2021.12.15(水 13:10 ～)

兵庫高校

15　相関分析と回帰分析

（この章のポイント）

1）2 次元の連続型データの分析には、相関と回帰がある。
2）相関分析とは、2 つの変量に関係があるかどうかを調べる手法である。
3）回帰分析とは、1 つの変量で、もう 1 つの変量を説明する手法である。

15.1　相関分析

相関分析は、6 章で紹介した **2 次元正規分布** にしたがうデータの分析方法です。

15.1.1　相関係数の分布

標本相関係数 r は、母相関係数が 0 のとき、次の定理が成り立ちます。

定理 15.1　相関係数の分布は、母相関係数が 0 かどうかで異なる。
1) $\rho = 0$ のとき、以下の t_0 は自由度 $(n-2)$ の t 分布に従う。

$$t_0 = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}} \tag{15.1}$$

2) $\rho \neq 0$ のとき、$Z(r)$ が 近似的に $N(Z(\rho), \frac{1}{n-3})$ に従う。これを標準化して、次の式が近似的に成り立つ。

$$u_0 = \sqrt{n-3}\{Z(r) - Z(\rho)\} \sim N(0,1) \tag{15.2}$$

ただし、$Z(t)$ は、Z 変換 という関数で、次の式で定義される。

Z 変換：$-1 < t < 1$ を満たす t に対して、$(-\infty, \infty)$ を動く次の $Z(t)$ のこと。

$$Z(t) := \frac{1}{2} \ln \frac{1+t}{1-t} = \tanh^{-1} t \tag{15.3}$$

15.1.2　無相関の検定

無相関の検定 とは、2 つの変量に相関があるかどうかを、(15.1) 式を用いて調べる検定です。

15.1.3　母相関係数の推定

母相関係数に関する点推定値と、Z 変換と (15.2) 式を用いることで、信頼率 95% の信頼区間を計算できます。

1) 点推定

$$\widehat{\rho} = r$$

2) 区間推定

1) r の Z 変換を行う。

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

2) $\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ の信頼区間を求める。

$$(\zeta_1, \zeta_2) = \left(Z - \frac{1.960}{\sqrt{n-3}}, Z + \frac{1.960}{\sqrt{n-3}} \right)$$

3) 逆変換の式を用いて、ρ の信頼区間を求める。

$$(\rho_1, \rho_2) = \left(\frac{e^{2\zeta_1} - 1}{e^{2\zeta_1} + 1}, \frac{e^{2\zeta_2} - 1}{e^{2\zeta_2} + 1} \right)$$

15.1.4　母相関係数の検定 *

母相関係数 ρ が、特定の相関係数 ρ_0 と等しいかどうか、Z 変換と (15.2) 式を用いて検定することができます。

従来の相関係数 0.7 と一致するかどうかの検定を行う。（有意水準は、$\alpha = 0.05$ とする。）

検定：$\rho_0 = 0.7$ として、帰無仮説 $H_0 : \rho = \rho_0$, 対立仮説 $H_1 : \rho \neq \rho_0$ を設定する。検定統計量は、(15.2) 式の u_0 を用い、棄却域は両側検定なので $R : |u_0| \geqq u(0.025) = 1.960$ である。

$$u_0 = \sqrt{n-3}\{Z(r) - Z(\rho_0)\}$$

で判定する。

15.2　単回帰分析

2 次元のデータ (x, y) について、x を用いて y についての情報を知りたいと考えるのは自然なことだと思います。その目的としては、予測と制御が主な事柄として考えられます。予測 とは、将来の値について、どのような値かを知りたいと思うことで、制御 とは、将来の値がこの範囲に収めたいと思うことです。

このとき、予測や制御される変数を 目的変数 といい、これを説明するのに用いる変数を 説明変数 といいます。このため、データは 2 次元正規分布ではなく、x は定数として扱われ、15.2.2 項で説明する「直線性の仮定」を分析の前提とします。

15.2.1　最小二乗推定量

2 次元のデータ $(x_i, y_i),\ i = 1, 2, \cdots, n$ に対して、直線関係 $y = a + bx$ を想定します。ここで、直線による予測値と実測値のズレを最小にするため、ズレの二乗和を最小にする係数 a, b を求める方法が 最小二乗法 です。

この結果、次の推定量が得られます。導出方法は 15.3 節で述べます。

$$\widehat{a} = \overline{y} - \widehat{b}\overline{x}, \quad \widehat{b} = \frac{S_{xy}}{S_{xx}}$$

15.2.2　直線性の仮定

また、得られたデータには、次の直線性の仮定をします。

直線性の仮定：

$$Y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \cdots, n \tag{15.3}$$

このとき、$\beta_0,\ \beta_1$ は 回帰係数、特に β_1 のことを 偏回帰係数 と呼びます。また、ε_i は 誤差 と呼ばれます。

これらの偏回帰係数と a, b の関係は、$b = \beta_1,\ a = \beta_0 - \beta_1\overline{x}$ ですから、$\beta_0,\ \beta_1$ の推定量は、次のようになります。

$$\widehat{\beta}_0 = \overline{y} \tag{15.4}$$

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{15.5}$$

これらの推定量は、最小二乗推定量 と呼ばれます。

15.2.3　予測値とは

説明変数 $x = x_i$ のときの回帰直線上の点の y 座標の値は、予測値 と呼ばれ、次の式で求められます。

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1(x_i - \overline{x}) \tag{15.6}$$

