最小二乗近似

樋口さぶろお

龍谷大学理工学部数理情報学科

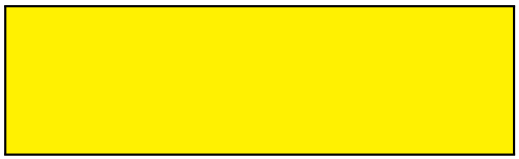
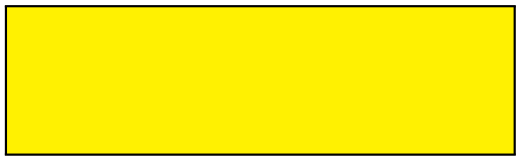
数値計算法 L11(2010-07-02)

今日の目標

1. 共分散・相関係数を計算して 2 つの量の関係をイメージできるようになろう.
2. 最小二乗近似の仕組みを理解して世の中のデータの分析に使えるようになろう.

hig3.net

2 変量データ

2 変量データ (x_i, y_i) $(i = 1, 2, 3, \ldots, n)$.
例: x_i:指輪のサイズ, y_i:身長, 出席番号 $i = 1, 2, 3, \ldots, n$.
以下, $\sum = \sum_{i=1}^{n}$.
一方のデータだけ見て考えると,

,

など

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)^2$$

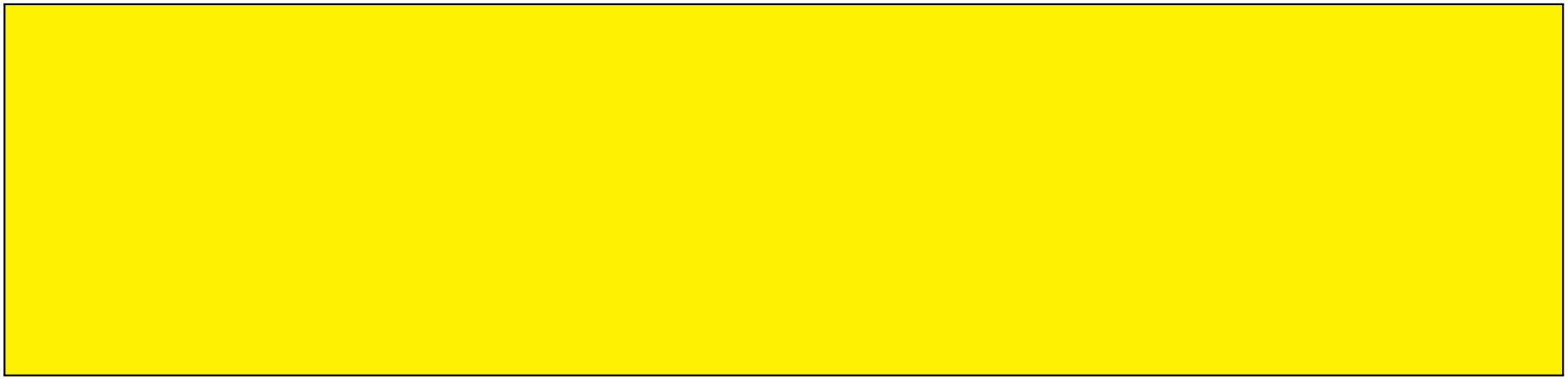
$$\mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad \sigma_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu_y)^2$$

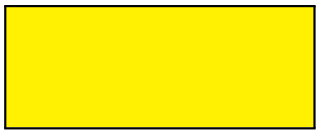
相関:x_i が大きいとき y_i は大きい (小さい) 傾向があるのか?

統計学入門, 東大出版会 (1991) 図 3.3-3.6(p.44) より引用

PDF 版では図省略

相関を判定するための量





$$C_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$x(y)$ を A 倍すると A 倍される. 同じ大きさなら同じ相関とはいえない.
そこで.

相関係数

正なら正の相関, 負なら負の相関. 絶対値が大きいと強い相関.

$$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{\frac{1}{n-1}\sum(x_i - \mu_x)(y_i - \mu_y)}{(\frac{1}{n-1}\sum(x_i - \mu_x)^2)^{1/2}(\frac{1}{n-1}\sum(y_i - \mu_y)^2)^{1/2}}$$

実は $-1 \leq r_{xy} \leq +1$. 以下 $\sum = \sum_{i=1}^{n}$.

相関係数の便利な式

$$
\begin{aligned}
r_{xy} &= \frac{\frac{1}{n-1}\sum(x_i-\mu_x)(y_i-\mu_y)}{(\frac{1}{n-1}\sum(x_i-\mu_x)^2)^{1/2}(\frac{1}{n-1}\sum(y_i-\mu_y)^2)^{1/2}} \\
&= \frac{\left(\frac{1}{n}\sum x_i y_i\right)-\mu_x\mu_y}{\left(\left(\frac{1}{n}\sum x_i^2\right)-\mu_x^2\right)^{1/2}\left(\left(\frac{1}{n}\sum y_i^2\right)-\mu_y^2\right)^{1/2}}
\end{aligned}
$$

次の量がわかればいい.

$$
B = \sum x_i^1 y_i^0, \quad E = \sum x_i^0 y_i^1,
$$
$$
D = \sum x_i^2 y_i^0, \quad F = \sum x_i^1 y_i^1, \quad G = \sum x_i^0 y_i^2.
$$

Quiz

次の 2 変量データから相関係数を求めよう.
$(x, y) = (1, 8), (3, 8), (4, 10), (4, 14)$

近似 1 次式

x, y 間の関係を表すいちばん安易な策:

y は x の 近似 1 次式 $a + bx$ で
と書けるとしたら?

