データ解析の基礎（II）

ーその手法はどうして考えるのかー

安田晃

先週の復習（1）

- n個のデータの平均値は，各データ $(x_i \mid i=1,\cdots,n)$ からの誤差の2乗和が最小となる値であった．
- それを導くために最小2乗法という方法を使った．
- 平均値 $\bar{x}$ は，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

だけではなく，$\bar{x}_G=(x_1x_2\cdots x_n)^{1/n}$

や

$$\bar{x}_H=\frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}$$

のようなものもある．

先週の復習（1）

- n個のデータの平均値は，各データ $(x_i \mid i=1,\cdots,n)$ からの誤差の2乗和が最小となる値であった．
- それを導くために最小2乗法という方法を使った．
- 平均値 $\bar{x}$ は，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

だけではなく，$\bar{x}_G=(x_1x_2\cdots x_n)^{1/n}$

や

$$\bar{x}_H=\frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}$$

のようなものもある．

先週の復習（1）

- n個のデータの平均値は，各データ $(x_i \mid i=1,\cdots,n)$ からの誤差の2乗和が最小となる値であった．
- それを導くために最小2乗法という方法を使った．
- 平均値 $\bar{x}$ は，$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

だけではなく，$\bar{x}_G=(x_1x_2\cdots x_n)^{1/n}=\left(\prod_{i=1}^{n}x_i\right)^{1/n}$

や

$$\bar{x}_H=\frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

のようなものもある．

先週の復習（2）

- n個のデータの $(x_i \mid i=1,\cdots,n)$ の平均からの散らばりは，以下のように順序だてて考えればよかった．
  先ず，平均値から各データまでを1辺とする正方形の和を考える．これは各データから平均値までの全体の散らばりを考えたことになり，

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

と書ける．ここで，S を偏差平方和を呼ぼう．

先週の復習（2）

- n個のデータの $(x_i \mid i=1,\cdots,n)$ の平均からの散らばりは，以下のように順序だてて考えればよかった．
  先ず，平均値から各データまでを1辺とする正方形の和を考える．これは各データから平均値までの全体の散らばりを考えたことになり，

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

と書ける．ここで，S を偏差平方和を呼ぼう．

先週の復習（3）

- 偏差平方和を n で割り算すれば，平均値からの平均的乖離を表現できる．即ち，

$$V = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

である．この V を分散という．

分散は s^2 と表現することもある．

先週の復習（4）

- しかし，偏差平方和，分散の単位はもとのデータと異なり2乗となっている．平均的な正方形から1辺を求めるためには，

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

となる．この s を標準偏差と呼ぼう．

先週の復習（5）

- n 組の2変量 $\left(x_i,\ y_i \mid i = 1, \cdots, n\right)$ があり，これらを2次元平面にプロットしたとき，ある直線を，以下のような拘束条件下で考える．
  - 簡単な直線 $y = ax + b$ を考える．
  - そのとき各点が直線の上にくるか下にくるかの確率は50%とする．
  - 各点からこの直線からまでの距離の合計が最小となるような傾きと切片を求める．

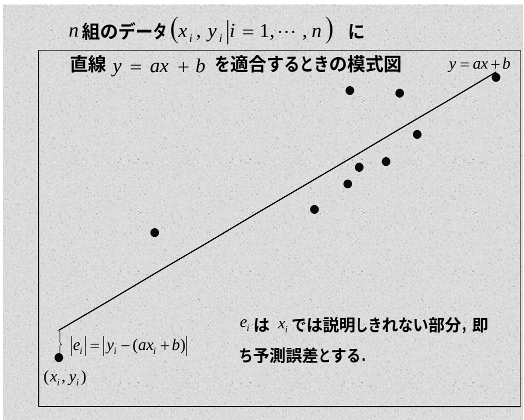
先週の復習（5）

- n 組の2変量 $\left(x_i,\ y_i \mid i = 1, \cdots, n\right)$ があり，これらを2次元平面にプロットしたとき，ある直線を，以下のような拘束条件下で考える．
  - 簡単な直線 $y = ax + b$ を考える．
  - そのとき各点が直線の上にくるか下にくるかの確率は50%とする．
  - 各点からこの直線からまでの距離の合計が最小となるような傾きと切片を求める．

