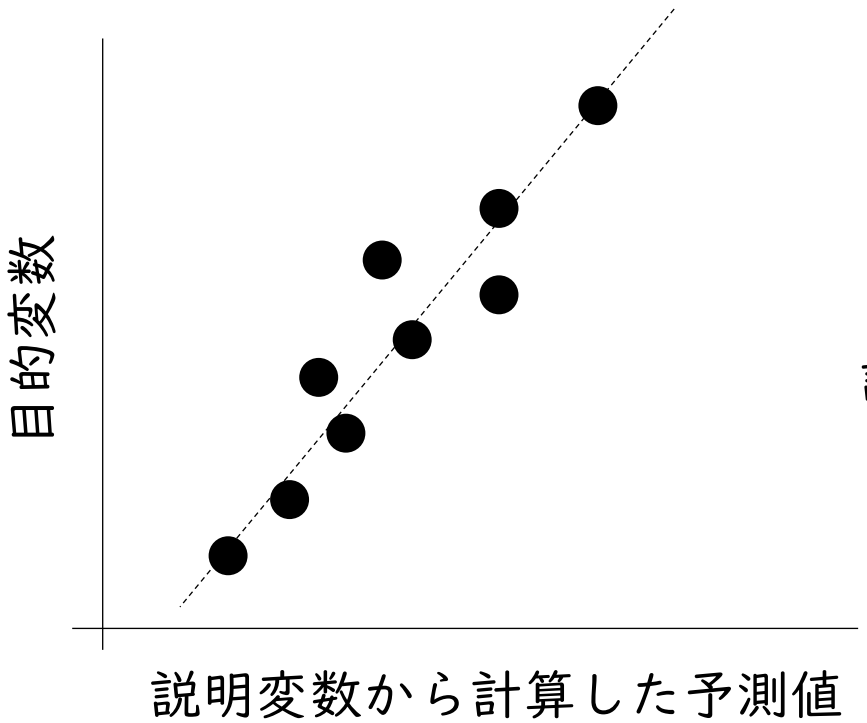
回帰分析

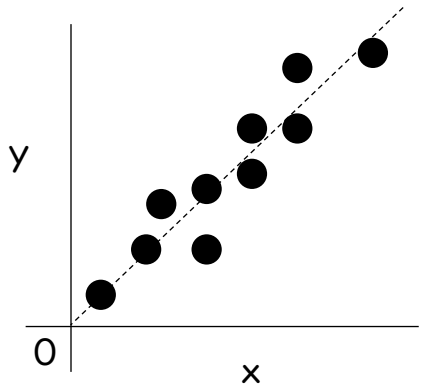
回帰分析のイメージ



目的変数：データとは別に
観測した指標(連続値)
説明変数：観測データ

例えば、測定コストが掛かる変数(目的変数)を、簡易的に測定できる変数(説明変数)から予測値を計算することが出来れば、データを収集するコストを削減出来る

回帰分析のモデル



- 単回帰分析

目的変数　　1変数　　誤差

$$\begin{bmatrix} y \end{bmatrix} = \overset{n(\text{サンプル})}{\left\{ \begin{bmatrix} x \end{bmatrix} \times b + \begin{bmatrix} e \end{bmatrix} \right.}$$

連続値のベクトル

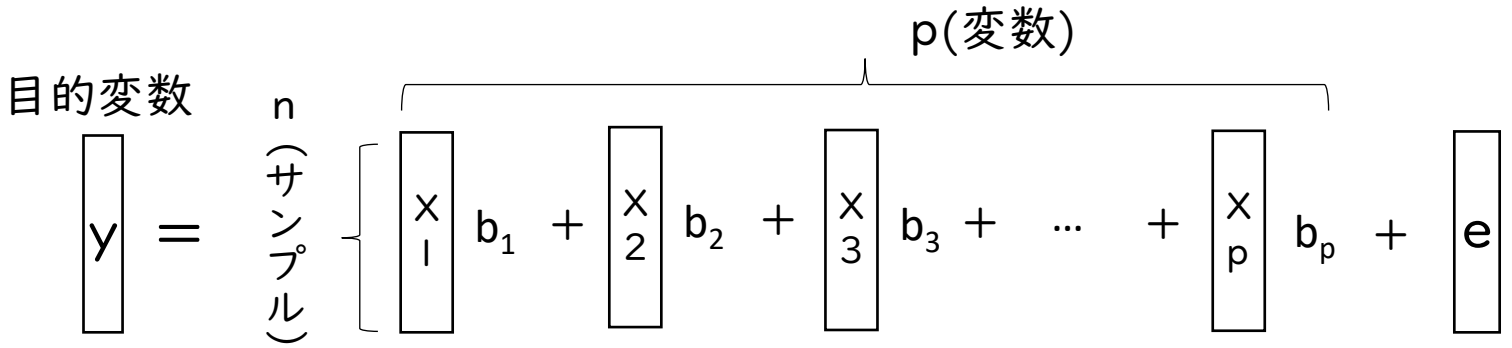
(原点を通るとき)

- 重回帰分析(2変数の場合)

