レポート提出票

科目名:	情報工学実験2
実験テーマ:	実験テーマ2 Webアプリケーション
実施日:	2020年 10月 5,12日
学籍番号:	4619055
氏名:	辰川力駆
共同実験者:	

1　はじめに

下全部までです

2　目的

1. 単回帰分析の考え方と手順

   単回帰分析の目的、考え方、手順を理解する

2. 単回帰分析における行列表現

   単回帰分析における行列表現 (線形回帰モデル、正規方程式、最小二乗推定量など) を理解する

3. 実際のデータ解析

   実際のデータに回帰分析を適用することで、解析法を実践的に利用・応用できるようにする

3　実験方法

3.1　実験 1 単回帰分析の考え方と手順

6 つの市町村の人口と行政職員数の仮想データを表 1 に示す。また、各市町村の人口を x_i, 職員数を $y_i(i = 1, ..., n(= 6))$ と表記する。

表 1: 市町村の人口と行政職員数

市町村	人口 x(千人)	職員数 y(人)
A	1	10
B	2	20
C	3	20
D	3	40
E	5	40
F	1	5
合計	15	135

1. 次の統計量を計算する。

$$\sum_{i=1}^{n} x_i, \quad \sum_{i=1}^{n} y_i, \quad \sum_{i=1}^{n} {x_i}^2, \quad \sum_{i=1}^{n} {y_i}^2, \quad \sum_{i=1}^{n} x_i y_i$$

2. 次式が整理することを証明する。

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}{x_i}^2 - n\bar{x}^2 \tag{1}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}{y_i}^2 - n\bar{y}^2 \tag{2}$$

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y} \tag{3}$$

3. 人口 x と職員数 y の基本統計量 (データ数、平均、標準偏差、最小値、最大値) を計算する。

$$\begin{aligned}
\text{人口 } x \text{ の平均} &= \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \\
\text{人口 } x \text{ の標準偏差} &= \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}{x_i}^2-n\bar{x}^2}{n-1}} \\
\text{職員数 } y \text{ の平均} &= \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \\
\text{職員数 } y \text{ の標準偏差} &= \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}{y_i}^2-n\bar{y}^2}{n-1}}
\end{aligned}$$

4. 人口 x と職員数 y の Pearson の積率相関係数 r を計算する。

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{4}$$

5. 人口 x を横軸, 職員数 y を縦軸にした散布図を作成して、両者の関係を調べる。

6. 単回帰モデル $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i (i = 1, ..., n)$ をあてはめる。β_0 と β_1 の推定量を $\hat{\beta}_0$ と $\hat{\beta}_1$ とすると、目的変数 (応答変数) である職員数の予測値は $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ で与えられる。次式の残差平方和 S_e を $\hat{\beta}_0$ と $\hat{\beta}_1$ でそれぞれ偏微分する。

$$S_e = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}(y_i-(\hat{\beta}_0+\hat{\beta}_1x_i))^2 \tag{5}$$

7. 正規方程式を作成する。

8. 正規方程式を解き、β_0 と β_1 の最小二乗推定量を数式で表現する。

9. 最小二乗推定量 $\hat{\beta}_0$,$\hat{\beta}_1$ の値を求める。

10. 得られた回帰直線 ($\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) を手順 5 で作成した散布図に図示して、結果を考察する。

11. データ分析 [回帰分析] を用いて、これまでに得られた結果と同様の結果が得られることを確認する。

3.2　実験 2 単回帰分析における行列表現

データ数を n とする。目的変数ベクトル $\boldsymbol{Y}$ と説明変数を含む定数行列 $\boldsymbol{X}$ を次式で定義する。

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

このとき、実験 1 の単回帰モデルは

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

で与えられる。ここで $\boldsymbol{\beta}$ は母回帰係数,$\boldsymbol{\varepsilon}$ は誤差ベクトルであり、次式で定義される。

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\boldsymbol{\beta}$ の推定量を $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ とする。