先週の復習（5）

- n 組の2変量 $\left(x_i,\ y_i \mid i = 1, \cdots, n\right)$ があり，これらを2次元平面にプロットしたとき，ある直線を，以下のような拘束条件下で考える．
  - 簡単な直線 $y = ax + b$ を考える．
  - そのとき各点が直線の上にくるか下にくるかの確率は50%とする．
  - 各点からこの直線からまでの距離の合計が最小となるような傾きと切片を求める．

先週の復習（5´）



先週の復習 (6)

- これらの拘束条件から，平均値のところで用いた最小2乗法を用い，

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

$$b = \bar{y} - a\bar{x}$$

を得る．

先週の復習 (7)

- この直線の周りにデータが集中しているときは大きく，散らばっているときは小さくなるような係数を，データから求めたい．

このとき，以下のように与えられる．

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

この r を相関係数と呼ぼう．

先週の復習 (8)

- n 個のデータ $\left(x_i \mid i = 1, \cdots, n\right)$ があるとき， x_i に関して新たな関数 x_i^* を，

$$x_i^* = \frac{x_i - \bar{x}}{s}$$

$$\left(\text{但し，} \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

とすれば，

$$\bar{x}^* = 0, \quad s^* = 1$$

である．

先週の復習 (8´)

- 偏差値は，この x_i^* を使って，

$$50 + 10x_i^*$$

のような関数とし，あらゆる標本集団に対して平均を50，

標準偏差を10と標準化した値であった．

今日の概要 (1)

	変量1	変量2	…	変量 p	y_i
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{22}	…	x_{2p}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{np}	y_n

のとき，y_i の推定値 $\hat{y}_i$ を p 個の変量を使い線形式で表したい．

そのとき，線形式で示された $\hat{y}_i$ と y_i の相関係数は示せないか．

各変量の単位が異なっているとき，各変量が y_i に対してどの程度影響を及ぼしているかを知りたい．

今日の概要

	変量1	変量2	…	変量 p	y_i
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{22}	…	x_{2p}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{np}	y_n

のとき，y_i の推定値 $\hat{y}_i$ を p 個の変量を使い線形式で表したい．

そのとき，線形式で示された $\hat{y}_i$ と y_i の相関係数は示せないか．

各変量の単位が異なっているとき，各変量が y_i に対してどの程度影響を及ぼしているかを知りたい．

今日の概要

	変量1	変量2	…	変量 p	y_i
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{22}	…	x_{2p}	y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{np}	y_n

のとき，y_i の推定値 $\hat{y}_i$ を p 個の変量を使い線形式で表したい．

そのとき，線形式で示された $\hat{y}_i$ と y_i の相関係数は示せないか．

各変量の単位が異なっているとき，各変量が y_i に対してどの程度影響を及ぼしているかを知りたい．

重回帰分析の基礎 (1)

具体的に

一昨年3月から12月までの日本の国内総生産 (GDP) は，鉄鋼生産量，自動車生産台数，コンビニ全店売上高の3項目でどのように表現できるだろうか．

自動車生産台数
http://www.jama.or.jp/stats/m_report/pdf/2003_04.pdf

四半期ごとのGDP
http://www.esri.cao.go.jp/jp/sna/qe031-2/gaku-jg0312.csv

鉄鋼生産量，コンビニ売上
http://www.econ-jp.com/

重回帰分析の基礎 (2)

具体的に

次のような表を得る．

年月	鉄鋼生産量 (千トン)	自動車生産台数 (台)	コンビニ売上 (億円)	GDP (10億円)
2002年3月	8756	799979	5575	129605.50
4月	8756	660561	5384	131466.40
5月	9390	675343	5540	131466.40
6月	9141	698499	5525	131466.40
7月	8977	772036	6097	134600.70
8月	9241	598626	6116	134600.70
9月	9124	763973	5523	134600.70
10月	9457	765872	5582	141594.10
11月	9305	767857	5345	141594.10
12月	9299	703189	5733	141594.10

