# 【単回帰解析/重回帰分析/過学習の防止について】

単回帰分析とは、以下のようなモデル（式）を用いて予測する手法。

## 【単回帰直線の式】

Y = a \* x + b

|  |  |
| --- | --- |
| 変数名 | 説明 |
| Y | 目的変数（予測値） |
| y | 目的変数（観測値） |
| x | 説明変数（観測地 |
| a | 相関係数（重み、回帰係数） |
| b | 切片（定数） |

例えば「景気動向指数」から「売り上げ」を予測しようと考えたとき

「目的変数＝売り上げ」「説明変数＝景気動向指数」となる。

相関係数（重み）は目的変数が説明変数に対してどのくらい影響を及ぼすかを示す。

**よって、目的変数と説明変数の相関が強くないと上手く学習が出来なくなる。**

## 【具体的な計算手順】

1. 目的変数と説明変数の観測地のセット（x,y）を用意する。
2. (x,y)から最小二乗法で単回帰直線の式を求める
3. 求めた単回帰直線の式を決定係数Rで評価する。
4. 評価の結果が良ければ生成したモデル（単回帰直線の式）に説明変数の観測値をxに入れて予測値Yを求める。

## ## 単回帰分析の例

単回帰の式のイメージを掴むために、地価の予測を例に考える。

【式】　　地価　＝　重み　＊　立地条件　+　定数

|  |  |
| --- | --- |
| パラメータ | 項目 |
| 目的変数 | 地価 |
| 説明変数 | 立地条件 |
| 相関係数 | 重み（立地条件が地価に与える影響） |
| 切片（定数） | 立地条件以外の要因からなる数値（治安、景気など |

## 回帰直線の求め方

単回帰分析では、まずデータの散布図から回帰直線を求める。  
回帰直線は観測データ(x, y)の各点との残差が最も小さくなるようにする。

挿絵, 時計 が含まれている画像

自動的に生成された説明

各点の残差の合計が最も小さくなるように直線を求めるために「最小2乗法」を用います。  
最小2乗法では、傾きa（相関係数、重み）と切片bを以下の式で求める。

a = Sxy/Sx

b = Y - aX

|  |  |
| --- | --- |
| 変数 | 説明 |
| Sxy | (x,y)の共分散 |
| Sx | xの分散 |
| X | ｘの平均 |
| Y | yの平均 |

事前に用意したデータセット（x,y）からX,Y,aを求める。

それが決まればbも計算でき、単回帰直線の式が求まる。

## 【決定係数（寄与率）】

単回帰の式の正当性を確かめるための指標。

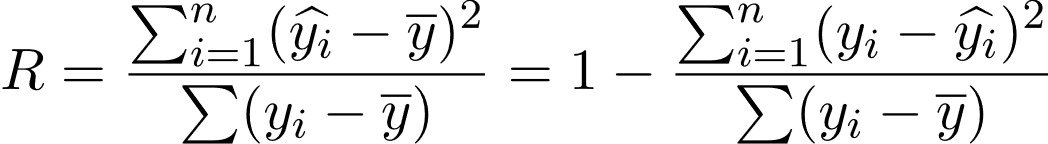
回帰式がどのくらい実際のデータ全体に対して近いのかを示すために決定係数Rを計算する。

決定係数Rは０～１の値を取る。

全データが回帰直線の近くにあるほど1に近づく。

寄与率は単に決定係数をパーセント表示にしたもの。（決定係数１＝寄与率100％）

決定係数Rが１（寄与率が100％）に近いほど、回帰直線は説明力のあるものであると言える。



つまり、決定係数は「１―（偏差の平方和）/(偏差の全平方和)」で計算できる。

## 【重回帰分析】

単回帰分析は説明変数が一つであったが、重回帰分析は説明変数を複数持つ。

先ほどの例で言うと、「目的変数＝売り上げ」「説明変数＝景気動向指数、若者の平均年収」

のようなもの。

単回帰分析よりも、複数の要素を説明変数として取るため推測精度は単回帰分析よりも高くなる。

## 【過学習とは】

同じようなデータを使って学習し続けると、汎化性能が下がってしまうという現象のこと。

未知のデータに対する誤りを「汎化誤差」といい、汎化誤差が小さいことを「汎化能力が高い」という。

## 【過学習防止手法】

1. 相関値の高い説明変数を除外する。

相関が強いデータを使うと「多重共線性」が生じてします。同じような意味を持つデータが別々のデータとして与えられることで、同じ学習が重複して行われてしまうため、過学習を起こしやすくなる。

→相関ヒートマップを使うと各説明変数同士の相関係数を見る事が出来るため、説明変数を選択する際に役にたつ。

1. 正則加工を加える事で抵抗を付与する。

正則化項には、Lasso回帰（L1正則化）リッジ回帰（L2正則化）がある。