多変量解析–重回帰分析–

2019.11.21 Taichi Katsumoto

多変量解析

多変量解析とは、複数の変数に関するデータをもとに、これらの変数間の相互関連を分析する統計的技法の総称。重回帰分析やクラスター分析など様々な分析手法が含まれる。

目的

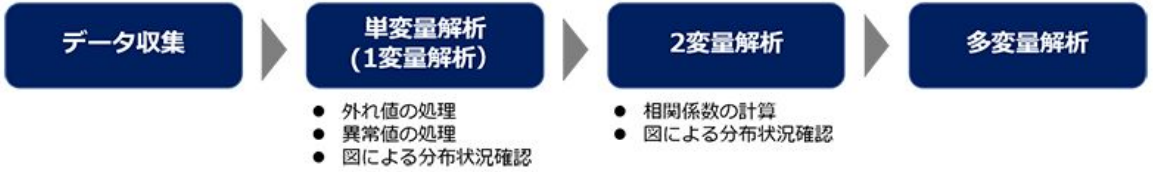
予測	-原因側- 説明変数	→	-結果側- 目的変数	ex.）回帰分析
要約	複数の変数	→	新しい変数	ex.）主成分分析

扱うデータ

量的データ	間隔尺度	ex.）気温、西暦
	比例尺度	ex.）身長、睡眠時間
質的データ	名義尺度	ex.）性別、住所
	順序尺度	ex.）順位、検定の等級

「扱うデータ」と「目的」によって分析手法を決める必要あり！！

多変量解析

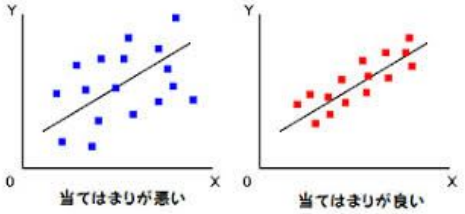


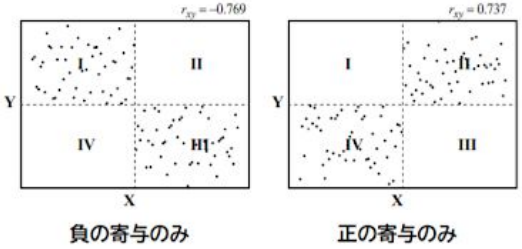
予測の手法		目的変数	
		量的変数	質的変数
説明変数	量的変数	**重回帰分析**	判別分析 ロジスティック回帰
		身長から体重を予測	身長、体重から性別を予測
	質的変数	数量化I類	数量化II類
		体格の良し悪しと食べ物の好き嫌いからマラソンのタイムを予測	体格の良し悪しと食べ物の好き嫌いから体育大学の合否を予測

要約の手法	手法
量的変数	主成分分析、因子分析、クラスター分析
	・身長、体重、腕の長さ、首回り、胸囲の関係を調べる ・様々な価値観やライフスタイルで生活者を分類する
質的変数	コレスポンデンス分析、数量化III類 MDS(多次元尺度構成法)
	・好きなお酒の種類とつまみの関係を調べる ・クロス集計表から属性と設問間の関係をプロットする。

https://udemy.benesse.co.jp/ai/multivariate-analysis.html
https://www.albert2005.co.jp/knowledge/statistics_analysis/multivariate_analysis/multivariate_method

単回帰分析(2変数)





データが右肩上がりに分布しているとき→　共分散は正
データが右肩上がりに分布しているとき→　共分散は負

説明変数： $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$
目的変数： $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]$

平均：データとの差の二乗和を最小にする値

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$$

標準偏差：データのスケールと比較可能

$$s_x^2 = \frac{1}{n}\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right)$$

$$s_y^2 = \frac{1}{n}\left((y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\right)$$

共分散

$$s_{xy} = \frac{1}{n}\left((x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})\right)$$

相関係数：共分散を規格化したもの

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$r_{xy} = 1$　完全な正の相関

$y = ax + b\ (a > 0)$ 全ての点が直線上に乗る

$r_{xy} = -1$　完全な負の相関

$y = ax + b\ (a < 0)$

$r_{xy} \approx 0$　無相関

http://ruby.kyoto-wu.ac.jp/~konami/Text/PartII.pdf

重回帰分析

多変量となったときはベクトルとして扱う！

説明変数： $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$

目的変数： $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]$

$$\boldsymbol{x}_1 = [x_{1,1}, x_{1,2}, \ldots, x_{10,p}]$$

こんな感じになる．これは記法上ややこしい．そこで・・・

生の(Raw)データを $x_{R1}, x_{R2}, \ldots$ のように書いて，$\boldsymbol{x}$ を

$$\boldsymbol{x} = \boldsymbol{x}_R - \bar{\boldsymbol{x}} = [x_{R1} - \bar{x}, x_{R2} - \bar{x}, \ldots]$$

