の共通正規分布 $\mathcal{N}(0, \sigma^2)$ に従う。標本偏回帰係数 $\hat{\beta}_2$ は、正規分布に従っている ε_i の線形関数であるから、正規標本論で述べたように、その標本分布はふたたび正規分布となる。

このとき、 $\hat{\beta}_2$ は正規分布に従う：

$$\hat{\beta}_2 \sim \mathcal{N}\left(\beta_2, \frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right) \quad (12)$$

ただし、 σ^2 は未知であるため、回帰残差

$$\hat{e}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

を用いて、式(8)のごとく $s^2 = \frac{\sum \hat{e}_i^2}{n-2}$ のように標本から推定する。また推定値の標準誤差 $s.e.$ は以下となる：

$$s.e. = \sqrt{s^2} = \sqrt{\frac{\sum \hat{e}_i^2}{n-2}} \quad (13)$$

回帰係数 $\hat{\beta}_2$ の標準誤差 (standard error) は式(11b)を用いて、次のように与えられる：

$$s.e.(\hat{\beta}_2) = \frac{s}{\sqrt{\sum(X_i - \bar{X})^2}} \quad (14)$$

このとき標準化した $(\hat{\beta}_2 - \beta_2)/\sqrt{\sigma^2 / \sum(X_i - \bar{X})^2}$ は標準正規分布に従うが、 $\sqrt{\cdot}$ 内で σ^2 を標本からの s^2 で置き換えて、検定統計量は

$$t = \frac{\hat{\beta}_2 - \beta_2}{s.e.(\hat{\beta}_2)} \quad (15)$$

であり、 $t(n-2)$ 分布 (自由度 $n-2$) に従う。

例：気圧の関連

東京と福岡の気圧データにおける推定値より、

$$s.e. = 3.148, \quad s.e.(\hat{\beta}_2) = \frac{3.148}{\sqrt{336.6}} = 0.1718$$

である。

また、あまり用いないが、 $\hat{\beta}_1$ の標本分布は：

$$\hat{\beta}_1 \sim \mathcal{N} \left(\beta_1, \frac{\sigma^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2} \right)$$

$\hat{\beta}_1$ の標準誤差は：

$$\text{s.e.}(\hat{\beta}_1) = s \cdot \sqrt{\frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}}$$

そして、 β_2 に関する検定統計量は：

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{s.e.}(\hat{\beta}_1)} \quad (16)$$

も自由度 $n - 2$ の t 分布に従う。

1.3.2 偏回帰係数の検定

標本回帰方程式において、 X が Y をどのように説明しているかは、その傾き $\hat{\beta}_2$ によって表される。この傾きの有意性について検定を行う。

帰無仮説を次のようにおく：

$$H_0 : \beta_2 = a \quad (a \text{は指定された定数})$$

対立仮説は次のいずれかである：

$$H_1 : \beta_2 \neq a \quad (\text{両側検定}), \quad \text{または} \quad H_1 : \beta_2 > a \quad (\text{片側検定})$$

両側検定と片側検定のどちらを採用するかは、仮説検定で述べたように検定の目的によって決まる。

標本から得られた回帰係数 $\hat{\beta}_1$ とその標準誤差 $\text{s.e.}(\hat{\beta}_1)$ を用いて、以下のような検定統計量を計算する：

$$t_2 = \frac{\hat{\beta}_2 - a}{\text{s.e.}(\hat{\beta}_2)} \quad (17)$$

次に、自由度 $n - 2$ の t 分布 $t(n - 2)$ におけるパーセント点 $t_{\alpha/2}(n - 2)$ (両側検定)または $t_{\alpha}(n - 2)$ (片側検定)を付表2より調べる。

- 両側検定では、 $|t_2| > t_{\alpha/2}(n - 2)$ のとき、 H_0 を棄却する。
- 片側検定では、 $t_2 > t_{\alpha}(n - 2)$ のとき、 H_0 を棄却する。

Y の X 上への回帰方程式は、 X で Y を説明するための分析方法である。したがって、 X によって Y を説明できるかどうかの検定、すなわち

$$H_0 : \beta_2 = 0$$

の検定がとくに重要である。この帰無仮説が採択されるとき、 X によって Y を説明できないことを意味し、回帰式が適切でないと考えられる。

例：気圧の関連の有無

- 表13.1の東京と福岡の気圧のデータを用いて、福岡の気圧が東京の気圧を説明するかどうかを検定する。

帰無仮説は $H_0 : \beta_2 = 0$ 、対立仮説は $H_1 : \beta_2 > 0$ とする(正の相関を仮定)。

標本値より：

$$\hat{\beta}_2 = 0.9822, \quad \text{s.e.}(\hat{\beta}_2) = 0.1718$$

検定統計量は：

$$t_2 = \frac{0.9822}{0.1718} = 5.717$$

標本サイズは $n = 20$, よって自由度は $n - 2 = 18$. 有意水準 $\alpha = 0.01$ のとき

$$t_{0.01}(18) = 2.552$$

であり,

$$t = 5.717 > 2.552$$

したがって, 有意水準1%で帰無仮説 H_0 は棄却される. ゆえに, 回帰方程式は有意であり, 福岡の気圧は東京の気圧を有意に説明しているといえる.

