最小二乗法による単回帰の回帰係数の推定

クリスカ*

2011 年 7 月 27 日

1　回帰分析とは

身長と体重の間には, 一方が増加すれば他方も増加するという関係があることが分かります。このような関係が成り立っているとき, この間の関係式を推定しようというのが **回帰分析** です. 回帰分析は複数種類のデータを分析するための 1 つであり, あるデータを他のデータの関数によって説明するものです.

2　データの統計量

2 種類のデータ x, y について n 個の観測値 $(x_i, y_i)\quad (i = 1, 2, \ldots, n)$ が与えられていたとします.

データの **平均**：$\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$

データの **分散**：$(s_x)^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2 = \overline{x_i^2} - (\overline{x})^2$

データの **共分散**：$s_{xy} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\ \overline{y} = \overline{xy} - \overline{x}\ \overline{y}$

データの **相関係数**：$r_{xy} = \dfrac{s_{xy}}{s_x s_y}$

Remark

データから平均を差し引き標準偏差で割って, 位置, 尺度の調整をすると, 平均は 0, 標準偏差は 1 に揃ったことになります. この操作をデータの **標準化** といいます.

x_i を $z_i = \dfrac{x_i - \overline{x}}{s_x}$, y_i を $w_i = \dfrac{y_i - \overline{y}}{s_y}$ と標準化したデータ z_i, w_i を考えると, 定義より,

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n} z_i w_i$$

となります. これは標準化したデータ z_i, w_i の共分散に等しいことが分かります.

* k-nakagawa@h6.dion.ne.jp

このことから, $-1 \leq r_{xy} \leq 1$ が言えます. 証明は以下の通りです.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(z_i \pm w_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(z_i^2 \pm 2z_iw_i + w_i^2\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}z_i^2 \pm \frac{2}{n}\sum_{i=1}^{n}z_iw_i + \frac{1}{n}\sum_{i=1}^{n}w_i^2\\ &= 1 \pm 2r_{xy} + 1\\ &= 2(1 \pm r_{xy})\end{aligned}$$

左辺は常に非負であるので $1 \pm r_{xy} \geq 0$ より $-1 \leq r_{xy} \leq 1$ が言えます.
また, $x' = ax + b,\ y' = cx + d \quad (ac > 0)$ のように線型変換した相関係数を r'_{xy} とすると線型不変性 $r'_{xy} = r_{xy}$ も言えます.

3 回帰分析

相関係数が 1 (もしくは -1) に近く, 正の相関 (負の相関) が認められるとき, 一方の変数を他方の変数の一次関数として表すことが考えられます.
つまり, y を x の関数として考えるならば, $y = \alpha + \beta x$ というモデルになりますが, このモデルを **(線型) 回帰モデル**, このモデルを用いた分析を **回帰分析** といいます. また, パラメータ α, β を **回帰係数** といいます.
上記のように y を x の関数として考えるとき, x を **説明変数** (または **独立変数**), y を **被説明変数** (または **従属変数**) といい, y を x によって説明することを 「**y を x に回帰する**」 といいます.
ここでは説明変数が 1 つの場合を考えていますが, 観測データが $(x_{1i}, x_{2i}, \ldots, x_{mi}, y_i) \quad (i = 1, 2, \ldots, n)$ として m 種類あり, $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$ というモデル (y を $x_1,\ x_2, \ldots, x_m$ に回帰するモデル) を考えることもできます. このように説明変数が複数ある場合を **重回帰** と呼び, これに対して説明変数が 1 つである場合を **単回帰** と呼びます.
$\alpha,\ \beta_1,\ \beta_2, \ldots$ は観測値を用いて推定することになりますが, 推定された $\hat{\alpha},\ \hat{\beta_1},\ \hat{\beta_2}, \ldots$ を用いた関係式を **回帰式** (もしくは **回帰直線**) といいます.

Remark
相関係数 r_{xy} は x と y の直線的関係を測る尺度であり, この値から非線形の関係を読みとることはできません.

