Introduzione alla chemiometria e disegno sperimentale

Modulo 8: PCR, PLS e modelli QSAR/QSPR in R

Docente: Dr. Sabina Licen (slicen@units.it)

Regressione multilineare

OLS: *Ordinary Least Squares regression*

```
Model<-lm(vettoreY~vettoreX1+vettoreX2+vettoreX3)
```

PCR: *Principal Component Regression*

PLS: *Partial Least Squares regression*

pls package

https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf

Dividere il dataset in *Training* e *Test*

Si può dividere il dataset in *Training* e *Test* a seconda di indicazioni sperimentali o rispetto a studi precedenti oppure in proporzione circa 70 % - 30 % in diversi modi, tra cui "lineare":

```
Divisione<-round((nrow(Dataset)*70/100),digit=0)
Train<-Dataset[1:Divisione,]
Test<-Dataset[(Divisione+1):nrow(Dataset),]
```

o "random" usando la funzione sample (ricordando di usare il seed):

```
NumTest<-round((nrow(Dataset)*30/100),digit=0)
set.seed(7)
Train<-Dataset[-sample(c(1:nrow(Dataset)),NumTest),]
set.seed(7)
Test<-Dataset[sample(c(1:nrow(Dataset)),NumTest),]
```

PCR in R

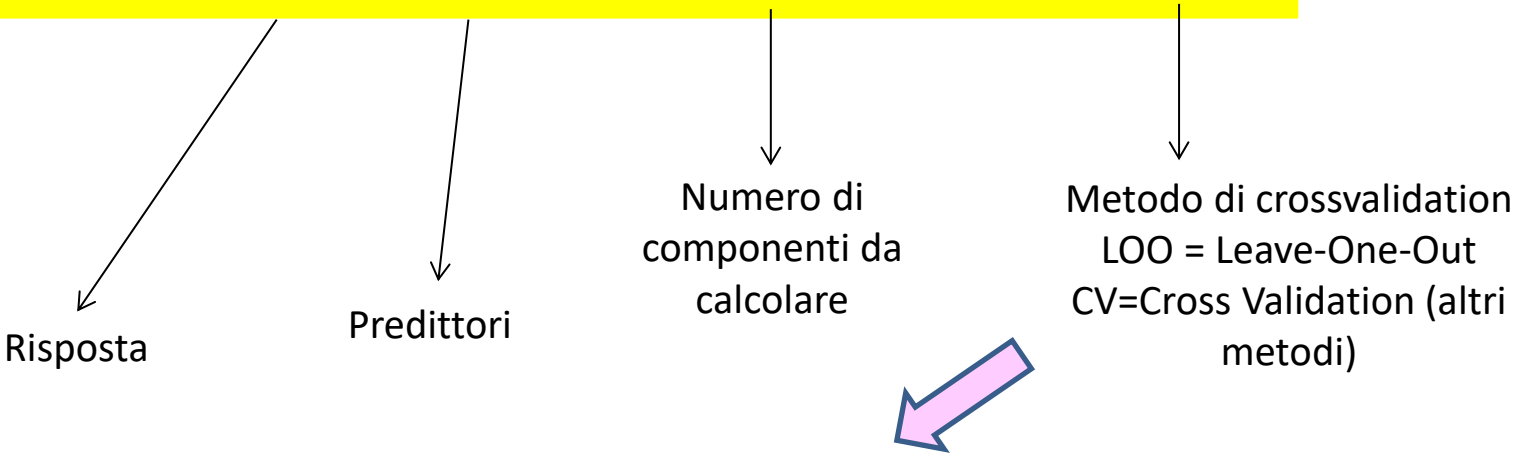
pls package

Siano:

TrainP la matrice dei Predittori (per il *training*)

TrainR il vettore Risposta (per il *training*)

Model<-pcr(TrainR ~ TrainP, ncomp = 10, validation = "LOO")



Se si usa "**CV** " si aggiungono i seguenti argomenti:

