データ解析 (直線によるデータの当てはめ)

学籍番号 045713C:大城和也

提出日:平成 18 年 02 月 13 日 (月)

実験共同者
045709E：上原直久
045734B：知念栄作
045739G：友寄雄一朗

1　目的

本実験では，与えられたデータに直線を当てはめるアルゴリズムを C++ 言語を用いてプログラミングする事を目的とす.

2　課題内容

課題 1. 3 つの補足プログラムの動作を確認せよ.

課題 2. 線形回帰，最小二乗法について調べて簡潔にまとめよ.

課題 3. カイ 2 乗関数について調べて簡潔にまとめよ.

課題 4. 線形回帰当てはめプログラム (lin-reg.cpp) を完成させよ.

課題 5. octave(gnuplot) を用いて結果をプロットし考察せよ.

3　報告事項

3.1　動作の確認

これについての報告は省略する.

3.2　線形回帰，最小二乗法について

3.2.1　線形回帰

線形回帰とは回帰分析において従属変数 (以下の y) と説明変数 (以下の x) との関係式を一次式として当てはめる方法である．つまりは，2 つの変数からなる点の分布を最もらしい直線で現そうとするもので，その直線を求めるには最小二乗法がよく知られている．この求められた直線からは 2 つの変数の関係を読みとる事や，1 つの値から未知のもう一つの値をある程度予測できたりする.

$$y = a_1 + a_2 x$$

ちなみに，線形回帰は直線回帰，単回帰などとも呼ばれる．また，関係式を一次式以外で回帰分析を行うことを一般的に非線形回帰と呼ばれる.

3.2.2 最小二乗法

最小二乗法とは想定した特定の関数とデータとを近似させるとき，その関数がデータの値に対して最も近似するように，関数の理想値とデータの実際の値との誤差の 2 乗和を求め，それが最小となるように係数を変化させる方法である．

具体的な方法 (線形回帰の場合) は以下のようになる．

まず，図 1 を見る．この時，直線 $y = ax + b$ の理想の値はそれぞれの点に対して，d_1，d_2，d_3，d_4 の誤差がある．この誤差の合計 (S) が最小になればいいので次のような式が導きだされる．ちなみに二乗するのはマイナスの値を出さないようにするため．

$$S = {d_1}^2 + {d_2}^2 + {d_3}^2 + {d_4}^2$$

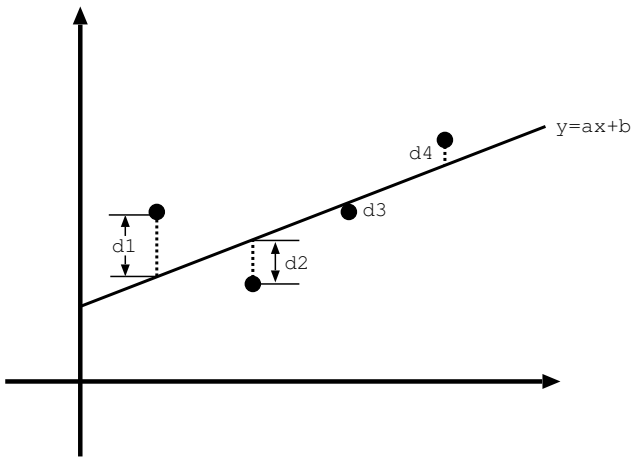

図 1: 線形回帰

この式を最小にすればいい．さきほど導いた式は二次関数であり，その最小値を求めればいいので，偏微分した値が 0 になるように係数を調整すればいい．

線形回帰の場合の一般式は以下のようになる．

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

$$b = \frac{\sum_{i=1}^{n} {x_i}^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} {x_i}^2 - (\sum_{i=1}^{n} x_i)^2}$$

今回の実験ではこれのさらに推定誤差範囲を用いた形式であるカイ 2 乗関数を扱っている.

3.3 カイ 2 乗関数について

カイ二乗関数は誤差を現す関数である．カイ二乗関数では推定誤差範囲をもうけることにより，以下のような関数となっている.

$$\text{誤差} = \sum_{i=1}^{N} \left(\frac{\triangle_i}{\sigma_i}\right)^2$$

この関数はカイ二乗検定として使える．また，誤差が正規分布に属すればその当てはめの妥当性などが調べることが可能．これらの理由により，カイ二乗の関数は当てはめ関数として一般的に使われることが多い.

3.4 lin-reg.cpp の完成プログラム

今回，以下の内容を追加し lin-reg.cpp のプログラムを完成させた.

```
-------------------------------------------------------------------------------
#include "NumMeth.h"

void lin_reg(Matrix x, Matrix y, Matrix sigma,
             Matrix &a_fit, Matrix &sig_a, Matrix &yy, double &chisqr) {

  //* Evaluate various sigma sums
  int i, nData = x.nRow();
  double sigmaTerm;
  double s = 0.0, sx = 0.0, sy = 0.0, sxy = 0.0, sxx = 0.0;
  for( i=1; i<=nData; i++ ) {
    //式 (5.10) の計算を行う
    sigmaTerm = sigma(i)*sigma(i);
    s  += 1.0/sigmaTerm;
    sx += x(i)/sigmaTerm;
    sy += y(i)/sigmaTerm;
    sxy += x(i)*y(i)/sigmaTerm;
    sxx += x(i)*x(i)/sigmaTerm;
  }

  //* 切片 a_fit(1) と 傾き a_fit(2) の計算
  a_fit(1) = ((sy*sxx)-(sx*sxy))/((s*sxx)-(sx*sx));
  a_fit(2) = ((s*sxy)-(sy*sx))/((s*sxx)-(sx*sx));

  //* 切片と傾きの誤差範囲を計算
  sig_a(1) = sqrt(sxx/(s*sxx-sx*sx));
  sig_a(2) = sqrt(s/(s*sxx-sx*sx));

  //* 推定値とカイ二乗関数の計算
  chisqr = 0.0;
```

```
  for( i=1; i<=nData; i++ ){
    sigmaTerm = sigma(i) * sigma(i);
    yy(i) = a_fit(1) + (a_fit(2)*x(i));
    chisqr += (1/sigmaTerm)*
      (a_fit(1)+(a_fit(2)*x(i))-y(i))*(a_fit(1)+(a_fit(2)*x(i))-y(i));
  }
}
-------------------------------------------------------------------------------
```