中学校的テクニック

散布図で上下へのずれが小さくなるように直線 $y = a + bx$ をひいて傾き b と切片 a を読み取る.

本来 y が x の関数 $y = f(x)$ であるはずのとき, $y = a + bx$ を[　　]式, [　　]直線, a, b のことを[　　]係数, 式を求める分析のことを[　　]分析という.

x 独立変数
y 従属変数

最小二乗近似: 真剣にずれ最小化

ずれが小さいって何が小さいこと? 栗原 §4.1(p.94-100)
近似 1 次式を $y = a + bx$ とする.

最小二乗近似

次の量を最小化するように a, b を定めよう

$$Q(a,b) = \sum ((a + bx_i) - y_i)^2$$

極値を与える (a,b) を探そう $\rightsquigarrow$ $\frac{\partial Q}{\partial a}(a,b) = \frac{\partial Q}{\partial b}(a,b) = 0.$ 微積分 I

$$0 = \frac{\partial Q}{\partial a} = \sum 2(a + bx_i - y_i) = a \cdot 2n + b \cdot 2\sum x_i - 2\sum y_i.$$
$$0 = \frac{\partial Q}{\partial b} = \sum 2(a + bx_i - y_i)x_i = a \cdot 2\sum x_i + b \cdot 2\sum x_i^2 - 2\sum x_i y_i.$$

係数が複雑だけど未知数 a, b の連立 1 次方程式じゃん.

未知数 a, b の連立 1 次方程式.

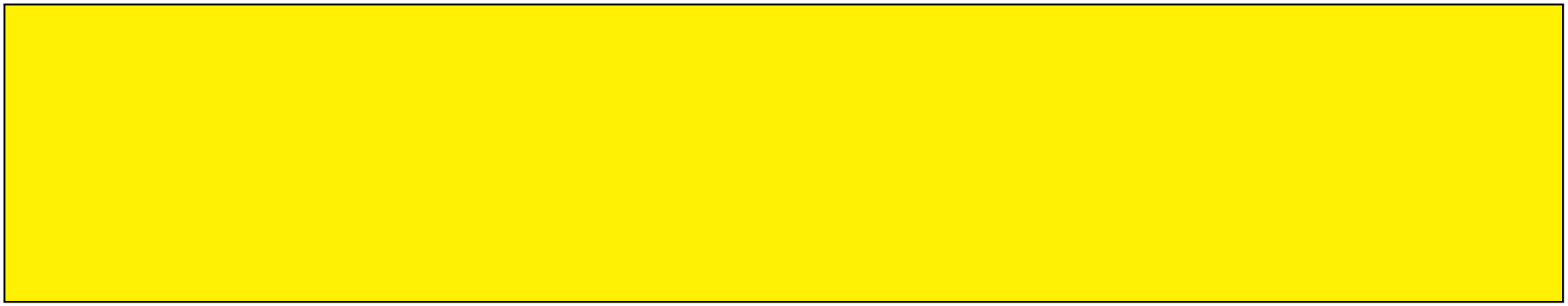
$$\begin{aligned} Aa + Bb - E &= 0 \\ Ca + Db - F &= 0 \end{aligned}$$

係数

$$\begin{aligned} A = n &= \sum x_i^0 y_i^0 = \sum 1 \\ B = C &= \sum x_i^1 y_i^0 = \sum x_i \\ D &= \sum x_i^2 y_i^0 = \sum x_i^2 \\ E &= \sum x_i^0 y_i^1 = \sum y_i \\ F &= \sum x_i^1 y_i^1 = \sum x_i y_i \end{aligned}$$

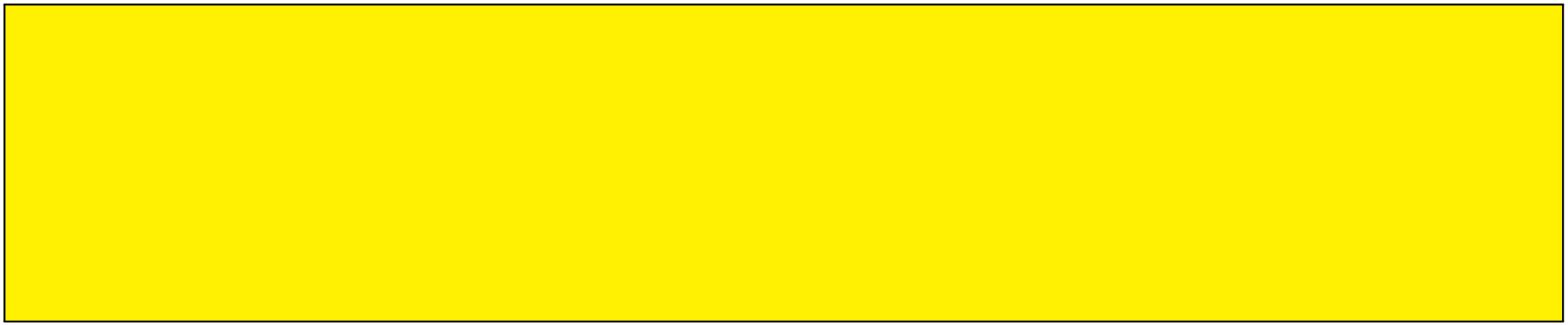
相関係数のところと変数名あわせた.