- "**segments**" = numero di segmenti in cui dividere i campioni oppure "**length.seg**" = numero di campioni per segmento;
- e "**segment.type**"= può essere "*random*", "*consecutive*", "*interleaved*"

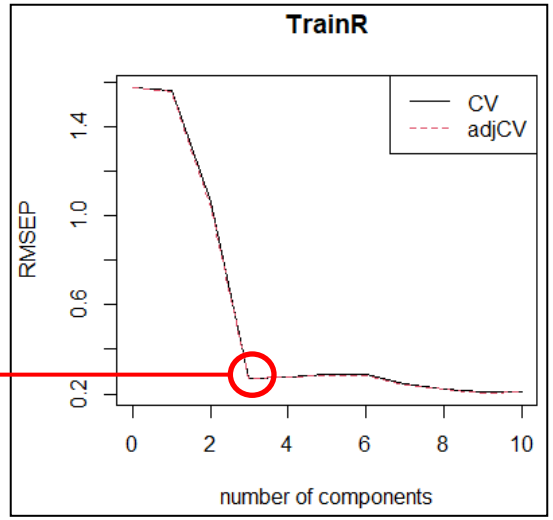
PCR in R

summary(Model) → mostra i risultati

RMSEP(Model) → mostra i valori di RMSEP in base al numero di componenti

plot(RMSEP(Model), legendpos = "topright") → mostra i valori di RMSEP in grafico vs. le componenti

Si osserva la zona del "gomito" della curva



PLS in R

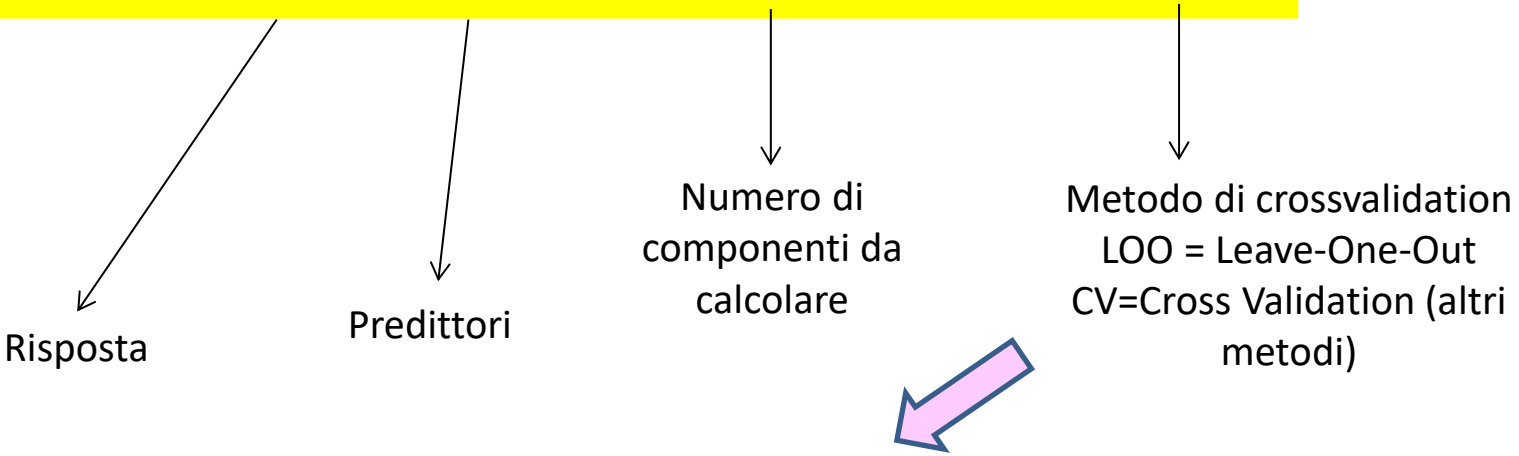
pls package

Siano:

TrainP la matrice dei Predittori (per il *training*)

TrainR la matrice delle Risposte (per il *training*)

Model<-plsr(TrainR ~ TrainP, ncomp = 10, validation = "LOO")



Se si usa "**CV** " si aggiungono i seguenti argomenti:

