

回帰分析の基礎 1

crimsonbach

2004 年 8 月 31 日

計量経済学の重要な部分、回帰分析 (regression analysis) の基礎について勉強する。ある変数 Y を別の
ある変数 X で定量的に説明する回帰方程式 (regression equation) を求めることが目的である。 Y を従属変
数、被説明変数、内生変数、内挿値といい、 X を独立変数、説明変数、外生変数、外挿値という。ここでは説
明変数が 1 つの単回帰分析 (simple regression analysis) の話をしよう。説明変数 X によって説明できる部
分を X の関数として

$$\beta_1 + \beta_2 X_i. \quad (1)$$

これを回帰関数 (regression function) あるいは回帰方程式という。とくに (1) 式がそうであるように回帰関
数が線型の場合は線型関数、線型回帰 (linear regression) という。非線型関数であっても、対数をとるなど
の関数変換あるいはテーラー展開などで線型関数に近似できる。したがって、線型回帰は回帰分析の中でもよ
く扱われる手法らしい。では、 i 番目の観測値を (X_i, Y_i) 、ばらつき (X によって説明できない部分) を ϵ_i
とすると

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

これが母集団において成立する場合、母回帰方程式 (population regression equation) といい、 β_1 、 β_2 を
母 (偏) 回帰係数 (population (partial) regression coefficient) という。母回帰係数は未知のパラメータであ
る。 ϵ_i は誤差項 (error term) 攪乱項 (disturbance term) という。説明変数 X_i および誤差項 ϵ_i には以下
の 4 つの仮定をおく。

- $X_i \neq$ 確率変数、すでに確定した値
- ϵ_i は確率変数で、 $E(\epsilon_i) = 0$, $i = 1, \dots, n$
- 異なった誤差項は無相関 $Cov(\epsilon_i, \epsilon_j) = E(\epsilon_i \epsilon_j) = 0$, $i \neq j$
- 誤差項の分散が一定 σ^2 、 $V(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$, $i = 1, \dots, n$

この条件のもとで

$$E(Y_i) = \beta_1 + \beta_2 X_i.$$

が成り立つ。また、とくに上記の仮定がおかれた誤差項をホワイト・ノイズ、ショック、標準乱数という。す
でに決まった X に対応して変数 Y が誤差項 ϵ_i を含んで Y_i という値をとるが、その確率変数のとりうる値
の期待値が $\beta_1 + \beta_2 X_i$ になるというのが上の式の意味である。

それでは回帰係数 β_1 および β_2 という未知のパラメータをデータをもとに推定する。当然求める回帰方
程式は当てはまりが良い方がいい。したがって、説明できない部分

$$\epsilon_i = Y_i - (\beta_1 + \beta_2 X_i).$$

は小さいほうが望ましい。符号の影響を取り除くため誤差項の二乗和

$$S = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n \{Y_i - (\beta_1 + \beta_2 X_i)\}^2. \quad (3)$$

S を最小にして $\hat{\beta}_1$ 、 $\hat{\beta}_2$ （推定量あるいは理論値、回帰値、予測値）を求める方法を最小二乗法（OLS：Ordinary Least Squares）という。 $\hat{\beta}_1$ および $\hat{\beta}_2$ は最小二乗推定量である。 S を最小にするために β_1 および β_2 にて偏微分すると

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

$$\frac{\partial S}{\partial \beta_2} = \sum_{i=1}^n \frac{\partial}{\partial \beta_2} \{Y_i - (\beta_1 + \beta_2 X_i)\}^2 = -2 \sum_{i=1}^n (Y_i - \beta_1 - \beta_2 X_i) X_i = 0. \quad (5)$$

少し整理すると

$$n\hat{\beta}_1 + \left(\sum_{i=1}^n X_i\right)\hat{\beta}_2 = \sum_{i=1}^n Y_i \quad (6)$$

$$\left(\sum_{i=1}^n X_i\right)\hat{\beta}_1 + \left(\sum_{i=1}^n X_i^2\right)\hat{\beta}_2 = \sum_{i=1}^n X_i Y_i. \quad (7)$$

これらを回帰係数を求めるための正規方程式（normal equation）という。以上から回帰係数を求める公式は

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (8)$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}. \quad (9)$$

