§５　重回帰解析

5.1　行列

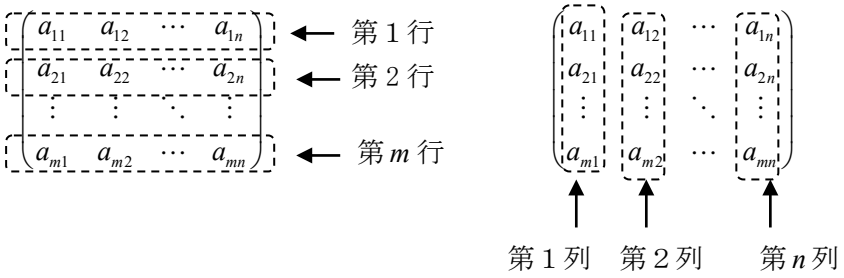
行列とは？

数値を長方形の形に並べて，括弧（　）でくくったものを
行列といい，記号 A で表す。

［行列］　$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$

また，行列内の数値を**成分**という。

上から順に，　第１**行**，第２**行**，…，第 m **行**　といい，　[横が行]
左から順に，　第１**列**，第２**列**，…，第 n **列**　という。　[縦が列]



m 行 n 列からなる行列を　$\boldsymbol{m\times n}$ **型行列**　と言います。
i 行 j 列にある成分 a_{ij} を　$\boldsymbol{(i,\ j)}$ **成分**　と言います。

特に，$m=n$ である行列を，$\boldsymbol{n}$ **次正方行列**という。

例）$A=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$　；２次正方行列

$A=\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$　；３次正方行列

※３年で，行列に四則演算を導入し，いろいろな問題を解いていきます。

5.2　標準化

標準化とは？

変数 x のデータの平均値が $\mu_x = E(x)$，分散が $\sigma_x^2 = V(x)$ とする。

これを，変数変換　$X = \dfrac{x-\mu}{\sigma}$　によって書き直す作業を

標準化と言います。

標準化された変数 X の平均値は 0，標準偏差は 1 になります。

※通常は，データ毎に分布の状況が異なります。しかし，
標準化することにより，全てが同じ分布状況として
考えることができるようになると言っています。

説明）変数変換の性質(公式)より[第 03 回を参照]

(1) 平均値　$\mu_X = E(X) = \dfrac{E(x)-\mu_x}{\sigma} = \dfrac{\mu_x-\mu_x}{\sigma} = 0$

(2) 分散　$\sigma_X^2 = V(X) = \dfrac{V(x)}{\sigma_x^2} = \dfrac{\sigma_x^2}{\sigma_x^2} = 1$

⇒　標準偏差　$\sigma_X = 1$

＝＝＝

【研究：偏差値】

平均値が 0，標準偏差が 1 であるよう変数に変換する作業を**標準化**と言います。

［標準化］　$X = \dfrac{x-\mu}{\sigma}$

これと似たようなものに，**偏差値**というものがあります。

偏差値は，平均値が 50，標準偏差が 10 である変数にするものです。

［偏差値］　$X = 50 + \dfrac{10(x-\mu)}{\sigma}$

これは，様々な試験の点数を画一的に判断するときなどに用いられます。

資料の整理_TEXT02 の**チェビシェフの不等式**［$\sigma = 2.5$ の場合］を用いると，

25 点～75 点の間に少なくとも 84%が存在することになります。

つまり，75 点以上の場合は受験者の上位 8%以内にいることになります。

＝＝＝

5.3　分散共分散行列と相関行列

分散共分散行列，相関行列とは？

成分が，次の様に配列された行列を**分散共分散行列**と言います。

例）２つの変数 x, y のデータに対して　$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$

３つの変数 x, y, z のデータに対して　$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{xy} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{xz} & \sigma_{yz} & \sigma_z^2 \end{pmatrix}$

