応用統計学 2018 第3回

2018年10月17日 (水) ver. 1.1

清 智也 sei@mist.i.u-tokyo.ac.jp

http://www.stat.t.u-tokyo.ac.jp/~sei/lec-j.html

3 最小二乗法，重回帰分析

3.1 最小二乗法

ある変数 y を別の変数 x の一次式で近似したい場面が応用上しばしばある。たとえば x が模試の点数，y がセンター試験の点数の場合を想定するとよい。

近似として求めたい一次式を

$$y = a + bx, \quad a, b \in \mathbb{R}, \tag{1}$$

とおき，これを回帰式 (regression equation) という。x のことを説明変数 (explanatory variable)，y のことを目的変数 (objective variable) あるいは応答変数 (response variable) と呼ぶ。また a, b を回帰係数 (regression coefficient) という。

実際に観測値 $\boldsymbol{x} = (x_1, \ldots, x_n)', \boldsymbol{y} = (y_1, \ldots, y_n)'$ が得られたとき，式 (1) による近似の「悪さ」を

$$f(a, b) = \sum_{t=1}^{n} (y_t - a - bx_t)^2 \tag{2}$$

という量で測り，これを最小にするように a, b を定めることが考えられる。この方法を最小二乗法 (least squares method) と呼ぶ。

定理 3.1. $\boldsymbol{x}$ の標本分散が正であれば最小二乗法の解 $(\hat{a}, \hat{b})$ は存在し，

$$\hat{a} = \bar{y} - \hat{b}\bar{x}, \quad \hat{b} = \frac{\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})}{\mathrm{V}(\boldsymbol{x})} = r_{xy}\frac{s_y}{s_x}$$

で与えられる。 ただし $\bar{x}, \bar{y}$ は $\boldsymbol{x}, \boldsymbol{y}$ の平均，s_x, s_y は標準偏差，r_{xy} は相関係数を表す。

証明は演習問題である。定理より，各変数があらかじめ標準化されているとき，回帰式の切片は0，傾きは相関係数となることが分かる。

回帰式に $\hat{a}, \hat{b}$ を代入して得られる

$$\hat{y} = \hat{y}(x) = \hat{a} + \hat{b}x$$

のことを予測値 (predicted value) という。また観測値から予測値を引いた値

$$\begin{aligned} r_t &= y_t - \hat{y}_t \\ &= y_t - (\hat{a} + \hat{b}x_t) \end{aligned}$$

のことを残差 (residual) という。

3.2 適用例

たとえば表 1 のデータを考える。これは第 1 回に実施した講義アンケートの結果（10 月 8 日現在）のうち，通学時間と睡眠時間の項目のみ示したものである[1]。いま，睡眠時間を通学時間の一次式で表すことに興味があるとしよう。

表 1: 学生の通学時間と睡眠時間

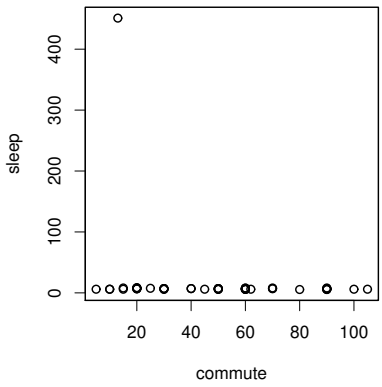
No.	通学時間 [分]	睡眠時間 [時間]	No.	通学時間 [分]	睡眠時間 [時間]
1	40	7	26	40	7
2	20	8	27	30	7
3	20	7	28	100	6
4	70	8	29	20	6.5
5	60	7.5	30	60	6
6	70	7	31	60	6
7	60	7	32	30	6
8	105	6	33	45	6
9	50	7	34	10	6
10	80	5.5	35	90	7
11	5	6	36	20	7
12	20	8.5	37	70	7
13	60	6	38	30	7
14	50	6	39	20	8
15	62	6	40	90	6
16	60	8	41	25	7.5
17	30	7	42	50	6
18	15	6	43	90	5.5
19	30	6	44	50	6
20	10	6	45	15	8
21	90	6	46	50	7
22	13	451	47	50	7
23	60	6	48	90	6.5
24	60	7	49	90	8
25	60	6			

