第９回　データ解析
（多変量解析）

目標

・多変量解析の基礎を学ぶ

0. 準備

今日の作業をするディレクトリを作成しなさい.

```
% mkdir 20131205
% cd 20131205
```

1. 多変量解析

多変量解析とは，多くの変数データの間に隠れた傾向を統計的に抽出する方法である．ここでは多変量解析の導入を学ぶ.

1.1. 重回帰分析

以前，２変量（例えば，y と x）データに対して，目的変数（y）の値を説明変数（x）の関係を一次式で表わせると仮定した際の最小二乗法フィッティングを学んだ（第三回「DO ループの応用」参照）．次は多変量データについて考える.

例えば，「テストの点数」と「テスト直前の勉強時間」との関係を考えると，おそらく両者には相関関係があると思われる．もし両者が下記のような一次式で関係付けられるとすると，

（テストの点数）= a + b x（テスト直前の勉強時間）

複数人のデータから a と b の値を最小二乗法で決定することで，勉強時間からテストの点数を予測することができるようになる．しかし実際には，「テスト直前の勉強時間」だけでなく，「講義の出席回数」や「課題の提出回数」，「IQ」によっても点数は変わるだろう．その場合，よりよいテストの点数の予測は

$$\text{（テストの点数）} = a_0 + a_1 \times \text{（テスト直前の勉強時間）} + a_2 \times \text{（講義の出席回数）} + a_3 \times \text{（課題の提出回数）} + a_4 \times \text{（IQ）} + \cdots$$

から得られる．このように１つの目的変数（この場合，テストの点数）を複数の説明変数で説明・予測できると仮定し，a_0〜a_nを決定する分析手法を重回帰分析という.

ここでは一次式でのフィッティングを考える.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = a_0 + \sum_{i=1}^{n} a_i x_i \quad (1)$$

関数のフィッティングの条件は，観測データ（ここでデータ数は p 個，データの番号を j で表す）と関数との差の自乗和

$$\begin{aligned} Q &= \sum_{j=1}^{p}\{y_j - (a_0 + \sum_{i=1}^{n} a_i x_{ij})\}^2 \\ &= \sum_{j=1}^{p}\{y_j^2 - 2y_j(a_0 + \sum_{i=1}^{n} a_i x_{ij}) + (a_0 + \sum_{i=1}^{n} a_i x_{ij})^2\} \end{aligned} \tag{2}$$

が最小になることである．よって，それぞれの a_i で Q を偏微分して 0 になる場合を考えればよい．

$$\begin{cases} \dfrac{\partial Q}{\partial a_0} = -2\displaystyle\sum_{j=1}^{p}\left\{y_j - \left(a_0 + a_1 x_{1j} + \ldots + a_n x_{nj}\right)\right\} = 0 \\ \dfrac{\partial Q}{\partial a_1} = -2\displaystyle\sum_{j=1}^{p} x_{1j}\left\{y_j - \left(a_0 + a_1 x_{1j} + \ldots + a_n x_{nj}\right)\right\} = 0 \\ \quad \cdots \\ \dfrac{\partial Q}{\partial a_n} = -2\displaystyle\sum_{j=1}^{p} x_{nj}\left\{y_j - \left(a_0 + a_1 x_{1j} + \ldots + a_n x_{nj}\right)\right\} = 0 \end{cases} \tag{3}$$

これを整理すると，

$$\begin{cases} a_0 m + a_1\displaystyle\sum_{j=1}^{p} x_{1j} + \cdots + a_n\sum_{j=1}^{p} x_{nj} = \sum_{j=1}^{p} y_j \\ a_0\displaystyle\sum_{j=1}^{p} x_{1j} + a_1\sum_{j=1}^{p} x_{1j}^2 + \cdots + a_n\sum_{j=1}^{p} x_{1j}x_{nj} = \sum_{j=1}^{p} x_{1j}y_j \\ \quad \cdots \\ a_0\displaystyle\sum_{j=1}^{p} x_{nj} + a_1\sum_{j=1}^{p} x_{1j}x_{nj} + \cdots + a_n\sum_{j=1}^{p} x_{nj}^2 = \sum_{j=1}^{p} x_{nj}y_j \end{cases} \tag{4}$$

となる．第一式を $a_0 = \overline{y} - a_1\overline{x}_1 - \cdots - a_n\overline{x}_n$ としたものを第二式以降に代入し，

$$\begin{aligned} S_{ik} &= \sum_{j=1}^{p}(x_{ij} - \overline{x}_i)^2 && (i = k = 1,2,\ldots,n) \\ S_{ik} &= \sum_{j=1}^{p}(x_{ij} - \overline{x}_i)(x_{kj} - \overline{x}_k) && (i \neq k, i = 1,2,\ldots,n, k = 1,2,\ldots,n) \\ S_{iy} &= \sum_{j=1}^{p}(x_{ij} - \overline{x}_i)(y_j - \overline{y}) && (i = 1,2,\ldots,n) \end{aligned} \tag{5}$$