1. 残差平方和 S_e を行列で表現する。
2. 残差平方和 S_e を $\hat{\boldsymbol{\beta}}$ で微分し、正規方程式を導く。
3. 正規方程式から最小二乗推定量が

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{6}$$

   で得られることを確認する。
4. ベクトル $\boldsymbol{Y}$ と行列 $\boldsymbol{X}$ を定義する。

5. 次の値を計算する。

   (a) x の平均 $\bar{x}$

   (b) y の平均 $\bar{y}$

   (c) x の偏差平方和 $\bar{x} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

   (d) y の偏差平方和 $\bar{y} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

6. 次の値を計算する。

   (a) 最小二乗推定量 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

   (b) 予測値 $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$

   (c) 残差 $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$

7. 射影行列 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ を計算し、次式が成り立つことを確認する。

   (a) 対称性 $\boldsymbol{H}^T = \boldsymbol{H}$

   (b) べき等性 $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$

   (c) $trace(\boldsymbol{H}) = 2$（パラメータ数）

8. 得られた最小二乗推定量のもとで、総平方和、モデル平方和、残差平方和を計算し

$$\text{総平方和} = \text{モデル平方和} + \text{残差平方和} \tag{7}$$

   が成り立つことを確認する。

9. 寄与率（決定係数）= モデル平方和/総平方和を計算し、モデルの当てはまりを評価する。

4　結果・考察・課題

4.1　実験 1 単回帰分析の考え方と手順

$\boxed{\text{課題 1}}$ 実験 1 の結果をまとめる。

1. 計算すると次のようになった。

$$
\begin{aligned}
\sum_{i=1}^{n} x_i &= 15 \\
\sum_{i=1}^{n} y_i &= 135 \\
\sum_{i=1}^{n} {x_i}^2 &= 49 \\
\sum_{i=1}^{n} {y_i}^2 &= 4125 \\
\sum_{i=1}^{n} x_i y_i &= 435
\end{aligned}
$$

2. 式 (1),(2),(3) が成り立つことを示す。式 (1) の左辺を変形すると、

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

となり、右辺と一致する。同様にして、式 (2) の左辺を変形すると、下記のようになり右辺と一致する。

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i^2-2y_i\bar{y}+\bar{y}^2) \\
&= \sum_{i=1}^{n} y_i^2 - 2n\bar{y}^2 + n\bar{y}^2 \\
&= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2
\end{aligned}
$$

式 (3) も同様の考え方より、

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}(x_iy_i-x_i\bar{y}-y_i\bar{x}+\bar{x}\bar{y}) \\ &= \sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}-n\bar{x}\bar{y}+n\bar{x}\bar{y} \\ &= \sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}\end{aligned}$$

となり、証明できた。

3. データの数は x,y どちらも、6 つである。残りの基本統計量 (平均、標準偏差、最小値、最大値) は以下のようになった。

表 2: 人口と行政職員数の基本統計量

基本統計量	人口 x(千人)	職員数 y(人)
平均	2.5	22.5
標準偏差	1.52	14.75
最小値	1	5
最大値	5	40

4. Pearson の積率相関係数 r は式 (4) より、次のようになった。

$$\begin{aligned}r &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \\ &\fallingdotseq 0.87\end{aligned}$$