標準化されたデータに対しての分散共分散行列を，**相関行列**と言います。

例）２つの変数 x, y の標準化された変数 X, Y に対して

$$\begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} 1 & \sigma_{XY} \\ \sigma_{XY} & 1 \end{pmatrix} = \begin{pmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{pmatrix}$$

但し，$r_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$　；２変数 x, y の相関係数

※実際　標準化されているので，

標準偏差　$\sigma_X = \sigma_Y = 1$　⇒　分散　$\sigma_X^2 = \sigma_Y^2 = 1$

共分散　$\sigma_{XY} = E(XY) - E(X)E(Y) = E(XY) = \dfrac{1}{N}\sum_{k=1}^{N} X_k Y_k$

[※標準化されているので　$E(X) = E(Y) = 0$]

$$= \frac{1}{N}\sum_{k=1}^{N} \frac{x_k - \mu_x}{\sigma_x} \cdot \frac{y_k - \mu_y}{\sigma_y}$$

$$= \frac{1}{\sigma_x \sigma_y} \times \frac{1}{N}\sum_{k=1}^{N} (x_k - \mu_x)(y_k - \mu_y)$$

$$= \frac{1}{\sigma_x \sigma_y} \times \sigma_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = r_{xy}$$

【復習】　［共分散］　$\sigma_{xy} = \dfrac{1}{N}\sum_{k=1}^{N} (x_k - \mu_x)(y_k - \mu_y)$

$= E(XY) - E(X)E(Y)$

例）３つの変数 x, y, z の標準化された変数 X, Y, Z に対して

$$\begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{pmatrix} = \begin{pmatrix} 1 & r_{xy} & r_{xz} \\ r_{xy} & 1 & r_{yz} \\ r_{xz} & r_{yz} & 1 \end{pmatrix}$$

但し，r_{xy}, r_{xz}, r_{yz} は各２変数の相関係数

5.4　単回帰分析

単回帰分析とは？

２変数 x, y に関して，線形モデル　$y = ax + b$　を
取り扱う手法を，**単回帰分析**と言います。

実は，第 04 回でお話しした**回帰直線**のことです。

[復習：**最小２乗法**]
変数 y の「実測値」 y_k と
変数 x から求められる変数 y の「予測値」　$ax_k + b$
との差 $\varepsilon_k = y_k - (ax_k + b)$ を「２乗」したものの和

$$\varepsilon = \sum_{k=1}^{N} \varepsilon_k^2 = \sum_{k=1}^{N} \{y_k - (ax_k + b)\}^2$$

を「最小」にするように，係数 a, b を定めます。

第 04 回では，証明を割愛しましたが，今回は掲載しておきます。
解析学では，「偏微分の極値問題」となります。
偏微分とは，指定された文字を変数とし，それ以外を定数とする
計算であると思ってください。偏微分の記号は ∂ を使いますが，
通常の微分(**常微分**という)の記号の d と思ってください。
[※∂ の読み：ラウンド　ディ　又は　パーシャル　ディ]

では，回帰直線を求めてみましょう。（※偏微分は３年生の内容です！）

$$\varepsilon = \sum_{k=1}^{N} (y_k - ax_k - b)^2 \quad \cdots①$$

①を a で偏微分すると，「合成関数の微分」より

$$\frac{\partial \varepsilon}{\partial a} = \sum_{k=1}^{N} 2(y_k - ax_k - b) \times (-x_k) = 2\sum_{k=1}^{N} (ax_k^2 + bx_k - x_k y_k) \cdots②$$

①を b で偏微分すると，

$$\frac{\partial \varepsilon}{\partial b} = \sum_{k=1}^{N} 2(y_k - ax_k - b) \times (-1) = 2\sum_{k=1}^{N} (ax_k + b - y_k) \cdots③$$

極値をとるところでは　$\dfrac{\partial \varepsilon}{\partial a} = \dfrac{\partial \varepsilon}{\partial b} = 0$　となるので，