4 最小二乗法による単回帰の回帰係数の推定

観測値 $(x_i, y_i) \quad (i = 1, 2, \ldots, n)$ を用いた単回帰の回帰係数の推定について考えることにします.
この場合, n 個の観測値に対して, $y_i = \alpha + \beta x_i \quad (i = 1, 2, \ldots, n)$ という回帰モデルをあてはめて回帰係数 α, β の推定を行うことになります. しかし, 観測値が 1 つの直線上に完全に乗ることはまずあり得ません. ですので, 回帰式の誤差である $y_i - (\alpha + \beta x_i)$ を小さくする α, β を推定値とすることになります.
最小二乗法 では, 誤差の平方和である下式 S を最小にする α, β を推定値とします.

$$S = \sum_{i=1}^{n}\{y_i - (\alpha + \beta x_i)\}^2$$

x_i, y_i は変数ではなく, 既に確定した値をとっていることに注意すると, S は α, β の関数だということが分かります. S は非負値の二次式であるので, これを最小化する α, β は, α と β についての偏導関数が 0 となることが必要です. よって, 次の連立方程式の解が求める推定量 $\hat{\alpha}$, $\hat{\beta}$ となります.

$$
\begin{aligned}
&\begin{cases} \dfrac{\partial S}{\partial \alpha} = 0 \\ \dfrac{\partial S}{\partial \beta} = 0 \end{cases} \\
\iff &\begin{cases} -2\displaystyle\sum_{i=1}^{n}\{y_i - (\alpha + \beta x_i)\} = 0 \\ -2\displaystyle\sum_{i=1}^{n} x_i\{y_i - (\alpha + \beta x_i)\} = 0 \end{cases} \\
\iff &\begin{cases} \displaystyle\sum_{i=1}^{n}\left\{y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right)\right\} = 0 \\ \displaystyle\sum_{i=1}^{n} x_i\left\{y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right)\right\} = 0 \end{cases} \\
\iff &\begin{cases} \hat{\alpha} n + \hat{\beta}\displaystyle\sum_{i=1}^{n} x_i &= \displaystyle\sum_{i=1}^{n} y_i \\ \hat{\alpha}\displaystyle\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i^2 &= \displaystyle\sum_{i=1}^{n} x_i y_i \end{cases}
\end{aligned}
$$

ここで, 式変形によって得られた最後の $\hat{\alpha}$, $\hat{\beta}$ に関する連立方程式を **正規方程式** といいます.
正規方程式の上式の両辺を n で割って整理すると

$$
\begin{aligned}
\hat{\alpha} + \hat{\beta}\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} y_i &\iff \hat{\alpha} + \hat{\beta}\,\overline{x} = \overline{y} \\
&\iff \hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x}
\end{aligned}
$$

となります. これを下式に代入して, 両辺を n で割って整理すると

$$
\begin{aligned}
(\overline{y} - \hat{\beta}\,\overline{x})\frac{1}{n}\sum_{i=1}^{n} x_i + \hat{\beta}\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\sum_{i=1}^{n} x_i y_i &\iff \left(\overline{y} - \hat{\beta}\,\overline{x}\right)\overline{x} + \hat{\beta}\overline{x^2} = \overline{xy} \\
&\iff \hat{\beta}\left(\overline{x^2} - (\overline{x})^2\right) = \overline{xy} - \overline{x}\,\overline{y} \\
&\iff \hat{\beta}(s_x)^2 = s_{xy} \\
&\iff \hat{\beta} = \frac{s_{xy}}{(s_x)^2} \\
&\iff \hat{\beta} = \frac{r_{xy} s_y}{s_x}
\end{aligned}
$$

$\overline{y} = \hat{\alpha} + \hat{\beta}\,\overline{x}$ より, 回帰直線は観測データの平均 $(\overline{x}, \overline{y})$ を通ることが分かります.
$y = \hat{\alpha} + \hat{\beta} x$ を y を被説明変数 (x を説明変数) とする回帰直線といいます.
y を x で回帰するともいいます.