- また, $\hat{\beta}_2 = 0.9822$ であることから, 母集団において $\beta_1 = 1$ ではないかという仮説も考えられる.

この仮説 $H_0 : \beta_2 = 1$ に対して, 対立仮説 $H_1 : \beta_2 \neq 1$ により両側検定を行う.

$$t = \frac{0.9822 - 1.0}{0.1718} = -0.1036$$

一方,

$$t_{0.005}(18) = 2.878$$

であるから,

$$|t| = 0.1036 < 2.878$$

したがって, 有意水準1%で $\beta_2 = 1$ であるという帰無仮説は棄却されない.

- なお, ここで行った $H_0 : \beta_2 = 0$ の有意性検定において計算された検定統計量

$$t_2 = \frac{\hat{\beta}_2}{s.e.(\hat{\beta}_2)}$$

は, t値 (t-ratio) と呼ばれることが多い.

以下は, TSP で表 13.1 のデータを使って回帰方程式を推定した例である.

TSP コマンド

```
1 ? SMPL 1 20;
2 ? LOAD(FILE="A:INPUT.DAT") X Y;
3 ? OLSQ Y C X;
```

推定結果

```
EQUATION 1
*****
METHOD OF ESTIMATION = ORDINARY LEAST SQUARES

DEPENDENT VARIABLE: Y

SUM OF SQUARED RESIDUALS = 178.403      1)
STANDARD ERROR OF THE REGRESSION = 3.14822    2)
MEAN OF DEPENDENT VARIABLE = 1009.11       3)
STANDARD DEVIATION = 5.14121      4)
R-SQUARED = 0.644763     5)
ADJUSTED R-SQUARED = 0.625028
DURBIN-WATSON STATISTIC = 1.1113
F-STATISTIC( 1, 18) = 32.6704      6)
LOG OF LIKELIHOOD FUNCTION = -50.2619
NUMBER OF OBSERVATIONS = 20        7)

VARIABLE      ESTIMATED COEFFICIENT 8)      STANDARD ERROR 9)      T-STATISTIC 10)
C             16.089          173.73          0.92609E-01
X             0.98221         0.17184         5.7158
```

- 1) 回帰残差の平方和 $S_r^2 = \sum e_i^2$ 2) 推定値(回帰)の標準誤差 $s.e.$ 3) 従属変数の標本平均 \bar{Y} 4) 同標本標準偏差 s_y 5) 決定係数 = (相関係数) 2 η^2 6) 回帰式の F 値 F (後述. $k=2, p=1$ の場合) 7) 標本の大きさ(観測値の個数) n 8) 標本偏回帰係数 $\hat{\beta}_i$ 9) 同標準誤差 $s.e.(\hat{\beta}_i)$ 10) $\beta_i = 0$ を検定する t 値 t_i

1.4 重回帰分析

これまででは、説明変数がただ一つである単回帰分析(単純回帰分析, simple regression analysis)を扱ってきた。しかし、多くの場合、複数の変数が被説明変数に影響することが考えられる。

たとえば、被説明変数 Y_i を消費とすれば、その説明変数としては収入、財産の保有高、性別、家族の人数など、いくつかの要因が考えられる。このように二つ以上の説明変数を考慮する回帰分析を、重回帰分析(multiple regression analysis)と呼ぶ。

母集団モデル 重回帰方程式は複数の説明変数を含むもので、母集団において

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (i = 1, 2, 3, \dots, n) \quad (18)$$

となるモデルである。 $X_{2i}, X_{3i}, \dots, X_{ki}$ は説明変数、 ε は誤差項で、式(2)に述べた単回帰分析の場合と同様の仮定を満足するものとする。また、 $\beta_1, \beta_2, \beta_3, \dots, \beta_k$ は、ある説明変数の係数、他の説明変数の影響を除いて、純粋の影響を表すものである。なお、式(18)で定数項 β_1 に対応する定数 $X_{1i} = 1$ は省略されている。