次の連立方程式が得られます（$2N$ で割った形で表記しています)。

$$\begin{cases} a\times\dfrac{1}{N}\displaystyle\sum_{k=1}^{N}x_k^2+b\times\dfrac{1}{N}\sum_{k=1}^{N}x_k=\dfrac{1}{N}\sum_{k=1}^{N}x_k y_k \\ a\times\dfrac{1}{N}\displaystyle\sum_{k=1}^{N}x_k+b\times\dfrac{1}{N}\sum_{k=1}^{N}1=\dfrac{1}{N}\sum_{k=1}^{N}y_k \end{cases}$$

$$\Rightarrow \quad \begin{cases} aE(x^2)+bE(x)=E(xy) & \cdots④ \\ aE(x)+b=E(y) & \cdots⑤ \end{cases}$$

⑤より　$b=E(y)-aE(x)$　　$(\Rightarrow \quad b=\mu_y-a\mu_x \cdots⑥)$

これを④に代入すると　$aE(x^2)+\{E(y)-aE(x)\}E(x)=E(xy)$

$$a[E(x^2)-\{E(x)\}^2]=E(xy)-E(x)E(y)$$

$$a\,\sigma_x^2=\sigma_{xy} \qquad \therefore\ a=\frac{\sigma_{xy}}{\sigma_x^2}$$

よって求める方程式は

$$y=ax+b \quad \Rightarrow \quad y=ax+\mu_y-a\mu_x \quad \Rightarrow \quad y-\mu_y=a(x-\mu_x)$$

［回帰直線］　$y-\mu_y=\dfrac{\sigma_{xy}}{\sigma_x^2}(x-\mu_x)$

5.5　重回帰分析

重回帰分析とは？

３変数以上に関して，線形モデルを取り扱う手法を，**重回帰分析**と言います。
今回は，３変数 x, y, z に関して，線形モデル $z=ax+by+c$ を取扱います。

［重回帰分析］　連立方程式

$$\begin{cases} \mu_z=a\mu_x+b\mu_y+c & \cdots① \\ a\,\sigma_x^2+b\,\sigma_{xy}=\sigma_{xz} & \cdots② \\ a\,\sigma_{xy}+b\,\sigma_y^2=\sigma_{yz} & \cdots③ \end{cases}$$

を解くことにより，係数 a, b, c を求めることができる。

【注意】②と③は行列を用いると次のように表現できます。

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}=\begin{pmatrix} \sigma_{xz} \\ \sigma_{yz} \end{pmatrix}$$

分散共分散行列

説明）主となる計算部分のみを取り扱います。
単回帰分析の証明の計算部分と比較してみてください。

各データの関係は　　$z_k = ax_k + by_k + c \cdots ④$

データの総和を取り，総数 N で割ると

$$\frac{1}{N}\sum_{k=1}^{N} z_k = a \times \frac{1}{N}\sum_{k=1}^{N} x_k + b \times \frac{1}{N}\sum_{k=1}^{N} y_k + c \times \frac{1}{N}\sum_{k=1}^{N} 1$$

$$E(z) = aE(x) + bE(y) + c \cdots ⑤$$

$$\Rightarrow \quad \mu_z = a\mu_x + b\mu_y + c$$

④の両辺に x_k を掛けて，総和を取り総数 N で割ると

$$\frac{1}{N}\sum_{k=1}^{N} x_x z_k = a \times \frac{1}{N}\sum_{k=1}^{N} x_k^2 + b \times \frac{1}{N}\sum_{k=1}^{N} x_k y_k + c \times \frac{1}{N}\sum_{k=1}^{N} x_k$$

$$E(xz) = aE(x^2) + bE(xy) + cE(x) \cdots ⑥$$

よって，⑤と⑥より

$$\sigma_{xz} = E(xz) - E(x)E(z)$$

$$= \{aE(x^2) + bE(xy) + cE(x)\} - E(x)\{aE(x) + bE(y) + c\}$$

$$= a[E(x^2) - \{E(x)\}^2 + b\{E(xy) - E(x)E(y)\}$$

$$= a\,\sigma_x^2 + b\,\sigma_{xy}$$

④の両辺に y_k を掛けて，総和を取り総数 N で割ると

$$\frac{1}{N}\sum_{k=1}^{N} y_x z_k = a \times \frac{1}{N}\sum_{k=1}^{N} x_k y_k + b \times \frac{1}{N}\sum_{k=1}^{N} y_k^2 + c \times \frac{1}{N}\sum_{k=1}^{N} y_k$$

