付録: 最小二乗法による直線回帰

直線とデータのずれを測る量として

$$E = \sum_i (y - ax_i - b)^2$$

を考える．これは，各点でデータ y 成分と直線上の点の y 成分の差を 2 乗したものである．E を最小にする a, b を求めることで，データ列を近似する直線を探す．

最小点を求めるには偏微分を用いるのが簡単であるが，未修の人もいるので計算は複雑になるが愚直に平方完成で求める．計算を少しだけ簡単にするために，データを平均と残りに分ける:

$$\begin{aligned} x_i &= \bar{x} + \xi_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ y_i &= \bar{y} + \eta_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \end{aligned}$$

こうすると

$$\begin{aligned} \sum\nolimits_i \xi_i &= 0 \\ \sum_i \eta_i &= 0 \\ \frac{1}{n}\sum\nolimits_i \xi_i^2 &= \frac{1}{n}\sum_i (x_i - \bar{x})^2 = c_{xx} \\ \frac{1}{n}\sum\nolimits_i \eta_i^2 &= \frac{1}{n}\sum_i (y_i - \bar{y})^2 = c_{yy} \\ \frac{1}{n}\sum\nolimits_i \xi_i\eta_i &= \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = c_{xy} \end{aligned}$$

が後から使えて見通しがよくなる (i の動く範囲を省略した．以下，同様)．ξ_i, η_i で E を表示して展開する．

$$\begin{aligned} E &= \sum_i (\eta_i - a\xi_i + \bar{y} - a\bar{x} - b)^2 \\ &= \sum_i (\eta_i - a\xi_i)^2 + (\bar{y} - a\bar{x} - b)^2 + \sum_i 2(\eta_i - a\xi_i)(\bar{y} - a\bar{x} - b) \end{aligned}$$

$\sum_i \eta_i = \sum_i \xi_i = 0$ だから，最後の項は 0 である．さらに展開を続けて，

$$\begin{aligned} (\eta_i - a\xi_i)^2 &= \xi_i^2 a^2 - 2\xi_i\eta_i a + \eta_i^2 \\ (\bar{y} - a\bar{x} - b)^2 &= (\bar{y} - b)^2 - 2a\bar{x}(\bar{y} - b) + a^2\bar{x}^2 \\ &= \bar{y}^2 - 2\bar{y}b + b^2 - 2\bar{x}\bar{y}a + 2\bar{x}ab + a^2\bar{x}^2 \end{aligned}$$

なので E は a^2,b^2,ab,a,b についての多項式である．各係数を集めると ($\sum_i$ (定数) $= n$ に注意)

$$E = (n\bar{x}^2 + \sum_i \xi_i^2)a^2 + nb^2 + 2\bar{x}ab - 2(n\bar{x}\bar{y} + \sum_i \xi_i\eta_i)a - 2\bar{y}b + Const.$$

となる．分散・共分散を使って係数を書き換えると (ただし全体を n で割る)，

$$E/n = (\bar{x}^2 + c_{xx})a^2 + nb^2 + 2\bar{x}ab - 2(\bar{x}\bar{y} + c_{xy})a - 2\bar{y}b + Const.$$

適当に「平行移動」すると一次の項 a,b は消せる．ただし，クロスターム ab があるので a, b を同時に動かす必要がある．そこで，

$$E/n = (\bar{x}^2 + c_{xx})(a - a_0)^2 + (b - b_0)^2 + 2\bar{x}(a - a_0)(b - b_0) + Const.$$

を仮定する．これを展開すると

$$E/n = \{-2(\bar{x}^2 + c_{xx})a_0 - 2\bar{x}b_0\}a + \{-2b_0 - 2\bar{x}a_0\}b + \cdots$$

となるので，a,b の係数の一致より，a_0,b_0 についての連立方程式

$$\begin{aligned}(\bar{x}^2 + c_{xx})a_0 + \bar{x}b_0 &= \bar{x}\bar{y} + c_{xy} \\ a_0\bar{x} + b_0 &= \bar{y}\end{aligned}$$

が得られる．特に，第 2 式から $(\bar{x}, \bar{y})$ を通過することがわかる．これを解くと

$$a_0 = \frac{c_{xy}}{c_{xx}}, \quad b_0 = \bar{y} - a_0 x$$

となる．

E/n には $(a-a_0)(b-b_0)$ という正にも負にもなる項があるので，本当に $(a, b) = (a_0, b_0)$ で最小になるか不安かもしれない．しかし，以下の補題により大丈夫である．

補題　$a, d > 0, ad > b^2$ とすると，すべての実数 x, y で

$$f(x, y) = ax^2 + dy^2 + 2bxy \geq 0 \quad (\text{等号成立} : x = y = 0)$$

(証明) $b > 0$ の場合を考える．$xy \geq 0$ の場合は自明なので，$xy < 0$ の場合を考える．このとき $ad > b^2$ なので $\sqrt{ad}xy < bxy$ だから

$$f(x, y) = ax^2 + dy^2 + 2bxy > ax^2 + dy^2 + 2\sqrt{ad}xy = (\sqrt{a}x + \sqrt{d}y)^2 \geq 0$$

$b < 0$ の場合は，$xy > 0$ の領域で，$0 > b > -\sqrt{ad}$ を使えば同じように示せる．

なお，条件 $ad > b^2$ に相当するものは，$(\bar{x}^2 + c_{xx}) \cdot 1 > \bar{x}^2$ は明らかに成立する．