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

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

を a_k で微分し、極値条件を解く。

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

よって求める答えは (2) より

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

である。

(2)

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

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

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 μ_i と置くと、

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

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。

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

(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。次の環境で解析した。

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

(1.1)　回帰分析

回帰分析とは、1 つ（単回帰分析）または複数（重回帰分析）の説明変数と、1 つの目的変数の関係を求め、説明変数から目的変数を推定する手法である。今回は単回帰分析を用いて解析を行った。

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

$$(\text{転出者数}) = a \cdot (\text{労働人口総数}) + b \tag{12}$$

という仮設を立て、回帰分析を行った。その結果、図 1を得た。

また、pandas の ols 関数で生成したモデルは次のようになった。

Listing 1　モデル 1

```
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
```

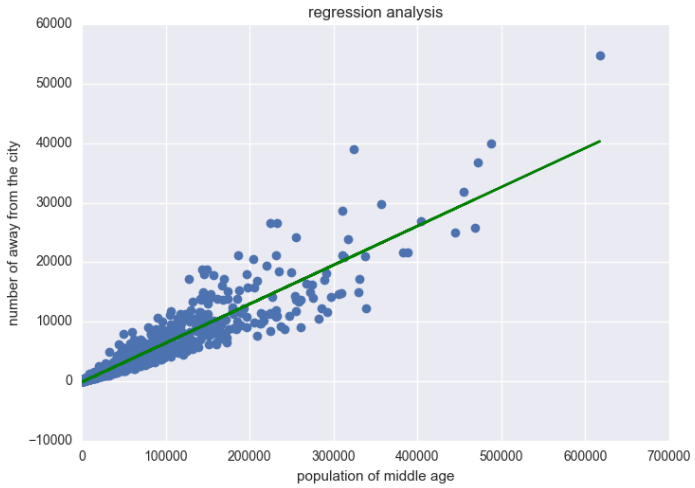

図1　15 歳から 64 歳までの人口総数と転出者

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。ゆえに、この仮設は妥当であると判断できる。

■65 歳以上の総人口と離婚件数の関係　さらに、65 歳以上の総人口（老年人口数）と離婚件数の関係について、

$$(\text{離婚件数}) = a \cdot (\text{老年人口数}) + b \tag{13}$$

という仮設を立て、回帰分析を行った。その結果、図 2を得た。

また、pandas の ols 関数で生成したモデルは次のようになった。

Listing 2　モデル 2

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

Number of Observations:         1870
Number of Degrees of Freedom:   2

R-squared:         0.9359
Adj R-squared:     0.9359

Rmse:              51.1998
```

```
F-stat (1, 1868): 27277.3908, 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      0.0096      0.0001      165.16      0.0000      0.0095      0.0097
intercept    -14.4010      1.4768       -9.75      0.0000    -17.2956    -11.5064
```

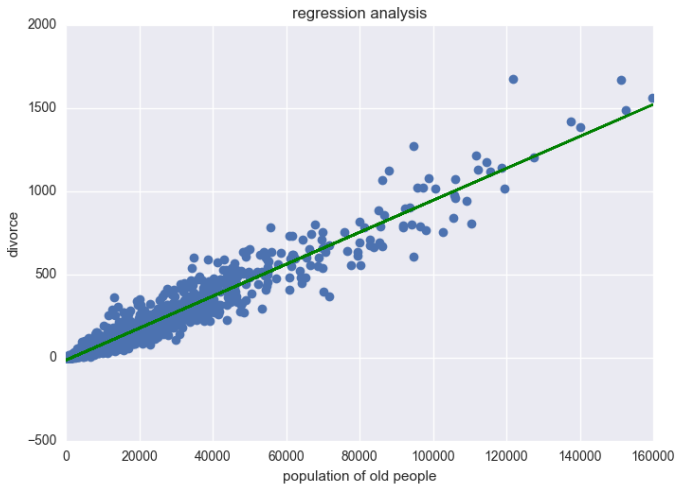

図2

今回の場合、決定係数が 0.9359 とあり、このモデルで 93% 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（27277.3908）、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 0.0000 であるため、両方の値は妥当であるといえる。ゆえに、この仮設は妥当であると判断できる。

(1.2)　主成分分析

主成分分析とは、もとのデータの情報の損失ができるだけ少ない軸を探す（射影したデータの分散が最大となる軸を探す）ための手法である。「市町村 2012estat.csv」に対して主成分分析を行い、どのパラメータが市区町村のデータに影響を与えているかを解析した。

まず、12 次元のデータに対し値を $1+\log(X)$ としてスケール変換した。そして、「村」は赤色、「町」は青色、「市」は緑色、「区」は黄色に塗り分けたところ、図 4が得られた。また、各変数の各主成分への影響力を表す因子負荷量をプロットしたら図 3が得られた。赤は第一主成分、青は第二主成分を表す。

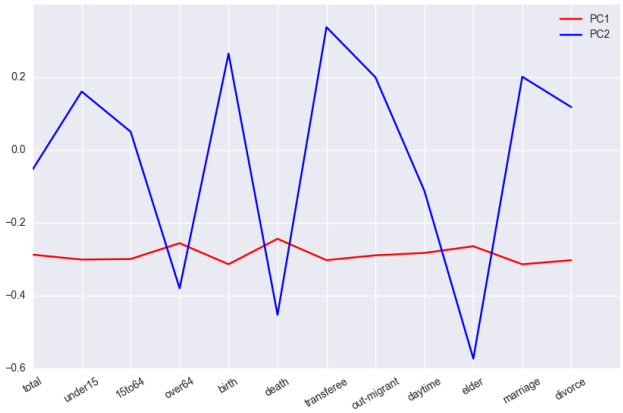

図3　因子負荷量

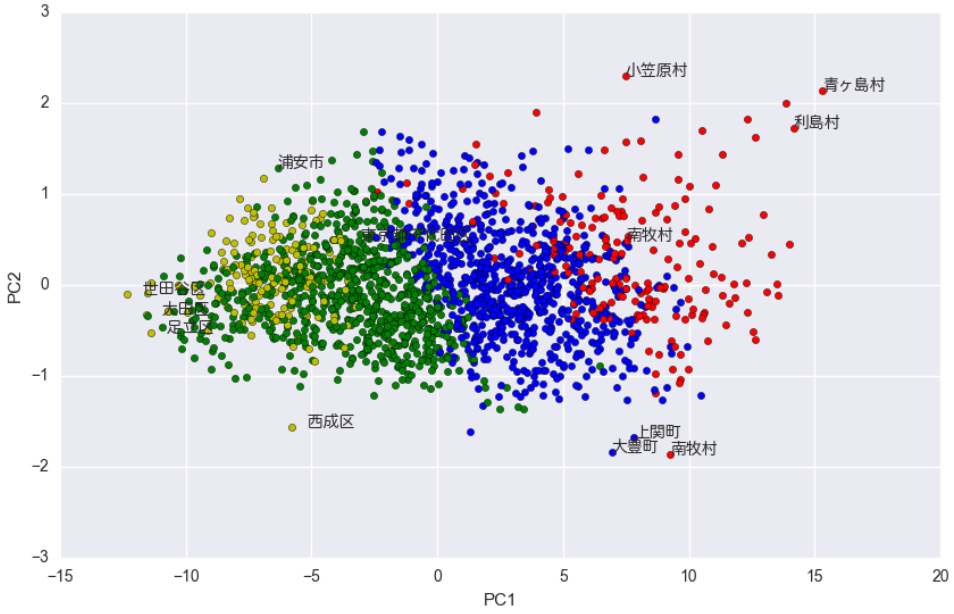

図4　主成分分析（スケール変換後）

累積寄与率は次のようになった。2 つの主成分で 98% も説明出来ているということになる。

Listing 3　累積寄与率

