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4 統計的推測と単回帰分析

実施日: 2020年 9月 28日

学籍番号: 4619055

氏名: 辰川力駆

共同実験者:

1　はじめに

本実験では、単回帰分析の考え方と手順を理解することを目標とする。

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。

$$\begin{aligned} \sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}{x_i}^2-n\bar{x}^2 &(1)\\ \sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}{y_i}^2-n\bar{y}^2 &(2)\\ \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y} &(3) \end{aligned}$$

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。

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

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 $y_i=\beta_0+\beta_1x_i+\varepsilon_i(i=1,...,n)$ 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 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$ 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。

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

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 ($\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$) 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

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. 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。

$$S_e = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}(y_i-(\hat{\beta}_0+\hat{\beta}_1x_i))^2 \tag{8}$$

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。

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。

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。

課題 2 行列を用いて統計演算を行う利点を考察する。

ぱ あ

ソースコード 1: read_2_1byte.c

```
#include <stdio.h>
```

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

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