xは平均偏差ということにする！

こうしておくと，分散や共分散がシンプルに書ける．

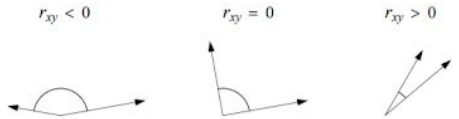
$$s_x^2 = \frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2)$$

$$s_{xy} = \frac{1}{n}(x_1 y_1 + x_2 y_2 + \cdots + x_n y_n)$$

相関係数の幾何学的な意味

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{(\boldsymbol{x}, \boldsymbol{y})}{||\boldsymbol{x}||\ ||\boldsymbol{y}||} = \cos\theta$$

相関係数は，2組のデータのベクトルの成す角度をθ とした時の$\cos\theta$ の値



ベクトルのノルム

$$||\boldsymbol{x}||^2 = (\boldsymbol{x}, \boldsymbol{x}) = x_1 \times x_1 + x_2 \times x_2 + \cdots + x_n \times x_n$$

$$s_x^2 = \frac{1}{n}||\boldsymbol{x}||^2,\ s_x = \sqrt{\frac{1}{n}}\ ||\boldsymbol{x}||$$

$||x||$: $\boldsymbol{x}$ のノルム（長さ）

2つのベクトルの内積

$$(\boldsymbol{x}, \boldsymbol{y}) = x_1 \times y_1 + x_2 \times y_2 + \cdots + x_n \times y_n$$

$$s_{xy} = \frac{1}{n}(\boldsymbol{x}, \boldsymbol{y})$$

http://ruby.kyoto-wu.ac.jp/~konami/Text/PartII.pdf

重回帰分析

ベクトルとして扱うことで連立方程式が解ける。

例えば・・・

$$y_1 = a_{11}x_1 + a_{12}x_2 + b_1$$
$$y_2 = a_{11}x_1 + a_{12}x_2 + b_2$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{b} \quad \Rightarrow \quad y = ax + b$$

$$\boldsymbol{x} = A^{-1}(\boldsymbol{y} - \boldsymbol{b}) \quad \Rightarrow \quad x = a^{-1}(y - b)$$

単純な一次方程式の解法と同じ形に持ち込める。

一般式化すると・・・

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

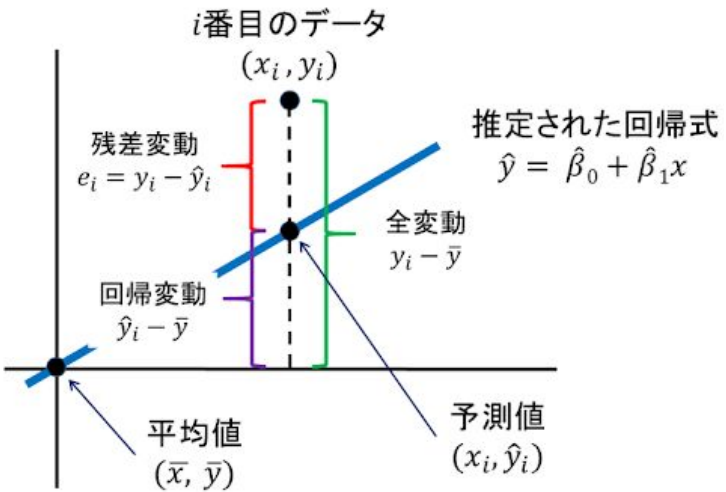
$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} = \begin{bmatrix} s_{x_1x_1} & s_{x_1x_2} & \cdots & s_{x_1x_p} \\ s_{x_2x_1} & s_{x_2x_2} & \cdots & s_{x_2x_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{x_px_1} & s_{x_px_2} & \cdots & s_{x_px_p} \end{bmatrix}^{-1} = \begin{bmatrix} s_{x_1y} \\ s_{x_2y} \\ \vdots \\ s_{x_py} \end{bmatrix}$$

