回帰分析

樋口さぶろお

龍谷大学理工学部数理情報学科

使える統計!　L06(2013-11-06 Wed)

今日の目標

1. 2 変量データから, 手で, 共分散, 相関係数が計算できる
2. 2 変量データから, 手で, 回帰直線を求められる
3. Excel で散布図が描ける
4. Excel で回帰分析できる
5. Excel でクロス集計表を作れる

http://hig3.net

L05-S4 Quiz 解答:共分散　　X の平均値 4, 分散 4, 標準偏差 $\sqrt{4}$.
Y の平均値 13, 分散 $122/5 = 24.4$, 標準偏差 $\sqrt{122/5} = 4.94$.
共分散 $C = \frac{1}{5}[(1-4)(5-13) + (3-4)(11-13) + (4-4)(14-13) + (5-4)(15-13) + (7-4)(20-13)] = 9.8$.
相関係数 $r = \frac{9.8}{\sqrt{4}\cdot\sqrt{122/5}} = 0.992$.
すみません問題文中で Y の分散の値が間違えてました.

ここまで来たよ

共分散

共分散 (covariance)

$$
\begin{aligned}
&X, Y \text{ の共分散 } C \\
&= \frac{1}{\text{データの個数 } N} \\
&\times [(X \text{ のデータ } 1 - X \text{ の平均値}) \times (Y \text{ のデータ } 1 - Y \text{ の平均値}) \\
&+ (X \text{ のデータ } 2 - X \text{ の平均値}) \times (Y \text{ のデータ } 2 - Y \text{ の平均値}) \\
&+ \cdots (\text{データすべて}) \cdots \\
&+ (X \text{ のデータ } N - X \text{ の平均値}) \times (Y \text{ のデータ } N - Y \text{ の平均値})]
\end{aligned}
$$

相関係数

(ピアソンの積率) 相関係数 (correlation coefficient)

$$X, Y \text{ の相関係数 } r = \frac{X, Y \text{ の共分散 } C}{X \text{ の標準偏差 } s_X \times Y \text{ の標準偏差 } s_Y}$$

- 相関係数は, 相関の正負, 強さを表す.
- $-1 \leq r \leq +1$.
- $r = +1 \Leftrightarrow$ 正の強い相関 右上がりの一直線上にのる
- $r = -1 \Leftrightarrow$ 負の強い相関 右下がりの一直線上にのる

にせの因果関係にだまされるな

被シュートと失点は正の相関

- 原因:被シュートが多い, 結果: 失点が多い?
- 原因:失点が多い, 結果: 被シュートが多い?
- 原因:???, 結果: 失点が多い, かつ, 被シュートが多い?

フリーキックと被シュートは負の相関

- 原因:フリーキックが多い, 結果:被シュートが少ない?
- 原因:被シュートが少ない, 結果:フリーキックが多い?
- 原因:???, 結果:被シュートが少ない, かつ, フリーキックが多い?

- 相関が強くても
- 因果関係があっても

回帰分析

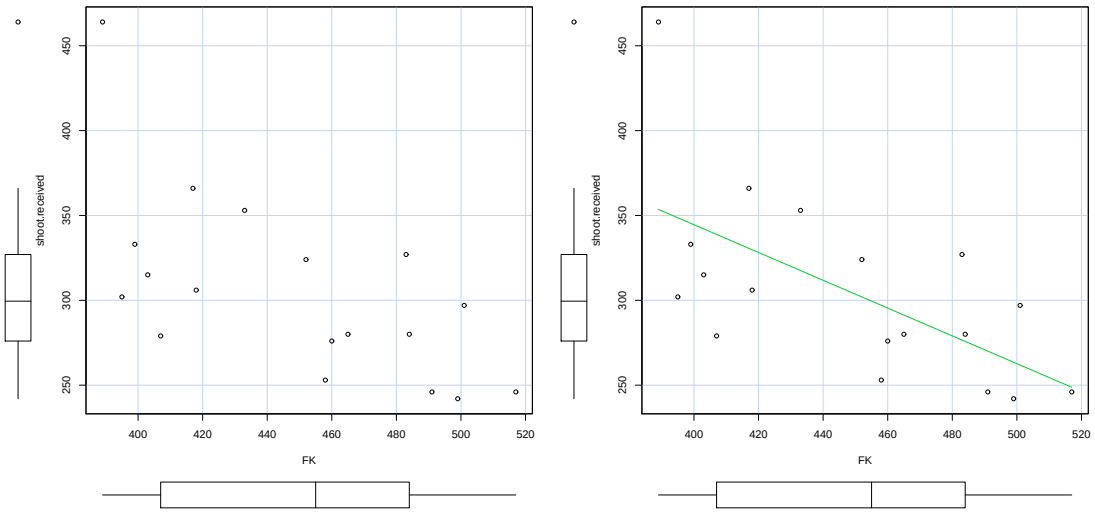
回帰 (regression), 単回帰分析=1 変数回帰分析

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, b を決めたい.