- "**segments**" = numero di segmenti in cui dividere i campioni oppure "**length.seg**" = numero di campioni per segmento;
- e "**segment.type**"= può essere "*random*", "*consecutive*", "*interleaved*"

PLS in R

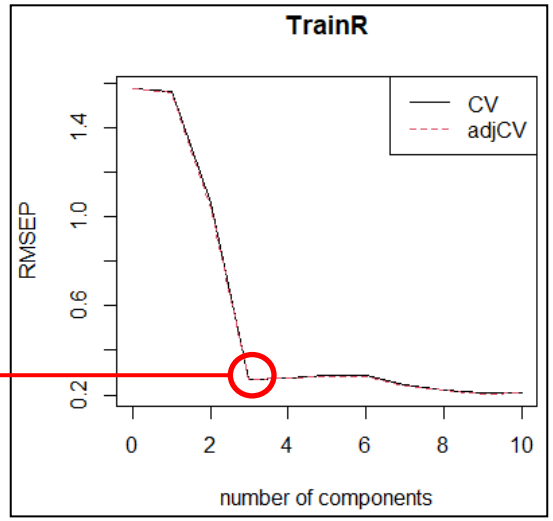
Analogamente alla PCR:

summary(Model) → mostra i risultati

RMSEP(Model) → mostra i valori di RMSEP in base al numero di componenti

plot(RMSEP(Model), legendpos = "topright") → mostra i valori di RMSEP in grafico vs. le componenti

Si osserva la zona del "gomito" della curva



Validazione esterna (sia per PCR che per PLS)

Siano:

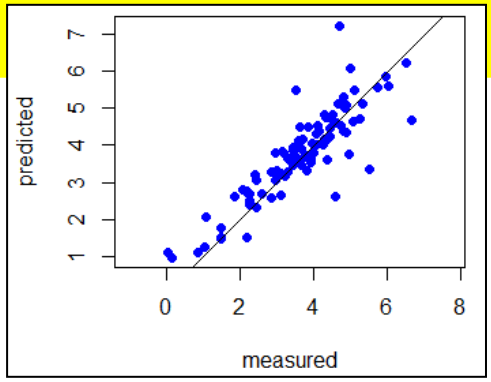
TestP la matrice dei Predittori (per il *test*)

TestR il vettore (in PCR) o la matrice (in PLS) delle Risposte (per il *test*)

```
PredizionePLS<-predict(Model, ncomp=  16  ,newdata=TestP)
```

Numero di componenti scelte in base al RMSEP

```
plot(TestR,PredizionePLS[,,1],pch=16,col="blue",xlab="measured",ylab="predicted",asp=1)
abline(a=0,b=1) # retta y=x
```




```
summary(lm(PredizionePLS[,,1]~TestR))
```

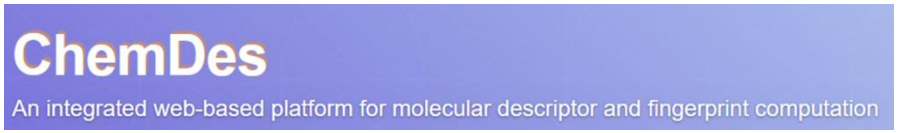
Da questo si ricava R^2 in predizione

QSAR e QSPR – Descrittori molecolari

Software free per la generazione di descrittori molecolari:

Mordred: a molecular descriptor calculator

https://github.com/mordred-descriptor
DOI: 10.1186/s13321-018-0258-y



http://www.scbdd.com/chemdes/
DOI: 10.1186/s13321-015-0109-z

PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints

https://www.download.io/padel-descriptor-download-windows.html
DOI: 10.1002/jcc.21707

Pacchetti di R dedicati per QSAR e QSPR

ezqsar package

camb package

} **Beginner users**

caret package **Advanced users**

Consentono di:

- calcolare alcuni tipi di descrittori molecolari
- di selezionare i descrittori più rilevanti per la Risposta
- di costruire modelli con vari tipi di algoritmi
- di visualizzare i risultati