この式を解いている。

重回帰分析をする上で特に気をつける必要があるのは多重共線性と過学習

http://ruby.kyoto-wu.ac.jp/~konami/Text/PartII.pdf

決定係数R^2



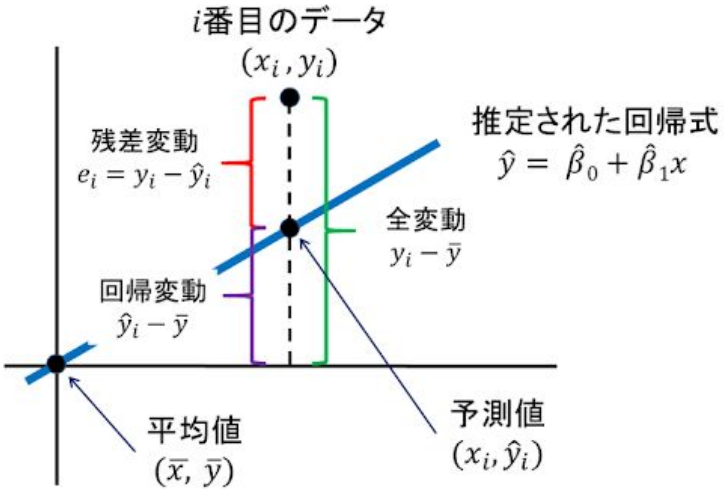
$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

決定係数は説明変数の数が増えるほど1に近づく性質があり、正しくモデルを評価できない。

- 「全変動」：実際のデータとデータ全体の平均値との差を表します（上の図の緑の部分）
- 「回帰変動」：推定された回帰式から得られた予測値とデータ全体の平均値の差を表します（上の図の紫の部分）
- 「残差変動」：実際のデータと推定された回帰式から得られた予測値との差を表します（上の図の赤の部分）

https://bellcurve.jp/statistics/course/9706.html

自由度調整済み決定係数



$$R_f^2 = 1 - \frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

n：データ点数
k：説明変数

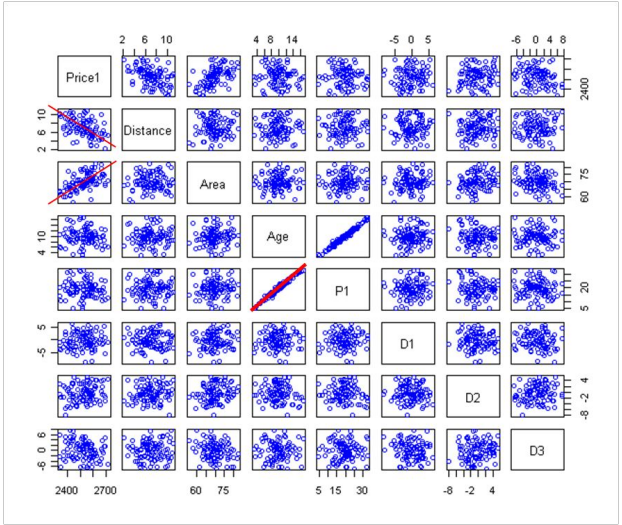
- 「全変動」：実際のデータとデータ全体の平均値との差を表します（上の図の緑の部分）
- 「回帰変動」：推定された回帰式から得られた予測値とデータ全体の平均値の差を表します（上の図の紫の部分）
- 「残差変動」：実際のデータと推定された回帰式から得られた予測値との差を表します（上の図の赤の部分）

https://bellcurve.jp/statistics/course/9706.html

多重共線性(マルチコ，Multicollinearity)

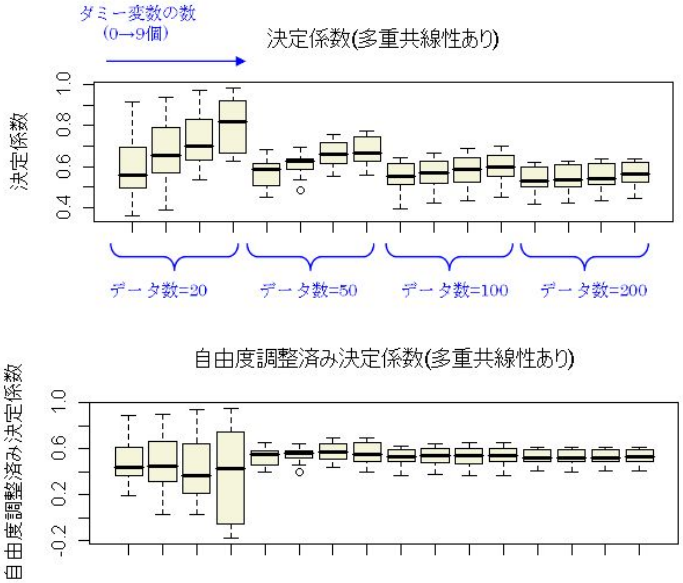
重回帰分析において目的変数の要因となる説明変数を増やすと予測精度が向上する。
説明変数間同士で相関関係が強いものが含まれていると、予測精度が低下する。⇒多重共線性