5. 散布図を作成すると、次のようになった (後述の手順 10 の直線を含む)。

6. 式 (5) の残差平方和 S_e を $\hat{\beta}_0$ と $\hat{\beta}_1$ でそれぞれ偏微分すると、

$$
\begin{aligned}
\frac{\partial S_e}{\partial \hat{\beta}_0} &= \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\
&= \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^{n} (-\hat{\beta}_0 + (y_i - \hat{\beta}_1 x_i))^2 \\
&= -2 \sum_{i=1}^{n} (-\hat{\beta}_0 + (y_i - \hat{\beta}_1 x_i)) \\
&= 2 \sum_{i=1}^{n} (\hat{\beta}_0 - y_i + \hat{\beta}_1 x_i) \\
&= 2(n\hat{\beta}_0 - \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i) \qquad (8)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial S_e}{\partial \hat{\beta}_1} &= \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\
&= \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^{n} (-\hat{\beta}_1 x_i + (y_i - \hat{\beta}_0))^2 \\
&= -2 \sum_{i=1}^{n} (-\hat{\beta}_1 x_i^2 + (y_i - \hat{\beta}_0) x_i) \\
&= 2 \sum_{i=1}^{n} (\hat{\beta}_1 x_i^2 - x_i y_i + \hat{\beta}_0 x_i) \\
&= 2(\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + \hat{\beta}_0 \sum_{i=1}^{n} x_i) \qquad (9)
\end{aligned}
$$

7. 式 (8),(9) において、

$$
\begin{aligned}
\frac{\partial S_e}{\partial \hat{\beta}_0} &= 0 \\
\frac{\partial S_e}{\partial \hat{\beta}_1} &= 0
\end{aligned}
$$

とするのでそれぞれ整理すると、

$$
\begin{aligned}
\frac{\partial S_e}{\partial \hat{\beta}_0} &= 0 \\
2(n\hat{\beta}_0 - \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i) &= 0 \\
n\hat{\beta}_0 - \sum_{i=1}^{n} y_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= 0 \\
\sum_{i=1}^{n} y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \qquad (10)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial S_e}{\partial \hat{\beta}_1} &= 0 \\
2(\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + \hat{\beta}_0 \sum_{i=1}^{n} x_i) &= 0 \\
\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + \hat{\beta}_0 \sum_{i=1}^{n} x_i &= 0 \\
\sum_{i=1}^{n} x_i y_i &= \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i \qquad (11)
\end{aligned}
$$

よって正規方程式ができた。

8. 式 (10),(11) の連立方程式を解く。式 (10) を変形すると、

$$
\begin{aligned}
\sum_{i=1}^{n} y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \\
n\hat{\beta}_0 &= \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\
\hat{\beta}_0 &= \frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \qquad (12)
\end{aligned}
$$

となるので、式 (12) を式 (11) に代入すると、

$$
\begin{aligned}
\sum_{i=1}^{n} x_i y_i &= \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i \\
\sum_{i=1}^{n} x_i y_i &= \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + (\frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n}) \sum_{i=1}^{n} x_i \\
n\sum_{i=1}^{n} x_i y_i &= n\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + (\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i) \sum_{i=1}^{n} x_i \\
n\sum_{i=1}^{n} x_i y_i &= n\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - \hat{\beta}_1 (\sum_{i=1}^{n} x_i)^2 \\
n\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \hat{\beta}_1 (\sum_{i=1}^{n} x_i)^2 &= n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \\
\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \hat{\beta}_1 (\sum_{i=1}^{n} x_i)^2/n &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i/n \\
\hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n) &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i/n \\
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i/n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n} \qquad (13)
\end{aligned}
$$

また、$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ を用いて、

$$\begin{aligned}
\sum_{i=1}^{n} \hat{y}_i &= \sum_{i=1}^{n} \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\beta}_1 x_i \\
n\bar{y} &= n\hat{\beta}_0 + n\hat{\beta}_1 \bar{x} \\
\bar{y} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \qquad (14)
\end{aligned}$$

となり、最小二乗推定量 $\hat{\beta}_0$,$\hat{\beta}_1$ を数式で表現することができた。

9. 式 (13),(14) に求めた統計量を代入すると

$$\begin{aligned}
\hat{\beta}_0 &= 1.3043 \\
\hat{\beta}_1 &= 8.4783
\end{aligned}$$

