

# 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

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

を  $a_k$  で微分し、極値条件を解く。

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

よって求める答えは (2) より

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

である。

(2)

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

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

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  $\mu_i$  と置くと、

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

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

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

となる。したがって、求める分散共分散行列  $\Sigma$  は、(5) と (6) より

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

と求まる。

(4)

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

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

(5)

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

とおく。まず、 $\mathcal{A}$  を計算する。

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

であることを用いて、以下ようになる。

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

よって、 $\mathcal{A}$  の期待値は、

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

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。環境は

- macOS Sierra 10.12.2
- Python 3.5.2
- numpy, scipy, pandas, scikit-learn, matplotlib

である。

### (1.1) 回帰分析

まず、15 歳から 64 歳までの人口総数と転出者数の関係性を調べるため、この二者に対して

$$[ ] = a \cdot [ ] + b \quad (12)$$

という仮説を立て、回帰分析を行った。その結果、次のようなグラフを得た。



図1 15 歳から 64 歳までの人口総数と転出者

また、モデルは次のようになった。

```
Formula: Y ~ <x> + <intercept>
```

```
Number of Observations:      1870
```

```
Number of Degrees of Freedom:  2
```

R-squared: 0.8835

Adj R-squared: 0.8834

Rmse: 1505.4458

F-stat (1, 1868): 14160.3938, p-value: 0.0000

Degrees of Freedom: model 1, resid 1868

-----Summary of Estimated Coefficients-----						
Variable	Coef	Std Err	t-stat	p-value	CI 2.5%	CI 97.5%
x	0.0655	0.0006	119.00	0.0000	0.0644	0.0666
intercept	-116.7771	41.7843	-2.79	0.0052	-198.6743	-34.8800

今回の場合、決定係数が0.8835とあり、このモデルで88%以上説明できているということになる。また、F値が十分に大きく(14160.3938)、p値も0.0000と99%以上妥当であるといえる。さらに、係数 $a$ (上の表における $x$ )と $b$ (上の表におけるintercept)の優位確率はそれぞれ0.0000と0.0052であるため、両方の値は妥当であるといえる。