[b] 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

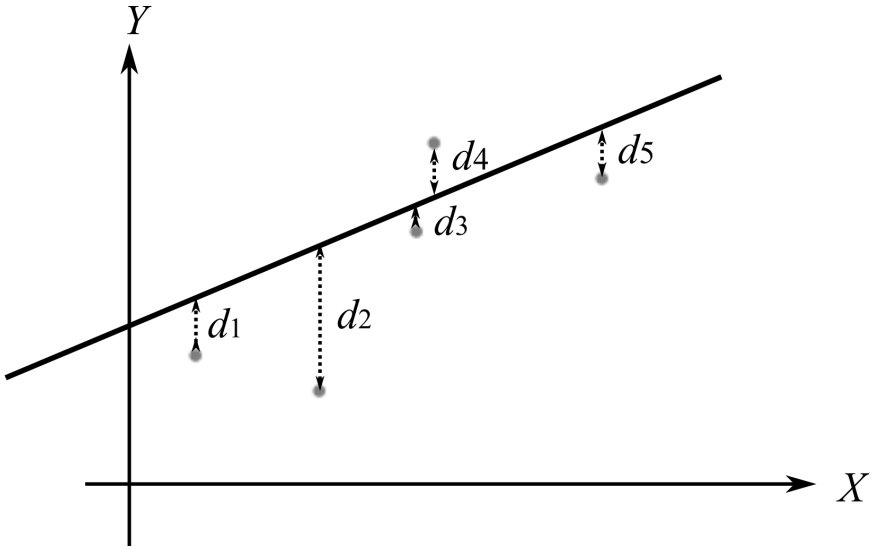
回帰直線の決め方

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 $f = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2$ がなるべく小さくなるように a, b を決める.

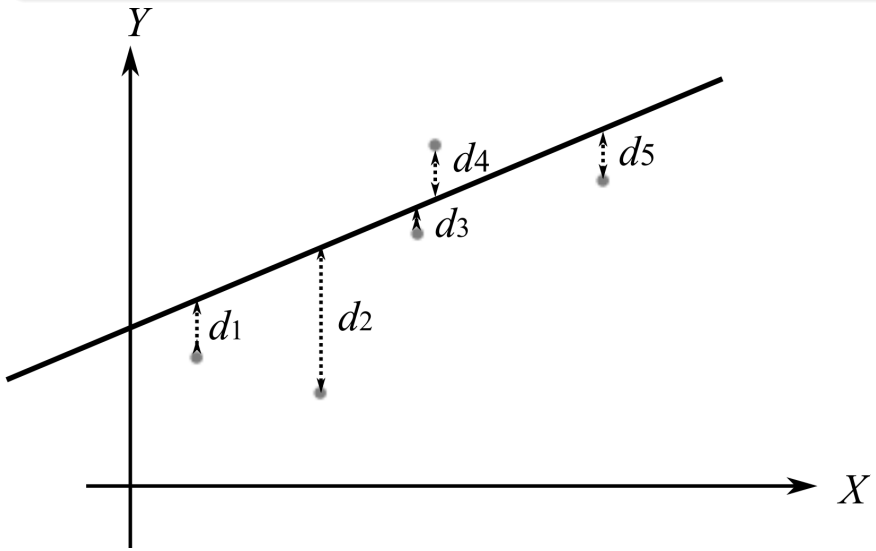
大学で微積分をやった人への注: 2 変数関数 $f(a, b)$ の a, b についての偏微分係数 $= 0$ から最小値を求めます.



1 次関数と直線の式の復習

直線の式

傾き a, 点 (c, d) を通る直線 $Y = a(X - c) + d$
傾き a, 切片 b (点 $(0, b)$ を通る) の直線 $Y = aX + b$



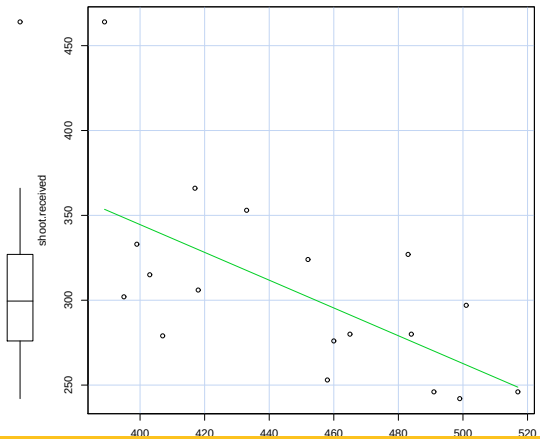
最小 2 乗法の公式

回帰直線

X, Y の平均値を m_X, m_Y, 標準偏差を s_X, s_Y, 相関係数を r とする.
回帰直線は,

- 傾き $\frac{r \times s_Y}{s_X}$ で,
- 点 (m_X, m_Y) を通る

$$Y = \frac{r \times s_Y}{s_X} \times (X - m_X) + m_Y$$



回帰直線の傾きのおぼえ方

(X, Y) が (m,kg) だとする.
傾きは r みたいなもの.

r は無次元の数 ( ) だが

傾き a の単位は .

