回帰分析

樋口さぶろお

龍谷大学理工学部数理情報学科

確率統計☆演習 I L06(2015-10-23 Fri)

最終更新: Time-stamp: "2015-10-23 Fri 14:55 JST hig"

今日の目標

- 2 変量データから, 回帰直線を手で求められる.
- 2 変量データがファイルとして与えられたとき, Excel を使って, 平均値, 分散, …, 回帰直線を求められる. この科目や他の科目のレポート作成に利用できる.

http://hig3.net

L4-Q7 の解答の訂正
Quiz 解答:共分散と相関係数

1. x の平均値は $\overline{x} = 18\text{cm}$, y の平均値は $\overline{y} = 4\text{g}$.
   共分散は
$$C_{xy} = \frac{1}{6}[(13-18)(2-4)+\cdots] = \frac{13}{3}\text{cm}\cdot\text{g}.$$
2. x の分散は $s_x^2 = 9\ \text{cm}^2$, y の分散は $s_y^2 = 4\ \text{g}^2$. よって,
$$r = \frac{\frac{13}{3}\text{cm}\cdot\text{g}}{\sqrt{9\text{cm}^2}\sqrt{4\text{g}^2}} = \frac{13}{18}.$$

L05-Q1
Quiz 解答:離散的な確率変数の母平均・母分散・母標準偏差

1. 期待値 $\mathrm{E}[\mathrm{e}^X] = \frac{4}{12}\cdot\mathrm{e}^{-1} + \frac{5}{12}\cdot\mathrm{e}^0 + \frac{3}{12}\cdot\mathrm{e}^2$.

2. 母平均値 $\mathrm{E}[X] = \frac{4}{12} \cdot (-1) + \frac{5}{12} \cdot 0 + \frac{3}{12} \cdot 2 = \frac{1}{6}$.
3. 母分散
   $\mathrm{V}[X] = \mathrm{E}[(X-m)^2] = \frac{4}{12} \cdot (-1-\frac{1}{6})^2 + \frac{5}{12} \cdot (0-\frac{1}{6})^2 + \frac{3}{12}(2-\frac{1}{6})^2 = \frac{47}{36}$.
4. 母標準偏差 $\sqrt{\mathrm{V}[X]} = \sqrt{\frac{47}{36}}$.
5. 確率 $\mathrm{E}[\mathbf{1}_{[a]}(X)] = \frac{4}{12} \cdot 1 + \frac{5}{12} \cdot 1 + \frac{3}{12} \cdot 0 = \frac{9}{12} = \frac{3}{4}$.

L05-Q2

Quiz 解答:離散的な確率変数の母平均値・母分散・母標準偏差・確率

1. $\mathrm{E}[X] = -\frac{1}{5}$
2. $\mathrm{E}[2X+1] = \frac{3}{5}$
3. $\mathrm{E}[X^2] = \frac{11}{5}$

L05-Q4

Quiz 解答:離散的な確率変数の母平均値・母分散・母標準偏差・確率

1. $\mathrm{E}[\mathbf{1}_{[X\leq 50]}(X)] = \sum_{x=0}^{100} \frac{x}{5050}\mathbf{1}_{[X\leq 50]}(x) = \sum_{x=1}^{50} \frac{x}{5050} = \frac{\frac{1}{2}\cdot 100\cdot(100+1)}{5050} = \frac{51}{202}$
2. $\mathrm{E}[X] = \sum_{x=0}^{100} \frac{x}{5050}\cdot x = \frac{\frac{1}{6}\cdot 100\cdot(100+1)(2\cdot 100+1)}{5050} = 67.$
3. $\mathrm{V}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \sum_{x=0}^{100} \frac{x}{5050}\cdot x^2 - 67^2 = (\frac{1}{2}\cdot 100\cdot(100+1))^2 - 67^2 = 25498011.$

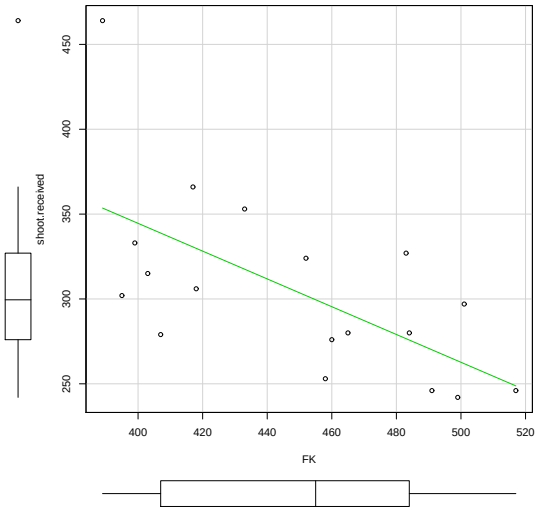
回帰分析

回帰 (regression), 直線回帰=単回帰分析=1 変数回帰分析

2 変量データ (x, y) が
相関係数 $r = \pm 1$ に近い $\Leftrightarrow$ 散布図上のデータ点 (x, y) がほぼ直線に載っている

その直線 ( ) の式 $y = ax + b$ を知りたい!

