

第 12 回 データ解析 2 (補間法, 多変量解析)

目標

- ・データの補間法を学ぶ (前回の残り分)
- ・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てください.

```
% mkdir 20121227
```

```
% cd 20121227
```

1. 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

$$P_n(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{n+1}x^n = \sum_{i=0}^n a_i x^i \quad (1)$$

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.

$$\begin{aligned} P_n(x) &= y_0 \frac{(x-x_1)(x-x_2)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)\cdots(x_0-x_n)} + y_1 \frac{(x-x_0)(x-x_2)\cdots(x-x_n)}{(x_1-x_0)(x_1-x_2)\cdots(x_1-x_n)} \\ &\quad \cdots + y_n \frac{(x-x_0)(x-x_1)\cdots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1})} \\ &= \sum_{i=0}^n y_i \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)} \end{aligned} \quad (2)$$

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 (2) 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 1

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.

2. 多変量解析

多変量解析とは, 多くの変数データの中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.

2.1. 重回帰分析

以前, 2 変量データに対して一次式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 x_i の一次式でのフィッティングを考える.

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n = a_0 + \sum_{i=1}^n a_ix_i \quad (3)$$

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 m) と関数との差の自乗和

$$Q = \sum_{j=1}^m \{y_j - (a_0 + \sum_{i=1}^n a_ix_{ij})\}^2 \quad (4)$$

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.

$$\begin{cases} \frac{\partial Q}{\partial a_0} = -2 \sum_{j=1}^m \{y_j - (a_0 + a_1x_{1j} + \cdots + a_nx_{nj})\} = 0 \\ \frac{\partial Q}{\partial a_1} = -2 \sum_{j=1}^m x_{1j} \{y_j - (a_0 + a_1x_{1j} + \cdots + a_nx_{nj})\} = 0 \\ \cdots \\ \frac{\partial Q}{\partial a_n} = -2 \sum_{j=1}^m x_{nj} \{y_j - (a_0 + a_1x_{1j} + \cdots + a_nx_{nj})\} = 0 \end{cases} \quad (5)$$

これを整理すると,

$$\begin{cases} a_0m + a_1 \sum_{j=1}^m x_{1j} + \cdots + a_n \sum_{j=1}^m x_{nj} = \sum_{j=1}^m y_j \\ a_0 \sum_{j=1}^m x_{1j} + a_1 \sum_{j=1}^m x_{1j}^2 + \cdots + a_n \sum_{j=1}^m x_{1j}x_{nj} = \sum_{j=1}^m x_{1j}y_j \\ \cdots \\ a_0 \sum_{j=1}^m x_{nj} + a_1 \sum_{j=1}^m x_{1j}x_{nj} + \cdots + a_n \sum_{j=1}^m x_{nj}^2 = \sum_{j=1}^m x_{nj}y_j \end{cases} \quad (6)$$

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

$$\begin{aligned}
S_{ii} &= \sum_{j=1}^m (x_{ij} - \bar{x}_i)^2 & (i = 1, 2, \dots, n) \\
S_{ik} &= \sum_{j=1}^m (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k) & (i \neq k, i = 1, 2, \dots, n, k = 1, 2, \dots, n) \\
S_{iy} &= \sum_{j=1}^m (x_{ij} - \bar{x}_i)(y_j - \bar{y}) & (i = 1, 2, \dots, n)
\end{aligned} \tag{7}$$

とすると、 $\partial Q / \partial a_i = 0$ の連立方程式は以下ようになる。

$$\begin{cases}
S_{11}a_1 + S_{12}a_2 + \dots + S_{1n}a_n = S_{1y} \\
S_{21}a_1 + S_{22}a_2 + \dots + S_{2n}a_n = S_{2y} \\
\vdots \\
S_{n1}a_1 + S_{n2}a_2 + \dots + S_{nn}a_n = S_{ny}
\end{cases} \tag{8}$$

これを行列で表わすと以下ようになる。

$$\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n1} & S_{n2} & \cdots & S_{nn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{ny} \end{pmatrix} \tag{9}$$

a_0 は $a_0 = \bar{y} - a_1\bar{x}_1 - \dots - a_n\bar{x}_n$ から計算できる。

練習問題 2

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/data02.dat> からダウンロード可能。

2.2. 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。

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (10)$$

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 $\mathbf{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$, $\mathbf{y} = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$ 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 $n \times m$ の行列で表現できる。

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

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。

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

この行列の転置行列をかけると相関行列が得られる.

$$\begin{aligned}
 R &= \frac{1}{n} X'^T X' \\
 &= \frac{1}{n} \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sigma_1} & \frac{x_{21}-\bar{x}_1}{\sigma_1} & \dots & \frac{x_{n1}-\bar{x}_1}{\sigma_1} \\ \frac{x_{12}-\bar{x}_2}{\sigma_2} & \frac{x_{22}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{n2}-\bar{x}_2}{\sigma_2} \\ \dots & \dots & \dots & \dots \\ \frac{x_{1m}-\bar{x}_m}{\sigma_m} & \frac{x_{2m}-\bar{x}_m}{\sigma_m} & \dots & \frac{x_{nm}-\bar{x}_m}{\sigma_m} \end{pmatrix} \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sigma_1} & \frac{x_{12}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{1m}-\bar{x}_m}{\sigma_m} \\ \frac{x_{21}-\bar{x}_1}{\sigma_1} & \frac{x_{22}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{2m}-\bar{x}_m}{\sigma_m} \\ \dots & \dots & \dots & \dots \\ \frac{x_{n1}-\bar{x}_1}{\sigma_1} & \frac{x_{n2}-\bar{x}_2}{\sigma_2} & \dots & \frac{x_{nm}-\bar{x}_m}{\sigma_m} \end{pmatrix} \\
 &= \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n \frac{(x_{i1}-\bar{x}_1)^2}{\sigma_1^2} & \sum_{i=1}^n \frac{(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sigma_1\sigma_2} & \dots & \sum_{i=1}^n \frac{(x_{i1}-\bar{x}_1)(x_{im}-\bar{x}_m)}{\sigma_1\sigma_m} \\ \sum_{i=1}^n \frac{(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1)}{\sigma_2\sigma_1} & \sum_{i=1}^n \frac{(x_{i2}-\bar{x}_2)^2}{\sigma_2^2} & \dots & \sum_{i=1}^n \frac{(x_{i2}-\bar{x}_2)(x_{im}-\bar{x}_m)}{\sigma_2\sigma_m} \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^n \frac{(x_{im}-\bar{x}_m)(x_{i1}-\bar{x}_1)}{\sigma_m\sigma_1} & \sum_{i=1}^n \frac{(x_{im}-\bar{x}_m)(x_{i2}-\bar{x}_2)}{\sigma_m\sigma_2} & \dots & \sum_{i=1}^n \frac{(x_{im}-\bar{x}_m)^2}{\sigma_m^2} \end{pmatrix} \quad (13)
 \end{aligned}$$

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.

■ 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/data_baseball01.dat

■ 惑星データ

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

項目は, 太陽間距離 [10⁸km], 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/cm³], 脱出速度 [km/s], 自転周期 [日], 公転周期 [日], 反射率である.

3. 宿題

練習問題 1 と 3 のプログラムと結果を下記アドレスに送ること.

宿題の提出先: 城野 (sirono@eps.nagoya-u.ac.jp)

諸田 (morota@eps.nagoya-u.ac.jp)

前島 (maeshima.naohiko@c.mbox.nagoya-u.ac.jp)

宿題のしめきり: 1 月 8 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してください.

・ ログアウト

「 マーク」 => 「…をログアウト」