調整するためには分子に $\frac{s_Y}{s_X}$ (単位) をかけておく.

なんで s_X, s_Y とかつくの? なかったら X, Y いれかえても同じ傾きになっちゃうじゃん.

Q1

Quiz(共分散と相関係数)

下のデータを考える.

X	Y
2	4
2	6
4	11
5	9
7	15

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 $m_X = 4, m_Y = 9$, 標準偏差 $s_X = 1.90, s_Y = 3.85$ (四捨五入してます) であることを使っていい.

Excel で

e ラーニングシステムからデータをダウンロード.
前回の実習室の日に参加した人は済んでる準備
Office ボタン > Excel のオプション > アドイン > 管理:Excel アドイン 設定… で分析ツールにチェック.
散布図挿入 > グラフ > 散布図 (点のみのものが趣味よい?)

クロス集計表表全体を選択した状態で,
挿入 > ピボットテーブル > ピボットテーブル
一定幅の階級を作るには, 行ラベル, 列ラベルにカーソルをおいた状態で,
ピボットテーブルツール > オプション > グループフィールド

共分散・相関係数データ > データ分析 > 共分散
データ > データ分析 > 相関係数

回帰分析データ > データ分析 > 回帰分析

連絡

- きょうは紙 1 枚, ファイル 1 個提出.
- 2013-11-13 水 は休講. だけど, すぐ e ラーニングで補講. 2013-11-11 月以降またはメールで連絡してから, 2013-11-20 水 までに受講してね. e ラーニングのコースの 2013-11-13 水 のところに指示を書きます.
- いつか 台風の分の補講
- 加減乗除と平方根 (ルート) の使える電卓持ってきてね. 関数電卓でなくてもいいです. 携帯電話の機能・アプリでもかまいません.

プチテスト計画

2013-11-20 水 プチテスト
(関数 or 通常) 電卓持込可. テストのときは携帯不可.
Excel の操作方法の問題はありません.

日時 2013-11-20 水 3 14:05-15:05(60 分).
場所 いつもと同じ
形式 ペーパーテスト. 計算問題中心. (関数 or 普通) 電卓使用可 (ただし過程を書いてもらうので電卓の統計機能だけでは答えられないでしょう). 携帯不可.
参照 公式外部記憶ペーパーのみ持込可 (今日も用紙配布してます). A4 $\times$ 1 枚両面. 手書き, コピー等何でも. ただし**縮小**コピー, 貼り付けは不可.
配点 100 点 30 ピーナッツ
公欠 基準と届が独自です. Web ページの病欠・公務欠席等の届出とそれを考慮する (しない) 方法参照.

プチテスト出題計画

Excel の問題はありません. 過去の問題例は `http://hig3.net` > 過去の授業 > 2012 > 生活の中の統計技術 などで参照できます.

- データから度数分布表, 箱ひげ図, ヒストグラム (L01), クロス集計表, 散布図 (L05) などを作ろう
- データから平均値, 最頻値, 中央値 (L02) を求めよう
- データから分散, 標準偏差 (L03), 変動係数 (L04) を求めよう
- 標準得点, 偏差値 (L04) を求めよう
- 共分散, 相関係数 (L05) を求めよう
- 回帰直線 (L06) を求めよう
- これらの量の性質についての選択肢問題もあるかも

新たなる課題

- 各追加 2 ピーナッツ=計 4 ピーナッツになる新たな課題.
  提出: 2013-11-06 水 の授業 or 2013-11-20 水 のテスト前
  1. 龍谷大学 e ラーニングシステム
     `https://moodle.media.ryukoku.ac.jp/` → リメディアルコース統計学 → 第 3 章修了テスト
  2. 龍谷大学 e ラーニングシステム
     `https://moodle.media.ryukoku.ac.jp/` → リメディアルコース統計学 → 第 5 章修了テスト

  このサイトには, `http://hig3.net` → 龍大 Moodle, や Info Seta → e ラーニングサイト → 新 e ラーニングシステム でも到達できます.
  すべてを送信して終了する → レビューを終了する の後に出る, 「あなたの前回受験の要約」ページ (下) を印刷して, 紙で提出. (スクリーンショットを課題にアップロードしてもいい)
- 今週は授業内で紙を 1 枚提出 (+修了テストも提出できます)