となり、最小二乗推定量 $\hat{\beta}_0$,$\hat{\beta}_1$ の値を求めることができた。

10. 図 1 の散布図に直線を図示した。データ数は少ないが、直線の傾きより、人口と行政職員数で正の相関があると言える。

11. データ分析 [回帰分析] を用いると、下記のようになり、同様の結果を得ることができた。

表 3: 人口と行政職員数の回帰分析

回帰統計	数値
重相関 R	0.87
	係数
切片	1.3043
X 値 1	8.4783

課題 2 公表されてるデータ (標本数が 50 以上) を集めて、回帰分析を適用し、結果を考察する。

表 4: 労働力人口のデータ

年月	男 (万人)	女 (万人)	年月	男 (万人)	女 (万人)
平成 28 年 1 月	3784	2896	平成 30 年 5 月	3816	3008
平成 28 年 2 月	3774	2874	平成 30 年 6 月	3816	2997
平成 28 年 3 月	3758	2871	平成 30 年 7 月	3809	3010
平成 28 年 4 月	3774	2869	平成 30 年 8 月	3811	3023
平成 28 年 5 月	3780	2871	平成 30 年 9 月	3815	3021
平成 28 年 6 月	3784	2900	平成 30 年 10 月	3822	3034
平成 28 年 7 月	3783	2910	平成 30 年 11 月	3843	3039
平成 28 年 8 月	3782	2905	平成 30 年 12 月	3830	3022
平成 28 年 9 月	3781	2900	平成 31 年 1 月	3810	3033
平成 28 年 10 月	3788	2897	平成 31 年 2 月	3829	3044
平成 28 年 11 月	3784	2899	平成 31 年 3 月	3836	3057
平成 28 年 12 月	3799	2915	平成 31 年 4 月	3822	3050
平成 29 年 1 月	3795	2916	令和元年 5 月	3820	3046
平成 29 年 2 月	3772	2901	令和元年 6 月	3824	3046
平成 29 年 3 月	3769	2898	令和元年 7 月	3829	3050
平成 29 年 4 月	3777	2918	令和元年 8 月	3834	3056
平成 29 年 5 月	3786	2938	令和元年 9 月	3827	3069
平成 29 年 6 月	3782	2947	令和元年 10 月	3834	3082
平成 29 年 7 月	3792	2947	令和元年 11 月	3839	3075
平成 29 年 8 月	3793	2953	令和元年 12 月	3836	3085
平成 29 年 9 月	3793	2951	令和 2 年 1 月	3835	3067
平成 29 年 10 月	3786	2946	令和 2 年 2 月	3839	3072
平成 29 年 11 月	3775	2963	令和 2 年 3 月	3839	3064
平成 29 年 12 月	3780	2968	令和 2 年 4 月	3810	2993
平成 30 年 1 月	3796	2971	令和 2 年 5 月	3802	3021
平成 30 年 2 月	3805	3000	令和 2 年 6 月	3803	3027
平成 30 年 3 月	3814	3019	令和 2 年 7 月	3828	3017
平成 30 年 4 月	3820	3024	令和 2 年 8 月	3831	3034

参考文献 [2] の日本の労働力人口についてのデータについて、表 4 にまとめた。データ分析 [回帰分析] を用いると、下記の表 5 のようになった。重相関は 0.89 であった。

よって、重相関が 1 に近いことから、男性と女性では労働者人口の比ははほぼ変わらなく、正の相関があるといえる。

表 5: 人口と行政職員数の回帰分析

回帰統計	数値
重相関 R	0.89
	係数
切片	-5643.7114
X 値 1	2.2673

4.2 実験 2 単回帰分析における行列表現

課題 1 実験 2 の結果をまとめる。

1. 単回帰モデルは、

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

と表現できることから、

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X\beta}$$

となるので、残差平方和はこれを二乗して、

$$\begin{aligned} S_e &= ||\boldsymbol{\varepsilon}||^2 \\ &= ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{y}^T - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T)(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{aligned} \tag{15}$$

となる。

2. 式 (15) を $\hat{\boldsymbol{\beta}}$ で微分すると、

$$\begin{aligned} \frac{dS_e}{d\hat{\boldsymbol{\beta}}} &= \frac{d}{d\hat{\boldsymbol{\beta}}}(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} + \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= -\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{y} + (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} \\ &= -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{aligned} \tag{16}$$