目的変数　　変数1　　変数2　　誤差

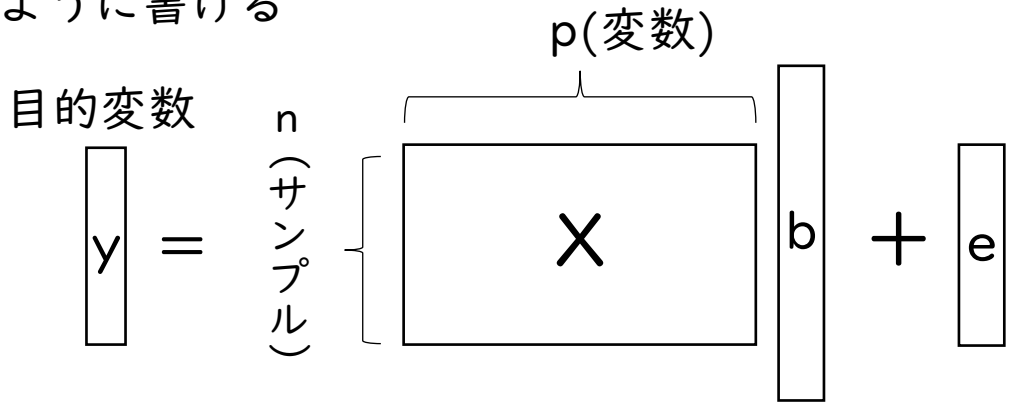
$$\begin{bmatrix} y \end{bmatrix} = \overset{n(\text{サンプル})}{\left\{ \begin{bmatrix} x_1 \end{bmatrix} \times b_1 + \begin{bmatrix} x_2 \end{bmatrix} \times b_2 + \begin{bmatrix} e \end{bmatrix} \right.}$$

重回帰分析のモデル



行列で書くと次のように書ける

$y=Xb$



変数の数が少ない場合($p \leqq n$)

$$b = (X^tX)^{-1}X^ty$$

変数の数が多い場合($p > n$)

逆行列$(X^tX)^{-1}$が計算できず、bが求まらない

逆行列が計算できないケース

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}^{-1}$$ について考える

例えば、

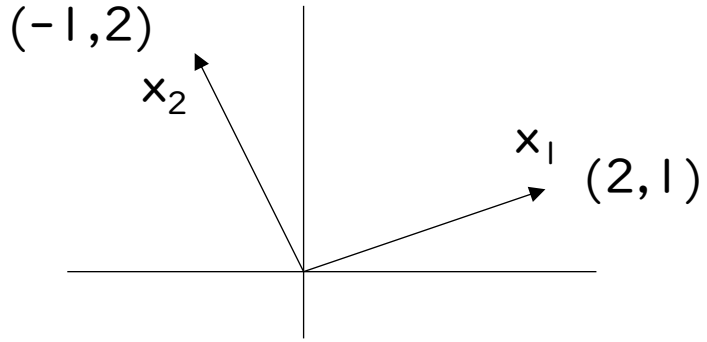
$$x_1 = x_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$ の場合 $$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}^{-1} = \frac{1}{2-2}\begin{bmatrix} 2 & -1 \\ -2 & 1 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad x_2 = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$$ の場合 $$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}^{-1} = \frac{1}{4-4}\begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}$$

いずれの場合も逆行列は計算できない

各変数の値が同じ、もしくは比例関係にあり相関係数が1になる場合は、逆行列の計算が出来ず、重回帰分析の回帰係数は計算できない。
変数同士の相関が高い場合、回帰係数を正しく推定できないケースがあることが知られている

ベクトルの直交と相関



$x_1=(2,1)$

$x_2=(-1,2)$

平均値を引いたベクトル

X_1の平均値は、$(2+1)/2=3/2$ ⟶ $x_1=(1/2,-1/2)$

X_2の平均値は、$(-1+2)/2=1/2$ ⟶ $x_2=(-3/2,3/2)$

x_1とx_2の共分散は、

$$\frac{1}{2}[1/2,-1/2]\begin{bmatrix}-3/2\\3/2\end{bmatrix}=\frac{1}{2}\{(1/2)\times(-3/2)+(-1/2)\times(3/2)\}=0$$

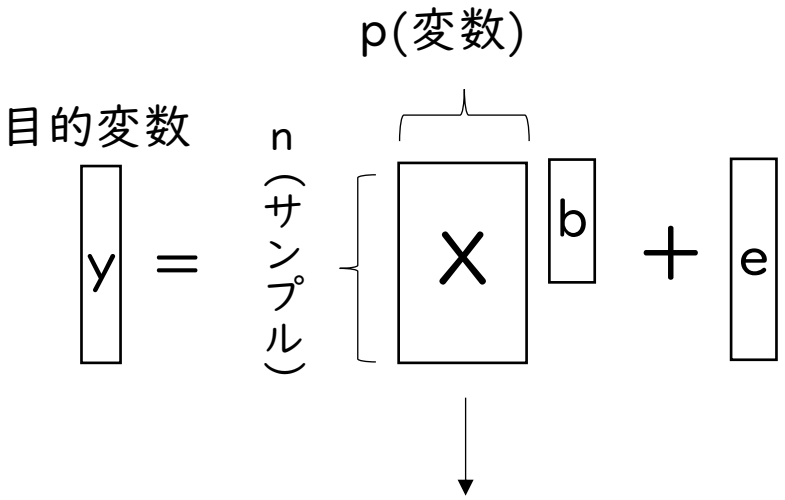
よって、x_1とx_2の相関係数は0

↓

(スコア)ベクトルが直交していれば
逆行列は計算でき、回帰係数は正しく推定可能

重回帰分析のモデル

- 一般の場合(n>p)