但し，GDPは四半期ごとのデータ

重回帰分析の基礎 (3)

具体的に

先ほどのデータから，

$$a_0,\ a_1,\ a_2,\ a_3$$

を定数として，

$$\mathrm{GDP} = a_0 + a_1 \times \text{鉄鋼生産量} + a_2 \times \text{自動車生産台数} + a_3 \times \text{コンビニ売上}$$

という式が欲しい

重回帰分析の基礎 (4)

一般化して

先ほどのデータから，説明される変数 (GDP) を目的変数，説明するための変数 (鉄鋼，自動車，コンビニ) を説明変数という．今，3つの説明変数と1つの目的変数を考える．

n 組のデータ，

$$\{(x_{11}, x_{21}, x_{31}, y_1), (x_{12}, x_{22}, x_{32}, y_2), \cdots, (x_{1n}, x_{2n}, x_{3n}, y_n)\}$$

がある．

重回帰分析の基礎 (5)

一般化して

n 組のデータから，求める式を，

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3$$

とするとき，$(y_i \mid i = 1, \cdots, n)$ と推定値 $\hat{y}_i$ との誤差 e_i の2乗和を最小とする係数 a_0, a_1, a_2, a_3 を求める問題を考えよう．

ここでも，最小2乗法で計算しよう．即ち，

重回帰分析の基礎 (6)

一般化して

$$f(a_0,a_1,a_2,a_3)=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left(y_i-a_0-a_1x_{1i}-a_2x_{2i}-a_3x_{3i}\right)^2$$

を，a_0,a_1,a_2,a_3 で偏微分し，0とおけば，

$$\frac{\partial f}{\partial a_0}=\sum_{i=1}^{n}\left(y_i-a_1x_{1i}-a_2x_{2i}-a_3x_{3i}-a_0\right)(-1)=0$$
$$\frac{\partial f}{\partial a_1}=\sum_{i=1}^{n}\left(y_i-a_1x_{1i}-a_2x_{2i}-a_3x_{3i}-a_0\right)(-x_{1i})=0$$
$$\frac{\partial f}{\partial a_2}=\sum_{i=1}^{n}\left(y_i-a_1x_{1i}-a_2x_{2i}-a_3x_{3i}-a_0\right)(-x_{2i})=0$$
$$\frac{\partial f}{\partial a_3}=\sum_{i=1}^{n}\left(y_i-a_1x_{1i}-a_2x_{2i}-a_3x_{3i}-a_0\right)(-x_{3i})=0$$

重回帰分析の基礎 (7)

一般化して

整理すれば，

$$\begin{aligned}
a_1\sum x_{1i} &+a_2\sum x_{2i} &+a_3\sum x_{3i} &+na_0 &&=\sum y_i\\
a_1\sum x_{1i}^2 &+a_2\sum x_{1i}x_{2i} &+a_3\sum x_{1i}x_{3i} &+a_0\sum x_{1i} &&=\sum x_{1i}y_i\\
a_1\sum x_{2i}x_{1i} &+a_2\sum x_{2i}^2 &+a_3\sum x_{2i}x_{3i} &+a_0\sum x_{2i} &&=\sum x_{2i}y_i\\
a_1\sum x_{3i}x_{1i} &+a_2\sum x_{3i}x_{2i} &+a_3\sum x_{3i}^2 &+a_0\sum x_{3i} &&=\sum x_{3i}y_i
\end{aligned}$$

求める a_0,a_1,a_2,a_3 は高校時代に経験したクラメールの方法などで解けばよい．

重回帰分析の基礎 (7′)

