PLS 回帰分析

PLS 回帰分析と重回帰分析の違い

MLR (Multi Linear Regression)
応答が 1、m 因子、n サンプルのデータでは以下で表す。

$$Y_i = \beta_0 + \sum_{k=1}^{m} \beta_k x_{ik} + E_i \quad (i = 1,2,...,n)$$

PLS (Partial Least Square)
因子 X を説明変数として直接回帰に用いず、潜在変数 T を説明変数として従属変数 Y への回帰を行う。

$$Y_i = \beta_0 + \sum_{k=1}^{r} \beta_k T_{ik} + E_i \quad (i = 1,2,...,n)$$

ここで、

$$T_{i1} = c_{11}x_{i1} + c_{12}x_{i2} + ... + c_{1m}x_{im}$$
$$T_{i2} = c_{21}x_{i1} + c_{22}x_{i2} + ... + c_{2m}x_{im}$$
...
$$T_{ir} = c_{r1}x_{i1} + c_{r2}x_{i2} + ... + c_{rm}x_{im}$$

T_1 の係数 $c_{11}, c_{12}, ..., c_{1m}$ は T_1 と Y の共分散が最大になるように決められる。
T_2 の係数は、T_1 と無相関で T_1 により説明されない部分と T_2 との共分散が最大になるように決められる。以下 r 個の T について同様に繰り返し係数 c を決定する。
ここで、T を主成分得点、c を固有ベクトルとすると PCR (Principle Component Regression)となる。
潜在変数の数 r を決めるために、観測データを推定と検証用に分け、検証時の予測誤差が最小になるように決める方法がある。クロスバリデーション(cross-validation)を用いた方法では、

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

$$Q^2 = 1 - \frac{PRESS}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

として、Q^2 が最大になるように潜在変数の数 r を決める。$\hat{y}_i$ は i 番目のデータを除いてパラメータの推定を行い、その結果から y_i を予測した値、$\bar{y}$ は平均値をあらわす。

PLS 回帰分析のアルゴリズム

サンプル数を I 、独立変数（因子）の数を J 、従属変数（応答）の数を K とすると、因子 $\mathbf{X}$ は $I \times J$ 行列、応答 $\mathbf{Y}$ は $I \times K$ 行列で表される。PLS では $\mathbf{X}$ と $\mathbf{Y}$ を共通する直交因子と、特定の因子負荷量の積として分解する。すなわち、

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T \quad \text{ただし、} \mathbf{T}\mathbf{T}^T = \mathbf{I} \text{（}\mathbf{I}\text{ は単位行列）}$$

PCR では $\mathbf{T}$ はスコア(score)行列、$\mathbf{P}$ は負荷量(loading)行列となる。なお、PLS 回帰では負荷量は直交しない。

同様に $\mathbf{Y}$ は $\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^T$ として推定される。ここで、$\mathbf{B}$ は回帰の重み(regression weight)を対角要素して持つ対角行列である。$\hat{\mathbf{Y}}$ は $\mathbf{Y}$ の推定量であり、完全には一致しない。

上述した条件を満たす $\mathbf{T}$ はいくらでも存在し、$\mathbf{T}$ を特定するためにはさらに制約条件を追加する必要がある。PLS では、$\mathbf{X}$ と $\mathbf{Y}$ の共分散が最大になるような 1 次結合の重み $\mathbf{w}$ 、$\mathbf{c}$ を見つけるようにする。すなわち $\mathbf{t} = \mathbf{X}\mathbf{w}$ 、$\mathbf{u} = \mathbf{Y}\mathbf{c}$ とした時に、$\mathbf{w}^T\mathbf{w} = 1$ 、$\mathbf{t}^T\mathbf{t} = 1$ の条件の下で $\mathbf{t}^T\mathbf{u}$ を最大にするような $\mathbf{w}$ と $\mathbf{c}$ を見つける。

ここでは基本的なアルゴリズムを紹介する。まず初めに $\mathbf{X}$ と $\mathbf{Y}$ の各列を平均 0、分散 1 に正規化した（Z 変換した）行列 $\mathbf{E}$ 、$\mathbf{F}$ を作成する。これらの行列の平方和を SS_X と SS_Y と表す。ベクトル $\mathbf{u}$ を乱数で初期化した後、以下のプロセスを繰り返す。なお、記号 $\propto$ は記号右側の演算結果を正規化して左側に代入するという意味で用いている。

ステップ 1	$\mathbf{w} \propto \mathbf{E}^T\mathbf{u}$	$\mathbf{X}$ の重みの計算
ステップ 2	$\mathbf{t} \propto \mathbf{E}\mathbf{w}$	$\mathbf{X}$ のスコアの計算
ステップ 3	$\mathbf{c} \propto \mathbf{F}^T\mathbf{t}$	$\mathbf{Y}$ の重みの計算
ステップ 4	$\mathbf{u} = \mathbf{F}\mathbf{c}$	$\mathbf{Y}$ のスコアの計算

上記ステップを $\mathbf{t}$ が収束するまで繰り返す。$\mathbf{t}$ が収束した場合には、$\mathbf{t}$ から $\mathbf{Y}$ を予測するための「回帰の重み」 b を $b = \mathbf{t}^T\mathbf{u}$ をして計算する。また $\mathbf{X}$ の因子負荷量 $\mathbf{p}$ も $\mathbf{p} = \mathbf{E}^T\mathbf{t}$ として計算する。続いて $\mathbf{t}$ の効果を $\mathbf{E}$ と $\mathbf{F}$ から減じる。すなわち、$\mathbf{E} \leftarrow \mathbf{E} - \mathbf{t}\mathbf{p}^T$ 、$\mathbf{F} \leftarrow \mathbf{F} - b\mathbf{t}\mathbf{c}^T$ とする。ここで i 番目の成分についてのベクトル $\mathbf{t}, \mathbf{u}, \mathbf{w}, \mathbf{c}, \mathbf{p}$ はこれらのベクトルを i 番目の列に持つ行列 $\mathbf{T}, \mathbf{U}, \mathbf{W}, \mathbf{C}, \mathbf{P}$ に保持される。また i 番目の成分についてのスカラー b は、対角行列 $\mathbf{B}$ の i 番目の対角成分となる。各成分により $\mathbf{X}$ と $\mathbf{Y}$ の平方和がそれぞれ $\mathbf{p}^T\mathbf{p}$ 、b^2 だけ説明される。したがってこれらの値を SS_X と SS_Y で割ったものが、各成分の寄与率となる。

上記のプロセスを、最終的に $\mathbf{E}$ が 0 行列になるまで（$\mathbf{X}$ の変動をすべて説明するまで）繰り返す。

最終的に $\mathbf{X}$ と $\mathbf{Y}$ の関係は以下のような多変量回帰のかたちで表すことができる。すなわち

$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^T = \mathbf{X}\mathbf{B}_{PLS}$$

ここで $\mathbf{B}_{PLS} = \left(\mathbf{P}^{T+}\right)\mathbf{B}\mathbf{C}^T$ 、$\mathbf{P}^{T+}$ は $\mathbf{P}^T$ についての Moore-Penrose の一般化逆行列。