一般に正方行列ではないので、
そのままでは逆行列が計算できない

$y=Xb$ に左から$(X^tX)^{-1}X^t$を掛けて

$$(X^tX)^{-1}X^t y= (X^tX)^{-1}X^t Xb=b$$

解決策(Ⅰ)　リッジ回帰

- 重回帰分析

$$\mathbf{b} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'y}$$

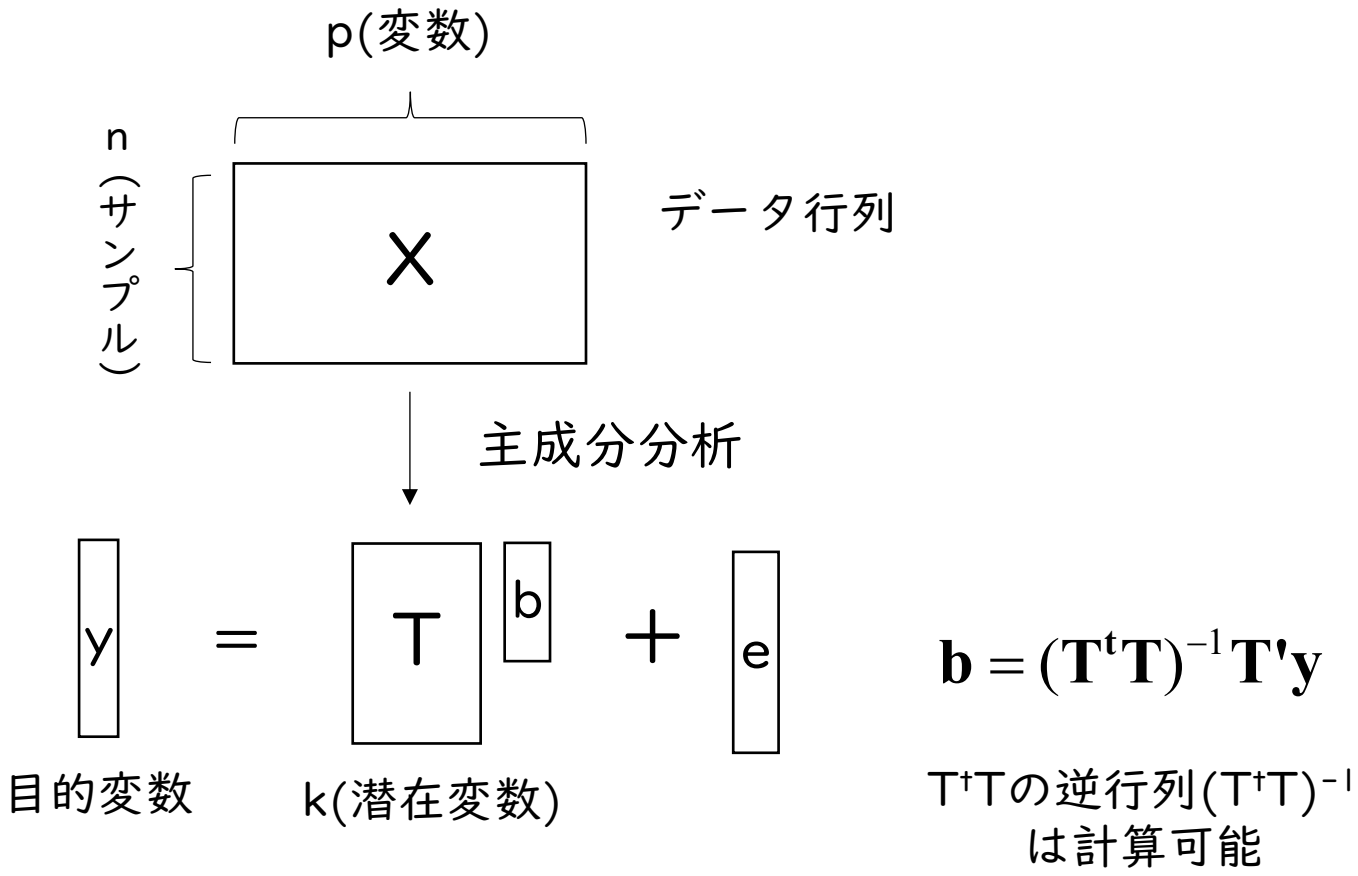
- リッジ回帰

$$\mathbf{b} = \left\{(1-k)\mathbf{X'X} + k\mathbf{I}\right\}^{-1}\mathbf{X'y}$$

kが1に近づけば近づくほど、この行列は単位行列に近づく。
単位行列の逆行列は単位行列なので、回帰係数bは問題なく
計算することができる

解決策(2)　主成分回帰

- 主成分回帰



解決策(3)　PLS回帰

PLS回帰は、PLSによる次元の圧縮と重回帰分析を組み合わせた方法

NIRS

説明変数
PLSスコア

目的変数
(エタノール濃度)

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & x_{24} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \cdots & x_{np} \end{bmatrix} \rightarrow \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{bmatrix} \leftrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

PLSによる次元の圧縮(p変数→k変数)　　重回帰分析

重回帰分析

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = a_1 \begin{bmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{bmatrix} + a_2 \begin{bmatrix} t_{12} \\ t_{22} \\ \vdots \\ t_{n2} \end{bmatrix} + b \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

