データ解析レポート課題・第二

14_01043 伊澤侑祐

計算問題 1

(1)

計算すると以下のようになる。

$$
\begin{aligned}
K\left(p(k|a) \mid\mid p(k|b)\right) &= \sum_{k=0}^{\infty} p(k|a) \log \frac{\frac{a^k}{k!}\exp(-a)}{\frac{b^k}{k!}\exp(-b)} \\
&= \sum_{k=0}^{\infty} p(k|a) \log\Big\{ \Big(\frac{a}{b}\Big)^k \exp(-a+b) \Big\} \\
&= \sum_{k=0}^{\infty} \frac{a^k}{k!}\exp(-a)\Big\{ k(\log a - \log b) - a + b \Big\} \\
&= \sum_{k=0}^{\infty} \Big\{ a\frac{a^{k-1}}{(k-1)!}\exp(-a)(\log a - \log b) - \frac{a^k}{k!}(a-b) \Big\} \\
&= a(\log a - \log b) - (a-b)\exp(a)
\end{aligned}
$$

(2)

尤度関数は統計モデルのサンプルの積で表される。

$$
\begin{aligned}
p(X|a) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i-a)^2}{2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-a)^2\right)
\end{aligned} \tag{1}
$$

a は正規分布に従う。

$$
p(a) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right) \tag{2}
$$

以上より、尤度と事前分布の積を考えると、

$$
\begin{aligned}
p(a|X) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-a)^2\right) \times \frac{1}{\sqrt{2n}} \exp\left(-\frac{1}{2}a^2\right) \\
&\propto \exp\left(-\frac{a^2}{2} - \frac{\sum_{i=1}^{n}(x_i-a)^2}{2}\right)
\end{aligned} \tag{3}
$$

となる。カーネルをまとめると、x_i の平均を $\bar{x}$ と置いて、

$$\begin{aligned}\sum_{i=1}^{n}(x_i-a)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-a)^2 \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-a)^2 \end{aligned} \tag{4}$$

となる。したがって、

$$\begin{aligned} p(a|X) &\propto \exp\left[-\frac{a^2}{2}-\frac{n(\bar{x}-a)^2}{2}\right] \\ &\propto \exp\left[-\frac{a^2+n(\bar{x}-a)^2}{2}\right] \\ &\propto \exp\left[-\frac{(n+1)\{a-\frac{n}{n+1}\bar{x}\}^2-\frac{\bar{x}^2}{n+1}}{2}\right] \\ &\propto \exp\left[-\frac{(a-\frac{n}{n+1}\bar{x})^2}{2(n+1)}\right] \end{aligned} \tag{5}$$

よって、

$$a|X \sim N\left(\frac{n}{n+1}\bar{x}, \frac{1}{\sqrt{n+1}}\right) \sim N\left(\frac{\sum_{i=1}^{n}x_i}{n+1}, \frac{1}{\sqrt{n+1}}\right) \tag{6}$$

となる。

(3)

2 つのモデルについて **AIC** と **BIC** を求める。
$p_1(x|a_1)$、$p_2(x|a_2)$ の **AIC** と **BIC** を $AIC_1, BIC_1, AIC_2, BIC_2$ と置くと、以下のようになる。

$$AIC_1 = -2\times 36.1+2\times 1 = -70.2 \tag{7}$$
$$BIC_1 = -2\times 36.1+\log 500\times 1 = -65.985391901577813\cdots \simeq -66.0 \tag{8}$$
$$AIC_2 = -2\times 37.8+2\times 2 = -71.6 \tag{9}$$
$$BIC_2 = -2\times 37.8+\log 500\times 2 = -63.170783803155615\cdots \simeq -63.2 \tag{10}$$

これらより、以下のことが考察できる。単に $p_1(x|a_1)$、$p_2(x|a_2)$ を p_1, p_2 と書く。

- **AIC** は一致性は持たないが有効せを持つケースがあることが知られており、最も汎化誤差を小さくするモデルを選ぶことができる。p_2 のほうが p_1 より **AIC** の値が小さいので、p_2 のモデルは汎化誤差が少なくするモデルであると言える。
- **BIC** は有効性を持たないが一致性を持つケースがあると知られており、パラメータ数が一番少ないモデルを選ぶことができる。結果、p_1 のほうが p_2 よりパラメータ数が少ないため、**BIC** の値は p_1 のほうが p_2 より小さくなっている。

(4)

何度も繰り返せば一度は帰無仮説が棄却されてしまうので、多角的な観点から検証する必要がある。

応用問題 1

(1)

