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 L04(2018-10-17 Wed)

最終更新: Time-stamp: "2018-10-17 Wed 06:50 JST hig"

今日の目標

- Excel で代表値・分散が求められる
- 2 変量データから, 手で回帰直線が求められる
- 2 変量データから, Excel で散布図が描け共分散と相関係数と回帰直線が求められる

L03-Q1

Quiz 解答:平均値・分散・標準偏差の換算

1.6m, 0.0025m^2, 0.05m.

L03-Q2

Quiz 解答:分散の意味

1

L03-Q3 Quiz 解答:標準得点と偏差値

平均値 $\bar{x} = 90$, 分散 $S_x^2 = 4$, 標準偏差 $S_x = 2$.

標準得点 $z = (87 - 90)/2 = -1.5$.

偏差値 $w = (-1.5) \times 10 + 50 = 35$.

L03-Q4

Quiz 解答:偏差値の性質

1. 誤り
2. もっともらしいが正しいとは断定できない
3. 誤り
4. もっともらしいが正しいとは断定できない

L03-Q5 Quiz 解答:共分散

$\overline{x} = 4,\ s_x^2 = 4,\ s_x = 2.$

$\overline{y} = 13,\ s_x^2 = 122/5 = 24.4,\ s_y = \sqrt{122/5} = 4.94.$

共分散 $s_{xy} = \frac{1}{5}[(1-4)(5-13) + (3-4)(15-13) + (4-4)(14-13) + (5-4)(11-13) + (7-4)(20-13)] = 41/5 = 8.2.$

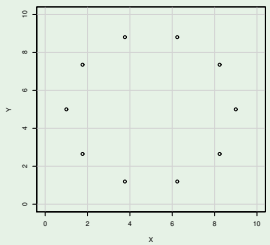
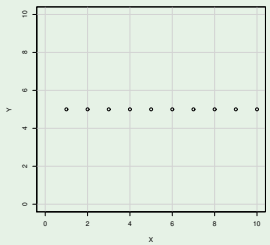
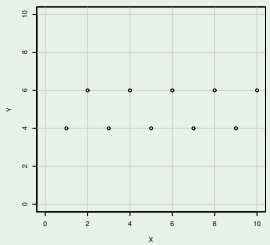
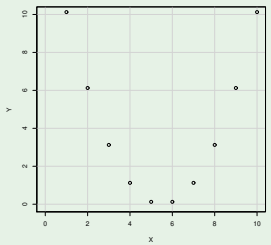
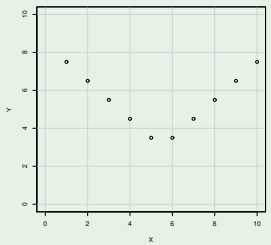
相関係数 $r = \dfrac{41/5}{2 \cdot \sqrt{122/5}} = 0.83.$

だまされたくない相関の性質

L03-Q6

Quiz(相関係数)

次のうち, 相関係数 r がもっとも大きいものはどれ?

Anscombe(1973)

ここまで来たよ

3. 略解:データの変換 (標準得点, 偏差値)・2 変量データと相関

4. 回帰分析
   - Excel で統計
   - 回帰分析

Excel 使用の準備

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 表計算. 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.

起動 スタートボタン >Excel 2016
準備 (データ分析の有効化)
ファイル > オプション > アドイン > Excel のアドイン > 設定 > データ分析 に
チェックを入れて OK する.
Excel によるグラフ描画 挿入 > グラフ > (グラフの種類)
題名や軸の変数名の追加
挿入 > グラフ > グラフのデザイン > グラフ要素を追加
使用するデータの調整
挿入 > グラフ > グラフのデザイン > グラフデータの選択

表計算ソフトウェア (Excel) による分析 高校 数学 I

メニューからデータ範囲を指定, 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	

メニューベースのデータ分析 > 基本統計量の分散は, さらに $\frac{n-1}{n}$ 倍しないと, 「データの分散」var.p にならない.

メニューベースでデータ分析をするときの注意

- 列=縦, または 行=横 (線形代数と同じ) にデータを N 個並べる. 多変量の時は, 直交する方向に p 個を並べる.
- 「ラベル」は, 1 行目 (または 1 列目) に書かれている変数名 (身長) (データ (60 点) でなく). ラベルを範囲に含めるか含めないか, チェックボックスがあることが多い.
- $p = 2$ 変量の統計量である, 共分散 S_{xy} や相関係数 r_{xy} の出力は $p \times p$ の正方行列状.

$$\begin{matrix} S_{xx} = S_x^2 & S_{yx} \\ S_{xy} & S_{yy} = S_y^2 \end{matrix} \text{ , } \qquad \begin{matrix} r_{xx} = 1 & r_{yx} \\ r_{xy} & r_{yy} = 1 \end{matrix}$$