係数a_1,a_2,b(切片)を
最小二乗法を用いて推定

PLSとPLS回帰の違い

- PLS　　本セミナーで紹介してきたPLS

$$\frac{1}{(n-1)^2}\mathbf{X'YY'Xw} = \lambda\mathbf{w}$$

スコア　$\mathbf{t} = \mathbf{Xw}$

- スコアの直交化

繰り返し

$$\mathbf{X} = \mathbf{X} - \mathbf{tp'}$$

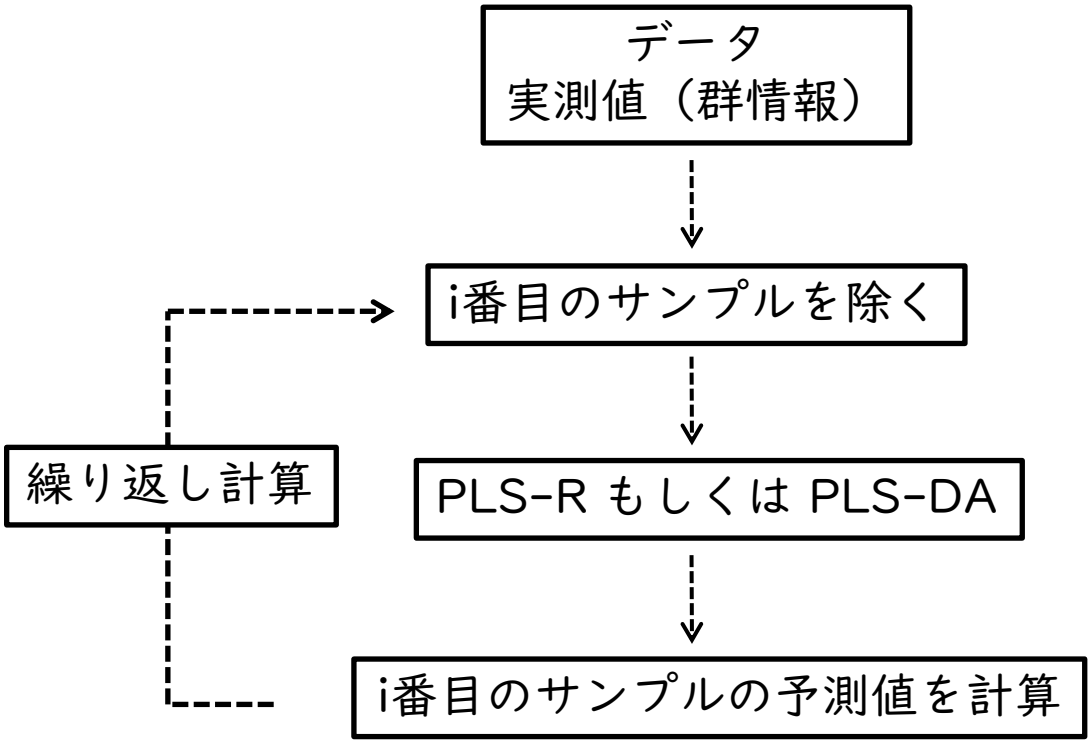
PLS回帰

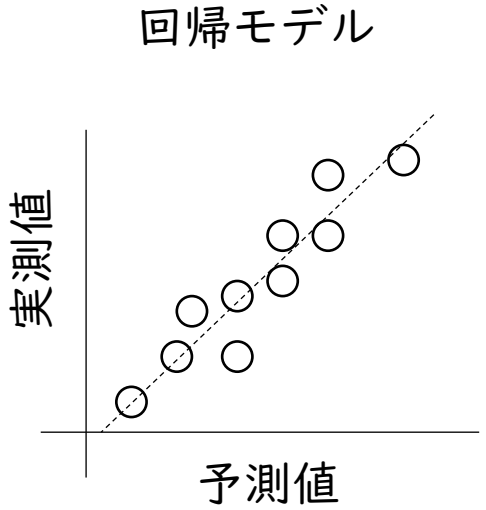
ただし　$\mathbf{p} = \dfrac{\mathbf{X't}}{\mathbf{t't}}$

PLSとスコアの直交化の操作を繰り返し計算して得られるスコアを用いて、回帰分析を行うのがPLS回帰

回帰(判別)モデルの性能評価

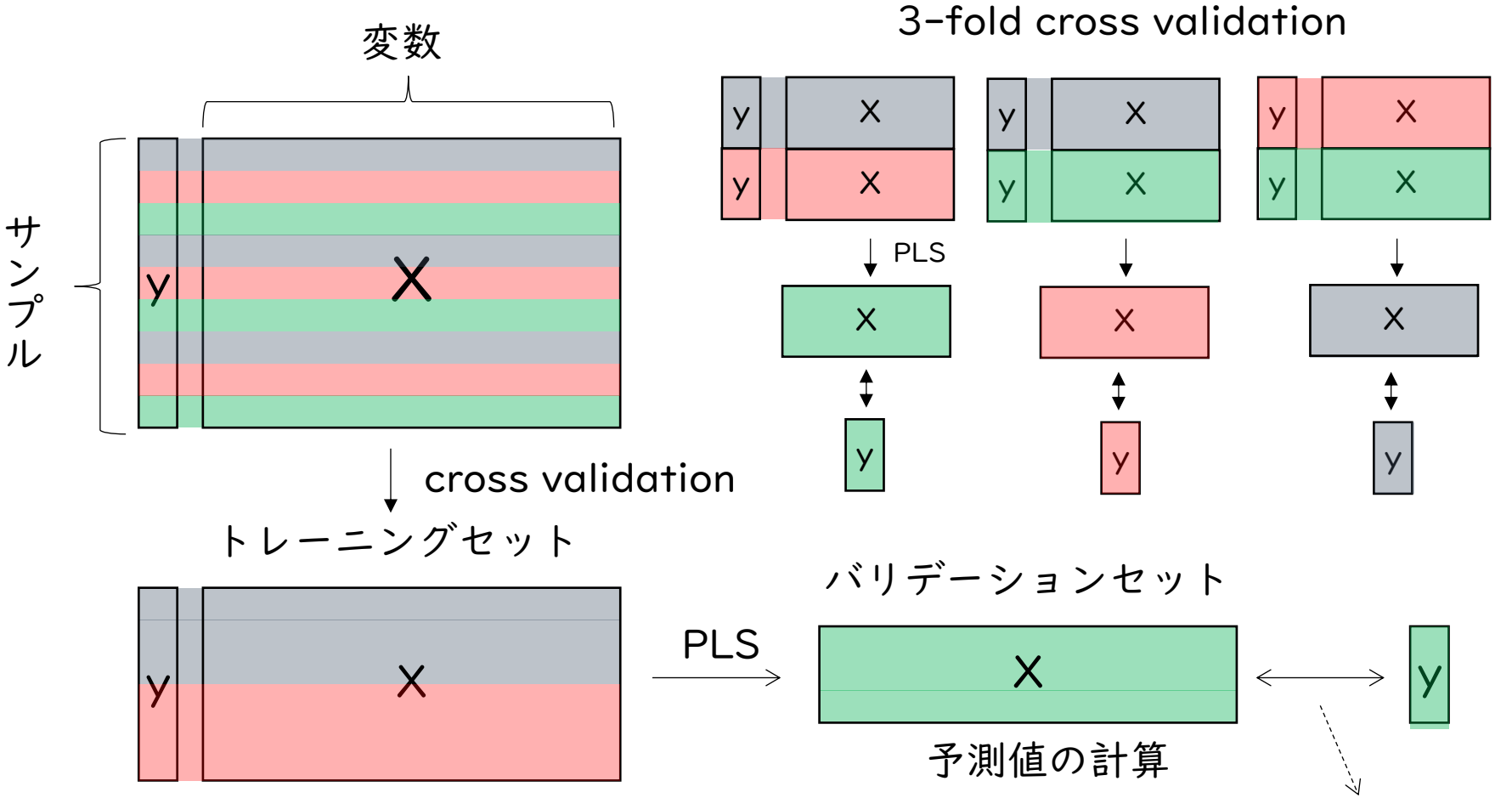
例. Leave-one-out cross validation (LOOCV)の場合





