重回帰分析

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　L05(2018-10-22 Mon)

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

今日の目標

- 重回帰分析のあてはまりのよさ/わるさを評価できる

ここまで来たよ

3 略解:回帰分析

4 重回帰分析
- 回帰分析
- 説明変数の選択

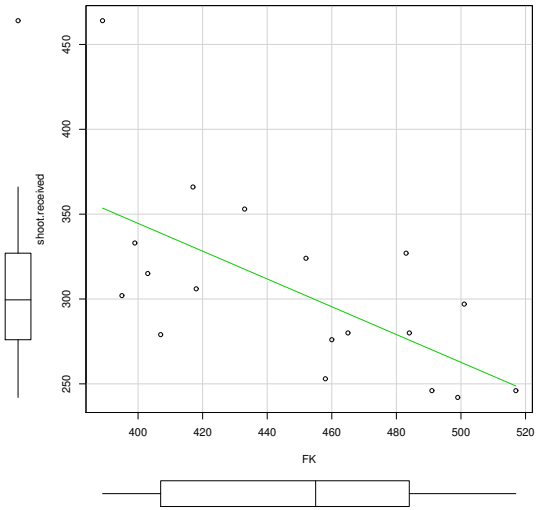
回帰分析

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

相関についてご注意

- x を説明変数, y を目的変数にしたときの回帰直線 $y = ax + b$ と, x,y を入れ替えたときの回帰直線は
- 決定係数 R^2 は, 結果としては相関係数の 2 乗だが, 意味としては,

$$R^2 = \frac{\text{回帰直線上の } y \text{ の分散}}{\text{データの } y \text{ の分散}} = \frac{\frac{1}{N}\sum(((ax+b)-\overline{y})^2}{\frac{1}{N}\sum((y-\overline{y})^2}.$$

変動のうちどれだけの割合を, 回帰直線で説明できるかの比. 1 に近いほどよい.

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.

L05-Q2

Quiz(回帰係数と回帰直線)

ある 2 変量データ (x, y) を Excel の分析ツールで回帰分析したところ, 次のような結果になった.
ただし, 目的変数が $y =$ 期末試験の点数, 説明変数が $x =$ 中間試験の点数 である.

回帰統計

重相関 R	0.918984208
重決定 R2	0.844531974
補正 R2	0.792709299
標準誤差	11.60771105
観測数	5

分散分析表

	自由度	変動	分散
回帰	1	2195.783133	2195.783133
残差	3	404.2168675	134.7389558
合計	4	2600	

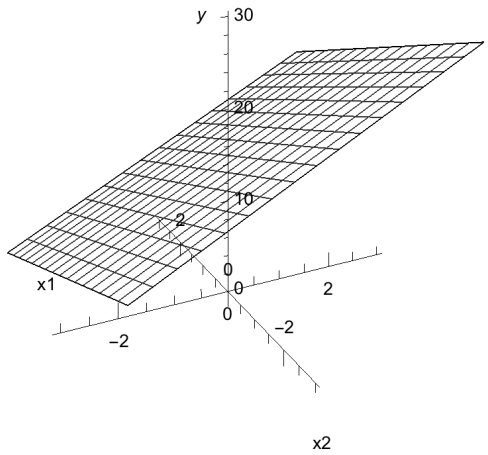
	係数	標準誤差	t
切片	14.45783133	12.41850582	1.164216657
中間試験	0.813253012	0.201454766	4.036901322

1. 回帰直線の式を書こう.
2. 中間試験が 50 点のときの期末試験の点数を予想しよう.

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)

$p = 1$　　$y = a_1 x_1 + b$
↓
$p = 2$　　$y = a_1 x_1 + a_2 x_2 + b.$　　3 次元空間の中の平面.
$p \geq 2$　　$y = a_1 x_1 + a_2 x_2 + \cdots a_p x_p + b.$



重回帰のときも, 決定係数 R^2 が 1 個だけある.

L05-Q3

Quiz(回帰係数と回帰直線)

ある 2 変量データ (x, y) を Excel の分析ツールで回帰分析したところ, 次のような結果になった.
ただし, 目的変数が $y =$ 期末試験の点数, 説明変数が $x =$ 中間試験の点数 である.

回帰統計

重相関 R	0.919106444
重決定 R2	0.844756656
補正 R2	0.689513312
標準誤差	14.20620805
観測数	5

分散分析表

	自由度	変動	分散
回帰	2	2196.367306	1098.183653
残差	2	403.6326942	201.8163471
合計	4	2600	

	係数	標準誤差	t
切片	13.25933401	26.96722561	0.491683283
レポート	0.031281534	0.581427257	0.053801285
中間試験	0.812310797	0.247173536	3.286398738

1. 重回帰の式を書こう.
2. レポートが 40 点, 中間試験が 50 点のときの期末試験の点数を予想しよう.

ここまで来たよ

3 略解:回帰分析

4 重回帰分析
- 回帰分析
- 説明変数の選択

問 単回帰 $(p = 1)$, $p = 2$ 重回帰, $p = 3$ 重回帰, $\cdots$ どれがいい?

仮の答 決定係数 $0 \leq R^2 \leq 1$ で勝負つければ?

$\rightarrow$

特に　　とき決定係数は 1 になってしまう.

いい予測モデルとは

簡単 (説明変数の個数 (自由度) が少ない) $\leftrightarrow$ 正確 (R^2 が大きい)

自由度調整済決定係数

$$\tilde{R}^2 = \frac{R^2}{p \text{ が大きいと大きくなるペナルティ}} = \text{「補正 } R^2\text{」 in Excel}$$

どの説明変数を使う?

目的変数との相関の強さ, $\tilde{R}^2$, その他のハイテクな量をみながら,
0 個から大事なものを増やしていく
全部入りから不要そうなものを減らしていく

多重共線性 (multi colinearity) I

こういうときって回帰係数決まる?
説明変数のどれかが, 他の説明変数の 1 次式で書けてしまうとき, 多重共線性がある, という.

x_1	x_2	y
5	10	55
7	14	75
9	18	95
2	4	25
⋮		

このとき,

- 回帰係数が不定になる (逆行列がない, みたいなもの)　線形代数
- ちょっとの差で, 回帰係数の符号が変わったり, 大きくなったりする.
  - 相関係数 $r_{x_k y}$ と 回帰係数 a_k の符号が違うときは要警戒

多重共線性への対処方法

- 意味を考えて, 役目の重複する変数のうち 1 個 x_k を取り除く
- 数値を見て, 役目の重複する変数のうち 1 個 x_k を取り除く

ダミー変数

ネコの 体長と体長から体重を予想しようとしたとき,
x_2 を オス=0, メス=1 のようにとるとき, ダミー変数という.
これは男女差別ではないし, 予測結果に影響しない.
血液型のときは?

お知らせ

- 中間試験計画
  - 30 ピーナッツ/科目 100 ピーナッツ
  - 60 分?
  - 2018-11-12 月 でどう?
  - 出題計画

    60% 計算問題. データが与えられたときに, 平均値, q-分位数, 中間値, 四分位数, 分散, 標準偏差, 共分散, 相関係数, 単回帰の回帰直線, データ中の 1 個の数値の偏差値が求められる.

    30% これらの量の性質や意味についての正誤判定問題

    10% 上記にあてはまらないかもしれない問題 (ワイルドカード)
    - ★ Excel の操作方法については出題しない
  - 持込 紙はコピーを含め何でも. 電子機器は単機能電卓 (平方根まで) のみ.