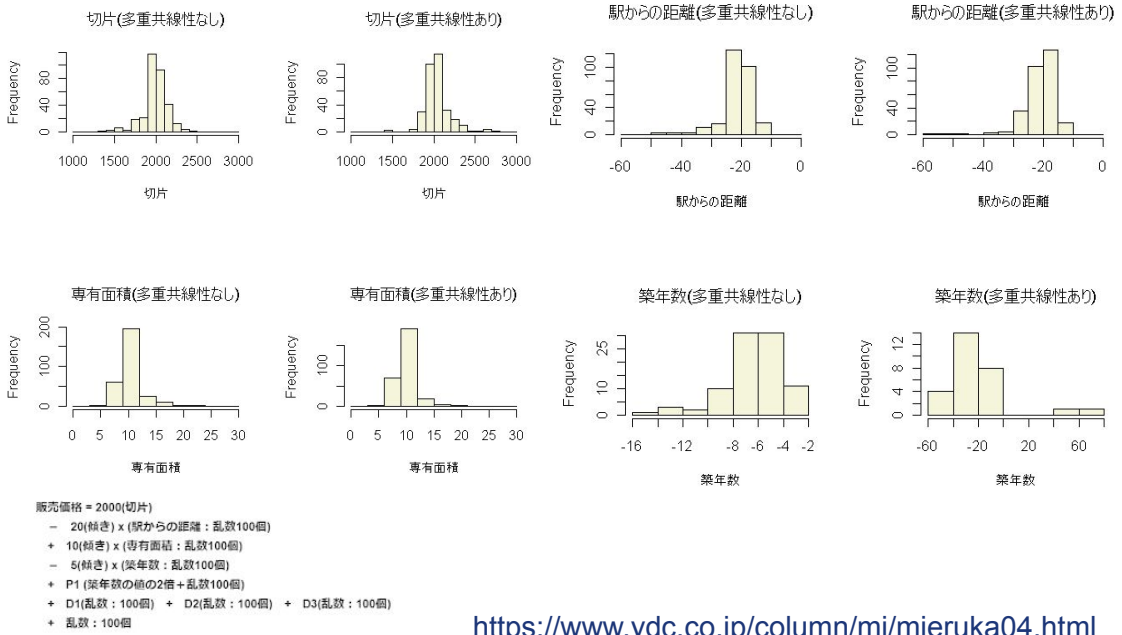
販売価格 = 2000(切片)
- 20(傾き) x (駅からの距離：乱数100個)
+ 10(傾き) x (専有面積：乱数100個)
- 5(傾き) x (築年数：乱数100個)
+ P1 (築年数の値の2倍＋乱数100個)
+ D1(乱数：100個) + D2(乱数：100個) + D3(乱数：100個)
+ 乱数：100個



https://www.ydc.co.jp/column/mi/mieruka04.html

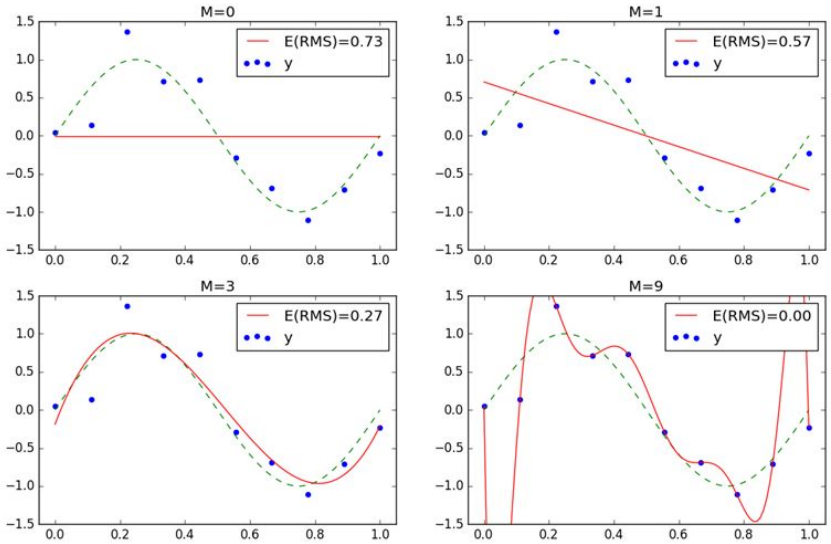
多重共線性(マルチコ，Multicollinearity)





https://www.ydc.co.jp/column/mi/mieruka04.html

過学習（過剰適合）



http://sonickun.hatenablog.com/entry/2016/07/18/191656

避ける方法

多重共線性

過学習

http://sonickun.hatenablog.com/entry/2016/07/18/191656

正則化

解析