統計解析のための行列計算（１）
～平均と分散の行列を用いた表現～

小松 邦岳

2019/1/16

統計の中で行列計算がどう生かされているのかを自己学習中です。基礎として平均と分散 (標準偏差) の算出を行列を用いて表現してみます。また、SAS、R、Python によるプログラム例も示しました。あくまで私の勉強ノートですから、ご自身の判断でご利用ください。なお、本稿においては、行列で表現することが重要なのであって、平均値についての議論をしたいわけではないことを、ご了承ください。

1　今回用いる計算

今回用いる行列計算を表記法の説明も兼ねて、簡単に紹介いたします。なお、細かい説明はしていませんので、詳しくは清書をご参照ください。

1.1　行列の転置

今回用いるのは、N 行 1 列の行列の転置のみです。転置行列は右上に「T」を付けて表現します。つまり、

$$\mathbb{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

に対して、その転置行列は以下のようになります。[1]

$$\mathbb{X}^{\mathrm{T}} = \begin{pmatrix} x_1 & x_2 & \cdots & x_N \end{pmatrix}$$

[1]外国の講義動画（英語）にて、$\mathbb{X}^{\mathrm{T}}$ のことは「エックスプライム」と呼ばれていました。これが一般的かはわかりませんが … 。

1.2 １行 N 列の行列と N 行１列の行列の内積

１行 N 列の行列と N 行１列の行列の内積を今回は利用します。以下の通り、積の和をとった一つのスカラー量となります。[2]

$$\mathbb{X}_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1N} \end{pmatrix} \quad , \quad \mathbb{X}_2 = \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2N} \end{pmatrix}$$

$$\begin{aligned} \mathbb{X}_1^{\mathrm{T}}\mathbb{X}_2 &= x_{11}x_{21} + x_{12}x_{22} + \cdots + x_{1N}x_{2N} \\ &= \sum_{i=0}^{N} x_{1i}x_{2i} \end{aligned}$$

$\mathbb{X}_1$ は、転置をしていることに注意をしてください。

内積は重要です。ここでは、ほんの一部の性質にしか触れていませんので、わからない、よく知らないという方は必ず清書をご確認ください。

1.3 二乗和の行列表現

N 行１列の行列について、その転置行列との内積をとると二乗和となります。

$$\mathbb{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

$$\begin{aligned} \mathbb{X}^{\mathrm{T}}\mathbb{X} &= x_1^2 + x_2^2 + \cdots + x_N^2 \\ &= \sum_{i=0}^{N} x_i^2 \end{aligned}$$

今回は N 行 1 列ですが、あらゆる行列において、$\mathbb{X}^{\mathrm{T}}\mathbb{X}$ の形は重要です。何度も出てくることになると思います。

[2]内積の表現として ”· ”を付けることもありますが、本稿では採用していません。

2　平均・分散の行列を用いた表現

2.1　平均

観測したN個データを

$$\mathbb{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

とした時、その平均値は1をN個並べた行列$\mathbb{A}$との内積と、データの個数を表すスカラー量Nを用いて、以下のように表現できます。

$$\mathbb{A} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\text{平均} : \overline{X} = \frac{1}{N}\mathbb{A}^{\mathrm{T}}\mathbb{X}$$

計算をしてみますと、以下のように平均値の定義式を導出できることがわかります。

$$\begin{aligned} \text{平均} : \overline{X} = \frac{1}{N}\mathbb{A}^{\mathrm{T}}\mathbb{X} &= \frac{1}{N}\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} \\ &= \frac{1}{N}(x_1 + x_2 + \cdots + x_N) \\ &= \frac{1}{N}\sum_{i=1}^{n} x_i \end{aligned}$$

とても簡単な例ですが、行列計算が統計に活きてくることが実感できたと思います。

2.2　分散（標準偏差）

上記のＮ個のデータの標準偏差は、残差の行列 $\mathbb{R}$ を用いて、以下のように表現できます。

$$\mathbb{R} = \begin{pmatrix} x_1 - \overline{X} \\ x_2 - \overline{X} \\ \vdots \\ x_N - \overline{X} \end{pmatrix}$$

$$\text{分散} : S^2 = \frac{1}{N-1}\mathbb{R}^{\mathrm{T}}\mathbb{R}$$
$$\text{標準偏差} : V = \sqrt{S^2}$$

