回帰分析

樋口さぶろお http://hig3.net

龍谷大学理工学部数理情報学科

生活の中の統計技術　L04(2018-10-15 Mon)

最終更新: Time-stamp: "2018-11-05 Mon 14:31 JST hig"

今日の目標

- 2 変数の量的データから, Excel で散布図が描ける
- 2 変数の量的データから, Excel で共分散と相関係数と回帰直線が求められる

L03-Q1 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 $r = \frac{41/5}{2 \cdot \sqrt{122/5}} = 0.83.$

ここまで来たよ

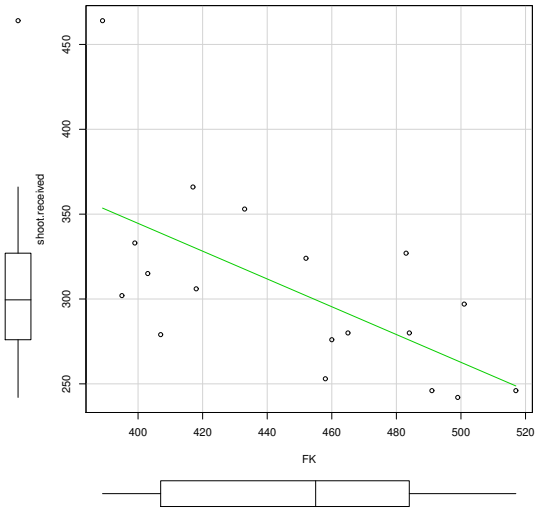
回帰分析

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
- 中間テストの点数 x から期末テストの点数 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.

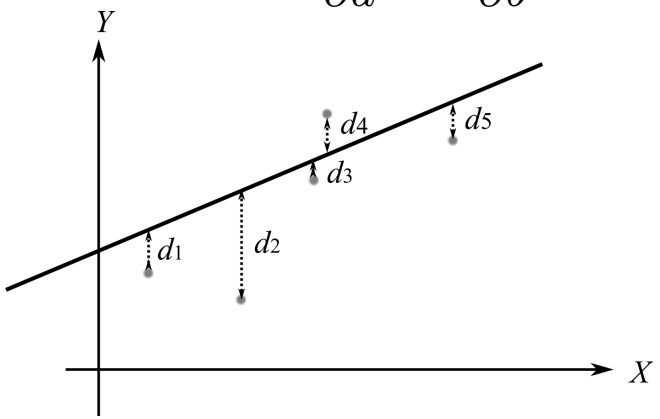
最小 2 乗法

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.

微積分 I



物理実験

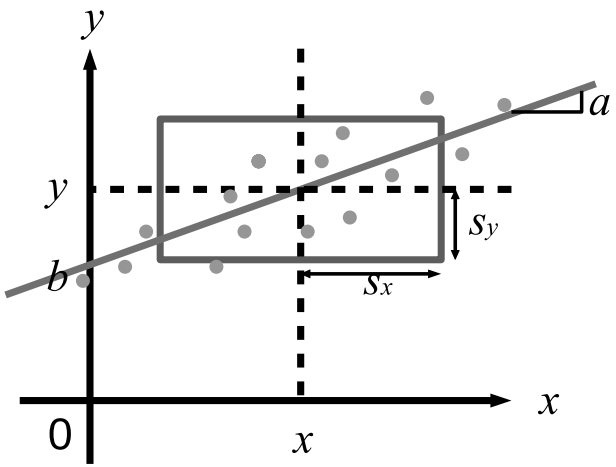
直線回帰の公式

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

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, x を説明変数, y を目的変数とする回帰直線の式を, x, y の式で書こう. 整理しなくてよい.

重回帰

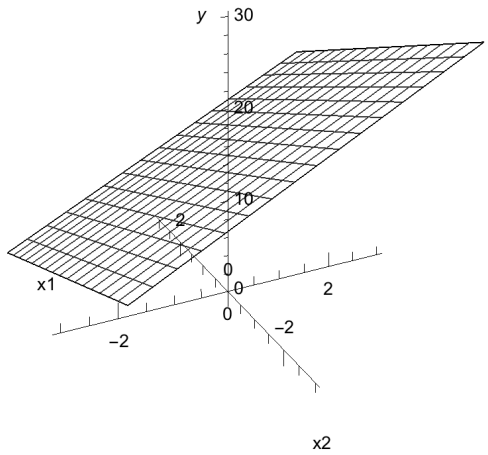
説明変数の個数が $p \geq 2$ になっただけ.
目的変数 y (期末試験の点数)
説明変数 $x_1, \cdots, x_p$ (小テスト 1 の点数, ..., 小テスト p の点数)

$p = 1$　　$y = a_1 x_1 + b$

$\downarrow$

$p = 2$　　$y = a_1 x_1 + a_2 x_2 + b.$　　3 次元空間の中の平面.

$p \geq 2$　　$y = a_1 x_1 + a_2 x_2 + \cdots a_p x_p + b.$



ここまで来たよ

2 略解:複数のテストの点数の相関

3 回帰分析
- 回帰分析
- 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. 社会学部向け.
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 2016
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.

表計算ソフトウェア (Excel) による主な分析 高校 数学 I

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	

行=横のセル

の並び, 列=縦のセルの並び

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

- 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 (p 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

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

のように行列状にになっている. S_{yy} や r_{yy} は, $y = x$ であるときの S_{xy}, r. よく考えると, $S_{yy} = S_y^2, r_{yy} = 1$ であることに気づく. $p \geq 3$ のときは $p \times p$ 行列になる (正方形状に並ぶ).
- 「ラベル」は, 1 行目 (または 1 列目) に書かれているのがデータ (60 点) でなく, 変数名 (小テスト 1) であることを表す.

メニューベースの回帰分析，重回帰分析

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 (複数個あれば重回帰になる)

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, a_1.
- X 値 2,$\cdots$(またはラベルで指定した変数名) = 重回帰の係数 a_2 などとなっていく.