最尤推定を用いて各都市ごとの回帰直線を求めると、次のような結果が得られた。

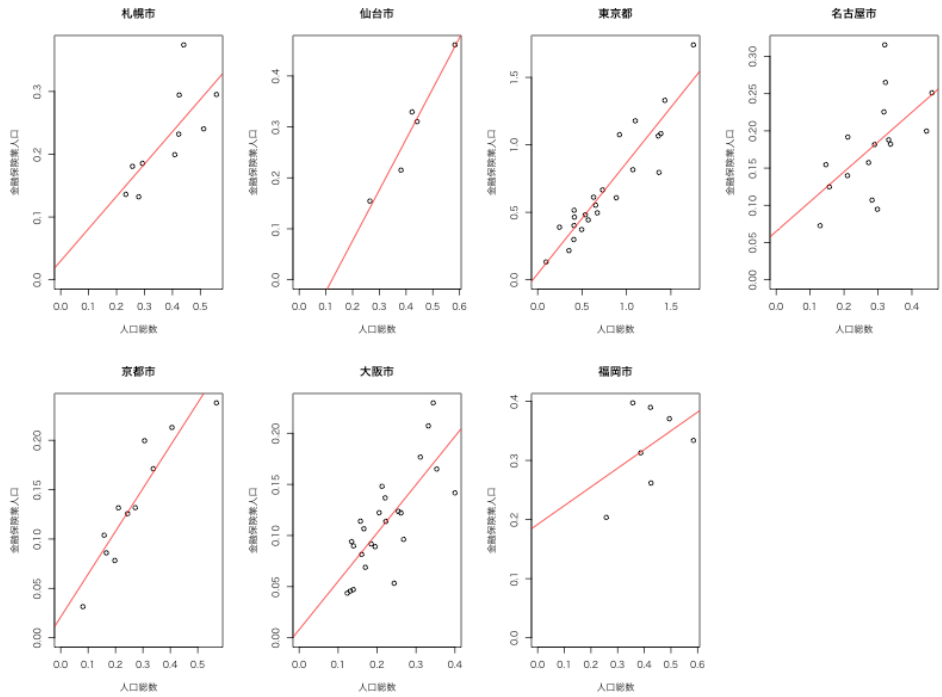

図1　最尤推定法による推定

これは次のスクリプトによって得られたものである。

```
d <- read.csv("input/data_example1_fit.csv")
a.likelihood <- list()
b.likelihood <- list()

likelihood <- function(x) {
  prefecture <- subset(d, prefectures == x)
  H1 <- sum(prefecture$population^2)
  H2 <- sum(prefecture$population)
  H3 <- H2
  H4 <- length(prefecture$population)

  v1 <- sum(prefecture$population * prefecture$finance)
  v2 <- sum(prefecture$finance)

  H <- matrix(c(H1, H2, H3, H4), 2, 2)
  v <- matrix(c(v1, v2), 2, 1)

  HInv <- solve(H)
```

```
  HInv %*% v
}

for (x in 1:7) {
  a.likelihood[[x]] <- likelihood(x)[1]
  b.likelihood[[x]] <- likelihood(x)[2]
}
```

(2)

階層ベイズ法を適用する際、まず Stan という MCMC サンプラーを用いて推定した。

データから σ_Y、a、b、σ_a、σ_b を推定する。Y は各年ごとのパラメータである。a, b の平均を $\hat{a}, \hat{b}$ とし、σ_a と σ_b は a と b を決めるハイパーパラメータとする。また、$PRE[n]$ は各都市における値である。Stan で書いたモデルのモデル式は次の通り。

$$Y[n] \sim \mathbf{Normal}(b[PRE[n]] + a[PRE[n]] \times X[n], \sigma_Y) \tag{11}$$
$$a[k] \sim \mathbf{Normal}(\hat{a}, \sigma_a) \tag{12}$$
$$b[k] \sim \mathbf{Normal}(\hat{b}, \sigma_b) \tag{13}$$

この書き方は、「すべての都市の平均を $\hat{a}$, であり、各都市の $a[k]$ は $\hat{a}$ を平均とした正規分布から生成された」という書き方になっている。$\hat{b}$ と $b[k]$ についても同様である。