また,

$$\begin{aligned} y = \hat{\alpha} + \hat{\beta}x &\iff y = (\overline{y} - \hat{\beta}\,\overline{x}) + \hat{\beta}x \\ &\iff y - \overline{y} = \hat{\beta}(x - \overline{x}) \\ &\iff y - \overline{y} = \frac{r_{xy}s_y}{s_x}(x - \overline{x}) \\ &\iff \frac{y - \overline{y}}{s_y} = r_{xy} \times \frac{x - \overline{x}}{s_x} \end{aligned}$$

と変形できることより (y の標準化) = (相関係数) $\times$ (x の標準化) という関係があることも分かります.
次に, 回帰式に x_i を代入して得られる値 $\hat{y_i}$ を y_i の **内挿値** といい, y_i と $\hat{y_i}$ の差 $e_i = y_i - \hat{y_i}$ を **残差** といいます. これを用いると

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \left\{ y_i - \left(\hat{\alpha} + \hat{\beta}x_i \right) \right\} = 0 \\ \displaystyle\sum_{i=1}^{n} x_i \left\{ y_i - \left(\hat{\alpha} + \hat{\beta}x_i \right) \right\} = 0 \end{cases} \iff \begin{cases} \displaystyle\sum_{i=1}^{n} e_i = 0 \\ \displaystyle\sum_{i=1}^{n} e_i x_i = 0 \end{cases}$$

が分かります. これは, e_i の平均が 0 であり, また, $\{e_i\}$ と $\{x_i\}$ はベクトルとして **直交** していることを示しています.

5　多重線型回帰と正規方程式

単回帰のときと同様な方法で求めます.
観測値 $(x_{1i}, x_{2i}, \ldots, x_{mi}, y_i) \quad (i = 1, 2, \ldots, n)$ を用いた多重線型回帰の回帰係数の推定について考えることにします.
この場合, n 個の観測値に対して, $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} \quad (i = 1, 2, \ldots, n)$ という回帰モデルをあてはめて回帰係数 $\alpha,\ \beta_1, \beta_2, \ldots, \beta_m$ の推定を行うことになります. 回帰式の誤差である $y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})$ を小さくする $\alpha,\ \beta_1, \beta_2, \ldots, \beta_m$ を推定値とすることになります.
最小二乗法では, 誤差の平方和である下式 S を最小にする $\alpha,\ \beta_1, \beta_2, \ldots, \beta_m$ を推定値とします.

$$S = \sum_{i=1}^{n} \{ y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi}) \}^2$$

S は非負値の二次式であるので, これを最小化する $\alpha,\ \beta_1, \beta_2, \ldots, \beta_m$ は, 各々の変数についての偏導関数が 0 となることが必要です. よって, 次の連立方程式の解が求める推定量 $\hat{\alpha},\ \hat{\beta_1}, \hat{\beta_2}, \ldots, \hat{\beta_m}$ となります.

$$\frac{\partial S}{\partial \alpha} = \frac{\partial S}{\partial \beta_1} = \frac{\partial S}{\partial \beta_2} = \cdots = \frac{\partial S}{\partial \beta_m} = 0$$

$$\iff \begin{cases} -2\displaystyle\sum_{i=1}^{n}\{y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})\} = 0 \\ -2\displaystyle\sum_{i=1}^{n} x_{1i}\{y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})\} = 0 \\ -2\displaystyle\sum_{i=1}^{n} x_{2i}\{y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})\} = 0 \\ \qquad\qquad \vdots \\ -2\displaystyle\sum_{i=1}^{n} x_{mi}\{y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})\} = 0 \end{cases}$$

$$\iff \begin{cases} \displaystyle\sum_{i=1}^{n}\{y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})\} = 0 \\ \displaystyle\sum_{i=1}^{n} x_{1i}\{y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})\} = 0 \\ \displaystyle\sum_{i=1}^{n} x_{2i}\{y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})\} = 0 \\ \qquad\qquad \vdots \\ \displaystyle\sum_{i=1}^{n} x_{mi}\{y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi})\} = 0 \end{cases}$$