つまり a, 定数項 b を決めたい.



y: 目的変数 (従属変数)
x: 説明変数 (独立変数)
何でそんなことしたいの?

- 法則を見つけたい
- x から y を予測したい

回帰直線の決め方

1. 定規をあてて ‘真ん中’ を通るように
2. 最小 2 乗法で.

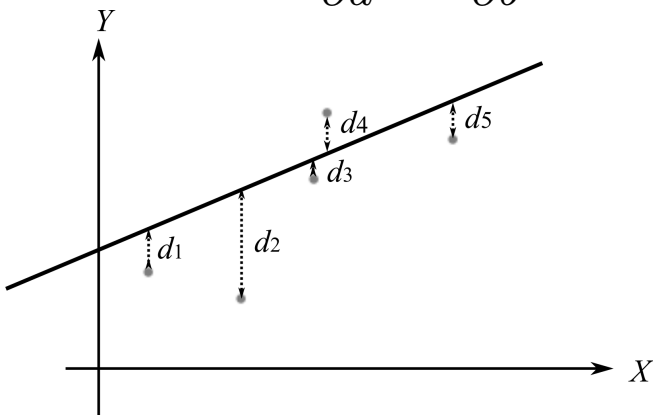
最小 2 乗法

直線からのずれの 2 乗 d^2 の合計

$$f(a,b) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$$

の最小条件 $\frac{\partial f}{\partial a} = \frac{\partial f}{\partial b} = 0$ で a, b を決める.

微積分 I



物理実験

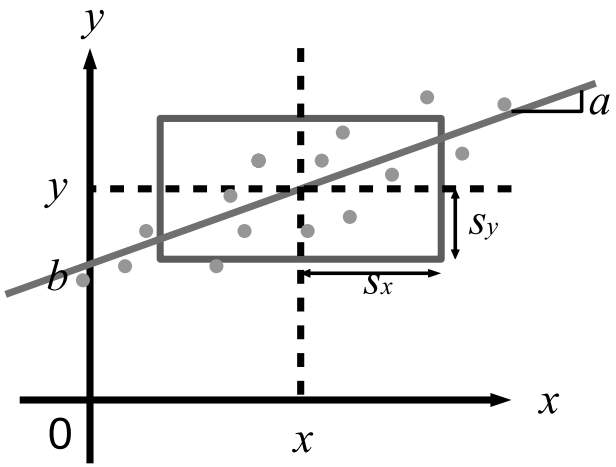
直線回帰の公式

回帰直線

x_i, y_i $(i = 1, \ldots, n)$ の平均値を $\overline{x}, \overline{y}$, 標準偏差を s_x, s_y, 相関係数を r とする. このとき回帰直線は,

$$y = \frac{r \times s_y}{s_x} \times (x - \overline{x}) + \overline{y} = ax + b.$$

傾きは $a = \frac{r \times s_y}{s_x}$, 切片は $b =$ (点 $(\overline{x}, \overline{y})$ を通るような値)



a: 回帰係数 (x を 1 だけ変えたときの y の変化量)
r^2: 決定係数 (あてはまりのよさ)

回帰直線の傾きのおぼえ方 I

広がり方

散布図上のデータ点の分布は, 横 $2s_x$, 縦 $2s_y \to$ 傾き $\frac{s_y}{s_x}$ くらい?
しか〜し, 傾きには正負があるし, 相関がなかったら傾きを 0 にしたいので, 相関係数 r をかけ算しておく.

単位チェック

(x, y) の単位が (m,kg) だとする.
r は無次元. 単位無し.
左辺 y (kg).
右辺 $r \times \frac{s_y(\mathrm{kg})}{s_x(\mathrm{m})} \times x(\mathrm{m}) + b(\mathrm{kg})$
で, s_x/s_y かけると単位があう.

L06-Q1 来週の非参照 Quiz 1 はこんな感じ

Quiz(共分散と相関係数)

下のデータを考える.

x	y
1	3
2	7
4	10
5	9
8	16

1. 共分散を求めよう.
2. 相関係数を求めよう.
3. 回帰直線の式を求めよう.

ただし, 平均値 $\overline{x} = 4, \overline{y} = 9$, 分散 $s_x^2 = 6, s_y^2 = 18$ であることを使っていい.