2 変数の量的データの記述でもっとも大事なのは散布図を描くことであると第 1 回の講義で述べた。上記のデータに対する散布図を描くと，図 1 (a) が得られる。明らかに 1 点の外れ値が認められる[2]。外れ値の原因を探ることも大事だが，ここでは単に外れ値を除いて考えることにする。すると散布図は同図 (b) のようになる。二つの変数の間にはあまり関係がなさそうである。しかし少しだけ右下がりになっているようにも見える。

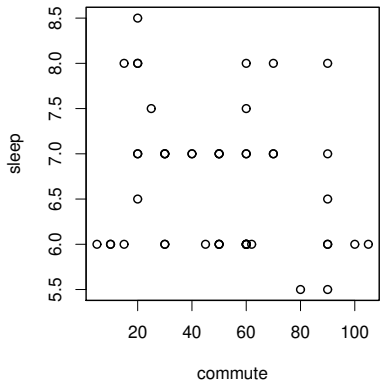
実際に表 1 のデータ（外れ値を除く）に対して回帰係数を計算すると表 2 および図 2 のようになった。得られた直線の傾きは -0.0066 であるから，通学時間が 1 分増えると睡眠時間が 0.0066 時間（$= 0.4$ 分）短くなる傾向がある。ただし，この結論にはあまり「説明力」がないことを次節で確認する。

[1] アンケートにおける質問は，それぞれ「おおよその通学時間を分単位で記入してください。」,「1 日あたりのおおよその睡眠時間を時間単位で記入してください。」であった。

[2] このような外れ値の個体番号を知りたいとき，R では identify という関数が便利である。



(a) 散布図



(b) 外れ値を除いたあとの散布図

図 1: 通学時間に対する睡眠時間の散布図

表 2: 通学時間 x と睡眠時間 y に対する回帰分析

統計量	データから求めた値
n	48
$\bar{x}$	50.25
$\bar{y}$	6.68
s_x	26.69
s_y	0.77
r_{xy}	-0.23
$\hat{a}$	7.01
$\hat{b}$	-0.0066

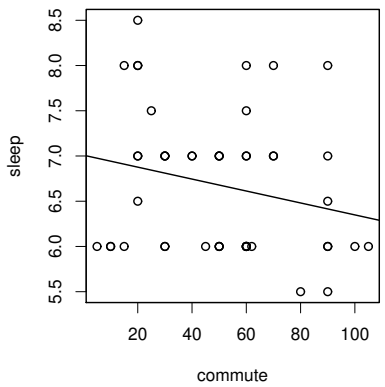


図 2: 最小二乗法による当てはめ結果（説明力は低い）

3.3 決定係数

定理 3.1 で求まる予測値 $\hat{y}_t$ に対して

$$\sum_{t=1}^{n}(y_t-\bar{y})^2=\sum_{t=1}^{n}(y_t-\hat{y}_t)^2+\sum_{t=1}^{n}(\hat{y}_t-\bar{y})^2 \tag{3}$$

という分解が成り立つ。これを平方和の分解という[3]。導出は演習問題とする。右辺第 2 項は予測値の平方和を表しており，第 1 項は残差の平方和を表している。左辺を全平方和という。

そこで，全平方和のうち説明変数によって説明できた割合を

$$R^2=\frac{\sum_{t=1}^{n}(\hat{y}_t-\bar{y})^2}{\sum_{t=1}^{n}(y_t-\bar{y})^2}$$

とおき，これを決定係数 (coefficient of determination) という。決定係数は観測値と予測値の相関係数の 2 乗となる：

$$R^2=\{\mathrm{Corr}(\boldsymbol{y},\hat{\boldsymbol{y}})\}^2. \tag{4}$$

その理由は次節で説明する。また $R=\mathrm{Corr}(\boldsymbol{y},\hat{\boldsymbol{y}})$ は非負となり，これを重相関係数 (multiple correlation coefficient) という。

表 1 の例では，決定係数はおよそ 0.05 となり，説明力があるとは言い難い。一般に決定係数がどのくらいあれば説明力があるか，という明確な基準はないが，0.5 を超えるかどうかが一つの目安にはなるだろう。

応用上，特に 3.5 節で述べる重回帰分析においては，決定係数だけでなく残差プロット（残差を縦軸，予測値を横軸にとった散布図）を検討することも重要である。

また別の観点として，傾き $\hat{b}$ の標準誤差も考慮すべきである。もし標準誤差が絶対値 $|\hat{b}|$ に比べて大きいようであれば，母集団レベルでは 2 つの変数がそもそも無相関である可能性がある。この点については第 10 回の講義で詳しく扱うが，前回学んだブートストラップ法を適用することも可能である。

