第１２回　データ解析
（多変量解析，補間法）

目標

・多変量解析の基礎を学ぶ

・データの補間法を学ぶ

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 20141218
% cd 20141218
```

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.

1.2. 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.

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

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.

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

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.

$$X' = \begin{pmatrix} \frac{x_{11}-\overline{x}_1}{\sigma_1} & \frac{x_{12}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{1m}-\overline{x}_m}{\sigma_m} \\ \frac{x_{21}-\overline{x}_1}{\sigma_1} & \frac{x_{22}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{2m}-\overline{x}_m}{\sigma_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{x_{n1}-\overline{x}_1}{\sigma_1} & \frac{x_{n2}-\overline{x}_2}{\sigma_2} & \cdots & \frac{x_{nm}-\overline{x}_m}{\sigma_m} \end{pmatrix} \quad (3)$$

この行列の転置行列をかけると相関行列が得られる.

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

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

$$= \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} \frac{(x_{i1}-\bar{x}_1)^2}{\sigma_1^2} & \sum_{i=1}^{n} \frac{(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sigma_1\sigma_2} & \cdots & \sum_{i=1}^{n} \frac{(x_{i1}-\bar{x}_1)(x_{im}-\bar{x}_m)}{\sigma_1\sigma_m} \\ \sum_{i=1}^{n} \frac{(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1)}{\sigma_2\sigma_1} & \sum_{i=1}^{n} \frac{(x_{i2}-\bar{x}_2)^2}{\sigma_2^2} & \cdots & \sum_{i=1}^{n} \frac{(x_{i2}-\bar{x}_2)(x_{im}-\bar{x}_m)}{\sigma_2\sigma_m} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} \frac{(x_{im}-\bar{x}_m)(x_{i1}-\bar{x}_1)}{\sigma_m\sigma_1} & \sum_{i=1}^{n} \frac{(x_{im}-\bar{x}_m)(x_{i2}-\bar{x}_2)}{\sigma_m\sigma_2} & \cdots & \sum_{i=1}^{n} \frac{(x_{im}-\bar{x}_m)^2}{\sigma_m^2} \end{pmatrix} \quad (4)$$

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１

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.

http://www.eps.nagoya-u.ac.jp/~morota/NumericalAnalysis/Data141218.html

■惑星データ

項目は，太陽距離 [10^8km]，赤道半径 [km]，赤道重力（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]である.

■2013 年名古屋市気象データ

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

の一部のデータをテキストデータにしたもの.

「ホームランバッター＝足が遅い」は本当か？「ホームランバッター＝四球が多い」は本当か？

2. データの補間法

離散データの解析において, データ点の内挿や外挿を行う場合や解析的に解けない問題を扱う場合に補間法が重要となるケースがある. ここでは代表的な補間法としてプログラミングが比較的容易であるラグランジュの補間法を学ぶ.

n+1 個のデータ点 (x_i , y_i) が与えられ，未知関数 $y = f(x)$ で関係づけられているとき，点 x_i 以外における y の値を求める場合を考える．ラグランジュの補間法は全てのデータ点を通る n 次多項式

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n+1} x^n = \sum_{i=0}^{n} a_i x^i \qquad (5)$$

を決定し，離散的に与えられた点 x_i 以外における $f(x)$ に近似しようというものである．$P_n(x)$ は次のラグランジュの補間多項式で与えられる．

$$\begin{aligned} P_n(x) &= y_0 \frac{(x-x_1)(x-x_2)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)\cdots(x_0-x_n)} + y_1 \frac{(x-x_0)(x-x_2)\cdots(x-x_n)}{(x_1-x_0)(x_1-x_2)\cdots(x_1-x_n)} \\ &\quad \cdots + y_n \frac{(x-x_1)(x-x_2)\cdots(x-x_{n-1})}{(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})} \\ &= \sum_{i=0}^{n} y_i \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)} \end{aligned} \qquad (6)$$

ラグランジュの補間法の問題点は新たにデータ点 (x_{i+1}, y_{i+1}) が加えられた時に，式 (6) の多項式の算出を最初からやり直さないといけないことである．また，データ点の数が非常に多い場合に大きな誤差を伴うことがある．

練習問題２

x が 0˚から 90˚の間で 30˚，10˚，2˚間隔で $\cos x$ の値が与えられたとき，ラグランジュの補間法によって 45˚と 1˚での値を計算し，真の値と比較せよ．

3. 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)
野上 (nogami.tatsuhiko@g.mbox.nagoya-u.ac.jp)

宿題のしめきり： 12 月 23 日（火曜日）

4. 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」