一般化して

例えば a_1 は，

$$a_1=\frac{\begin{bmatrix}\sum y_i & \sum x_{2i} & \sum x_{3i} & n\\ \sum x_{1i}y_i & \sum x_{1i}x_{2i} & \sum x_{1i}x_{3i} & \sum x_{1i}\\ \sum x_{2i}y_i & \sum x_{2i}^2 & \sum x_{2i}x_{3i} & \sum x_{2i}\\ \sum x_{3i}y_i & \sum x_{3i}x_{2i} & \sum x_{3i}^2 & \sum x_{3i}\end{bmatrix}}{\begin{bmatrix}\sum x_{1i} & \sum x_{2i} & \sum x_{3i} & n\\ \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \sum x_{1i}x_{3i} & \sum x_{1i}\\ \sum x_{2i}x_{1i} & \sum x_{2i}^2 & \sum x_{2i}x_{3i} & \sum x_{2i}\\ \sum x_{3i}x_{1i} & \sum x_{3i}x_{2i} & \sum x_{3i}^2 & \sum x_{3i}\end{bmatrix}}$$

のように．

重回帰分析の基礎 (8)

一般化して

一般化して説明変数が p 個の場合，各係数 $a_0,a_1,a_2,\cdots,a_p$ を自身で解いてほしい．書き損じた紙の裏と，鉛筆があればいつでも，どこでもできるので．

今までの一連の分析を重回帰分析という．

重回帰分析の基礎 (おまけ I)

皆さんは線形代数を履修しているので，

$$f(a_0,a_1,a_2,\cdots,a_p)=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left\{y_i-(a_0+a_1x_{1i}+a_2x_{2i}\cdots+a_px_{pi})\right\}^2$$

は，

$$f(a_0,a_1,a_2,\cdots,a_p)=(\boldsymbol{y}-X\boldsymbol{a})^T(\boldsymbol{y}-X\boldsymbol{a})$$

と書けることは分かるだろう．ここで，

$$\boldsymbol{y}=\begin{bmatrix}\sum y_i\\ \sum y_ix_{1i}\\ \sum y_ix_{2i}\\ \vdots\\ \sum y_ix_{pi}\end{bmatrix},\quad X=\begin{bmatrix}n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{pi}\\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{2i}x_{1i} & \cdots & \sum x_{pi}x_{1i}\\ \sum x_{2i} & \sum x_{2i}x_{1i} & \sum x_{2i}^2 & \cdots & \sum x_{pi}x_{2i}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \sum x_{pi} & \sum x_{pi}x_{1i} & \sum x_{pi}x_{2i} & \cdots & \sum x_{pi}^2\end{bmatrix},\quad \boldsymbol{a}=\begin{bmatrix}a_0\\ a_1\\ a_2\\ \vdots\\ a_p\end{bmatrix}$$

で，添え字 T は転置を示す．

重回帰分析の基礎 (おまけ II)

先ほどの要素の形のように f を $\boldsymbol{a}$ で偏微分すれば

$$\frac{\partial f}{\partial \boldsymbol{a}}=-2X^T(\boldsymbol{y}-X\boldsymbol{a})=0$$
$$X^T(\boldsymbol{y}-X\boldsymbol{a})=0$$
$$X^T\boldsymbol{y}=(X^TX)\boldsymbol{a}$$

従って，

$$\boldsymbol{a}=(X^TX)^{-1}X\boldsymbol{y}=\begin{bmatrix}a_0\\ a_1\\ a_2\\ \vdots\\ a_p\end{bmatrix}$$

として求めることができる．要素の形より見通しがきいて分かりやすい．掃き出し法など使って逆行列を求め，実際に $\boldsymbol{a}$ を求めてほしい．

重回帰分析の基礎 (9)

一般化して

先ほどのGDP，鉄鋼，自動車，コンビニデータを解析してみよう．各値を計算すれば，

$$a_1 = 13.24$$
$$a_2 = 2.314\times10^{-2}$$
$$a_3 = 0.977$$
$$a_0 = -8037$$

よって，

$$\text{GDP} = 13.24\times\text{鉄鋼} + (2.314\times10^{-2})\times\text{自動車} + 0.977\times\text{コンビニ} - 8037$$

