回帰分析

樋口さぶろお　https://hig3.net

龍谷大学理工学部数理情報学科

確率統計☆演習　L05(2020-10-26 Mon)

最終更新: Time-stamp: "2020-10-24 Sat 14:15 JST hig"

今日の目標

- 単回帰の回帰係数, 切片を手で求められる

  岩薩林 確率・統計 §9 を先取り
- Excel で回帰分析ができる

L04-Q1
Quiz 解答:分散共分散をモーメントから

1. $S_x^2 = \overline{x^2} - \overline{x}^2 = 34 - 5^2 = 9.$
2. $S_{xy} = \overline{xy} - \bar{x}\bar{y} = -11.$

L04-Q2 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 $r = \frac{41/5}{2 \cdot \sqrt{122/5}} = 0.83.$

L04-Q3
Quiz 解答:相関係数の性質

1. かわらない.
2. かわらない.

3. -1 倍になる
4. かわらない.

L04-Q4 Quiz 解答:相関係数

相関係数はすべて 0.

ここまで来たよ

4 2 次元データと相関

5 回帰分析
- 統計量の単位・次元
- 回帰分析
- Excel で 2 変量統計

単位 (物理量の次元)

物理量は次元を持つ (m, kg, s). 物理量のデータ, 統計量も次元を持つ.

- 両辺の単位は等しい
- 加減は同じ単位の量の間でしかできない
- 積/商の量は単位もそうなる

質量 $x = 10, 20$, kg, 速度 $y = 1, 2$, m/s の例で.

- $\overline{x}$ kg$= \frac{1}{n}[10\text{kg} + \cdots]$
- S_x^2 kg$^2 = \frac{1}{n}[(10\text{kg} - \overline{x}\text{kg})^2 + \cdots]$
- S_x kg $= \sqrt{S_x^2 \text{kg}^2}$
- $\overline{y}$ m/s, S_y^2 (m/s)^2, S_y^2 m/s
- S_{xy} kg m/s $= \frac{1}{n}[(10\text{kg} - \overline{x}\text{kg})(1\text{m/s} - \overline{y}\text{m/s}) + \cdots]$
- r_{xy} (単位なし (無次元)) $= \frac{S_{xy}\text{kg m/s}}{S_x\text{kg}S_y\text{m/s}}$.

ここまで来たよ

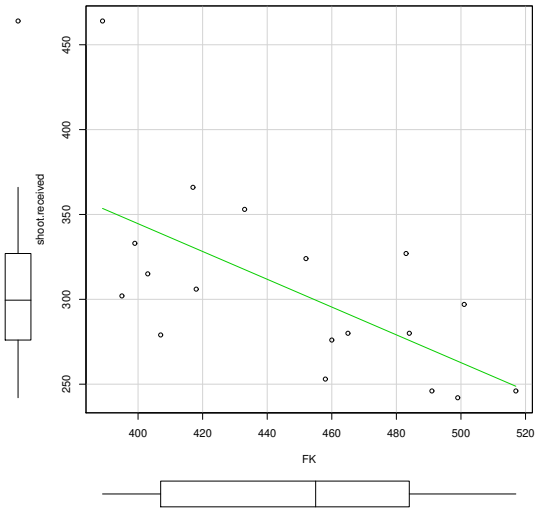
4 2 次元データと相関

5 回帰分析
- 統計量の単位・次元
- 回帰分析
- Excel で 2 変量統計

回帰分析 岩薩林 確率・統計 §9

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 $r_{xy} = \pm 1$ に近い $\Leftrightarrow$ 散布図で (x, y) がほぼ直線に載る
その直線 (回帰直線) の式 $y = \beta x + \alpha$ を知りたい! 岩薩林 確率・統計 (9.2)
つまり 回帰係数 β, 定数項 (切片) α 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　　小中学校
2. 最小 2 乗法で　　数値計算法, 物理実験

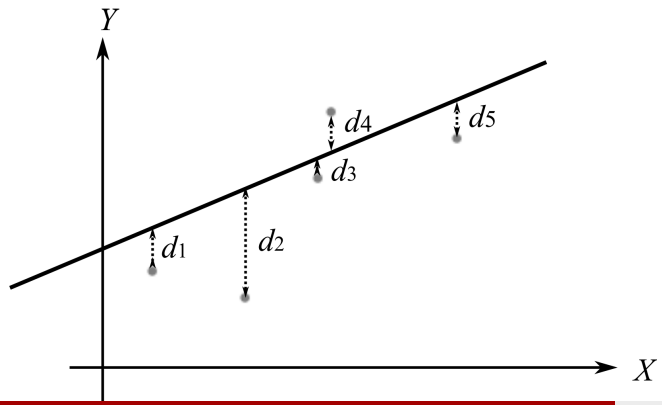
最小 2 乗法

直線からのずれの 2 乗 d^2 の合計

$$L(\alpha, \beta) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - (\beta x_i + \alpha))^2$$

の最小条件 $\frac{\partial L}{\partial \alpha} = \frac{\partial L}{\partial \beta} = 0$ で α, β を決める.