3.4.1 追加部分の説明

今回追加した部分はs, sx, sy, sxy, sxx, a_fit(1), a_fit(2), sig_a(1), sig_a(2), yy(i)，chisqr など，線形回帰のための計算を行う場所である．

s,sx,sy,sxy,sxx

これらの変数は後々扱う，

$$\sum_{i=1}^{N}\frac{1}{\sigma^2},\sum_{i=1}^{N}\frac{x_i}{\sigma^2},\sum_{i=1}^{N}\frac{y_i}{\sigma^2},\sum_{i=1}^{N}\frac{x_i^2}{\sigma^2},\sum_{i=1}^{N}\frac{x_iy_i}{2}$$

らとそれぞれ対応しており，for ループを用いることによって実現している．

a_fit(1)，a_fit(2)

この値は先ほどの値を用いて，一次関数の傾きと切片を求める．その関数は最小二乗法で述べた一般公式と同等である．

sig_a(1)，sig_a(2)

これらの値は求めた切片と傾きの推定誤差を求める計算である．その計算式は以下のようになっている．

$$\begin{aligned} sig\_a(1) &= \sqrt{\frac{\sum x^2}{s\sum x^2-(\sum x)^2}} \\ sig\_a(2) &= \sqrt{\frac{s}{s\sum x^2-(\sum x)^2}} \end{aligned}$$

yy(i)

これは線形回帰として当てはめた式であり，一次式の形となって扱われる．

chisqr

これがカイ二乗の公式に当てはまる変数となっている．対応している式は次のようになっている．

$$chisqr=\sum_{i=1}^{N}\frac{1}{\sigma_i^2}(a_1+a_2x_i-y_i)^2$$

$\sum$ の部分は先と同様，for ループで足し続けることによって成り立っている．

3.5　結果のプロットとその考察

実行結果は以下の図 2〜図 5 のようになった．この時，データ作成のために C(1)，C(2)，C(3) と Errbar の数値を入力するが，基本値を 10 とし，何か一つの値だけ変化させたのが今回の図である．

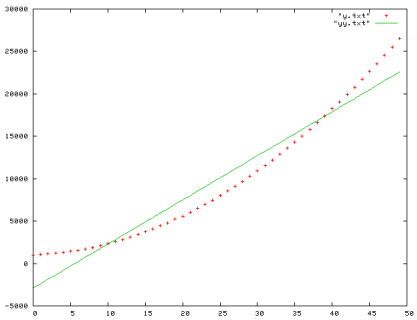

図 2: C(1)=1000 の時のグラフ

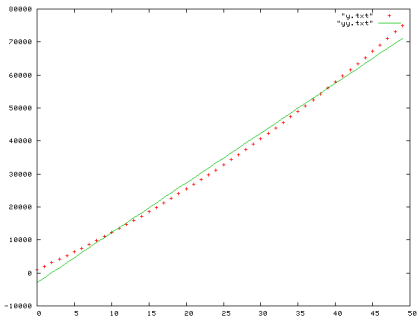

図 3: C(2)=1000 の時のグラフ

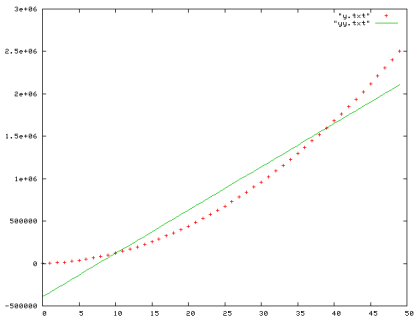

図 4: C(3)=1000 のグラフ

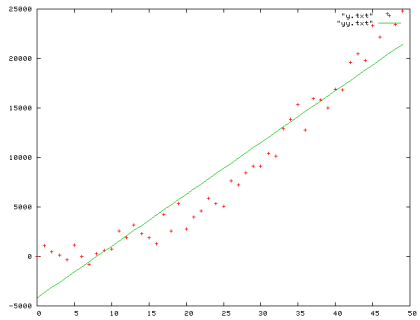

図 5: Errbar=1000 のグラフ

C(1)，C(2)，C(3)，Errbar は今回，データを作成する際のパラメータとして使用している．データの式は以下のようになっている．

$$y(x) = c(1) + c(2) * x + c(3) * x^2 + Errbar * randam$$

図 3 では他の図と比べデータが直線的である．つまり一次の係数が大きくなったためこうなったと思われる．また，図 4 では x^2 の係数のため急激に値が上昇していることがわかる．Errbar はノイズの大きさを変化させるので Errbar を大きくすると 5 のようにバラツキがでてくることがわかる．

これらのデータを用いた時も大体中心くらいを通る形で直線で近似できていることがわかる．

参考文献

[1] 最小二乗法について
”http://szksrv.isc.chubu.ac.jp/lms/lms1.html”

[2] Wikipedia
”http://ja.wikipedia.org/wiki/メインページ”

[3] 統計用語 2
”http://w3.cc.nagasaki-u.ac.jp/contrib/Excel/yougo2.html”

[4] カイ二乗を学ぶ
”http://w3.cc.nagasaki-u.ac.jp/contrib/Excel/yougo2.html.”