3.4 幾何学的な理解

最小二乗法はユークリッド空間 $\mathbb{R}^n$ における直交射影である。これを説明する。

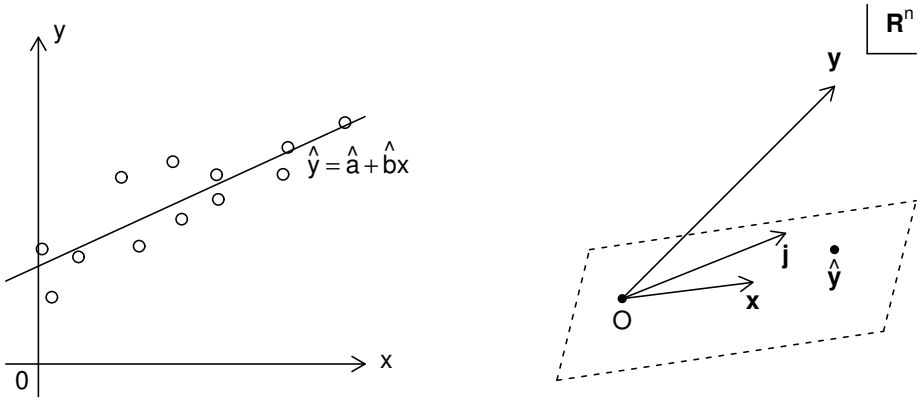
まず，各変量 $\boldsymbol{x}=(x_1,\ldots,x_n)',\boldsymbol{y}=(y_1,\ldots,y_n)'$ は $\mathbb{R}^n$ の点と見なせる。このように，各変量が属す空間 $\mathbb{R}^n$ を変量空間と呼ぶことがある[4]。対照的に，散布図のように各個体が属す空間を個体空間と呼ぶ。

[3]データから平均を引いた値（偏差）の平方和を分解しているので，正確には偏差平方和の分解という。
[4]柴田里程「データ分析とデータサイエンス」，近代科学社。

変量空間で考えると，式 (2) は

$$f(a,b) = \|\boldsymbol{y} - a\boldsymbol{j} - b\boldsymbol{x}\|^2$$

と書き換えられる。ただし $\boldsymbol{j} = (1,\ldots,1)' \in \mathbb{R}^n$ とおいた。また $\|\cdot\|$ はユークリッドノルムを表す。したがって $f(a,b)$ を最小にするということは変量空間においてベクトル $\boldsymbol{y}$ を $\boldsymbol{j}$ と $\boldsymbol{x}$ の張る 2 次元線形部分空間 $\mathrm{span}(\boldsymbol{j},\boldsymbol{x})$ に直交射影させることと同じである。そして直交射影された点は予測値 $\hat{\boldsymbol{y}} = \hat{a}\boldsymbol{j} + \hat{b}\boldsymbol{x}$ となる（図 3）。



(a) 個体空間　　(b) 変量空間

図 3: 個体空間と変量空間

同じように，ベクトル $\boldsymbol{y}$ を $\mathrm{span}(\boldsymbol{j})$ に直交射影すると $\bar{y}\boldsymbol{j}$（全ての成分が $\bar{y}$ であるようなベクトル）となる。すると，直交射影の性質から

$$\|\boldsymbol{y} - \bar{y}\boldsymbol{j}\|^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 + \|\hat{\boldsymbol{y}} - \bar{y}\boldsymbol{j}\|^2$$

が導かれる（いわゆる三垂線の定理）。これが式 (3) に対応する。

また，変量空間では 2 つの変数の相関係数は「余弦」を用いて表される。実際，$\boldsymbol{y}$ と $\boldsymbol{x}$ の相関係数は

$$\mathrm{Corr}(\boldsymbol{y},\boldsymbol{x}) = \frac{(\boldsymbol{y} - \bar{y}\boldsymbol{j})'(\boldsymbol{x} - \bar{x}\boldsymbol{j})}{\|\boldsymbol{y} - \bar{y}\boldsymbol{j}\|\|\boldsymbol{x} - \bar{x}\boldsymbol{j}\|}$$