最小二乗推定 重回帰方程式では、 k 個の未知の母集団回帰係数 $\beta_1, \beta_2, \dots, \beta_k$ を含むが、その推定には、式(4)同様、最小二乗法が用いられる。

$$\varepsilon_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki})$$

であるが、その平方和

$$S = \sum \varepsilon_i^2$$

を考えそれを最小にする S を最小にする $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ は、式(4)に準じてその一次の偏微分を 0 と置いた k 個の連立方程式

$$\frac{\partial S}{\partial \beta_1} = 0, \quad \frac{\partial S}{\partial \beta_2} = 0, \dots, \frac{\partial S}{\partial \beta_k} = 0 \quad (19)$$

を解くことによって求められる。この解を $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ とすれば、これらが求めるものであって、これらを標本(値)回帰係数という。

標本重回帰方程式 推定された係数を用いて、標本重回帰方程式は次のように表される：

$$Y = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \cdots + \hat{\beta}_k X_k \quad (20)$$

単回帰分析の式(6)に相当する。また、各 i の $E(Y_i)$ は、

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} \quad (i = 1, 2, 3, \dots, n) \quad (21)$$

で推定される。これを回帰値という。

決定係数と重相関係数 単回帰分析と同様、 Y_i の変動 $\sum(Y_i - \bar{Y})^2$ は、 X_1, X_2, \dots, X_k による重回帰方程式(式21)で説明できる変動 $\sum(\hat{Y}_i - \bar{Y})^2$ とそれ以外の説明できない変動 $\sum \hat{e}_i^2$ の和として、

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum \hat{e}_i^2$$

と表すことができる。ここに、 $\hat{e}_i = Y_i - \hat{Y}_i$ は回帰残差である。モデルの当てはまりの良さを表す決定係数 η^2 は、

$$\eta^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum \hat{e}_i^2}{\sum(Y_i - \bar{Y})^2}$$

で定義される。決定係数の正の平方根を重相関係数 multiple regression coefficient といい、 R で表す。これは、単回帰分析の通常の相関係数 r の一般化であるが、 $\eta^2 = r^2$ であったのに対して、 $\eta^2 = R^2$ となる。

ただし、説明変数を増やすにつれて残差二乗和が小さくなり、それに伴い決定係数が大きくなるという問題がある。そこで、自由度修正済み決定係数を用いる：

$$\text{adjusted } R^2 = 1 - \frac{\sum \hat{e}_i^2 / (n-k)}{\sum (Y_i - \bar{Y})^2 / (n-1)}, \text{ ここの } k \text{ は回帰係数の数である。}$$

推定量の性質と検定 重回帰分析では、最小二乗推定量 $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ は単回帰分析と同様、最良不偏線形定量である(ガウス・マルコフの定理)。

単回帰分析に準じて、誤差項 ε_i の分散 σ^2 を、式(8)にならって推定量 s^2 は

$$s^2 = \sum \hat{e}_i^2 / (n - k)$$

で推定する。この s を推定値の標準誤差 $s.e.$ として、これから $\hat{\beta}_i$ の標準誤差 $s.e.(\hat{\beta}_i)$ などが求められる(残念ながら、ここでは範囲を超えるので、説明できない。コンピューターのパッケージ・プログラムはこれを出力する)。

$$t_i = \frac{\hat{\beta}_i - \beta_i}{s.e.(\hat{\beta}_i)} \quad (22)$$

は、自由度 $n - k$ の t 分布 $t(n - k)$ に従う。したがって、一つの回帰係数についての仮説 $H_0 : \beta_i = a$ の検定は、単回帰分析と同様に行うことができる。

重回帰分析における同時検定とF検定 ところで、重回帰分析では複数の説明変数があるので、いくつかの回帰係数についての仮説を同時に検定したい場合がある。たとえば、二つの薬品 A, B の散布の植物への成長の影響を調べるとしよう。成長量を Y 、薬品 A, B それぞれの散布量を X_2, X_3 とすると、それらに効果がないという帰無仮説は

$$H_0 : \beta_2 = 0 \quad \text{かつ} \quad \beta_3 = 0$$

であり、少なくともどちらかの効果があるという対立仮説は

$$H_1 : \beta_2 \neq 0 \quad \text{または} \quad \beta_3 \neq 0$$

となる。このように帰無仮説が複数の制約式からなる場合、回帰係数ごとの t 検定では不充分であり、次に述べる F 検定を用いる。

- (i) H_0 が正しいとして、重回帰方程式(上記の例では説明変数に X_2, X_3 を含まないもの)を推定し、回帰残差の平方和 $\sum \hat{e}_i^2$ を S_0 とする。
- (ii) すべての説明変数を含む重回帰方程式を推定し、その回帰残差の平方和 $\sum \hat{e}_i'^2$ を S_1 とする。これは H_0 が成立していない場合に相当する。
- (iii) 帰無仮説に含まれる制約式の数(説明変数の数)を p とすると、統計量