を得る．

重回帰分析の基礎 (10)

相関係数のようなものはないか・・・

ある．しかし，考え方を少し変えよう．先ほどの

$$\text{GDP} = 13.24\times\text{鉄鋼} + (2.314\times10^{-2})\times\text{自動車} + 0.977\times\text{コンビニ} - 8037$$

を考える．例えば2002年3月の鉄鋼，自動車，コンビニを入れて計算する．計算によって得られえたGDPの予測値を $\hat{y}_1$，3月のGDPの実測値を y_1 とすれば，

$$\hat{y}_1 = 131850.7$$

$$y_1 = 129605.5$$

重回帰分析の基礎 (11)

相関係数のようなものはないか・・・

このような計算を，2002年12月まで行う．即ち，

i	月	y_i	$\hat{y}_i$
1	3月	129605.5	131800.7
2	4月	131466.4	128388.0
3	5月	131466.4	137276.6
4	6月	131466.4	134501.0
5	7月	134600.7	134590.2
6	8月	134600.7	134091.4
7	9月	134600.7	135789.1
8	10月	141594.1	140299.6
9	11月	141594.1	138101.5
10	12月	141594.1	136904.7

重回帰分析の基礎 (12)

相関係数のようなものはないか・・・

ここで，実測値 $(y_i \mid i = 1,\cdots,n)$ と予測値$(\hat{y}_i \mid i = 1,\cdots,n)$の相関係数を考えよう．

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

これを計算すれば，

$$R = 0.72020$$

を得る．この R を重相関係数と呼ぶ．

重回帰分析の基礎 (13)

このような分析を評価する値はないか・・・

実測値 y_i の変動は次のように分解される．

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

全変動 (S_T)　回帰による変動 (S_R)　回帰からの残差変動 (S_e)

重回帰分析の基礎 (14)

このような分析を評価する値はないか・・・

$$S_T = S_R + S_e$$

から，S_R が大きくなれば，回帰によって多くの部分が説明される．S_T は式から一定である．そこで，

$$R^2 = \frac{S_R}{S_T} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

を考える．これによって，全体の変動のうち回帰によって説明される部分の割合を示している．この R^2 を決定係数，あるいは寄与率という．

重回帰分析の基礎 (15)

このような分析を評価する値はないか・・・

この R^2 は先程計算した R の2乗である．計算の仕方が違っているように思えるが，最終的には，導くことができる．———— 時間があれば自身で解いてみよう．

重相関係数は説明変数によって予測するのにどの程度有効かを示す指標であり，決定係数は，説明変数に依存した説明される部分を示す指標である．

重回帰分析の基礎 (15)

このような分析を評価する値はないか・・・

この R^2 は先程計算した R の2乗である．計算の仕方が違っているように思えるが，最終的には，導くことができる．———— 時間があれば自身で解いてみよう．

重相関係数は説明変数によって予測するのにどの程度有効かを示す指標であり，決定係数は，説明変数に依存した説明される部分を示す指標である．

重回帰分析の基礎 (16)

一般化して

ここで，GDPを3つの変数で表した回帰式ができた．しかし，この3変数はすべて単位が異なる．一体，どの変数がGDPを変化させるのに貢献しているのだろうか．この式だけではワカラナイ．

しかし，異なった単位を統一すればいいだろう．そのためには，・・・

重回帰分析の基礎 (17)

一般化して

偏差値のところで用いたデータの標準化を行ってみよう．各パラメータの平均値を $\bar{x}$ ，標準偏差を s として， i 番目のデータ x_i の新たな関数を $\bar{x}_i^*$ とすれば，

$$x_i^* = \frac{x_i - \bar{x}}{s}$$

と書けた．

このように計算すれば， x_i^* の単位は無名数であり，すべてのパラメータが平均0，分散1となる．

重回帰分析の基礎 (18)