と表され，これは 2 つの単位ベクトル

$$\boldsymbol{e}_y = \frac{\boldsymbol{y} - \bar{y}\boldsymbol{j}}{\|\boldsymbol{y} - \bar{y}\boldsymbol{j}\|} \quad \boldsymbol{e}_x = \frac{\boldsymbol{x} - \bar{x}\boldsymbol{j}}{\|\boldsymbol{x} - \bar{x}\boldsymbol{j}\|}$$

の内積になっている。この考えをもとに，決定係数に関する等式 (4) を証明しよう。まず $\mathrm{Corr}(\boldsymbol{y},\hat{\boldsymbol{y}})$ は，$\boldsymbol{e}_y$ と $\boldsymbol{e}_{\hat{y}}$ の内積であるから，この 2 つのベクトルの成す角を θ とすれば $\mathrm{Corr}(\boldsymbol{y},\hat{\boldsymbol{y}}) = \cos\theta$ である。一方，$(\boldsymbol{y} - \hat{\boldsymbol{y}}) \perp (\hat{\boldsymbol{y}} - \bar{y}\boldsymbol{j})$ であるから

$$\|\hat{\boldsymbol{y}} - \bar{y}\boldsymbol{j}\| = \|\boldsymbol{y} - \bar{y}\boldsymbol{j}\|\cos\theta$$

が成り立つ。よって決定係数の定義から式 (4) が成り立つ。

3.5 重回帰分析

説明変数が複数個あっても基本的な考え方は同じである。目的変数 y を p 個の説明変数 $x_1, \dots, x_p$ の一次式

$$y = b_1 x_1 + \cdots + b_p x_p$$

で近似する問題を考える。ここで定数項を省略しているのは，必要に応じて $x_1 = 1$（定数）とおけばよいからである。慣習として，説明変数が 2 個以上の場合の回帰分析を重回帰分析 (multiple regression analysis)，説明変数が 1 個の場合を単回帰分析 (simple regression analysis) という。

目的変数の観測値を $\boldsymbol{y} \in \mathbb{R}^n$，説明変数の観測値を並べてできる行列を $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ とおく。この $\boldsymbol{X}$ を計画行列 (design matrix) と呼ぶ。

一次式による近似の悪さは

$$\begin{aligned} f(\boldsymbol{\beta}) &= \sum_{t=1}^{n} \Big(y_t - \sum_{i=1}^{p} \beta_i x_{ti} \Big)^2 \\ &= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \end{aligned}$$

と表される。

定理 3.2. $\boldsymbol{X}$ は列フルランクとする。このとき $f(\boldsymbol{\beta})$ を最小にする $\boldsymbol{\beta}$ は

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{5}$$

で与えられる。

式 (5) は数学的には扱いやすいが，数値計算上は注意が必要である（章末問題）。

Proof. $\partial f / \partial \boldsymbol{\beta} = (\partial f / \partial \beta_i)_{i=1}^{p}$ と記せば，簡単な計算から

$$\frac{\partial f}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

となる。これが 0 になるとき f は最小となる。方程式で表すと

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y} \tag{6}$$

となり，正規方程式 (normal equation) と呼ばれる。正規方程式の解は (5) となる。 □

重回帰における予測値は

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

となる。ここに現れる行列 $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ は直交射影行列である（章末問題）。残差や決定係数，重相関係数の定義は単回帰のときと同じである。平方和の分解 (3) も成り立つ。

重回帰において注意すべき点として，説明変数の追加や削除によって他の回帰係数の符号が変わることが挙げられる。これは本質的にはシンプソンのパラドックスと同じである（章末問題）。どの変数を回帰式に残すべきか，という変数選択の問題は講義の最終回に扱う予定である。

3.6 ダミー変数

説明変数が質的変数であるとき，これを数値に対応させ，重回帰分析を適用することができる。数値化された変数のことをダミー変数 (dummy variable) と呼ぶ。質的変数の取り得る値が 2 つしかない場合，片方の値を 0，もう片方の値を 1 に対応させることが多い。ただし数値の対応のさせ方は本質的ではない。

質的変数の取り得る値（水準と呼ぶ）が K 個あるとき，これを数値化するためには $(K-1)$ 次元のベクトルが必要である。対応させる値としては，数学的にはアフィン独立な K 個の点であれば何でもよいが，簡単な方法の一つとして

$$(0,0,\ldots,0),\ (1,0,\ldots,0),\ (0,1,\ldots,0),\ \cdots,\ (0,0,\ldots,1)\ \in\ \mathbb{R}^{K-1}$$

