

回帰分析の基礎 3

crimsonbach

2004 年 9 月 8 日

2 つ以上の説明変数を扱う場合、重回帰分析 (multiple regression analysis) という。重回帰方程式は母集団において

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \quad i = 1, \cdots, n. \quad (1)$$

β_1, \dots, β_k は未知のパラメータである。説明変数および誤差項について以下のことを仮定する。

- X_{2i}, \dots, X_{ki} は確率変数ではなく、決まった値をとる。
- u_i は確率変数で、 $E(u_i) = 0$
- 異なった誤差項は無相関 $Cov(u_i u_j) = E(u_i u_j) = 0, \quad i \neq j$
- $V(u_i) = E(u_i^2) = \sigma^2 \quad i = 1, \dots, n$ (分散の均一性)
- 説明変数は他の説明変数の線型関数では表されない。 $\alpha_1 + \alpha_2 X_{2i} + \dots + \alpha_k X_{ki} = 0$ となる $\alpha_1, \dots, \alpha_k$ は $\alpha_1 = \dots = \alpha_k = 0$ 以外には存在しない (変数間に完全な多重共線性がない)。

最小二乗法にて回帰方程式を推定する。

$$u_i = Y_i - (\beta_1 + \beta_2 X_{2i} + \cdots + \beta_{ki} X_{ki}).$$

この二乗和は

$$S = \sum u_i^2 = \sum \{Y_i - (\beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})\}^2.$$

S をそれぞれ β_j で偏微分して 0 とおいた k 次の連立方程式は

$$\begin{aligned} \frac{\partial S}{\partial \beta_1} &= -2 \sum \{Y_i - (\beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})\} = 0 \\ \frac{\partial S}{\partial \beta_2} &= -2 \sum X_{2i} \{Y_i - (\beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})\} = 0 \\ &\vdots \\ \frac{\partial S}{\partial \beta_k} &= -2 \sum X_{ki} \{Y_i - (\beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki})\} = 0 \end{aligned}$$

最小二乗推定量 $\hat{\beta}_1, \dots, \hat{\beta}_k$ は標本（偏）回帰係数であり、ガウス・マルコフの定理から最小不偏推定量である。標本回帰方程式および当てはめ値はそれぞれ

$$y = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}.$$

誤差項 u_i の分散 σ^2 は回帰誤差 $e_i = Y_i - \hat{Y}_i$ から

$$s^2 = \frac{\sum e_i^2}{(n-k)}. \quad (2)$$

$n - k$ で除すのは k 個の最小二乗推定量で求めているからだ ($\sum e_i = 0, \sum e_i X_{2i} = 0, \dots, \sum e_i X_{ki} = 0$ が成立し、自由度が k 個失われている。)。 s^2 は σ^2 の不偏推定量になる。

検定の話に入るが、 t 検定、 F 検定、 チョウ検定の順に進めよう。 σ^2 の推定量 s^2 を使って標本回帰係数 $\hat{\beta}_j$ の標準誤差 $s.e.(\hat{\beta}_j)$ を求める。

$$t_j = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)}.$$

これは単回帰と同様に仮説検定を行う。次に F 検定であるが、これは以下の手順で行う。

- H_0 (たとえば、 $\beta_2 = \dots = \beta_{k-1} = 0$) が正しいとして、重回帰方程式を推定し、残差の平方和 S_0 を求める。
- 全ての説明変数を加えて、 H_1 のもとで重回帰方程式を推定し、残差の二乗和 S_1 を求める。
- H_0 に含まれる式の数 p とすると $F = \frac{(S_0 - S_1)/p}{S_1/(n-k)}$ は帰無仮説のもとで自由度 ($p, n - k$) の F 分布 $F(p, n - k)$ に従う。検定統計量 F と $F_\alpha(p, n - k)$ を比較し、 $F > F_\alpha(p, n - k)$ の場合、帰無仮説を棄却する。

特に、『説明変数のすべてが Y_i を説明しない』という帰無仮説とそれに対する対立仮説を設定した場合、

$$\begin{aligned} H_0 : \beta_2 = \dots = \beta_k = 0 \\ H_1 : \text{少なくとも1つは0でない} \end{aligned}$$

$$p = k - 1, \quad S_0 = \sum (Y_i - \bar{Y})^2, \quad S_1 = \sum e_i^2, \quad S_0 - S_1 = \sum (\hat{Y}_i - \bar{Y})^2.$$

帰無仮説の制約式が1つの場合

$$F = t^2 \quad (\text{右片側検定}).$$

チョウ検定 (Chow test) は時系列データの構造変化を調べるために使われる。期間 $t = 1, 2, \dots, T$ の時系列データから、モデルは

$$Y_t = \beta_1 + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + u_t.$$

誤差項の分散は σ^2 で一定である。検定の手順は以下の通り。

- たとえば、帰無仮説 H_0 : 構造変化がなく回帰係数 β_1, \dots, β_k が期間を通して一定
- H_0 のもとで、すべての期間のデータを使って回帰モデルを推定し、その残差の平方和 S_0 を求める。
- データで $1 \sim T_1$ と $T_1 + 1 \sim T$ までの2つの期間に分け、それぞれの期間で回帰モデルを推定し、それぞれの残差の平方和を求め、2つのサンプル期間の残差の平方和を加え、残差の平方和の合計 S_1 を求める。
- $F = \frac{(S_0 - S_1)/k}{S_1/(T - 2k)}$ とすると、 H_0 のもとで F は自由度 ($k, T - 2k$) の F 分布 $F(k, T - 2k)$ に従う。 $F > F_\alpha(k, T - 2k)$ の場合、帰無仮説は棄却される。

Y_i の変動和

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

であり、決定係数 R^2 は

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}.$$

R を重相関係数 (multiple correlation coefficient) という。 $n = k$ の場合、 $R^2 = 1$ になって、不要な変数を加えるとモデルの当てはまりが悪くなる。そこで補正 R^2 (adjusted R^2) \bar{R}^2 は説明変数の数の違いを考慮したもので、

$$\bar{R}^2 = 1 - \frac{\sum e_i^2 / (n - k)}{\sum (Y_i - \bar{Y})^2 / (n - 1)}.$$