ここまで来たよ

3. 略解:データの変換 (標準得点, 偏差値)・2 変量データと相関

4. 回帰分析
   - Excel で統計
   - 回帰分析

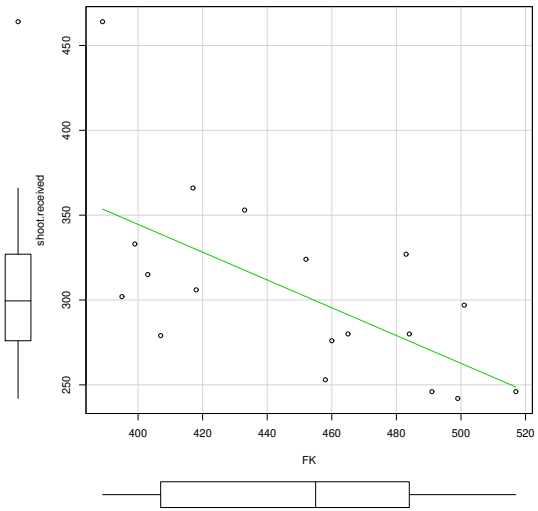
回帰分析 前園確率統計 §7.2

回帰 (regression), 直線回帰=単回帰分析=1 変数回帰分析 物理実験

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.

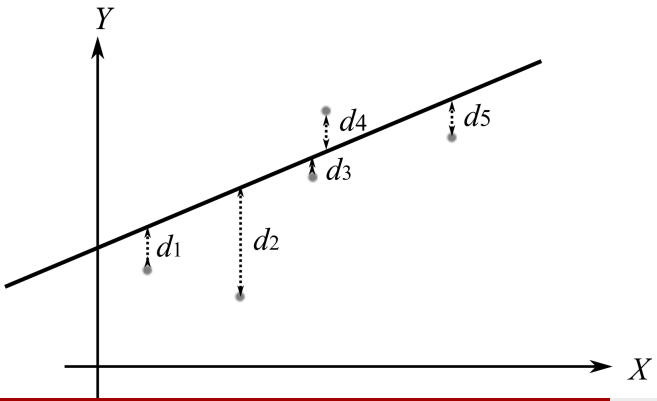
最小 2 乗法

直線からのずれの 2 乗 d^2 の合計

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

の最小条件 $\frac{\partial L}{\partial a} = \frac{\partial L}{\partial b} = 0$ で a, b を決める. $a = \beta_0, b = \beta_1$ in 前園確率統計 (7.3)

微積分 I



物理実験

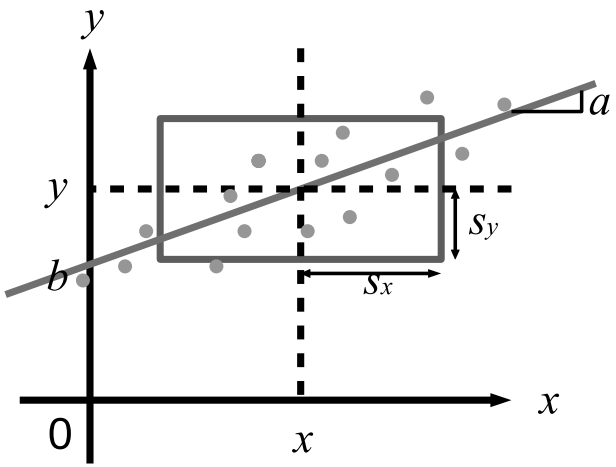
直線回帰の公式

回帰直線 前園確率統計 (7.4),(7.5)

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 $a = \frac{r \times S_y}{S_x} = \frac{C_{xy}}{S_x^2}$, 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)
誤差 $L(a, b) = N(1 - r^2)S_y^2$.

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 $2S_y$ $\rightarrow$ 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 $r \times \frac{S_y(\mathsf{kg})}{S_x(\mathsf{m})} \times x(\mathsf{m}) + b(\mathsf{kg})$
で, S_x/S_y かけると単位があう.

L04-Q1

Quiz(回帰係数と回帰直線)

ある 2 変量データ (x, y) について次のことがわかっている.

x の平均値 $\overline{x}$	9
y の平均値 $\overline{y}$	-4
x の分散 s_x^2	49
y の分散 s_y^2	36
x, y の共分散 s_{xy}	-25
(x, y) のデータの個数 n	16

このとき, 回帰直線の式を, x, y の式で書こう. 整理しなくてよい.

メニューベースの回帰分析

データ > データ分析 > 回帰分析

入力

入力 Y 範囲 = 目的変数

入力 X 範囲 = 説明変数

出力

- 重相関 R = 相関係数 r
- 重決定 R2 = 決定係数 r^2
- 切片 = 回帰直線の切片 b
- X 値 1(またはラベルで指定した変数名) = 回帰係数 a

連絡

- 次回は臨時教室変更で 4-209 講義室
- 樋口オフィスアワー火昼 (1-539) 金 14:40-15:40(1-502), Math ラウンジ月-木昼 (1-614)
- Trial 予告
- Learn Math Moodle の予習復習問題で来週の trial に備えてね.
- 来週から教科書をがんがん使います.

前園確率統計 §2.1 | 前園確率統計 §3.1 | 前園確率統計 §3.2 読んできてね.