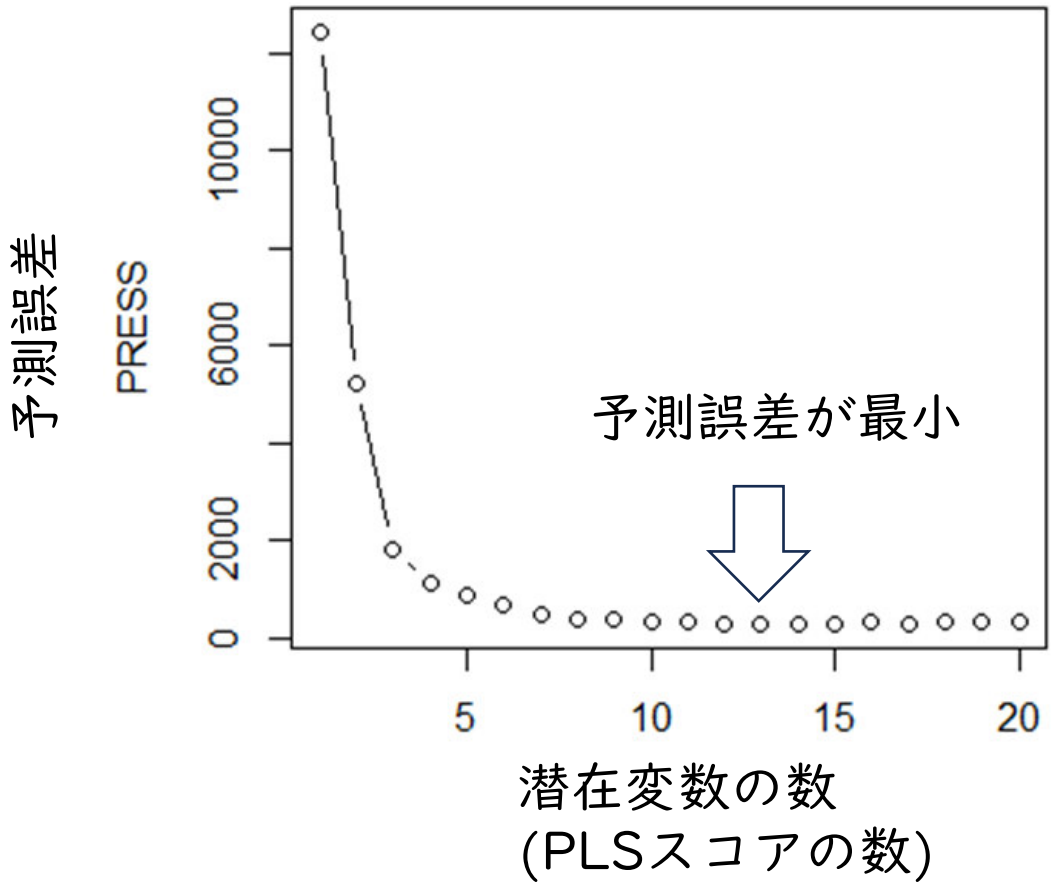
正しく性能評価を行うためには、さらに独立テストが必要

k-foldクロスバリデーション



クロスバリデーションによって、最適な潜在変数(スコア)の数を推定する

PLS回帰における最適な潜在変数の推定例



回帰分析や判別分析では、潜在変数の数に対して予測誤差が最小のモデルを選択するので、PLSではなくPLS回帰やPLS-DAを使う

回帰モデルの性能評価(R^2とQ^2)