一般化して

実際に計算してみる．

	鉄鋼	自動車	コンビニ	GDP
3月	8756	799979	5575	129605.5
4月	8756	660561	5384	131466.4
5月	9390	675343	5540	131466.4
6月	9141	698499	5525	131466.4
7月	8977	772036	6097	134600.7
8月	9241	598626	6116	134600.7
9月	9124	763973	5523	134600.7
10月	9457	765872	5582	141594.1
11月	9305	767857	5345	141594.1
12月	9299	703189	5733	141594.1
平均	9144.6	720593.5	5642	135258.9
標準偏差	223.04146	57550.621	241.4159	4229.916

重回帰分析の基礎 (19)

一般化して

標準化すれば，

	鉄鋼	自動車	コンビニ	GDP
3月	-1.74228	1.379403	-0.27753	-1.33653
4月	-1.74228	-1.04313	-1.0687	-0.89659
5月	1.100244	-0.78627	-0.42251	-0.89659
6月	-0.01614	-0.38391	-0.48464	-0.89659
7月	-0.75143	0.893865	1.884714	-0.15561
8月	0.432207	-2.11931	1.963417	-0.15561
9月	-0.09236	0.753762	-0.49293	-0.15561
10月	1.400636	0.78676	-0.24853	1.497711
11月	0.719149	0.821251	-1.23024	1.497711
12月	0.692248	-0.30242	0.376943	1.497711
平均	0	0	0	0
標準偏差	1	1	1	1

重回帰分析の基礎 (20)

一般化して

これより，

$$\mathrm{GDP} = 0.698\times \text{鉄鋼}^{*} + 0.315\times \text{自動車}^{*} + (5.574\times 10^{-2})\times \text{コンビニ}^{*}$$

GDPを変化させるため，コンビニの売上は0ではなく，関与はしているが，やはり鉄鋼生産量，自動車生産台数という長大産業の占める割合が大きいのである．

考えてほしいこと・・・上の式では，定数項がない．何故か．

重回帰分析の基礎 (20)

一般化して

これより，

$$\mathrm{GDP} = 0.698\times \text{鉄鋼}^{*} + 0.315\times \text{自動車}^{*} + (5.574\times 10^{-2})\times \text{コンビニ}^{*}$$

GDPを変化させるため，コンビニの売上は0ではなく，関与はしているが，やはり鉄鋼生産量，自動車生産台数という長大産業の占める割合が大きいのである．

考えてほしいこと・・・上の式では，定数項がない．何故か．

偏相関というものを考える (1)

こんなデータを見てみる．

ある健康診断した標本に対して，以下のデータを得た．

被験者	血 圧	年 齢	年 収
A	136	35	503
B	135	30	451
C	147	47	650
D	135	50	423
E	150	52	479
F	164	40	476
G	195	61	1221
H	133	36	306
I	144	47	689
J	155	56	570

偏相関というものを考える (2)

こんなデータを見てみる．

この相関行列は，

	血 圧	年 齢	年 収
血 圧	1		
年 齢	0.669	1	
年 収	0.712	0.816	1

これより，血圧が高いほど高給取りであることが分かった．

―――――― チョッと変ではないか．このような結論は．

偏相関というものを考える (2)

こんなデータを見てみる．

この相関行列は，

	血 圧	年 齢	年 収
血 圧	1		
年 齢	0.669	1	
年 収	0.712	0.816	1

これより，血圧が高いほど高給取りであることが分かった．

―――――― チョッと変ではないか．このような結論は．

偏相関というものを考える (3)

こんなデータを見てみる．

結論を急げば変である．血圧の上下は年齢と関係があり，年齢と年収にも関連がある．血圧と年収は見かけ上相関があるのではないか．

↓

このように年収と血圧を見るときに，年齢の要因を除いて相関を見たい．

―――――― ここで偏相関．

偏相関というものを考える (3)

こんなデータを見てみる.

結論を急げば変である. 血圧の上下は年齢と関係があり, 年齢と年収にも関連がある. 血圧と年収は見かけ上相関があるのではないか.