ここまで来たよ

3 離散型確率変数

4 回帰分析

5 Excel で統計
- Excel で統計

準備

統計ソフトウェア実習室にインストールされているのは

R 無料. オープンソース. 解説書が多い.
SPSS 伝統ある高級品.
Excel 機能は限られ怪しいところもあるが, 普及率高い. 龍大では Office365 で無料.

今日は Excel を使ってみます.
スタートボタン >Excel 2013
統計分析のための準備
ファイル > オプション > アドイン > Excel のアドイン > 設定 > 分析ツール にチェックを入れて OK する.

Excel による主な分析

どこかの段階でデータ範囲を指定, または関数の引数にデータ範囲を指定.

	メニューベース	関数ベース
平均値, 分散, 標準偏差	データ > 分析 > データ分析 > 基本統計量 > 統計情報	平均値 average, 分散 var.p, 標準偏差 stdev.p, 最頻値 mode
四分位数	データ > 分析 > データ分析 > 順位と百分位数	中央値 median, 四分位数 quartile
度数分布表, ヒストグラム	データ > 分析 > データ分析 > ヒストグラム > 入力範囲とデータ区間	frequency + グラフ
散布図	挿入 > グラフ > 散布図	
共分散, 相関係数	データ > 分析 > データ分析 > 共分散, 相関	covar=covariance.p, correl
回帰分析	データ > 分析 > データ分析 > 回帰分析	linest
クロス集計表	挿入 > テーブル > ピボットテーブル	

メニューベースの分析をするときの注意

- Excel は, 1 種類のデータは列方向 (縦方向) にならんでいるとデフォルトでは想定する. 分析の種類によっては, 列方向, 行方向のどちらに並んでいるかを指定できるものもある.
- 2 変量 (n 変量) の統計量である, 共分散 C_{xy} や相関係数 r_{xy} の出力は

$$\begin{matrix} C_{xx} & C_{yx} \\ C_{xy} & C_{yy} \end{matrix}, \qquad \begin{matrix} r_{xx} & r_{yx} \\ r_{xy} & r_{yy} \end{matrix}$$

  のように行列状にになっている. C_{yy} や r_{yy} は, $y = x$ であるときの C_{xy}, r. よく考えると, $C_{yy} = s_y^2, r_{yy} = 1$ であることに気づく. $n \geq 3$ のときは $n \times n$ 行列になる.
- 回帰分析の出力では
  - 重相関 R $=$ 相関係数 r
  - 従決定 R2 $=$ 決定係数 r^2
  - 切片の係数 $=$ 回帰直線の切片 b
  - X 値 1 の係数 $=$ 回帰係数 a
  - $n \geq 3$ の重回帰 $(x_1, x_2, \ldots, x_{n-1}, y)$ というものがあり, そのときは X 値 2,$\cdots$ などとなっていく.
- ここで紹介したメニューベースの分析では, 実はここまで学んだ「データの分散」すなわち var.p でなく, 今後学ぶ「不偏標本分散」 var.s を計算している$\cdots$ 両者の区別は考え方としては超重要だが, Excel で扱いたくなるようなデータ数が多いときは近い値になる.

次回の非参照 Quiz

- 2 変量データから回帰直線を求めよう
- 1 変量データから標準得点を求めよう (いまごろ L04 の内容)

連絡

- 2015-10-30 金 は全学休講
- Quiz L06 予習問題は 2015-11-05 木昼まで Math ラウンジで受けつけてます. ふだんの予習問題より大きな配点です.
- オフィスアワー月 4 木 6(1-502)

manaba 出席カード提出
https://attend.ryukoku.ac.jp

プチテスト計画!

- 2015-11-13 金 2, 90 分, 30 ピーナッツ, 参照相談なし. 紙のテスト.
- まず授業でやらなかったページに×つけましょう.
- 過去問公開してるけどあまり参考にはならないかも. 下の出題計画, 非参照 Quiz, 予習問題をやり直すことをお奨めします.
- 出題計画 (2015-11-06 金ごろ修正, 確定します). Excel 関係のものはありません.
  - データから平均値, 分散, 標準偏差を求める
  - データから (外れ値を考慮した大学レベルの) 箱ひげ図を描く
  - データから標準得点, 偏差値を求める (← 注意. 非参照 Quiz でカバーされてない)
  - データから共分散, 相関係数を求める
  - データから回帰係数, 回帰直線を求める
  - 離散型確率変数について, 確率, 母期待値, 母平均値, 母分散, 母標準偏差を求める
  - 連続型確率変数について, 確率, 母期待値, 母平均値, 母分散, 母標準偏差を求める (2015-11-06 にやります)
  - 選択肢的な問