とすると，$\partial Q/\partial a_i = 0$ の連立方程式は以下のようになる．

$$\begin{cases} S_{11}a_1 + S_{12}a_2 + \cdots + S_{1n}a_n = S_{1y} \\ S_{21}a_1 + S_{22}a_2 + \cdots + S_{2n}a_n = S_{2y} \\ \quad\cdots \\ S_{i1}a_1 + S_{i2}a_2 + \cdots + S_{in}a_n = S_{iy} \\ \quad\cdots \\ S_{n1}a_1 + S_{n2}a_2 + \cdots + S_{nn}a_n = S_{ny} \end{cases} \quad (6)$$

これを行列で表わすと以下のようになる.

$$\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1k} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2k} & \cdots & S_{2n} \\ \vdots & \vdots & \ddots & & & \vdots \\ S_{i1} & S_{i2} & & S_{ik} & & S_{in} \\ \vdots & \vdots & & & \ddots & \vdots \\ S_{n1} & S_{n2} & \cdots & S_{nk} & \cdots & S_{nn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_i \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{iy} \\ \vdots \\ S_{ny} \end{pmatrix} \quad (7)$$

よって，データ x_{ij}， y_j から S_{ik}， S_{iy} を計算し，(7)式の連立方程式を解けば，求めたい(1)式の係数 a_1 ～ a_n が計算できる.

a_0 は $a_0 = \overline{y} - a_1\overline{x}_1 - a_2\overline{x}_2 - \cdots - a_n\overline{x}_n$ から計算できる.

練習問題１

以下のようなデータがある．これを(3)式でフィッティングせよ.

y	x1	x2	x3	x4
-13.5D0	1.2D0	3.2D0	2.2D0	4.D0
10.3D0	3.5D0	4.D0	3.3D0	3.D0
-64.5D0	-5.D0	6.D0	5.5D0	6.D0
44.2D0	4.3D0	-3.D0	6.D0	2.D0
10.D0	2.D0	3.D0	2.D0	1.D0
4.8D0	5.5D0	4.3D0	5.D0	7.D0

http://www.eps.nagoya-u.ac.jp/~morota/NumericalAnalysis/data03.dat からダウンロード可能.

練習問題２

練習問題１の答えが正しいかを確認する為に， $Z_j = a_1x_{1j} + a_2x_{2j} + a_3x_{3j} + a_4x_{4j}$ を算出し，gnuplot で Z_j を横軸， y_j を縦軸とした図を描き，直線性を確認せよ.

1.2. 相関関係

まずは２変量解析を学ぶ．２変量の関係の強さを表わす指標として相関係数が使われる．相関係

数は以下のように与えられる.

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (8)$$

これは x と y の共分散を x と y のそれぞれの標準偏差で割ったものである．または n 次元における２つのベクトル $\mathbf{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})$，$\mathbf{y} = (y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})$ のなす角の余弦である.

相関係数は 1 から-1 の値をとり，一般的には相関係数の絶対値が 0.7 より大きいと高い相関，0.4～0.7 は相関があるとされる. しかし実際に有意であるか否かはデータ数によっても異なる. そのため, 得られた相関係数からデータ間に相関があるか否かの判断をする際には無相関検定を行う必要がある．無相関検定とは，母集団の相関係数が 0（無相関）であると仮定したとき，観測データの相関係数が起こりえる確率から仮定の妥当性を評価する検定法である.

有意水準ごと（10%，5%，2%，1%）のデータ数と相関係数の関係は滋賀大学の中川雅央氏の HP などを参照.

http://www.biwako.shiga-u.ac.jp/sensei/mnaka/ut/rtable.html

例えば，10 個の２変数の観測データの相関係数が 0.7 であったとき，5%の有意水準での相関係数は表より 0.63 であることから，相関関係は 95%の確率で有意であると言える.

