実験1 課題

学籍番号 035743A ：比嘉 雅樹

平成 16 年 4 月 30 日

1 課題

1. 与えられたデータに対する回帰直線を求める awk スクリプトを作成せよ。
2. gnuplot を用いて実データと回帰直線をプロットせよ。
3. リスト及びグラフを TEX に取り込んで最小自乗法の展開式と共に記し、最小自乗法によって得られる理由 (または、回帰直線の持つ意味) を説明せよ。
4. perl または C 言語を用いて今回の実験と同様の回帰直線を求めるプログラムを作成せよ。
5. 実は回帰直線は gnuplot 単体でも求めることができる。実際に gnuplot を使って回帰直線を求めよ。また、回帰曲線 ($y = ax^2 + bx + c$) で近似した場合のグラムを提示し、回帰直線の結果と比較せよ。

2 実験

2.1 回帰直線を求める awk スクリプトを作成せよ。

この回帰直線を求める awk スクリプトは別に提出した「j03043.awk」を参照してください。ちなみに傾きを a、切片を b とすると求めた結果は

$$\begin{aligned} a &= 0.763337 \\ b &= -63.8985 \end{aligned}$$

2.2 gnuplot を用いて実データと回帰直線をプロットせよ。

図 1 を参照してください。

2.3 最小自乗法の展開式と共に記し、最小自乗法によって得られる理由 (または、回帰直線の持つ意味) を説明せよ。

表 1: 身長・体重一覧表

	身長	体重		身長	体重		身長	体重		身長	体重
1	157.1	54.7	7	170.9	68.3	13	174.1	71.1	19	169.6	72
2	160.3	56.2	8	169.3	63.2	14	157.0	55.4	20	174.8	71.9
3	174.5	70.5	9	163	62.5	15	179.7	72.5	21	165.0	58
4	164.9	61.6	10	171.9	63.2	16	171.7	70.8	22	179.6	72.7
5	169.3	62.9	11	171.7	63.8	17	171.0	62.9	23	157.3	62.9
6	162.9	57.3	12	166.8	59.9	18	169.4	70.7	24	167	64.16

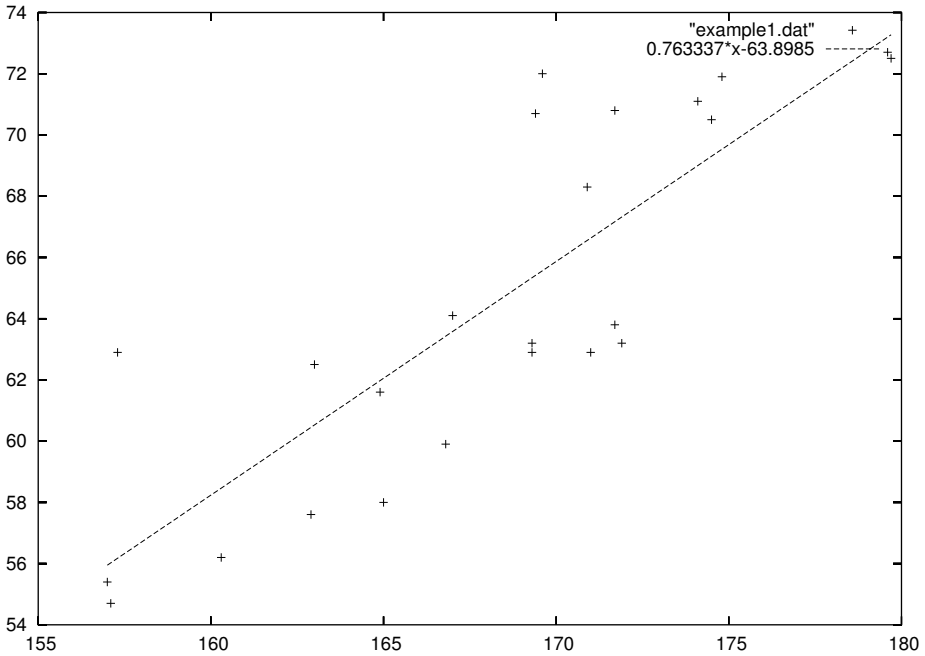

図 1: データ分布図と回帰直線

表 1 のデータで身長を x、体重を y とおいて図で表すと、図 1 のようなデータ分布になる。この図 1 のデータ分布から表 1 の身長・体重のデータには比例の関係があることが推測できる。

仮にデータの比例関係式を $y = ax + b$ とおく。
各データ (x_i, y_i) と直線上の点 $(x_i, ax_i + b)$ との距離の総和が最小になるような直線を回帰直線という。(図 1 参照)

この回帰直線の傾き及び切片は下記の式 f が最小になるような a, b を推定することで求まる。

$$f = \sum_{i=1}^{n} |y_i - (ax_i + b)|$$

上記 f の最小値は、a と b のそれぞれについて偏微分した時どちらも 0 になる場合の a,b を求めれば導き出せる。

ここで、f を偏微分したいのだが絶対値が付いている為、場合分をしなければならなく、面倒である。なので自乗を行い絶対値を外した後に偏微分を行う。

f を a で偏微分する

$$\begin{aligned} \frac{\partial f}{\partial a} &= \sum_{i=1}^{n} 2\,(-x_i)\,(y_i - ax_i - b) \\ &= \sum_{i=1}^{n} 2\left(-x_i y_i + ax_i^2 + bx_i\right) = 0 \end{aligned} \tag{1}$$