$$E(yz) = aE(xy) + bE(y^2) + cE(y) \cdots ⑦$$

よって，⑤と⑦より

$$\sigma_{yz} = E(yz) - E(y)E(z)$$

$$= \{aE(xy) + bE(y^2) + cE(y)\} - E(y)\{aE(x) + bE(y) + c\}$$

$$= a\{E(xy) - E(x)E(y)\} + b[E(y^2) - \{E(y)\}^2]$$

$$= a\,\sigma_{xy} + b\,\sigma_y^2$$

例題　右の資料は 10 人の大学生の身長(x)，体重(y)，ウエスト(z)を調べたものである。

このとき，z を x, y の式

$$z = ax + by + c$$

で表せ。

番号	身長(x)	体重(y)	ｳｴｽﾄ(z)
1	160	50	60
2	165	60	68
3	167	65	70
4	170	65	65
5	165	70	80
6	167	75	85
7	178	80	78
8	182	85	79
9	175	90	95
10	172	81	89

［解答］

各変数の平均値と分散共分散を求める。

平　均　値	$\mu_x = 170.1$	$\mu_y = 72.1$	$\mu_z = 76.9$
分散共分散	x	y	z
x	$\sigma_x^2 = 40.49$	$\sigma_{xy} = 62.99$	$\sigma_{xz} = 34.31$
y	$\sigma_{xy} = 62.99$	$\sigma_y^2 = 137.69$	$\sigma_{yz} = 109.91$

よって，次の連立方程式の解が求める係数 a, b, c である

$$\begin{cases} 170.1a + 72.1b + c = 76.9 \\ 40.49a + 62.99b = 34.31 \\ 62.99a + 137.69b = 109.91 \end{cases}$$

$$\Rightarrow \begin{cases} a = -1.4 \\ b = 1.4 \\ c = 206.9 \end{cases}$$

故に，求める線形モデルの方程式は

$$z = -1.4x + 1.4y + 206.9$$

（※この式での計算値は右図の通りです。）

番号	ｳｴｽﾄ(z)	計算値
1	60	59.2
2	68	66.6
3	70	71.0
4	65	66.9
5	80	80.9
6	85	85.3
7	78	77.3
8	79	79.0
9	95	95.7
10	89	87.0

課題　エクセルを用いて，計算結果を確認せよ。

5.6　行列式

行列式とは？

行列 $A=\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ に対して，次の計算式を**行列式**といい，

記号$|A|$ で表す。

［行列式］　$|A|=\begin{vmatrix} a & b \\ c & d \end{vmatrix}=ad-bc$

②　①

このとき，次のこと[**クラメルの公式**]が成り立つことが
知られています（※３年生で学習します)。

連立方程式 $\begin{cases} \boxed{a}x+\boxed{b}y=p \\ \boxed{c}x+\boxed{d}y=q \end{cases}$ の解は，次式で求めることができる。

x の係数 a, c と入替え　　y の係数 b, d と入替え

$$x=\frac{\begin{vmatrix} p & b \\ q & d \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}=\frac{pd-bq}{ad-bc},\qquad y=\frac{\begin{vmatrix} a & p \\ c & q \end{vmatrix}}{\begin{vmatrix} a & b \\ c & d \end{vmatrix}}=\frac{aq-pc}{ad-bc}$$

連立方程式の左辺の係数

係数行列という　　[※証明は演習に出題]