・RSS (residual sum of squares)： 残差平方和

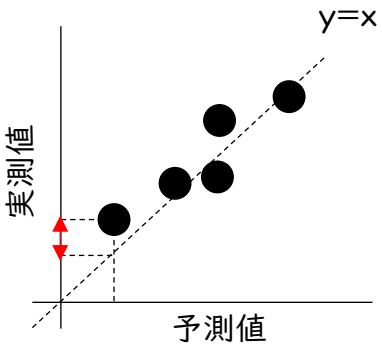
PLSで計算した予測値と実測値(目的変数y)の差の二乗和

・SS(sum of squares)： 平方和

実測値(目的変数y)と実測値の平均の差の二乗和

↓

$R^2 = 1 - RSS/SS$



・PRESS (predicted residual error sum of squares) ：予測残差平方和

PLSで計算した予測値(LOOCV)と実測値(目的変数y)の差の二乗和

↓

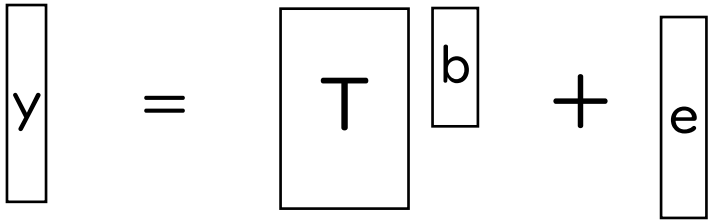
$Q^2 = 1 - PRESS/SS$

R^2とQ^2を性能評価の基準(〇〇以上だと予測精度が高い)としている論文もあるが、決まった基準は無い

R^2,Q^2の基準の例

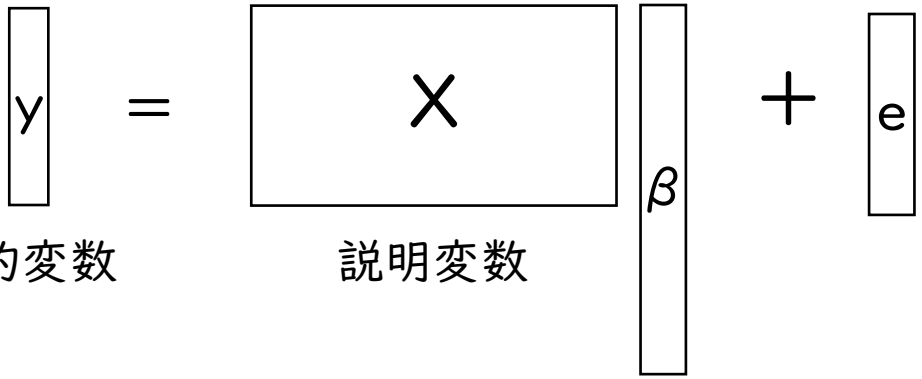
- The literature suggests that R^2 values of 0.67, 0.33, and 0.19 are substantial, moderate, and weak, respectively
- $Q^2 > 0$ implies the model has predictive relevance, whereas $Q^2 < 0$ represents a lack of predictive relevance.
  - Using partial least squares in operations management research: A practical guideline and summary of past research
  - Chin, W.W., 1998. The partial least squares approach to structural equation modeling. In: Marcoulides, G.A. (Ed.), Modern Methods for Business Research. Lawrence Brlbaum Associates, Mahwah, NJ, pp. 295–336.
- Because R^2 is embraced by a variety of disciplines, scholars must rely on a "rough" rule of thumb regarding an acceptable R^2,with 0.75, 0.50, 0.25, respectively, describing substantial, moderate, or weak levels of predictive accuracy
  - Partial least squares structural equation modeling (PLS-SEM) An emerging tool in business research
  - Hair, J.F., Ringle, C.M. and Sarstedt, M. (2011), "PLS-SEM: indeed a silver bullet", Journal of Marketing Theory and Practice, Vol. 19 No. 2, pp. 139-151.
  - Henseler, J., Ringle, C.M. and Sinkovics, R.R. (2009), "The use of partial least squares path modeling in international marketing", Advances in International Marketing, Vol. 20, pp. 277-320.
- A robust model should have $R^2>0.5$, $Q^2>0.5$, and $|R^2-Q^2|<0.2-0.3$
  - Predictive Approaches in Drug Discovery and Development: Biomarkers and In Vitro / In Vivo Correlations (Wiley Series on Technologies for the Pharmaceutical Industry)

PLSの回帰係数βを用いた変数選択



目的変数　　k(潜在変数)

$$\mathbf{b} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$$



目的変数　　説明変数

$$\boldsymbol{\beta} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{b}$$

ただしW、Pはそれぞれ
各列がベクトルwとpの行列

bではPLSの各スコアに対する係数であり、元の変数との関係が不明だが、βは各変数に対する係数になっており、この値が大きな変数が回帰モデルに重要な変数と判断できる

PLS回帰係数βの導出

$$\mathbf{X} = \mathbf{TP'}$$

右からWを掛けて

$$\mathbf{XW} = \mathbf{TP'W}$$

右から $(\mathbf{P'W})^{-1}$ を掛けて

$$\mathbf{XW}(\mathbf{P'W})^{-1} = \mathbf{TP'W}(\mathbf{P'W})^{-1} = \mathbf{T}$$