推定の結果、グラフと a, b の値は次のようになった。

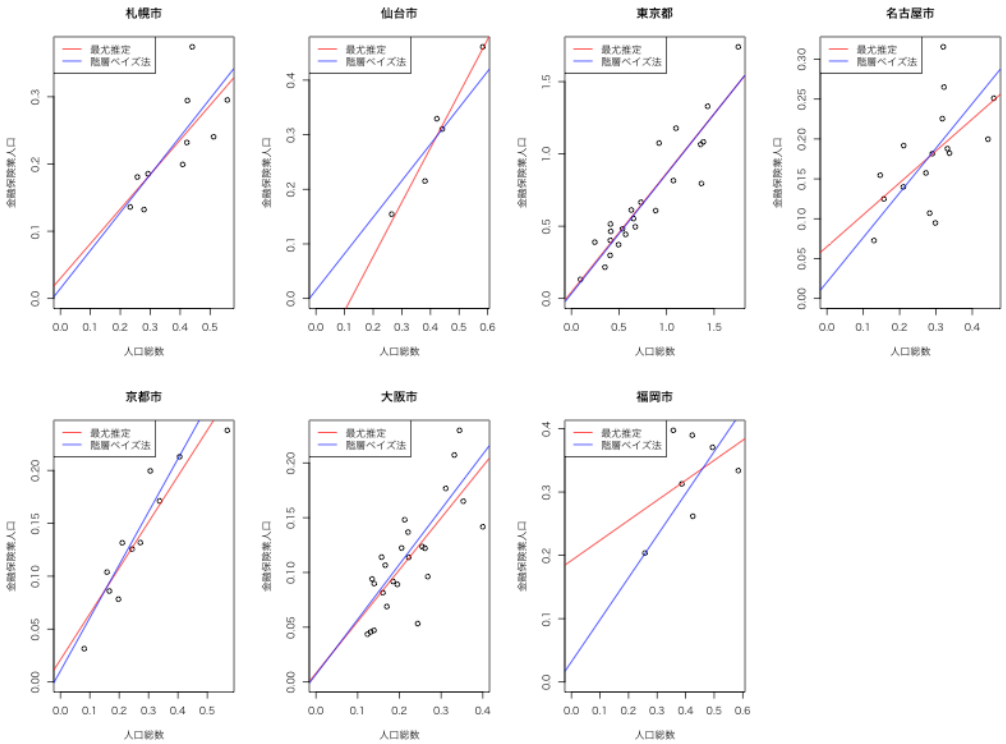

図2　階層ベイズ法 (ハミルトンモンテカルロ法) と最尤推定法の比較

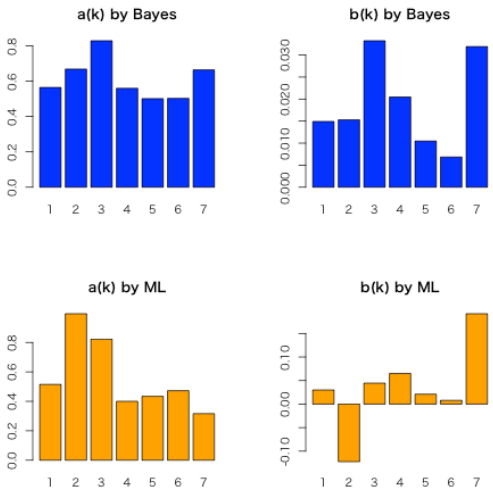

図3　ハミルトンモンテカルロ法と最尤推定法の a と b の値の比較

また、講義で説明されたギブスサンプラー法を用いたサンプリングも試みた。その結果は次のようになった。

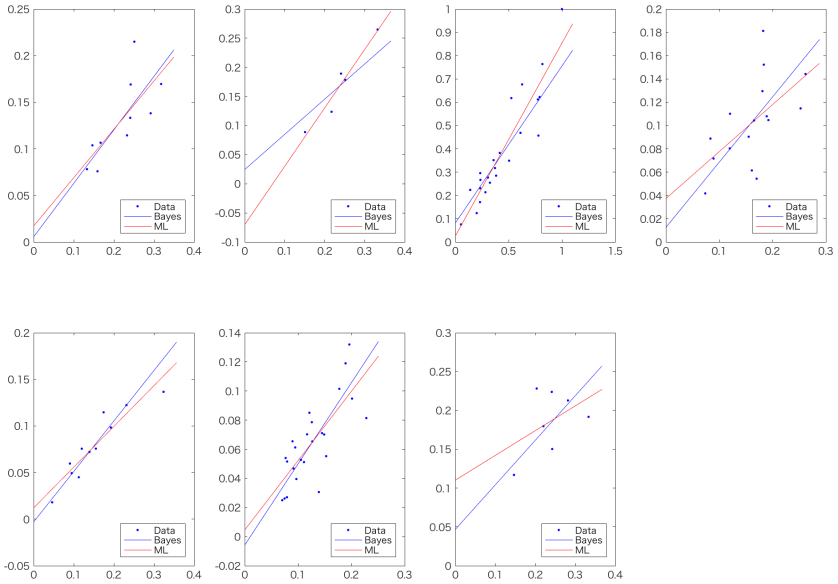

図4　ギブスサンプラー法と最尤推定法の比較

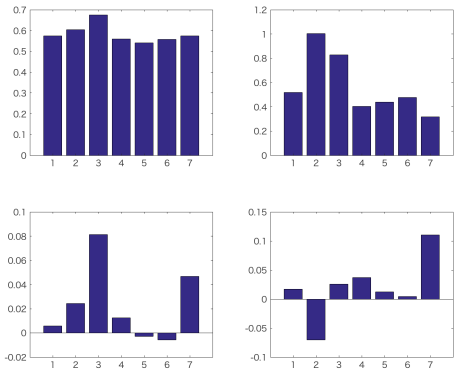

図5　ギブスサンプラー法と最尤推定法の a と b の比較

この 2 つの図を比較すると、ハミルトンモンテカルロ法の方がギブスサンプラー法より b の値が 0 に近づいていることから、ハミルトンモンテカルロ法による階層ベイズ推定の方が収束の度合いが良いと考えられる。

(3)

最も推定が大きく異なる都市は **福岡市**である。考えられる原因は以下の通り、

1. サンプル数が他の年に比べて少なく、ランダムネスによる影響が大きいため。
2. サンプルが少ない上に、点同士が反発しているような配置になっているため、よりランダムネスの影響を受けやすい（名古屋市も反発しているような配置になっているが、サンプル数が福岡市より多いため、最尤推定法と階層ベイズ法の差が福岡市に比べ小さいことより）。

また、もっとも $a(k)$ が大きい都市 k は **東京都**である。

応用問題 2

(1)

次の R スクリプトで計算した結果、

$$m_0 = 0.06476075 \tag{14}$$