例題　連立方程式 $\begin{cases} 2x+3y=12 \\ 5x-2y=11 \end{cases}$ をクラメルの公式を用いて解け。

［解答］

$$x=\frac{\begin{vmatrix} 12 & 3 \\ 11 & -2 \end{vmatrix}}{\begin{vmatrix} 2 & 3 \\ 5 & -2 \end{vmatrix}}=\frac{-24-33}{-4-15}=\frac{-57}{-19}=3$$

$$y=\frac{\begin{vmatrix} 2 & 12 \\ 5 & 11 \end{vmatrix}}{\begin{vmatrix} 2 & 3 \\ 5 & -2 \end{vmatrix}}=\frac{22-60}{-4-15}=\frac{-38}{-19}=2$$

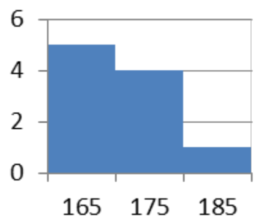
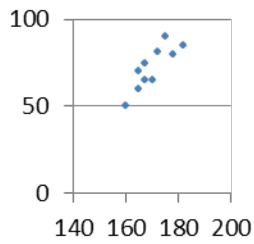
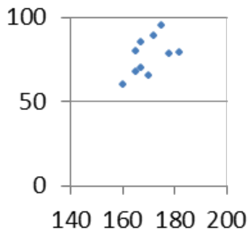
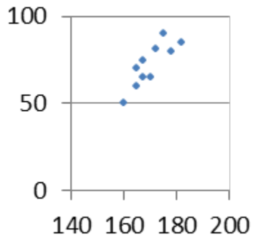
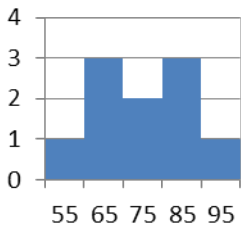
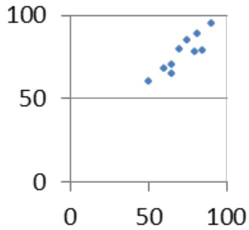
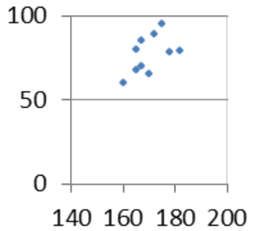
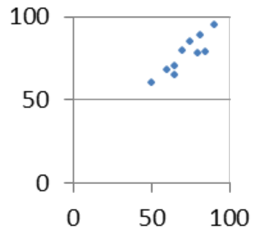
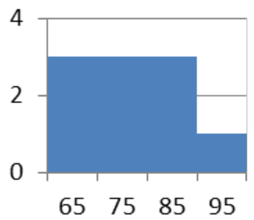
問 3.11　連立方程式 $\begin{cases} 3x-4y=6 \\ 2x+5y=1 \end{cases}$ をクラメルの公式を用いて解け。

5.7　散布図行列

散布図行列とは？

次のようなものを，**散布図行列**と言います。

[※データは 6 頁の例題のものを使用]

	身長(x)	体重(y)	ｳｴｽﾄ(z)
身長(x)			
体重(y)			
ｳｴｽﾄ(z)			

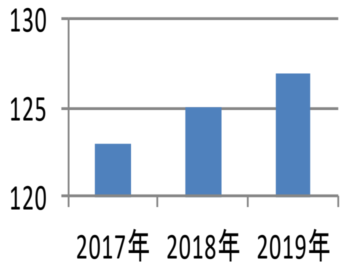
□□□□【参考】□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

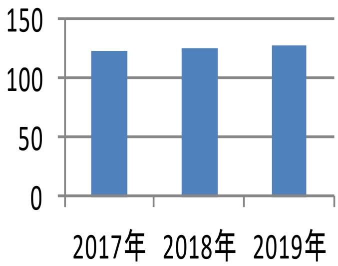
誇張表現の話をします。

データに関する数値やグラフを見る際には，

誇張表現等に騙されず，統計情報を正しく読み解くことが重要です。

例）軸目盛の設定を変えると印象が大きく変わります





□□