がある。

用語のまとめ

- 単回帰：目的変数，説明変数，最小二乗法，回帰式，回帰係数，予測値，残差。
- 重回帰：計画行列，正規方程式，決定係数，重相関係数，ダミー変数。
- 後日扱う内容：回帰モデル，変数選択，一般化線形モデル。

演習問題

問題 3.1. 定理 3.1 を証明せよ。

問題 3.2. 定理 3.2 を以下の方針で証明せよ。まず

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

と展開し，これを

$$(\boldsymbol{\beta}-\boldsymbol{\alpha})'\boldsymbol{A}(\boldsymbol{\beta}-\boldsymbol{\alpha}) + c$$

という形に平方完成させる。$\boldsymbol{A}$ が正定値であることを示せば，$f(\boldsymbol{\beta})$ は $\boldsymbol{\beta}=\boldsymbol{\alpha}$ で最小となることが分かる。

問題 3.3. 講義の web ページにある「LPGA データ」は 2017 年女子オープンゴルフ選手権競技の 3 日目までのスコア $\boldsymbol{x}$ と 4 日目のスコア $\boldsymbol{y}$ からなる CSV ファイルである[5]。これをダウンロードし，$\boldsymbol{y}$ を $\boldsymbol{x}$ で説明する回帰式を求めよ。また散布図を描き，回帰直線を重ね描きせよ。

問題 3.4. 2017 年度の東京の日最高気温（℃）の月平均値 y_t $(t = 1, \ldots, 12)$ は

$$10.8, 12.1, 13.4, 19.9, 25.1, 26.4, 31.8, 30.4, 26.8, 20.1, 16.6, 11.1$$

であった。目的変数を y_t，説明変数を $x_{t1} = \cos(2\pi t/12)$，$x_{t2} = \sin(2\pi t/12)$ として回帰式を求めよ。

なお，上のデータは気象庁のサイト http://www.jma.go.jp/から「各種データ・資料」，「過去の気象データ検索」と進み，地点を「東京都 東京」，年月日を「2017 年」，データの種類を「2017 年の月ごとの値を表示」とすれば得られる。

問題 3.5. 目的変数 $\boldsymbol{y}$ を p 個の説明変数 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ で説明する回帰式

$$\hat{y}(x_1, \ldots, x_p) = \hat{a} + \hat{b}_1 x_1 + \cdots + \hat{b}_p x_p, \quad (x_1, \ldots, x_p) \in \mathbb{R}^p,$$

は点 $(\bar{x}_1, \ldots, \bar{x}_p, \bar{y})$ を通ることを示せ。また，$\boldsymbol{j}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ が互いに直交しているとき，$\hat{a}, \hat{b}_1, \ldots, \hat{b}_p$ はどのように表されるか。

問題 3.6 (Simpson's paradox revisited)**.** $\boldsymbol{y}$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ がそれぞれ

$$\boldsymbol{y} = (4, 5, 6, 1, 2, 3)', \quad \boldsymbol{x}_1 = (1, 2, 3, 4, 5, 6)', \quad \boldsymbol{x}_2 = (1, 1, 1, -1, -1, -1)'$$

であるとする。$\boldsymbol{y}$ を $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ で説明する場合の回帰係数と，$\boldsymbol{x}_1$ だけで説明する場合の回帰係数を比較せよ。ただし回帰式には定数項も含めるものとする。

問題 3.7. ランク p の行列 $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ に対し，$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ と定義される行列 $\boldsymbol{P}$ は直交射影行列になること，すなわち $\boldsymbol{P}^2 = \boldsymbol{P}$ かつ $\boldsymbol{P}' = \boldsymbol{P}$ が満たされることを示せ。また $\boldsymbol{P}$ はどのような部分空間への射影になっているか説明せよ。

問題 3.8. 重回帰分析において，計画行列 $\boldsymbol{X}$ の QR 分解 $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ が得られているとき，回帰係数を数値的に求める上で有利な点を説明せよ。ただし $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ が $\boldsymbol{X}$ の QR 分解であるとは，$\boldsymbol{Q}$ が列直交（$\boldsymbol{Q}'\boldsymbol{Q}$ が単位行列）で，$\boldsymbol{R}$ が正の対角成分を持つ上三角行列となることをいう。

[5] 日本女子プロゴルフ協会ホームページ (https://www.lpga.or.jp) より作成。