よって

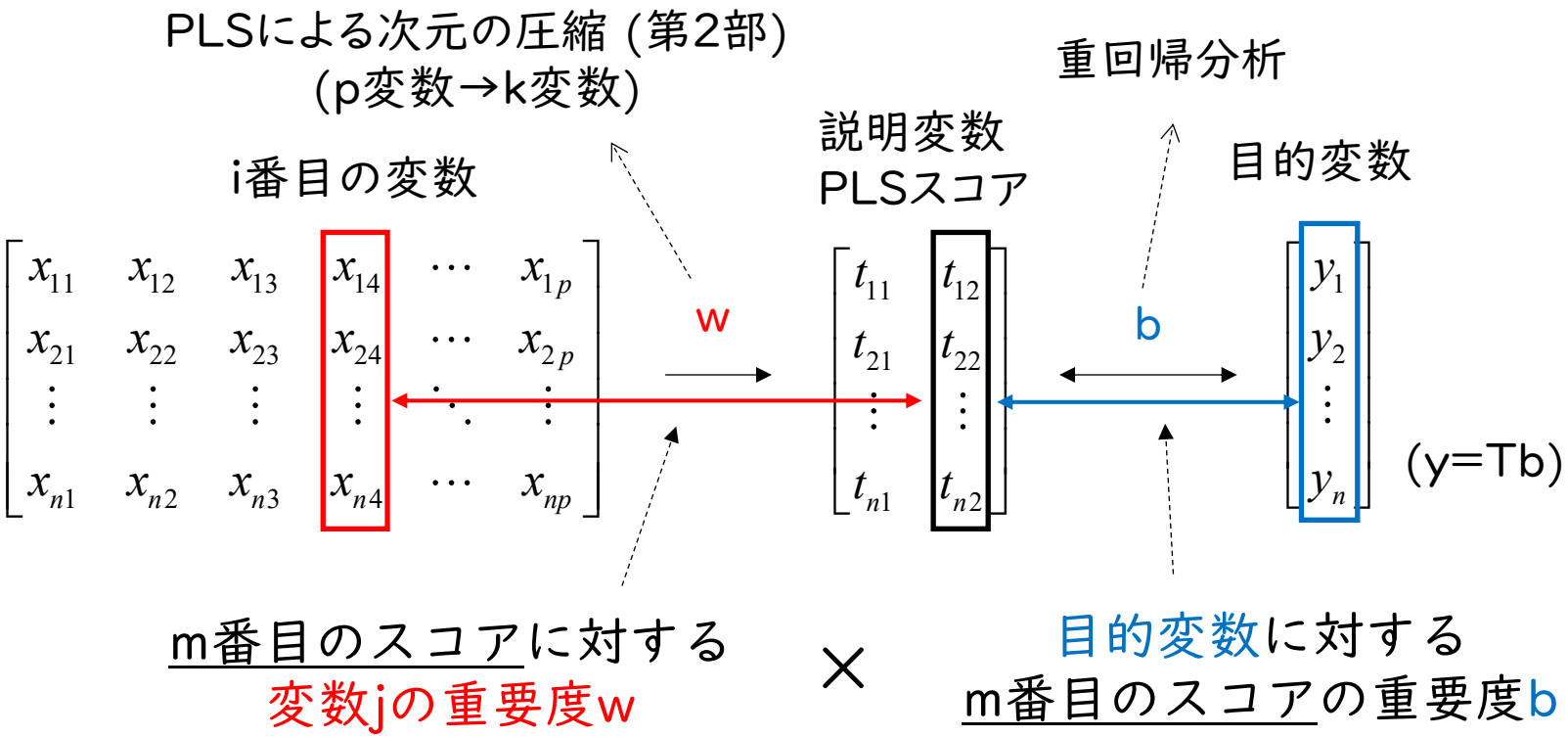
$$\mathbf{y} = \mathbf{Tb} = \mathbf{XW}(\mathbf{P'W})^{-1}\mathbf{b}$$

$\boldsymbol{\beta} = \mathbf{W}(\mathbf{P'W})^{-1}\mathbf{b}$ とすれば

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$$

VIP(Variable Importance in Projection)を用いた変数選択のイメージ

PLS回帰のイメージ



＝ 目的変数に対する変数jの重要度 VIP

VIPを用いた変数選択

- m番目のスコアの重要度 SS_m ： $b_m^2 t_m' t_m$

  回帰係数b_m：m番目スコアの全てのスコアに対する重要度(正 or 負)
  各スコアt_mの分散： m番目のスコアのバラツキ

×

- m番目のスコアにおける変数jの重要度 w_{mj}　　固有ベクトルw_m

j番目の変数の重要度

$$VIP_j = \sqrt{p \frac{\sum_{m=1}^{M} w_{mj}^2 SS_m}{\sum_{m=1}^{M} SS_m}}$$

m：潜在変数の数
w：PLSの固有ベクトル
p：変数の数
SS_m：$b_m^2 t_m' t_m$

例えば、Il-Gyo Chong et al, “Performance of some variable selection methods when multicollinearity is present”, Chemom. Intell. Lab. Sys., 78 (2005) 103–112

回帰(判別)モデルにおける変数の選び方と主成分またはPLS負荷量による変数の選び方の違い

- PCA・PLS負荷量を用いた変数の選び方
  - 特定の成分(例えばPC1、PC2)のスコアに着目する
  - 着目した成分(例えばPC1、PC2)に対して、主成分負荷量・PLS負荷量の値が大きな変数に着目する
    - 例えば、PC1ではAla, PC2ではGlyがそれぞれ重要な代謝物

- PLSモデリングのβやVIPによる変数の選び方
  - 回帰(または判別)モデルの予測に影響を及ぼす変数を選ぶ
    - 例えば、AlaとGlyがモデルにとって重要な代謝物
  - VIPに関する説明
    - 『VIP法ではPLS-beta法よりは根拠のある変数選択手法となっています．そのため，PLSでの変数選択ではVIP法を第一候補として用いることが多いです』
      - 藤原幸一、スモールデータ解析と機械学習、オーム社 (2022)より一部改変
    - “A major drawback of the VIP index is its lack of theoretical background. (VIP は理論的根拠に欠ける)
      - Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform. 2007 Jan;8(1):32-44. より引用

ケモメトリックスの研究例 (1)

- 試験内容の概要
  - 酵母によるバイオエタノール(サトウキビなどのバイオマスから生成されるエタノール)生産において、グルコースとエタノール濃度を、近赤外スペクトル(NIRS)データから推定する
  - グルコースは酵母の栄養源であり、発酵生成物がエタノールである
  - 詳しくは、B.Liebmann, A. Friedl, K.Varmuza, “Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics”, Analytica Chimica Acta, 642(1-2), pp 171-178(2009)を参照
- データの説明
  - 説明変数：ライ麦、小麦、トウモロコシを材料としてアルコール発酵を行い、波長領域が1115から2285nmを近赤外分光分析装置で測定し、得られた吸光度の値の1次微分をデータとしている
    - 166サンプル、 235変数のデータ
  - 目的変数：グルコースとエタノールの濃度(g/L)を液体クロマトグラフィーで測定した得られたデータ