次に多変量データの相関を考える．n 個の標本に対してデータ組が m 個ある場合を考える．そのような多変量データは以下のような n x m の行列で表現できる.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \qquad (9)$$

この行列の列ごとのデータは平均と分散が異なっている. そのため以下のようにデータ列の標準化を行う.

$$X' = \begin{pmatrix} \frac{x_{11}-\overline{x}_1}{\sigma_1} & \frac{x_{12}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{1m}-\overline{x}_m}{\sigma_m} \\ \frac{x_{21}-\overline{x}_1}{\sigma_1} & \frac{x_{22}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{2m}-\overline{x}_m}{\sigma_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{x_{n1}-\overline{x}_1}{\sigma_1} & \frac{x_{n2}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{nm}-\overline{x}_m}{\sigma_m} \end{pmatrix} \quad (10)$$

この行列の転置行列をかけると相関行列が得られる.

$$R = \frac{1}{n} X'^T X'$$

$$= \frac{1}{n} \begin{pmatrix} \frac{x_{11}-\overline{x}_1}{\sigma_1} & \frac{x_{21}-\overline{x}_1}{\sigma_1} & \cdots & \frac{x_{n1}-\overline{x}_1}{\sigma_1} \\ \frac{x_{12}-\overline{x}_2}{\sigma_2} & \frac{x_{22}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{n2}-\overline{x}_2}{\sigma_2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{x_{1m}-\overline{x}_m}{\sigma_m} & \frac{x_{2m}-\overline{x}_m}{\sigma_m} & \cdots & \frac{x_{nm}-\overline{x}_m}{\sigma_m} \end{pmatrix} \begin{pmatrix} \frac{x_{11}-\overline{x}_1}{\sigma_1} & \frac{x_{12}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{1m}-\overline{x}_m}{\sigma_m} \\ \frac{x_{21}-\overline{x}_1}{\sigma_1} & \frac{x_{22}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{2m}-\overline{x}_m}{\sigma_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{x_{n1}-\overline{x}_1}{\sigma_1} & \frac{x_{n2}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{nm}-\overline{x}_m}{\sigma_m} \end{pmatrix}$$

$$= \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} \frac{(x_{i1}-\overline{x}_1)^2}{\sigma_1^2} & \sum_{i=1}^{n} \frac{(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)}{\sigma_1\sigma_2} & \cdots & \sum_{i=1}^{n} \frac{(x_{i1}-\overline{x}_1)(x_{im}-\overline{x}_m)}{\sigma_1\sigma_m} \\ \sum_{i=1}^{n} \frac{(x_{i2}-\overline{x}_2)(x_{i1}-\overline{x}_1)}{\sigma_2\sigma_1} & \sum_{i=1}^{n} \frac{(x_{i2}-\overline{x}_2)^2}{\sigma_2^2} & \cdots & \sum_{i=1}^{n} \frac{(x_{i2}-\overline{x}_2)(x_{im}-\overline{x}_m)}{\sigma_2\sigma_m} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} \frac{(x_{im}-\overline{x}_m)(x_{i1}-\overline{x}_1)}{\sigma_m\sigma_1} & \sum_{i=1}^{n} \frac{(x_{im}-\overline{x}_m)(x_{i2}-\overline{x}_2)}{\sigma_m\sigma_2} & \cdots & \sum_{i=1}^{n} \frac{(x_{im}-\overline{x}_m)^2}{\sigma_m^2} \end{pmatrix} \quad (11)$$

この相関行列の各要素は２つの変量の相関係数を表わしている.

練習問題３

相関行列を求めるプログラムを作成し，下記のデータの相関関係を調べよ．どの項目に有意な相関があるか.

■惑星データ

http://www.eps.nagoya-u.ac.jp/~morota/NumericalAnalysis/data_planet01.dat

項目は，太陽間距離 [10^8km]，赤道半径 [km]，赤道重力（地球=1），体積（地球=1），衛星数，質量（地球=1），密度 [g/cm^3]，脱出速度 [km/s]，自転周期 [日]，公転周期 [日]，反

射率である.

■2013 年名古屋市気象データ

http://www.eps.nagoya-u.ac.jp/~morota/NumericalAnalysis/data_planet01.dat

項目は，月，平均気圧 [hPa]，降水量 [mm]，平均気温 [˚C]，平均湿度 [%]，平均風速 [m/s]，日照時間 [h]である.

■2012 年度セントラル・リーグ打撃成績（日本野球機構）

http://bis.npb.or.jp/2012/stats/bat_c.html

上記をテキストデータにしたものは以下から

http://www.eps.nagoya-u.ac.jp/~morota/NumericalAnalysis/data_baseball01.dat

「ホームランバッター = 足が遅い」は本当か？

2. 宿題

課題の進み具合で宿題を決めます.

宿題の提出先: 城野 (sirono@eps.nagoya-u.ac.jp)
諸田 (morota@eps.nagoya-u.ac.jp)
加藤 (katou.shinsuke@h.mbox.nagoya-u.ac.jp)

宿題のしめきり： 12 月 10 日（火曜日）

3. ログアウト

作業が終了したら必ずログアウトすること.

・ログアウト

「 マーク」=>「…をログアウト」