となった。$\frac{dS_e}{d\hat{\boldsymbol{\beta}}} = 0$ として整理すると下記のようになる。

$$\begin{aligned} \frac{dS_e}{d\hat{\boldsymbol{\beta}}} &= 0 \\ -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} &= 0 \\ \boldsymbol{X}^T\boldsymbol{y} &= \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{aligned} \tag{17}$$

3. 式 (17) の正規方程式より、

$$
\begin{aligned}
\boldsymbol{X}^T\boldsymbol{y} &= \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} \\
\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}^T\boldsymbol{y} \\
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \qquad (18)
\end{aligned}
$$

となる。式 (18) は確かに式 (6) と一致した。

4. ベクトル $\boldsymbol{Y}$ と行列 $\boldsymbol{X}$ は次のように定義した。

```
(Y <- matrix(c(10,20,20,40,40,5),nrow=6,ncol=1))

(X <- matrix(c(rep(1,6),1,2,3,3,5,1),nrow=6,ncol=2))
```

出力は次のようになった。

```
     [,1]
[1,]   10
[2,]   20
[3,]   20
[4,]   40
[5,]   40
[6,]    5
     [,1] [,2]
[1,]    1    1
[2,]    1    2
[3,]    1    3
[4,]    1    3
[5,]    1    5
[6,]    1    1
```

5. 手順 5 の R のソースコードは次のようになった。

```
#(a)
xbar <- mean(X[,2])
#(b)
ybar <- mean(Y)
##結果表示
data.frame(x.mean = xbar, y.mean = ybar)

#(c)
xSSD <- sum((X[,2]-xbar)^2)
#(d)
ySSD <- sum((Y-ybar)^2)
##結果表示
data.frame(x.SSD = xSSD, y.SSD = ySSD)
```

結果は次のようになった。

```
  x.mean y.mean
1    2.5   22.5
  x.SSD  y.SSD
1  11.5 1087.5
```

6. 手順６のＲのソースコードは次のようになった。

```r
#(a)
(beta <- solve(t(X) %*% X) %*% t(X) %*% Y)
#(b)
yhat <- X %*% beta
#(c)
e <- Y - yhat
##結果表示
data.frame(y=Y,yhat=yhat,e=e)
```

結果は次のようになった。

```
          [,1]
[1,] 1.304348
[2,] 8.478261
   y      yhat          e
1 10  9.782609  0.2173913
2 20 18.260870  1.7391304
3 20 26.739130 -6.7391304
4 40 26.739130 13.2608696
5 40 43.695652 -3.6956522
6  5  9.782609 -4.7826087
```

7. 射影行列を以下のように定義する。

```r
##射影行列の定義
(H <- X %*% solve(t(X) %*% X) %*% t(X))
```

結果は以下のようになる。

```
           [,1]       [,2]      [,3]      [,4]        [,5]       [,6]
[1,]  0.3623188 0.23188406 0.1014493 0.1014493 -0.15942029  0.3623188
[2,]  0.2318841 0.18840580 0.1449275 0.1449275  0.05797101  0.2318841
[3,]  0.1014493 0.14492754 0.1884058 0.1884058  0.27536232  0.1014493
[4,]  0.1014493 0.14492754 0.1884058 0.1884058  0.27536232  0.1014493
[5,] -0.1594203 0.05797101 0.2753623 0.2753623  0.71014493 -0.1594203
[6,]  0.3623188 0.23188406 0.1014493 0.1014493 -0.15942029  0.3623188
```

(a) 対称性 $\boldsymbol{H}^T = \boldsymbol{H}$ について、R のソースコードは次のようになった。

```
#(a)
t(H)
##対称性が成立しているか確認
sum((t(H)−H)^2)
```