PLS回帰(1)

- plsパッケージとデータの読み込み

```
library(pls)
library(chemometrics)
data(NIR)
```

- データの準備

```
X <- NIR$xNIR
y <- NIR$yGlcEtOH[,2] # エタノール
# y <- NIR$yGlcEtOH[,1] # グルコース
```

- データセットの分割(訓練3/4、テスト1/4)準備

```
index <- seq(1,length(y),4)
```

PLS回帰(2)

- 訓練データ

```
X_train <- X[-index,]
y_train <- y[-index]
```

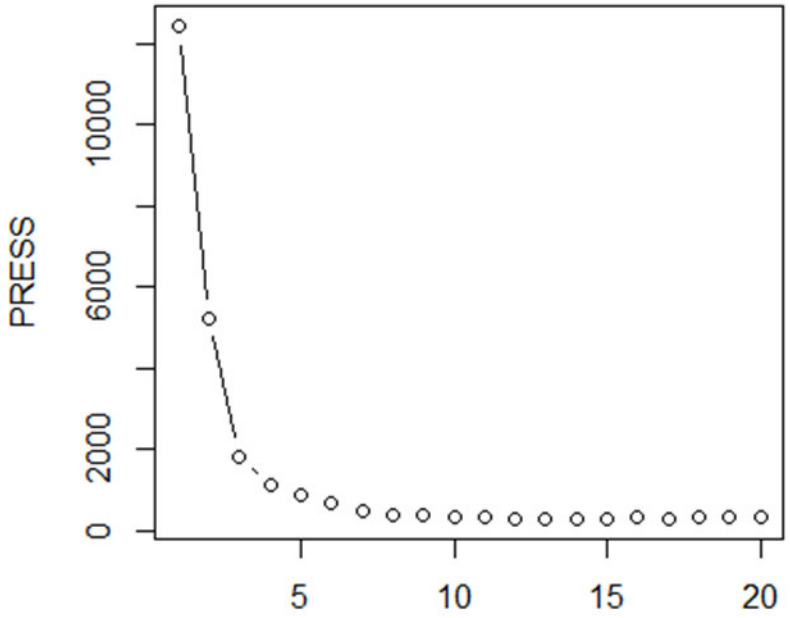
- PLS回帰の計算

```
pls <- plsr(y~., data=data.frame(X_train,y=y_train),20,
                                    validation="LOO", scale=TRUE)
```

- 潜在変数の数

```
PRESS <- as.vector(pls$validation$PRESS)
plot(PRESS, type="b")
```

PLS回帰(3)



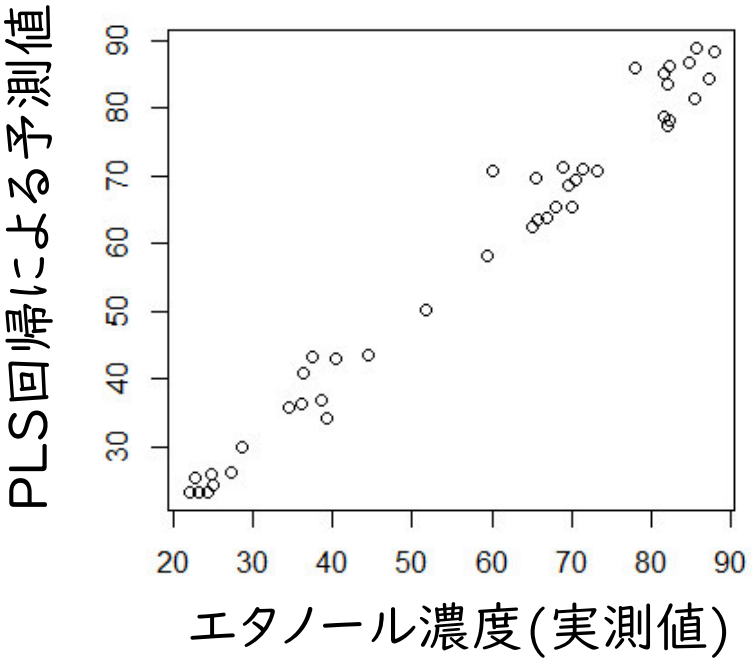
- 潜在変数の数

```
lvnum <- which.min(PRESS)  # 13
```

PLS回帰(4)

- テストデータ

```
X_test <- X[index,]
y_test <- y[index]
```



- テストデータの予測

```
predicted <- predict(pls, ncomp= lvnum,
                     newdata=data.frame(X_test))
plot(y_test,predicted)
```

PLS回帰(5)

- R^2、Q^2

```
R2(pls, estimate = "train")$val[,,lvnum+1]
R2(pls)$val[,,lvnum+1] # cross-validation
R2(pls, newdata=data.frame(X_test, y=y_test))
                                    $val[,,lvnum+1] # test set
```

- 回帰係数β

```
beta <- pls$coefficients[,,lvnum]
```

[参考] VIPの計算

- オリジナル

```
source("https://mevik.net/work/software/VIP.R")
        # plsパッケージ開発者のサイトより
pls <- plsr(y~., data=data.frame(X_train,y=y_train),
        20,validation="CV", scale=TRUE, method = "oscorespls")
vip <- VIP(pls)[lvnum,]
```

- 修正

```
source("C:/R/vip.R") # oscorespls をコメントアウト
pls <- plsr(y~., data=data.frame(X_train,y=y_train),
                                        20,validation="CV", scale=TRUE)
vip <- VIP(pls)[lvnum,]
```

グルコース濃度の予測

- データの準備、分割
- PLS回帰とクロスバリデーション
- 潜在変数の数の決定
- グルコース濃度の予測