↓

このように年収と血圧を見るときに, 年齢の要因を除いて相関を見たい.

―――― ここで偏相関.

偏相関というものを考える (4)

こんなデータを見てみる.

計算過程を除いて結論だけ言えば, 以下のようになる.

年齢の要因を除いたあとの血圧と年収の相関係数を, 血圧と年収の偏相関係数と呼び,

$$r_{\text{血圧年収}\cdot\text{年齢}} = \frac{r_{\text{血圧年収}} - r_{\text{血圧年齢}} r_{\text{年齢年収}}}{\sqrt{\left(1-r^2_{\text{血圧年齢}}\right)\left(1-r^2_{\text{年齢年収}}\right)}}$$

とする. 他の偏相関も同様に,

偏相関というものを考える (5)

こんなデータを見てみる.

$$r_{\text{血圧年齢}\cdot\text{年収}} = \frac{r_{\text{血圧年齢}} - r_{\text{血圧年収}} r_{\text{年齢年収}}}{\sqrt{\left(1-r^2_{\text{血圧年収}}\right)\left(1-r^2_{\text{年齢年収}}\right)}}$$

$$r_{\text{年齢年収}\cdot\text{血圧}} = \frac{r_{\text{年齢年収}} - r_{\text{血圧年齢}} r_{\text{血圧年収}}}{\sqrt{\left(1-r^2_{\text{血圧年齢}}\right)\left(1-r^2_{\text{血圧年収}}\right)}}$$

と計算できる. 4変数以上の場合も同様であるが, ・・・

偏相関というものを考える (6)

こんなデータを見てみる.

このような時, 行列表記がすごく役立つ.

相関行列を $R=[r_{ij}]$ とおき, その逆行列を $R^{-1}=[r^{ij}]$ とすれば, x_i と x_j 以外の変数を与えたときの x_i と x_j の偏相関係数 $r_{ij\bullet rest}$ は,

$$r_{ij\bullet rest} = \frac{r^{ij}}{[r^{ii}r^{jj}]^{1/2}}$$

である.

偏相関というものを考える (6´)

こんなデータを見てみる.

一般化して, y と x_j との偏相関は, x_j と残りの $(p-1)$個の変数の線形式で y を最小2乗近似したときの残差と, x_j を残り $(p-1)$個の変数の線形式で最小2乗近似したときの残差との相関係数である.

偏相関というものを考える (7)

実際に計算してみれば・・・

	血圧	年齢	年収
血圧	1		
年齢	0.669	1	
年収	0.712	0.816	1

↓

	血圧	年齢	年収
血圧	1		
年齢	0.685	1	
年収	0.385	0.816	1

偏相関というものを考える (7)

実際に計算してみれば･･･

	血圧	年齢	年収
血圧	1		
年齢	0.669	1	
年収	0.712	0.816	1

	血 圧	年 齢	年 収
血 圧	1		
年 齢	0.685	1	
年 収	0.385	0.816	1

今日話していないこと

- 重相関係数はどの程度あったら，有意か．
- 標準化したときの a_0 がなくなる証明．
- 重回帰分析の行列，ベクトル表示．
- $a_1, a_2, \cdots, a_p, a_0$ は一定の区間内にある．その計算方法を示していない．
- 相関と名がつくものは相関係数，重相関係数，偏相関，他に，積差率相関，相関比，相関指数･･･，説明していない相関もたくさんある．
- 非線形の回帰式．例えば $y = a_1 \sin b_1 x_1 + a_2 e^{-b_2 x_2} + a_0$

他にもたくさんあるが，残りはご自身でどうぞ．

今日話していないこと

- 重相関係数はどの程度あったら，有意か．
- 標準化したときの a_0 がなくなる証明．
- 重回帰分析の行列，ベクトル表示．
- $a_1, a_2, \cdots, a_p, a_0$ は一定の区間内にある．その計算方法を示していない．
- 相関と名がつくものは相関係数，重相関係数，偏相関，他に，積差率相関，相関比，相関指数･･･，説明していない相関もたくさんある．
- 非線形の回帰式．例えば $y = a_1 \sin b_1 x_1 + a_2 e^{-b_2 x_2} + a_0$