微積分 I



物理実験

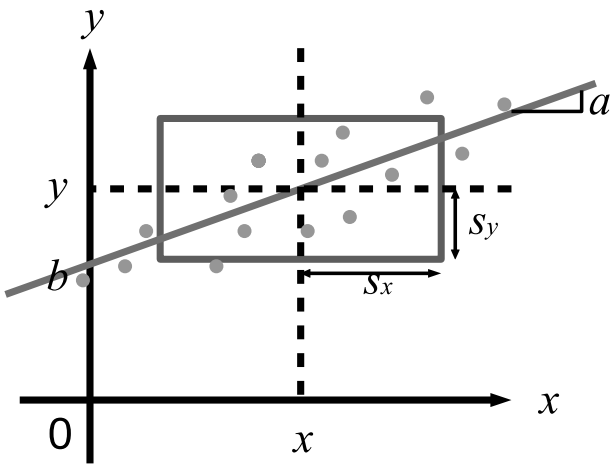
直線回帰の公式

回帰直線 岩薩林 確率・統計 (9.10)

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_{xy} とする. このとき回帰直線は,

$$y = \frac{r_{xy} \times S_y}{S_x} \times (x - \overline{x}) + \overline{y} = \beta x + \alpha.$$

傾きは $\beta = \frac{r \times S_y}{S_x} = \frac{S_{xy}}{S_x^2}$, 切片は $\alpha =$ (点 $(\overline{x}, \overline{y})$ を通るような値)



β: 回帰係数 (x を 1 だけ変えたときの y の変化量)
$0 \leq r_{xy}^2 \leq 1$: 決定係数 (あてはまりのよさ)
誤差 $L(\alpha, \beta) = N(1 - r_{xy}^2) S_y^2$.

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_{xy} 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_{xy} は無次元. 単位無し.
左辺 y (kg).
右辺 $r_{xy} \times \frac{S_y(\mathsf{kg})}{S_x(\mathsf{m})} \times (x(\mathsf{m}) - \bullet) + \alpha(\mathsf{kg})$
で, S_y/S_x かけると単位があう.

岩薩林 確率・統計 例題 9.2, 9.3, §9 問題 3,4,5, §9 練習問題 1

L05-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.

ここまで来たよ

4 2 次元データと相関

5 回帰分析
- 統計量の単位・次元
- 回帰分析
- Excel で 2 変量統計

Excel 使用の準備 (復習)

起動 スタートボタン > Excel

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

課題のデータで散布図を描こう

データの順序に意味がある場合以外は, データ点の間を直線や曲線で結ばない.

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 (分散は要 $(n-1)/n$ 倍)	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`, 百分位数 `percentile.inc`
順位, 分位	データ > 分析 > データ分析 > 順位と百分位数	順位 `rank`, 百分位 `percentrank.inc`
ヒストグラム, 箱ひげ図	挿入 > グラフ > ヒストグラム, 箱ひげ図	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 = 相関係数の絶対値 $|r_{xy}|$ 符号は表示されない
- 重決定 R2 = 決定係数 r_{xy}^2
- 切片 = 回帰直線の切片 α
- X 値 1(またはラベルで指定した変数名) = (X 値 1 の) 回帰係数 β

メニューベースでデータ分析をするときの注意

- 列=縦, データを n 個並べる.
  - 縦横を変えるときは, 形式を選択してペースト > 行列を入れ替える
- 「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 (160cm) 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.
- $p = 2$ 次元データの, 共分散 S_{xy} や相関係数 r_{xy} の出力は $p \times p$ の対称行列.

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

- なぜか, データ分析 > 共分散はそのままで正しい. $\frac{n-1}{n}$ する必要なし.
- なぜか, データ分析 > 相関はもちろんそのままで正しい.

重回帰分析

目的変数 価格 y 円
説明変数 質量 x_1 kg, 糖度 x_2
$y = \beta_1 x_1 + \beta_2 x_2 + \alpha$　　直線でなく平面の方程式
回帰係数は説明変数の個数だけ, 切片はひとつ

Excel の回帰分析の出力

- 切片 = 回帰直線の切片 α
- X 値 1(またはラベルで指定した変数名) = (X 値 1 の) 回帰係数 β_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 $\beta_2, \cdots$

Excel の共分散, 相関の出力

$p = 3$ 次元データの, 共分散 $S_{\bullet\bullet}$, 相関 $r_{\bullet\bullet}$ の出力は $p \times p$ の対称行列

$$\begin{bmatrix} S_{xx} = S_x^2 & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} = S_y^2 & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{bmatrix}, \qquad \begin{bmatrix} r_{xx} = 1 & r_{xy} & r_{xz} \\ r_{yx} & r_{yy} = 1 & r_{yz} \\ r_{zx} & r_{zy} & r_{zz} = 1 \end{bmatrix}.$$