$$F = \frac{(S_0 - S_1)/p}{S_1/(n - k)} \quad (23)$$

は、帰無仮説が正しい場合、自由度 $(p, n - k)$ の F 分布 $F(p, n - k)$ に従うことが知られている(上記の例では回帰係数の数 $k = 3$, 説明変数の数 $p = 2$)。したがって、この F 統計量 F -statistic を計算し、付表4の F 分布表からそのパーセント点を求め、

$$F \geq F_\alpha(p, n - k)$$

のときに帰無仮説を棄却し、それ以外は棄却しない。

- (iv) とくに、 X_2, X_3, \dots, X_k のすべてが Y を説明しないという帰無仮説

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$$

を X_2, X_3, \dots, X_k のどれか一つは説明しているという対立仮説

$$H_1 : \beta_2, \beta_3, \dots, \beta_k \text{ の少なくとも一つが } 0 \text{ でない}$$

に対し検定する場合は、 $S_0 = \sum (Y_i - \bar{Y})^2, p = k - 1$ であるから、これらを式(23)に代入して、 F 値を計算する。これによって、式(18)の重回帰方程式の妥当性の検定ができる。ただし、定数項は問題にしない。

<例> 英国保守党の得票率と社会経済指標 1987年のイギリス総選挙における保守党得票率(Y)の決定要因を分析する。633の選挙区のうち北アイルランドを除く616選挙区中からアルファベット順に127選挙区を等間隔抽出し回帰分析を行う。偏回帰係数の下のカッコ内は t 値である。

(a) 失業率(X_1)による単回帰

$$Y = 61.998 - 2.349X_1 \quad (R^2 = 0.461)$$

$$(-10.399)$$

この場合の、重回帰方程式の $F = 106.999$ である。一方、 $F_{0.05}(1, 125) \approx 3.920$ であるから、得票率は失業率で説明される($\beta_1 = 0$ という仮説は、有意水準5%で棄却される(重回帰方程式は採択される))。

(b) 失業率(X_1)および自動車保有率(X_2)による重回帰

$$Y = 36.867 - 1.511X_1 + 0.332X_2 \quad (R^2 = 0.478)$$

$$(-3.227) \quad (2.034)$$

重回帰方程式の $F = 113.697$, $F_{0.05}(2, 124) \approx 3.072$, ゆえに、この重回帰方程式も有意水準5%で採択される。

(c) 失業率(X_1)および地域ダミー(Z)による重回帰

いま、 $Z = 0$ (イングランド)、 1 (ウェールズ、スコットランド)という地域ダミー変数を導入すると、

$$Y = 68.580 - 2.057X_1 - 19.965Z \quad (R^2 = 0.728)$$

$$(-11.051) \quad (-12.543)$$

重回帰方程式の $F = 86.579$, $F_{0.05}(2, 124) \approx 3.072$, 重回帰方程式は有意水準5%で採択される。

補足 1.1 ガウス・マルコフの定理の証明

$\hat{\beta}_2$ が最良線型不偏推定量 (BLUE) であることを証明する。ここでは $\hat{\beta}_2$ のみについて扱う。 $\hat{\beta}'_2 = \sum c_i Y_i$ を任意の線型推定量とする。このとき：

$$\begin{aligned} \hat{\beta}'_2 &= \sum c_i (\beta_1 + \beta_2 X_i + \varepsilon_i) \\ &= \beta_1 \sum c_i + \beta_2 \sum c_i X_i + \sum c_i \varepsilon_i \end{aligned}$$

これが不偏推定量 $E(\hat{\beta}'_2) = \beta_2$ であるためには：

$$\sum c_i = 0, \quad \sum c_i X_i = 1$$

c_i がこの条件を満たす場合、下記が成り立つ：

$$\sum [c_i - \frac{X_i - \bar{X}}{\sum (X_j - \bar{X})^2}]^2 = \sum c_i^2 - \frac{1}{\sum (X_j - \bar{X})^2}$$

この条件下で分散を求める：

$$\begin{aligned} V(\hat{\beta}'_2) &= \sum c_i^2 V(Y_i) = \sigma^2 \sum c_i^2 = \sigma^2 (\sum [c_i - \frac{X_i - \bar{X}}{\sum (X_j - \bar{X})^2}]^2 + \frac{1}{\sum (X_j - \bar{X})^2}) \\ &= \sigma^2 \sum [c_i - \frac{X_i - \bar{X}}{\sum (X_j - \bar{X})^2}]^2 + \frac{\sigma^2}{\sum (X_j - \bar{X})^2} \end{aligned}$$

$V(\hat{\beta}'_2)$ を最小にするためには、

$$c_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}$$

となり、これは $\hat{\beta}_2$ を与える。