他にもたくさんあるが，残りはご自身でどうぞ．

今日話していないこと

- 重相関係数はどの程度あったら，有意か．
- 標準化したときの a_0 がなくなる証明．
- 重回帰分析の行列，ベクトル表示．
- $a_1, a_2, \cdots, a_p, a_0$ は一定の区間内にある．その計算方法を示していない．
- 相関と名がつくものは相関係数，重相関係数，偏相関，他に，積差率相関，相関比，相関指数･･･，説明していない相関もたくさんある．
- 非線形の回帰式．例えば $y = a_1 \sin b_1 x_1 + a_2 e^{-b_2 x_2} + a_0$

他にもたくさんあるが，残りはご自身でどうぞ．

今日話していないこと

- 重相関係数はどの程度あったら，有意か．
- 標準化したときの a_0 がなくなる証明．
- 重回帰分析の行列，ベクトル表示．
- $a_0, a_1, a_2, \cdots, a_p$ は一定の信頼区間，予測区間内にある．その計算方法を示していない．
- 相関と名がつくものは相関係数，重相関係数，偏相関，他に，積差率相関，相関比，相関指数･･･，説明していない相関もたくさんある．
- 非線形の回帰式．例えば $y = a_1 \sin b_1 x_1 + a_2 e^{-b_2 x_2} + a_0$

他にもたくさんあるが，残りはご自身でどうぞ．

今日話していないこと

- 重相関係数はどの程度あったら，有意か．
- 標準化したときの a_0 がなくなる証明．
- 重回帰分析の行列，ベクトル表示．
- $a_1, a_2, \cdots, a_p, a_0$ は一定の区間内にある．その計算方法を示していない．
- 相関と名がつくものは相関係数，重相関係数，偏相関，他に，積差率相関，相関比，相関指数･･･，説明していない相関もたくさんある．
- 非線形の回帰式．例えば $y = a_1 \sin b_1 x_1 + a_2 e^{-b_2 x_2} + a_0$

他にもたくさんあるが，残りはご自身でどうぞ．

今日話していないこと

- 重相関係数はどの程度あったら，有意か．
- 標準化したときの a_0 がなくなる証明．
- 重回帰分析の行列，ベクトル表示．
- $a_1, a_2, \cdots, a_p, a_0$ は一定の区間内にある．その計算方法を示していない．
- 相関と名がつくものは相関係数，重相関係数，偏相関，他に，積差率相関，相関比，相関指数･･･，説明していない相関もたくさんある．
- 非線形の回帰式．例えば $y = a_1 \sin b_1 x_1 + a_2 e^{-b_2 x_2} + a_0$

他にもたくさんあるが，残りはご自身でどうぞ．

今日話していないこと

- 重相関係数はどの程度あったら，有意か．
- 標準化したときの a_0 がなくなる証明．
- 重回帰分析の行列，ベクトル表示．
- $a_1, a_2, \cdots, a_p, a_0$ は一定の区間内にある．その計算方法を示していない．
- 相関と名がつくものは相関係数，重相関係数，偏相関，他に，積差率相関，相関比，相関指数･･･，説明していない相関もたくさんある．
- 非線形の回帰式．例えば $y = a_1 \sin b_1 x_1 + a_2 e^{-b_2 x_2} + a_0$

他にもたくさんあるが，残りはご自身でどうぞ．

参 考 文 献

- ブルーバックス『データ分析　はじめの一歩』
  清水誠．講談社．
  ―― ゴロッと横になってどうぞ．

- ブルーバックス『原因をさぐる統計学』
  豊田秀樹，前田忠彦，柳井晴夫．講談社．
  ――同じくゴロッと横になってどうぞ．

- 『基本演習確率統計』
  和田秀三．サイエンス社．
  ――少し構えてどうぞ．今回紹介しなかった相関が少しある．