\bar{X} および \bar{Y} は標本平均である。(8) 式は X と Y の標本共分散を X の標本分散で割っているのが分かる。因みに、2 階条件も必要だが、ここでは割愛して話を進めよう。 $\hat{\beta}_1$ および $\hat{\beta}_2$ は標本（偏）回帰係数（sample (partial) regression coefficient）という。またこれにより系統的に説明できる $y = \hat{\beta}_1 + \hat{\beta}_2 X_i$ を標本回帰方程式（sample regression equation）、標本回帰直線（sample regression line）と呼ぶ。 $E(Y_i)$ の標本回帰方程式による推定量を

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

\hat{Y} を回帰値、理論値、予測値という。また X で説明できずに残った部分は

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i.$$

回帰残差（residual）である。 $\hat{\epsilon}$ は誤差項の推定量をしめし、常に

$$\sum_{i=1}^n \epsilon_i = 0, \quad \sum_{i=1}^n \epsilon_i X_i = 0.$$

を満足する。2 つ目の式は X_i と ϵ_i が直交していて、 X_i と ϵ_i の相関がゼロということと同じだ。上の 2 つ以外の条件 $\sum \hat{Y}_i \hat{\epsilon}_i = 0$ を満たすが、これは上の 2 つから確認できるため省略する。回帰残差 $\hat{\epsilon}_i$ の分散 σ^2 は

$$s^2 = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n-2}. \quad (10)$$

から推定される。この値が小さいほど当てはまりがよい。分母が $n-2$ なのは自由度が2つ失われているからだ。 ϵ_i が正規分布に従えば、95 %の残差が $-2s < \hat{\epsilon}_i < 2s$ に入る。 $\hat{\beta}_1$ と $\hat{\beta}_2$ は

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

で不偏推定量となる。 s^2 も $E(s^2) = \sigma^2$ で σ^2 の不偏推定量になる。分散は

$$V(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}, \quad V(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}. \quad (11)$$

σ^2 は未知なので、 s^2 におきかえて推定する。 n が大きいと分散は小さくなる。 $\sum (X_i - \bar{X})^2$ 、 $\sum X_i^2$ は n 個の正の数である。同一の n に対しては $V(\hat{\beta}_1)$ は X_i のばらつきが大きいほど小さくなるが、 $\sum X_i^2 = \sum (X_i - \bar{X})^2 + n\bar{X}^2$ だから $V(\hat{\beta}_1)$ を小さくするために X_i のばらつきが大きいことだけでは不十分だ。 $\frac{n\bar{X}^2}{\sum (X_i - \bar{X})^2}$ が小さいことが必要になる。 $\hat{\beta}_1$ と $\hat{\beta}_2$ との共分散は

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\bar{X}V(\hat{\beta}_2). \quad (12)$$

$\hat{\beta}_1$ と $\hat{\beta}_2$ はガウス・マルコフの定理 (Gauss-Markov's theorem) によって線型不偏推定量のなかで最も分散の小さい推定量、最良線型不偏推定量 (BLUE: Best Linear Unbiased Estimator) であると知られている。線型不偏推定量は

$$\hat{\beta}_j = \sum_{i=1}^n c_{ij} Y_i, \quad E(\hat{\beta}_j) = \beta_j, \quad j = 1, 2.$$

を満足する。 ϵ_i が互いに独立で正規分布に従うと仮定した場合、クラメル・ラオの不等式 (Cramér-Rao's inequality) から、不偏推定量のなかで分散が最も小さくなる最良不偏推定量 (BUE: Best Unbiased Estimator) となる。

最後に回帰方程式の当てはまりの基準についての話をしよう。決定係数 (coefficient of determination) は R^2 という。 Y_i の変動和 $\sum (Y_i - \bar{Y})^2$ 、回帰方程式で説明できる部分 $\sum (\hat{Y}_i - \bar{Y})^2$ 、説明できない部分 $\sum \hat{\epsilon}_i^2$ から

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \hat{\epsilon}_i^2 \quad (13)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{回帰式で説明できる部分}}{\text{全変動}}. \quad (14)$$

実際に $\sum (\hat{Y}_i - \bar{Y})^2$ を計算する場合には $\hat{\beta}_2 \sum (X_i - \bar{X})(Y_i - \bar{Y})$ を利用するのが便利だ。決定係数は

$$0 < R^2 < 1.$$

の間にはいる。決定係数が1なら回帰方程式は当てはまりがよく、逆に0なら当てはまりが悪い。また、 r を X_i と Y_i の (標本) 相関係数とすると

$$r^2 = R^2.$$

因みに標本相関係数は

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$