[2018 前期火 5] 統計遺伝学 I: 課題（5 月 15 日）

Toru YOSHIYASU

2018 年 5 月 19 日

Least square regression

最小二乗法による線型回帰の公式を導出する。1 個の説明変数 X と 1 個の被説明変数 Y がある場合について考える: $Y \sim aX + b$。

n 個のサンプル $\{(x_i, y_i) \mid i = 1, \ldots, n\}$ が与えられた時、誤差関数

$$J(a,b) := \sum_{i=1}^{n} \bigl(y_i - (ax_i + b)\bigr)^2$$

を最小にする実数の組 (a, b) を求める。最小値問題だから臨界点とヘッシアンを計算すればよいが、誤差関数 J は変数 (a, b) の 2 次式であり、そのグラフは下に凸な放物面となっているから、臨界点は 1 つだけでそこが最小値となる。

偏微分を下付きの添字で表せば、

$$J_a = -2 \sum_{i=1}^{n} x_i (y_i - ax_i - b), \quad J_b = -2 \sum_{i=1}^{n} (y_i - ax_i - b)$$

となって、連立方程式 $J_a = J_b = 0$ の解が求めるもの。$J_b = 0$ より、

$$b = \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i) \tag{1}$$

がわかる。これを $J_a = 0$ に代入すると、

$$\begin{aligned}
0 &= \sum_{i=1}^{n} x_i(y_i - ax_i - b) = \sum_{i=1}^{n} x_i(y_i - ax_i) - b\sum_{j=1}^{n} x_j \\
&= \sum_{i=1}^{n} x_i(y_i - ax_i) - \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i)\sum_{j=1}^{n} x_j \\
&= \frac{1}{n}\sum_{i=1}^{n}(nx_i - \sum_{j=1}^{n} x_j)(y_i - ax_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(nx_i - \sum_{j=1}^{n} x_j)y_i - \frac{a}{n}\sum_{i=1}^{n} x_i(nx_i - \sum_{j=1}^{n} x_j)
\end{aligned}$$

と変形できる。a について解き直し、(1) に代入すれば、求める解

$$\begin{aligned}
a &= \frac{\sum_i (nx_i - \sum_j x_j)y_i}{\sum_i x_i(nx_i - \sum_j x_j)} = \frac{n\sum_i x_i y_i - \sum_{i,j} x_j y_i}{n\sum_i x_i^2 - \sum_{i,j} x_i x_j}, \\
b &= \frac{1}{n}\Big(\sum_{k=1}^{n} y_k - a\sum_{k=1}^{n} x_k\Big) \\
&= \frac{1}{n}\Big(\sum_{k=1}^{n} y_k - \frac{n\sum_i x_i y_i - \sum_{i,j} x_i y_j}{n\sum_i x_i^2 - \sum_{i,j} x_i x_j}\sum_{k=1}^{n} x_k\Big) \\
&= \frac{1}{n}\frac{n\sum_{i,k} x_i^2 y_k - \sum_{i,j,k} x_i x_j y_k - (n\sum_{i,k} x_i x_k y_i - \sum_{i,j,k} x_i x_k y_j)}{n\sum_i x_i^2 - \sum_{i,j} x_i x_j} \\
&= \frac{\sum_{i,k} x_i^2 y_k - \sum_{i,k} x_i x_k y_i}{n\sum_i x_i^2 - \sum_{i,j} x_i x_j}
\end{aligned}$$

を得る。

この導出からわかる誤差関数 J の定義の長所について述べる。定義式を 2 乗ではなく、より大きな指数や絶対値に変えた場合、臨界点の計算のみに帰着することはできず、解の個数についての議論や別のアプローチが必要になる可能性がある。さらに、2 次式の臨界点の計算は連立 1 次方程式で、これを解くことも平易となっている。2 乗の部分は単なる関数の平滑化ではなく、解の存在と一意性および計算の容易さに寄与している。