```
>>> np.cumsum(pca.explained_variance_ratio_)
>>> [0.97208879   0.98477677]
```

図 3より、第一主成分は「15 歳から 64 歳までの人口総数」、「出生率」、「転出者数」とそして「婚姻件数」の負荷量が高いことがわかる。よって、第一主成分は住民の文化的営みの度合いに関係があるといえる。また、第二主成分はどれも低い値を示している。これは、どの変数にも共通していることに負の関係があるといえる

(地区の規模などのようなもの)。

図 4を見ると、市区町村がクラスタにそれぞれ分割できているのがわかる。また、第一主成分が高い地区は住民の文化的営みの度合いが高く、第二主成分が低い地区は地区の規模が大きいといえる。

(1.3)　クラスタ分析

ラベル付けがなされていないデータに対して、近しい属性を持つデータをグループ化する手法である。市区町村のデータに対してクラスタ分析を行い、解析を行った結果図 5が得られた。

図5　クラスタ分析

この結果から次のように説明できる。まず、クラスタの分類に寄与しているのは人口総数と昼間人口、そして 15 歳から 64 歳までの人口総数である。その規模に従ってリニアーに 4 つのクラスタに分類され、クラスタ番号 2 に分類された地区が最も人口総数と昼間人口、そして 15 歳から 64 歳までの人口総数が大きく、クラスタ番号 3 に分類された地区が最も小さいと言える。

(2)

解析結果をもとにして日本全国の市区町村についてわかることを挙げる。

- 労働人口総数と転出者数は良い相関があり、仕事を求めていくという理由から労働人口が大きい地区ほど転出者数が増えるといえる。
- 老年人口総数と離婚件数は良い相関があり、老年人口数が大きい地区は離婚件数も増えるといえる。近

年「熟年離婚」が話題になっていることからも頷ける。

- 日本の市区町村は、大まかに 4 つのクラスタに分類され、それは「市区町村」という名前で区別されているものとほぼ一致する。
- 4 つのクラスタは住民の文化的営みと地区の規模という 2 つのパラメータによって分類される。地区の規模の大きさによって「区」「市」「町」「村」の順にリニアに分類される。
- 例えば「小笠原村」や「青ヶ島村」など、人口総数の小さい「村」でも文化的営みの度合いが高いところがある。そこは小さい集落で密度の高い生活が営まれていると推測できる。
- 「区」でも文化的営みが特に小さい地区があるのが興味深い。なお、それはドヤ街で有名な「西成区」である。一方で、同じ地区の規模でも「浦安市」のほうがずっと文化的営みの度合いが高いのは、「ディズニーランド」のようなレジャー施設があるからと推測できる。

参考

- 「Python で PCA」http://i.cla.kobe-u.ac.jp/murao/class/2015-SeminarB3/05_Python_de_PCA.pdf
- 「scikit-learn でクラスタ分析 (K-means 法)」http://pythondatascience.plavox.info/scikit-learn/
- 「Python を使ったデータ解析入門 3idea」https://openbook4.me/projects/183/sections/1366
- 「PyData 入門」http://www.tsjshg.info/intro_pydata/intro01.html