結果は次のようになり、最後の行で差を計算しているが R による丸め誤差しか存在しないので確かに対称性は成り立つ。

```
           [,1]       [,2]      [,3]      [,4]        [,5]       [,6]
[1,]  0.3623188 0.23188406 0.1014493 0.1014493 −0.15942029  0.3623188
[2,]  0.2318841 0.18840580 0.1449275 0.1449275  0.05797101  0.2318841
[3,]  0.1014493 0.14492754 0.1884058 0.1884058  0.27536232  0.1014493
[4,]  0.1014493 0.14492754 0.1884058 0.1884058  0.27536232  0.1014493
[5,] −0.1594203 0.05797101 0.2753623 0.2753623  0.71014493 −0.1594203
[6,]  0.3623188 0.23188406 0.1014493 0.1014493 −0.15942029  0.3623188
[1] 9.398538e−32
```

(b) べき等性 $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ について、R のソースコードは次のようになった。

```
#(b)
H %*% H
##べき等性が成立しているか確認
sum((H %*% H−H)^2)
```

結果は次のようになり、(a) と同様に最後の行で差を計算しているが R による丸め誤差しか存在しないので確かにべき等性は成り立つ。

```
           [,1]       [,2]      [,3]      [,4]        [,5]       [,6]
[1,]  0.3623188 0.23188406 0.1014493 0.1014493 −0.15942029  0.3623188
[2,]  0.2318841 0.18840580 0.1449275 0.1449275  0.05797101  0.2318841
[3,]  0.1014493 0.14492754 0.1884058 0.1884058  0.27536232  0.1014493
[4,]  0.1014493 0.14492754 0.1884058 0.1884058  0.27536232  0.1014493
[5,] −0.1594203 0.05797101 0.2753623 0.2753623  0.71014493 −0.1594203
[6,]  0.3623188 0.23188406 0.1014493 0.1014493 −0.15942029  0.3623188
[1] 2.000078e−31
```

(c) $trace(\boldsymbol{H}) = 2$ について、R のソースコードは次のようになった。

```
#(c)
sum(diag(H))
```

結果は以下のようになり、確かに $trace(\boldsymbol{H}) = 2$ は成り立つ。

```
[1] 2
```

8. 手順8のRのソースコードは次のようになった。

```
##モデル平方和
MSS <- sum((yhat - mean(yhat))^2)
##残差平方和
RSS <- sum(e^2)
##総平方和
TSS <- MSS + RSS
##結果表示
data.frame(ySSD = ySSD, TSS = TSS)
```

結果は以下のようになり、等しくなるので確かに式 (7) は成り立つ。

```
    ySSD     TSS
1 1087.5  1087.5
```

9. 寄与率（決定係数）に関して、Rのソースコードは次のようになった。

```
##寄与率
MSS/TSS
R2 <- MSS/TSS
##相関係数の絶対値
sqrt(MSS/TSS)
```

結果は以下のようになった。上が寄与率で、下が相関係数の絶対値である。

```
[1]  0.7601199
[1]  0.8718486
```

課題2 行列を用いて統計演算を行う利点を考察する。

今回は、行列を使わずに S_e を偏微分をすることで求める方法と、行列を用いて S_e を行列で微分する方法で回帰分析をしたが、最小二乗推定量 $\hat{\beta}_0,\hat{\beta}_1$ を求めるときに明らかに行列のほうが簡潔に楽に計算することができた。説明変数がさらに増えて、もっと複雑になった場合はさらに簡潔度において差が生まれると考える。

よって、簡潔に計算できることが行列を用いて統計演算を行う利点だと考える。

5 まとめ

1. 単回帰分析の考え方と手順を学んだ
   - 手計算やエクセルで分析を行った
2. 単回帰分析における行列表現を学んだ
   - 実験1の手順を行列表現した
   - Rを使い、単回帰分析を行った

6　感想

初めてRという言語を使ってみたが、少し触れる程度しかしなかったのでさらに自分でRを使ってみたい。統計解析に使われる他の言語として、PythonやStanが挙げられるがそれらとの違いについても調べてみようと考える。

参考文献

[1] 東京理科大学工学部情報工学科 情報工学実験2 2020年度東京理科大学工学部情報工学科出版

[2] 統計局ホームページ/労働力調査　長期時系列データ

http://116.91.128.50/data/roudou/longtime/03roudou.html#hyo_1

最終閲覧日 2020/10/3