15.2.4　平方和の分解

まず、次の等式を見て下さい。

$$y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}) \tag{15.7}$$

この左辺は全変動で、データのばらつきを表しています。これを、右辺第 1 項の予測値からのズレとしての誤差的なばらつきと、第 2 項の回帰によるばらつきに分解しています。実は、これらの二乗和には、次の等式が成り立ちます。この性質を 平方和の分解 といいます。（$\widehat{y}_i - \overline{y} = \widehat{\beta}_1(x_i - \overline{x})$ に着目せよ。詳細は章末問題 15.3）

$$S_T = S_R + S_e \tag{15.8}$$

$$\text{総平方和：} S_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = S_{yy} \tag{15.9}$$

$$\text{回帰平方和：} S_R = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2, \quad \text{残差平方和：} S_e = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

ここで、回帰平方和は、$\widehat{y}_i - \overline{y} = \widehat{\beta}_1(x_i - \overline{x})$ に着目すると

$$S_R = \frac{S_{xy}^2}{S_{xx}} \tag{15.10}$$

でも求められます（詳細は、章末問題 15.4）。

また各々の平方和にはその自由度があり、総自由度は $\phi_T = n - 1$ で、回帰の自由度は $\phi_R = 1$ で、残差の自由度は $\phi_e = n - 2$ です。これらの自由度の間にも平方和の分解と対応した、次の関係式が成り立ちます。

$$\phi_T = \phi_R + \phi_e \tag{15.11}$$

15.2.5　統計量の分布

単回帰分析における分析の中心は、傾き β_1 です。一方、回帰残差の平方和である S_e は、傾きの推定量と独立になることが示されます。そこで、下記のように傾きの推定量 $\widehat{\beta}_1$ の分布が導かれます。

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \tag{15.12}$$

$$\frac{S_e}{\sigma^2} \sim \chi^2(\phi_e), \quad \phi_e = n - 2 \tag{15.13}$$

したがって (15.12) 式を標準化し、(15.13) 式と併せると、次の性質が得られます。

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{V_e/S_{xx}}} \sim t(\phi_e) \tag{15.14}$$

定理 15.2　傾きの推定量 $\widehat{\beta}_1$ と残差平方和 S_e は互いに独立に、(15.12), (15.13) の分布に従う。よって、(15.14) 式が成り立つ。

15.2.6　単回帰分析の計算手順（その１）

目的変数を説明するのに、説明変数が影響しているかどうか調べたいとします。この分析は傾きがゼロかどうかの検定を行えばよく、結果は無相関の検定と一致します。

このとき、(15.14) 式で $\beta_1 = \beta_{10} = 0$ を代入した値は、

$$t_0 = \frac{\widehat{\beta}_1}{\sqrt{V_e/S_{xx}}} \sim t(\phi_e) \tag{15.15}$$

となり、(15.1) 式と一致することが確かめられます（両者を、S_{xx}, S_{yy}, S_{xy} で表現せよ)。また、従来の傾き β_{10} と同じかどうかの検定も行えます。

15.2.7　単回帰分析の計算手順（その２）*

前項と同じ結果をもたらす分散分析表による検定方法を紹介します。この方法は、説明変数が複数の重回帰分析においては、標準的な解法です。

例えば、傾きがゼロかどうかの検定は、以下のように分散分析表を用いて行うことができます。

解答 : まずは、$\beta_{10} = 0$ として、帰無仮説 $H_0 : \beta_1 = \beta_{10}$, 対立仮説 $H_1 : \beta_1 \neq \beta_{10}$, 有意水準は $\alpha = 0.05$ とする。

検定統計量は、$F_0 = V_R/V_e$ で、棄却域は、

$$R : F_0 \geqq F(\phi_R, \phi_e, 0.05)$$

とする。

ここで、平方和は、$S_T = S_{yy},\ S_R = S_{xy}^2/S_{xx},\ S_e = S_T - S_R$ となる。また、自由度は、$\phi_T = n-1,\ \phi_R = 1,\ \phi_e = n-2$ である。

さらに、分散は $V_R = S_R/\phi_R,\ V_e = S_e/\phi_e$ と得られる。

$$F_0 = \frac{V_R}{V_e} \tag{15.16}$$

これが、棄却域に入るかどうかで判定する。

補足 : (15.15) 式の t_0 と (15.16) 式の F_0 は、$F_0 = t_0^2$ が成り立つ。よって、これらの検定は同等（同じ結果をもたらす）である（詳細は、章末問題 15.5)。