$$\iff \begin{cases} \hat{\alpha} n + \hat{\beta_1}\displaystyle\sum_{i=1}^{n} x_{1i} + \hat{\beta_2}\sum_{i=1}^{n} x_{2i} + \cdots + \hat{\beta_m}\sum_{i=1}^{n} x_{mi} &= \displaystyle\sum_{i=1}^{n} y_i \\ \hat{\alpha}\displaystyle\sum_{i=1}^{n} x_{1i} + \hat{\beta_1}\sum_{i=1}^{n} x_{1i}^2 + \hat{\beta_2}\sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + \hat{\beta_m}\sum_{i=1}^{n} x_{1i}x_{mi} &= \displaystyle\sum_{i=1}^{n} x_{1i}y_i \\ \hat{\alpha}\displaystyle\sum_{i=1}^{n} x_{2i} + \hat{\beta_1}\sum_{i=1}^{n} x_{2i}x_{1i} + \hat{\beta_2}\sum_{i=1}^{n} x_{2i}^2 + \cdots + \hat{\beta_m}\sum_{i=1}^{n} x_{2i}x_{mi} &= \displaystyle\sum_{i=1}^{n} x_{2i}y_i \\ \qquad\qquad \vdots \\ \hat{\alpha}\displaystyle\sum_{i=1}^{n} x_{mi} + \hat{\beta_1}\sum_{i=1}^{n} x_{mi}x_{1i} + \hat{\beta_2}\sum_{i=1}^{n} x_{mi}x_{2i} + \cdots + \hat{\beta_m}\sum_{i=1}^{n} x_{mi}^2 &= \displaystyle\sum_{i=1}^{n} x_{mi}y_i \end{cases}$$

ここで, 式変形によって得られた最後の $\hat{\alpha},\ \hat{\beta_1}, \hat{\beta_2}, \ldots, \hat{\beta_m}$ に関する連立方程式を正規方程式といいます.
正規方程式は式の数が少ないときは容易に解けますが, 式の数が多いときは電子計算機が必要になります.

Remark

多重線形回帰の正規方程式は行列を用いることにより以下のように書き直すことが出来ます.

$$
\begin{cases}
\hat{\alpha}n + \hat{\beta_1}\sum_{i=1}^{n} x_{1i} + \hat{\beta_2}\sum_{i=1}^{n} x_{2i} + \cdots + \hat{\beta_m}\sum_{i=1}^{n} x_{mi} &= \sum_{i=1}^{n} y_i \\
\hat{\alpha}\sum_{i=1}^{n} x_{1i} + \hat{\beta_1}\sum_{i=1}^{n} x_{1i}^2 + \hat{\beta_2}\sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + \hat{\beta_m}\sum_{i=1}^{n} x_{1i}x_{mi} &= \sum_{i=1}^{n} x_{1i}y_i \\
\hat{\alpha}\sum_{i=1}^{n} x_{2i} + \hat{\beta_1}\sum_{i=1}^{n} x_{2i}x_{1i} + \hat{\beta_2}\sum_{i=1}^{n} x_{2i}^2 + \cdots + \hat{\beta_m}\sum_{i=1}^{n} x_{2i}x_{mi} &= \sum_{i=1}^{n} x_{2i}y_i \\
\vdots & \\
\hat{\alpha}\sum_{i=1}^{n} x_{mi} + \hat{\beta_1}\sum_{i=1}^{n} x_{mi}x_{1i} + \hat{\beta_2}\sum_{i=1}^{n} x_{mi}x_{2i} + \cdots + \hat{\beta_m}\sum_{i=1}^{n} x_{mi}^2 &= \sum_{i=1}^{n} x_{mi}y_i
\end{cases}
$$

$$
\Updownarrow
$$

$$
\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}
\begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{m1} \\ 1 & x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{mn} \end{pmatrix}
\begin{pmatrix} \hat{\alpha} \\ \hat{\beta_1} \\ \hat{\beta_2} \\ \vdots \\ \hat{\beta_m} \end{pmatrix}
=
\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}
$$

ここで, $X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{m1} \\ 1 & x_{12} & x_{22} & \cdots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{mn} \end{pmatrix}$ とおくと上式は

$$
X^T X \begin{pmatrix} \hat{\alpha} \\ \hat{\beta_1} \\ \hat{\beta_2} \\ \vdots \\ \hat{\beta_m} \end{pmatrix} = X^T \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}
$$

と変形できるので, 最小二乗法による推定値は

$$
\begin{pmatrix} \hat{\alpha} \\ \hat{\beta_1} \\ \hat{\beta_2} \\ \vdots \\ \hat{\beta_m} \end{pmatrix} = (X^T X)^{-1} X^T \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}
$$

となります.

単回帰のときのように, 各パラメータについて明示的な式を示すことはできませんが, 上式右辺の行列計算をした結果が最小二乗法による推定値となります.