第 2 章 回帰 regression

浅川 伸一

2006 年 10 月 19 日

1 復習

中学生レベル（たぶん？）

問題 1. 二点 $(1,4)$, $(3,2)$ を通る直線の方程式を求めよ。

問題 2. 二点 $A(x_a, y_a)$, $B(x_b, y_b)$ を通る直線の方程式を求めよ。

2 回帰 regression

2 変数 (x, y) のデータが n 個あるとする。変数 x から y を予測することを考える。このとき直線で近似できると仮定し、最小自乗法（最小二乗法ともいう） the least minimun method を用いる。各データ点からの距離の二乗を最小化する直線の方程式を求める。ガウスによって考案された。

2 次元上の直線は以下の式で表せる。

$$y = ax + b \tag{1}$$

n 個の点について

$$\begin{aligned} y_1 &= ax_1 + b && (2)\\ y_2 &= ax_2 + b && (3)\\ \vdots &= \vdots \\ y_i &= ax_i + b && (4)\\ \vdots &= \vdots \\ y_n &= ax_n + b && (5) \end{aligned}$$

$n > 3$ ではすべての点をとおる一次関数 (直線) を引くことができない。そこで、直線と各店との距離を最小化することを考える。i 番目のデータの y の値は $\hat{y_i} = ax_i + b$ と予測できるとする。このとき $|y_i - \hat{y_i}|$ は実測値と予測値

との距離の絶対値である。これを全てのデータについて集めて (距離の絶対値の総和) 最小化したい。距離の絶対値を最小する変りに差の自乗和を最小化しても同じ解を得る。これがガウスのアイデア。誤差を e_i と表記すれば

$$\begin{aligned} e_1^2 &= (y_1 - \hat{y_1})^2 = (y_1 - (ax_1 + b))^2 && (6)\\ e_2^2 &= (y_2 - \hat{y_2})^2 = (y_2 - (ax_2 + b))^2 && (7)\\ \vdots &= \qquad\qquad \vdots \\ e_n^2 &= (y_n - \hat{y_n})^2 = (y_n - (ax_n + b))^2 && (8) \end{aligned}$$

この誤差の自乗和

$$Q = \sum_{i=1}^{n} e_i^2 = \sum (y_i - \hat{y_i})^2 = \sum (y_i - (ax_i + b))^2 \qquad (9)$$

を最小化する。このとき得られたデータを固定して[1]、誤差の自乗和を a(直線の傾き) と b(切片) の関数と見なす。最小化すべき値（誤差の自乗和）は変数 a, b の 二次関数の形をしているので微分して 0 とおいた点が最小値である。

$$\begin{aligned} \frac{\partial \sum e_i^2}{\partial a} &= -2\sum x_i y_i + 2a\sum x_i^2 + 2b\sum x_i = 0 && (10)\\ \frac{\partial \sum e_i^2}{\partial b} &= -2\sum y_i + 2a\sum x_i + 2nb = 0 && (11) \end{aligned}$$

上の式 (10), (11) を整理すると

$$\begin{aligned} \sum y_i &= a\sum x_i + bn && (12)\\ \sum x_i y_i &= a\sum x_i^2 + b\sum x_i && (13) \end{aligned}$$

となる。この連立方程式を正規方程式と呼ぶ。この連立方程式を a および b について解けば

$$\begin{aligned} a &= r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} && (14)\\ b &= \bar{y} - a\bar{x} && (15) \end{aligned}$$

を得る。$a = r_{xy}\dfrac{s_y}{s_x}$ を y の x 上への回帰係数と呼ぶ。回帰係数は x によって y が予測できるか否かを推し量る意味で意味がある。

式 (15) $b = \bar{y} - a\bar{x}$ を元々の式 (1) $y = ax + b$ へ代入することで b を消去して

$$\begin{aligned} y &= ax + \bar{y} - a\bar{x} && (16)\\ (y - \bar{y}) &= a\,(x - \bar{x}) && (17)\\ (y - \bar{y}) &= r_{xy}\frac{s_y}{s_x}\,(x - \bar{x}) && (18) \end{aligned}$$

[1] これまではデータを確率変数とみなしてその性質を議論してきたが、ここでは変数と定数の関係が逆。

を公式として載せているテキストもある。この式から回帰直線は x が平均のとき y の平均を通ることを示している。

ほとんど同じ手順を踏めば x の y 上への回帰係数を計算することができる[2]。この時の回帰係数は、$r_{xy}\frac{s_x}{s_y}$ となる。y の x への回帰係数を a_y, x の y への回帰係数を a_x とすればこれらの回帰係数と相関係数との関係は

$$r_{xy} = \sqrt{a_y a_x} \tag{19}$$

と幾何平均（相乗平均）になっている[3]。

3 分散

式 (9) へ式 (14),(15) を代入して誤差の分散を求めることを考える。

$$\begin{aligned}
s_e^2 &= \sum \left(y_i - \hat{y_i}\right)^2 && (20)\\
&= \sum \left\{y_i - (ax_i + b)\right\}^2 && (21)\\
&= \sum \left\{(y_i - \bar{y}) - r_{xy}\frac{s_y}{s_x}(x_i - \bar{x})\right\}^2 && (22)\\
&= \sum (y_i - \bar{y})^2 - 2r\frac{s_y}{s_x}\sum (x_i - \bar{x})(y_i - \bar{y}) \\
&\qquad + r^2\frac{s_y^2}{s_x^2}\sum (x_i - \bar{x})^2 && (23)\\
&= s_y^2 - 2r\frac{s_y}{s_x}s_{xy} + r^2\frac{s_{y^2}}{s_x^2}s_x^2 && (24)\\
&= s_y^2 - 2\frac{s_{xy}}{s_x s_y}\frac{s_y}{s_x}s_{xy} + \left(\frac{s_{xy}}{s_x s_y}\right)^2\frac{s_{y^2}}{s_x^2}s_x^2 && (25)\\
&= s_y^2 - \frac{s_{xy}^2}{s_x^2} && (26)\\
&= s_y^2\left(1 - \frac{1}{s_y^2}\frac{s_{xy}^2}{s_x^2}\right) && (27)\\
&= s_y^2\left(1 - r^2\right) && (28)
\end{aligned}$$

これを誤差変動、あるいは回帰直線周りの分散と呼ぶことがある。

また y の予測値 $\hat{y}$ の分散は x の分散の a^2 倍であから

$$\begin{aligned}
s_{\hat{y}}^2 &= a^2 s_x^2 && (29)\\
&= \left(r\frac{s_y}{s_y x}\right)^2 s_x^2 && (30)\\
&= r^2 s_y^2 && (31)
\end{aligned}$$

[2] x と y とを入れ替えて考えてみれば良い

[3] 全てを足し合わせてデータの個数で割る平均を相加平均、もしくは算術平均という

となる。

式 (28) と (31) の和が全分散 s_y^2 になることから、y の分散が回帰で説明できる変動と誤差の変動とに分解できることが分かる。これを分散の分解定理というが次回以降でより一般的な形で取り上げる。