分散の式を計算してみると、確かに定義式が現れることを確認しましょう。

$$\begin{aligned} S^2 = \frac{1}{N-1}\mathbb{R}^{\mathrm{T}}\mathbb{R} &= \frac{1}{N-1}\begin{pmatrix} x_1 - \overline{X} & x_2 - \overline{X} & \cdots & x_N - \overline{X} \end{pmatrix}\begin{pmatrix} x_1 - \overline{X} \\ x_2 - \overline{X} \\ \vdots \\ x_N - \overline{X} \end{pmatrix} \\ &= \frac{1}{N-1}\left\{\left(x_1 - \overline{X}\right)^2 + \left(x_2 - \overline{X}\right)^2 + \cdots + \left(x_N - \overline{X}\right)^2\right\} \\ &= \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \overline{X})^2 \end{aligned}$$

ここで、$\mathbb{R}^{\mathrm{T}}\mathbb{R}$ は、残差平方和となっていることにも、注目してください。

3　SAS IML、R、Python での実施例

本稿で紹介した計算を SAS[3] 、R[4] 、Python[5]での実施例を示します。もっと上手いプログラム法や便利なパッケージやコマンドもあるかもしれませんが、それはここでは追い求めません。数式に実直に書くことだけ気にしています。4 つのデータ (1,2,3,4) の平均、分散、標準偏差を行列計算を用いて算出します。

3.1　SAS IML

SAS IML による実施例です。

【Program】

```
proc iml;
  reset ;
  N = 4 ;
  A = {1,1,1,1};
  X = {1,2,3,4};
  MEAN= t(A)*X / N ;      /* 1:Mean of X */
  Y   = X-MEAN ;
  VAR = t(Y)*Y / (N-1);   /* 2:Variance of X */
  SD  = sqrt(VAR);        /* 3:Standard deviation of X */
  print A ;
  print X ;
  print Y ;
  print MEAN VAR SD ;
quit;
```

【Output】

A
1
1
1
1

X
1
2
3
4

Y
-1.5
-0.5
0.5
1.5

MEAN	VAR	SD
2.5	1.6666667	1.2909944

[3]SAS Ondemand : Base SAS Software 9.4 M5、SAS/IML 14.3
[4]R version 3.4.0
[5]Python 3.6.1、Anaconda 4.4.0 (x86 64)。実行環境は Jupyter notebook。

3.2 R

R による実施例です。

【Program】

```
N <- 4
A <- c(1,1,1,1)
X <- c(1,2,3,4)
MEAN <- (t(A) %*% X) / N     # 1:Mean of X
Y <- X - c(MEAN)
VAR <- t(Y) %*% Y / (N-1)    # 2:Variance of X
SD  <- sqrt(VAR)             # 3:Standard deviation of X

print(X)
print(A)
print(MEAN)
print(VAR)
print(SD)
```

【Output】

```
> print(X)
[1] 1 2 3 4
> print(A)
[1] 1 1 1 1
> print(MEAN)
     [,1]
[1,]  2.5
> print(VAR)
         [,1]
[1,] 1.666667
> print(SD)
         [,1]
[1,] 1.290994
```

3.3 Python

Python による実施例です。

【Program】

```
import numpy as np
import math
N = 4
A = np.array([[1],[1],[1],[1]])
X = np.array([[1],[2],[3],[4]])
MEAN = np.dot(A.T, X) / N      # 1:Mean of X
Y = X-MEAN
VAR = np.dot(Y.T, Y) / (N-1)   # 2:Variance of X
SD  = math.sqrt(VAR)           # 3:Standard deviation of X

print("X",X,sep='\n')
print("A",A,sep='\n')
print("Mean",MEAN,sep='\n')
print("VAR",VAR,sep='\n')
print("SD",SD,sep='\n')
```

【Output】

```
X
[[1]
 [2]
 [3]
 [4]]
A
[[1]
 [1]
 [1]
 [1]]
Mean
[[ 2.5]]
VAR
[[ 1.66666667]]
SD
1.2909944487358056
```

4　参考文献

1. Rafael A Irizarry and Michael I Love 著「Data Analysis for the Life Sciences with R」Chapman and Hall/CRC

2. Gentle,JamesE. 著「NumericalLinearAlgebraforApplicationsinStatistics(Statistics and Computing)」Springer

5　その他

この文章は私のブログ (KMT92) の記事の一部として執筆しています。ブログは以下の URL となります。

http://kmt92.main.jp/

【連絡先】
info@kmt92.main.jp

【免責事項等】
http://kmt92.main.jp/about-access/