ここで $\{x_i\}$ の総和を T_x、平均値を $\bar{x}$、また $\{y_i\}$ についても同様に T_y、$\bar{y}$ とおくと (1) 式は

$$-T_{xy} + a\sum_{i=1}^{n} x_i^2 + bT_x = 0 \tag{2}$$

となる。次に b で偏微分すると

$$\begin{aligned} \frac{\partial f}{\partial b} &= \sum_{i=1}^{n} 2\,(-1)\,(y_i - ax_i - b) \\ &= \sum_{i=1}^{2} 2\,(-y_i + ax_i + b) \\ &= -T_y + aT_x + bn = 0 \end{aligned} \tag{3}$$

(3) 式を b について解くと

$$b = \frac{T_y - aT_x}{n} \tag{4}$$

(4) 式を (2) に代入

$$-T_{xy} + a\sum_{i=1}^{n} x_i^2 + bT_x = 0$$

$$
\begin{aligned}
-T_{xy} + a\sum_{i=1}^{n} x_i^2 + \frac{T_x T_y - aT_x^2}{n} &= 0 \\
a\left(n\sum_{i=1}^{n} x_i^2 - T_x^2\right) &= nT_{xy} - T_x T_y
\end{aligned}
\tag{5}
$$

よって、a は

$$
a = \frac{nT_{xy} - T_x T_y}{n\sum_{i=1}^{n} x_i^2 - T_x^2} \tag{6}
$$

そして (4) 式より

$$
\begin{aligned}
b &= \frac{1}{n}T_y - \frac{a}{n}T_x \\
b &= \bar{y} - a\bar{x}
\end{aligned}
\tag{7}
$$

2.4 perl,C を用いた回帰直線を求めるプログラム

別に提出した「j03043.c」と「j03043.pl」を参照してください。

2.5 gnuplot を使用して回帰直線、回帰曲線を求め比較する。

2.5.1 回帰直線を求める

gnuplot を使用して回帰直線を求めるには、下記のコマンドを入力します。

```
gnuplot > f(x)=a*x+b
gnuplot > fit f(x) "example1.dat" via a,b
```

2.5.2 回帰曲線を求める。

```
gnuplot > f(x)=a*x*x+b*x+c
gnuplot > fit f(x) "example1.dat" via a,b,c
```

2.5.3 比較

今回のデータの場合、回帰直線より回帰曲線の方がよりデータの値を近似している。ためしに 3 次の曲線〜5 次程まで曲線をプロットしてみたが、よりデータを近似していた。次ページに載せている図は図 1 の直線に 2 次の回帰曲線を加えたものです。

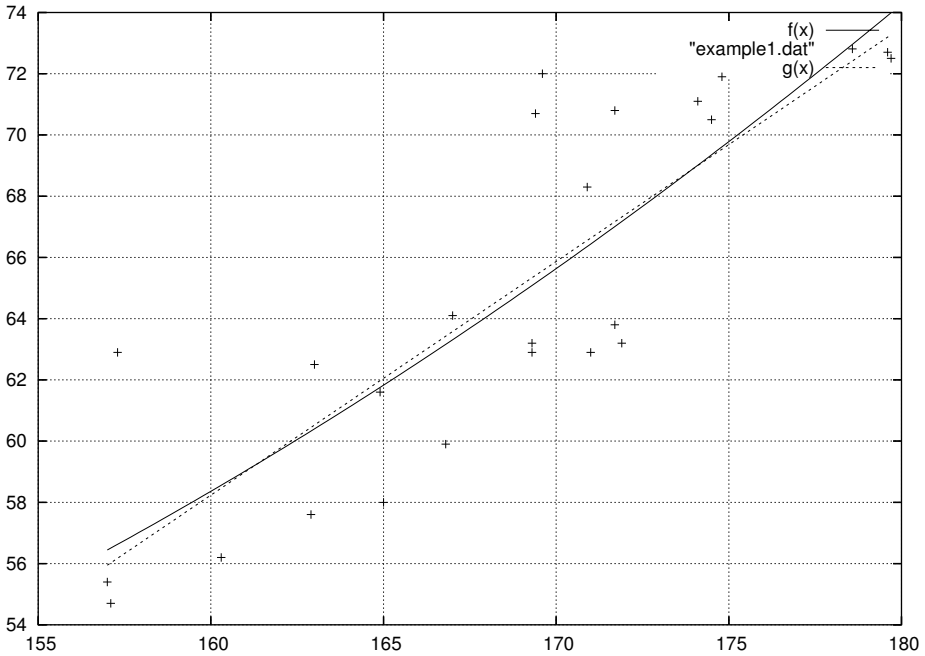

図 2: 回帰直線と回帰曲線

参考文献

[1] Perl 基礎講座 / 著：内田 保雄、富田 満 / 出版：オーム社 / pp.168-171

[2] http://www.kyoto-su.ac.jp/information/Guide/gnuplot/gnuplot.html