第 2 章 回帰 regression の続き

浅川 伸一

2006 年 10 月 26 日

1.4 相関係数に関するいくつかの注意点への追加

1. 外れ値の存在が不当に相関係数の値を歪めてしまう場合がある。
2. (時) 系列相関。

1 回帰係数のベクトル表現

変数 x と y との相関係数 r_{xy} は代数的には x と y との共分散をそれぞれの標準偏差で除したものと定義できる。一方ベクトル表現した場合、相関係数は平均偏差ベクトル $\boldsymbol{x}$ と $\boldsymbol{y}$ の長さに 2 つのベクトルのなす角の余弦 cosine をかけたものであった。x から y を予測する場合の回帰式に現れる回帰係数についてもベクトル表現することができる。相関係数は $r_{xy} = \cos\theta = \frac{(\boldsymbol{x},\boldsymbol{y})}{|\boldsymbol{x}||\boldsymbol{y}|} = \frac{s_{xy}}{s_x s_y}$ と表すことができた。$|\boldsymbol{y}|$, $|\boldsymbol{x}|$ はベクトルの長さを表している。ベクトルの長さは変数の分散とその平方根の正の値である標準偏差に関連する数であった。2 つのベクトルのなす角を θ とすると相関係数 r_{xy} は 2 つのベクトルのなす角の余弦 cosine と対応する。余弦のとる値は -1 から $+1$ であるので、相関係数の範囲と一致する。

変数 x の値から変数 y の値を予測するための回帰式を $y = ax + b$ としたとき、得られたデータから最小二乗法など[1]を使って回帰係数 regression coefficent を求めると前回の知識から、

$$a = r_{xy}\frac{s_y}{s_x} = \cos\theta\frac{|\boldsymbol{y}|}{|\boldsymbol{x}|} = \frac{|\boldsymbol{y}|\cos\theta}{|\boldsymbol{x}|} \tag{1}$$

であった。

(1) 中の分子は図 1 に図示したとおり、ベクトル $\boldsymbol{y}$ の端点からベクトル $\boldsymbol{x}$ へ垂線を下ろしたときの長さ $|\boldsymbol{y}|\cos\theta$ に相当する。x から y を予測する、つまり y を x へ回帰させるということは、ベクトル $\boldsymbol{x}$ から $\boldsymbol{y}$ を眺めたときの

[1]最小二乗法だけが回帰方程式を解く方法ではない。他にも同じ回にたどり着く方法があるが省略する

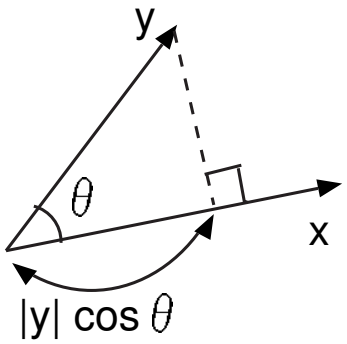

図 1: 回帰係数のベクトル表示。ベクトル $\boldsymbol{y}$ の $\boldsymbol{x}$ 上への射影と考える

長さ (射影) を計っていることに他ならない。$\boldsymbol{y}$ の中に $\boldsymbol{x}$ に含まれる要素がどのくらいあるかを問うていることとも言える。前回の (16)-(18) を再録すると

$$
\begin{aligned}
y &= ax + \bar{y} - a\bar{x} && (2)\\
(y - \bar{y}) &= a\,(x - \bar{x}) && (3)\\
(y - \bar{y}) &= r_{xy}\frac{s_y}{s_x}\,(x - \bar{x}) && (4)
\end{aligned}
$$

これをベクトル表現する。第一章の知識から素データから平均を引いた値を平均偏差ベクトルとすると $(y - \bar{y}) = r_{xy}\frac{s_y}{s_x}\,(x - \bar{x})$ は

$$
\begin{aligned}
\boldsymbol{y} &= r_{xy}\frac{s_y}{s_x}\boldsymbol{x} && (5)\\
\boldsymbol{y} &= \cos\theta\frac{|\boldsymbol{y}|}{|\boldsymbol{x}|}\boldsymbol{x} && (6)\\
\frac{\boldsymbol{y}}{|\boldsymbol{y}|} &= \cos\theta\frac{\boldsymbol{x}}{|\boldsymbol{x}|} && (7)
\end{aligned}
$$

標準偏差ベクトル $\boldsymbol{x}$, $\boldsymbol{y}$ をその長さで割ると言うことは各データから平均を引いて標準偏差で割ることを意味するので、標準得点化 (z スコア化) に変換していることを意味する。直上の式のとおり、標準化得点に変換した上での回帰係数は相関係数 ($\cos\theta$ と同義) に一致する。このときの回帰係数、つまり標準化得点に変換した上での相関係数のことを、標準回帰係数 standard regression coeffient と呼ぶ。

2　分散の分解、決定係数

図 1 のに示されているとおり、$\boldsymbol{y}$ は $\boldsymbol{x}$ への射影によって説明される部分と $\boldsymbol{x}$ では説明できない ($\boldsymbol{x}$ とは無関係な、$\boldsymbol{x}$ に直交する) 部分とに分解される。

$\boldsymbol{x}$ によって説明される部分を $\hat{\boldsymbol{y}} = \frac{|\boldsymbol{y}|\cos\theta}{|\boldsymbol{x}|}\boldsymbol{x}$ とし、$\boldsymbol{x}$ によっては説明できない部分を誤差ベクトル $\boldsymbol{e}$ と呼ぶことにすれば、 $\boldsymbol{y}$ の長さは

$$|\boldsymbol{y}|^2 = |\hat{\boldsymbol{y}}|^2 + |\boldsymbol{e}|^2 \tag{8}$$

と予測ベクトル $\hat{\boldsymbol{y}}$ の長さ (の二乗) と誤差ベクトル $\boldsymbol{e}$ の長さ (の二乗) とに分解される。これは $\boldsymbol{y}$(被説明変数ともいう) の分散が説明変数 $\boldsymbol{x}$ によって説明される分散と誤差分散 $\boldsymbol{e}$ とに分解できることを意味する。(8) 式の両辺を $|\boldsymbol{y}|$ で割れば、

$$1 = \frac{|\hat{\boldsymbol{y}}|^2}{|\boldsymbol{y}|^2} + \frac{|\boldsymbol{e}|^2}{|\boldsymbol{y}|^2} \tag{9}$$

をえる。$|\boldsymbol{y}|^2$ は分散に対応する。右辺第一項は

$$\frac{(\hat{\boldsymbol{y}}, \hat{\boldsymbol{y}})}{|\boldsymbol{y}|^2} = \left(r_{xy}\frac{s_y}{s_x}\right)^2 (\boldsymbol{x}, \boldsymbol{x})\,\frac{1}{s_y^2} = r_{xy}^2 \tag{10}$$

となる。従って右辺第二項、残差の分散は $1 - r^2$ となる。つまり相関係数の二乗は、y の全分散のうち x の分散で説明される割合を表していることになる。相関係数の二乗のことを決定係数 coefficient of determination と呼ぶ。