$$s_0 = 0.01153776 \tag{15}$$

となった。

```
d <- read.csv('input/data_example2.csv')
X <- d$under15man / d$population
K <- 47
d <- transform(d, ratio=X)

variance <- function(x) var(x) * (length(x) - 1) / length(x)

m0 <- mean(X)
s0 <- sqrt(variance(X))
```

(2)

$$\sum_{k=1}^{47} Y_i = \sum_{i=1}^{1901} X_i \tag{16}$$

であることから、

$$M = \frac{\sum_i X_i}{47} = \frac{\sum_k n(k)}{47} m_0 \tag{17}$$

となる。
また、

$$S = \sqrt{\frac{1}{1901}\sum_i (X_i - m_0)^2} \sum_k \sqrt{n(k)}$$
$$\therefore\ S = s_0 \sum_k \sqrt{n(k)} \tag{18}$$

となる。

(3)

(2) の帰無仮説が正しいとすると $Z_k = (Y_k - M)/S$ は平均 0、標準偏差 1 の正規分布に従う。したがって、$|Z_k| > 2.58$ となる確率は 0.01 である。Z_k を用いて有意水準 0.01 の検定を 47 個の都道府県すべてで行うと、その結果は以下のようになる。

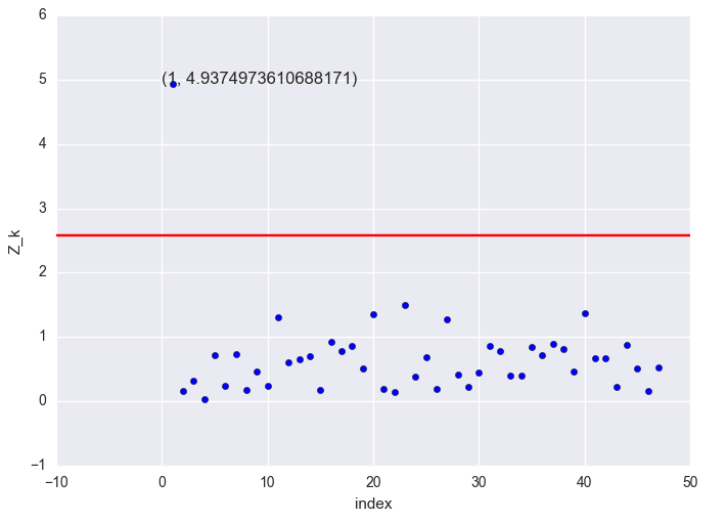

図6　検定の結果

帰無仮説が棄却された都道府県の名前は **都道府県番号 1：北海道**であり、その Z_k の値は **4.94** である。