連立方程式を解こう.
加減法で,



栗原 p.97,98

切片 $a =$



傾き $b = \dfrac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_j y_j)}{n\sum x_i^2 - (\sum x_i)^2}$

解無しとか不定とか桁落ち (分母に差がある) とかが気になる. $\rightsquigarrow$ 待て
次週

```
int main(void){
  int n=10,i;
  double x[10],y[10];
  double sx=0.0,sy=0.0;                  /* ∑x_i, ∑y_i  */
  double sxx=0.0,sxy=0.0,syy=0.0;  /* ∑x_i^2, ∑x_i y_i, ∑y_i^2 */

  x[],y[] をfopen, while(fscanf), fclose で読み込む.

  for(i=0;i<n;i++){
    sx=sx+x[i];
    sxx=sxx+x[i]*x[i];
    sxy=sxy+x[i]*y[i];
    syy=....;
  }
  a=...;
  b=...;
  r=...;

  printf("%f %f %f\n",a,b,r);
  return 0;
}
```

b と r_{xy} の関係

$$r_{xy} = b \times \frac{((\frac{1}{n}\sum x_i^2) - \mu_x^2)^{1/2}}{((\frac{1}{n}\sum y_i^2) - \mu_y^2)^{1/2}}$$

実は,
ぴったり直線に乗る $\Leftrightarrow$ $\boxed{Q = 0 \Leftrightarrow r_{xy} = \pm 1}$

$\boxed{\text{相関係数 } r \text{ を求める}}$と, どの程度よくあっているかがわかる.

他人をだますための最小二乗近似

Anscombe の例. 自然科学の統計学 p.53.

PDF 版では図省略

一般の近似多項式

近似多項式 $y = a + bx + cx^2 + dx^3$ で, ずれが小さくなるように a, b, c, d を決める.

$\rightsquigarrow$ 正規方程式は 4 元連立 1 次方程式.

待て次週 (n 元連立 1 次方程式の解)

レポート課題 R11(講義)

Web 上のある程度信頼できる 2 変量データの中から自分が興味あるものを探し出し, 演習課題 E11 ののりで最小二乗近似を用いて近似 1 次式を求め、グラフを描こう (E111 ができてればそのまま使えばいいでしょ). また、そのグラフから読み取れること説明しよう. ただし、テーマはなるべく他の人と違うものにすること。他の人とほとんど同じテーマである場合は, 後から提出した人は採点の対象にならない.

2 変量データの例

- 気温とビールの売上
- 気温と発電所の最大電力
- テレビ視聴時間と学力
- 男性の年収と既婚率
- 年収と学習費

レポートは Word で作成し, A4 の PDF でアップロード. 枚数は 1 枚以上で自由 (多いほどよいわけではない). 次の要素を含むこと.

- 氏名と学籍番号
- 何のデータかの説明. データの出典 (URL).
- 近似 1 次式と相関係数
- データおよび近似 1 次式のグラフ
- 考察 (読み取れること)
- プログラムのソース (e111.c をそのまま Word 文書に貼り付ければいい)

科目の成績 100 点中の日常活動点 20 点のうちの 5 点分. 提出期間 2010-07-18—2010-07-26. e ラーニングシステム (講義) に. レポートは受講者に公開されることがあります.

演習の課題チェックについての大事な連絡

- 演習での課題のチェックは, 時間が余っていても 12:15 までの予約の分で終了します. チェックの結果不完全であっても再度のチェックはしません.
- 演習での 12:15 以降に予約された質問は, 12:15 以前の予約すべてに対応した後, 余裕がある場合のみ対応します. ただし, 質問のみでチェックは行いません.
- 2010-07-16 金に最後の講義. 2010-07-17 土 に任意参加の演習の補講を行う予定. これ以降はチェックを行いません. 最後は課題完成が忙しくなっているので注意. 課題はすでに最後のものまで公開済み.
- 今回以降の演習の課題完成チェックの, 未チェック訂正リクエストは, チェックした TA の名前をあわせて報告するようにお願いしています. これまでは名前がなくても該当 TA を探して照会していましたが, 今回以降は, チェックした TA の名前が不明な場合は対応しません. かならず記録しておいてね.

介護実習, 教育実習などへの対応 $\longrightarrow$ 先週の配布資料.