データ解析レポート課題・第一

14_01043　伊澤　侑祐

問 1　計算問題

(1)

まず、k を固定する。二乗誤差

$$
\begin{aligned}
E(a) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i, a))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{k=1}^{K} a_k e_k(x))^2 \qquad (1)
\end{aligned}
$$

を a_k で微分し、極値条件を解く。

$$
\begin{aligned}
\frac{\partial E}{\partial a_k} &= \frac{2}{n}\sum_{i=0}^{n}(Y_i - a_k e_k(x)) \cdot e_k(x) = 0 \\
\Longleftrightarrow \ & \sum_{i=1}^{n} Y_i e_k(x) - n a_k = 0 \\
\Longleftrightarrow \ & a_k^* = \frac{1}{n}\sum_{i=1}^{n} Y_i e_k(x) \qquad (2)
\end{aligned}
$$

よって求める答えは (2) より

$$
a^* = \left\{ \left(\frac{1}{n}\sum_{i=1}^{n} Y_i e_k(x) \right)_k \right\} \qquad (3)
$$

である。

(2)

平均を $\mathscr{E}(\cdot)$ で表す。$\mathscr{E}(Y_i) = 0$ を用いて、

$$
\begin{aligned}
\mathscr{E}(a^*) &= \mathscr{E}\left(\frac{1}{n}\sum_{i=0}^{n} Y_i e_k(x) \right) \\
&= 0 \qquad (4)
\end{aligned}
$$

となる。

(3)

まず、一般に (k,l) 成分の場合の共分散を考える。i 成分の期待値を μ_i と置くと、

$$
\begin{aligned}
\sigma_{i,j} &= \mathscr{E}\left((a_k^* - \mu_i)\left(a_l^* - \mu_l\right)\right) \\
&= \mathscr{E}\left(a_k^* \cdot a_l^*\right) \\
&= \mathscr{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} Y_i e_k(x_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i e_l(x_i)\right)\right) \\
&= \frac{1}{n^2}\mathscr{E}\left(\sum_{i,j=1}^{n} Y_i Y_j e_k(x_i) e_l(x_j)\right) \qquad (5)
\end{aligned}
$$

となる。ここで、$\mathscr{E}(Y_i) = 0, \sqrt{\mathscr{E}(Y_i^2) - (E(Y_i))^2} = 1$, そして Y_i が独立であることより、

$$
\begin{aligned}
\mathscr{E}\left(\sum_{i,j=1}^{n} Y_i Y_j e_k(x_i) e_l(x_j)\right) &= \sum_{i,j=1}^{n} \mathscr{E}(Y_i)\mathscr{E}(Y_j) \sum_{i,j=1}^{n} e_k(x_i) e_l(x_j) \\
&= \begin{cases} \sum_{i=1}^{n} \mathscr{E}\left(Y_i^2\right) e_k(x_i) e_l(x_i) & (i = j) \\ 0 & (i \neq j) \end{cases} \\
&= \begin{cases} n & (i = j) \\ 0 & (i \neq j) \end{cases} \qquad (6)
\end{aligned}
$$

となる。したがって、求める分散共分散行列 Σ は、(5) と (6) より

$$
\Sigma = \frac{1}{n}\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \qquad (7)
$$

と求まる。

(4)

$E(a^*)$ の平均値 $\mathscr{E}(E(a^*))$ を求めると、次のようになる。

$$
\begin{aligned}
\mathscr{E}\left(E(a^*)\right) &= \mathscr{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{1}{n}\sum_{k=1}^{K} Y_i e_k(x_i) e_k(x_i)\right)^2\right) \\
&= \mathscr{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^2 - \frac{2}{n}\left(\sum_{k=1}^{K} Y_i^2 e_k(x_i) e_k(x_i)\right) + \left(\frac{1}{n}\sum_{k=1}^{K} Y_i e_k(x_i) e_k(x_i)\right)^2\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathscr{E}\left(Y_i^2\right) - \frac{2K}{n}\sum_{i=1}^{n}\mathscr{E}\left(Y_i^2\right) + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\left(\frac{1}{n}e_k(x_i)e_k(x_i)\right)^2 \mathscr{E}(Y_i^2) \\
&= 1 - \frac{2K}{n} + \frac{K}{n} \\
&= 1 - \frac{K}{n} \qquad (8)
\end{aligned}
$$

(5)

$$\mathscr{A} = \frac{1}{n}\int_{-\infty}^{\infty}\sum_{i=1}^{n}\left(y - f(x_i, a^*)\right)^2 q(y)dy \tag{9}$$

とおく。まず、$\mathscr{A}$ を計算する。

$$\int_{-\infty}^{\infty} q(y)dy = 1, \quad \int_{-\infty}^{\infty} yq(y)dy = 0, \quad \int_{-\infty}^{\infty} y^2 q(y)dy = 1 \tag{10}$$

であることを用いて、以下のようになる。

$$\begin{aligned}
\mathscr{A} &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(y^2 - \frac{2}{n}\sum_{k=1}^{K} yY_i e_k(x_i)e_k(x_i) + \left(\frac{1}{n}\sum_{k=1}^{K} Y_i e_k(x_i)e_k(x_i)\right)^2\right) q(y)dy \\
&= \frac{1}{n}\sum_{k=1}^{n}\int_{-\infty}^{\infty} y^2 q(y)dy \\
&\quad - \frac{2}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_{-\infty}^{\infty} yY_i e_k(x_i)e_k(x_i)q(y)dy + \frac{1}{n^3}\sum_{k=1}^{K}\sum_{i=1}^{n} Y_i^2\left(e_k(x_i)e_k(x_i)\right)^2 \cdot \int_{-\infty}^{\infty} yq(y)dy \\
&= \frac{1}{n}\cdot n \cdot 1 - \frac{2}{n^2}\sum_{k=1}^{K}\sum_{i=1}^{n} Y_i^2 e_k(x_i)e_k(x_i) \cdot \int_{-\infty}^{\infty} yq(y)dy + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} Y_i^2\left(\frac{1}{n}e_k(x_i)e_k(x_i)\right)^2 \\
&= 1 + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} Y_i^2\left(\frac{1}{n}e_k(x_i)e_k(x_i)\right)^2
\end{aligned}$$

よって、$\mathscr{A}$ の期待値は、

$$\begin{aligned}
\mathscr{E}(\mathscr{A}) &= 1 + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathscr{E}(Y_i^2)\left(\frac{1}{n}e_k(x_i)e_k(x_i)\right)^2 \\
&= 1 + \frac{K}{n}
\end{aligned} \tag{11}$$

となる。

問 2　応用問題

(1)

「市町村 2012estat.csv」に対し、回帰分析、主成分分析とクラスタ分析を用いて解析を行った。

(1.1)　回帰分析