

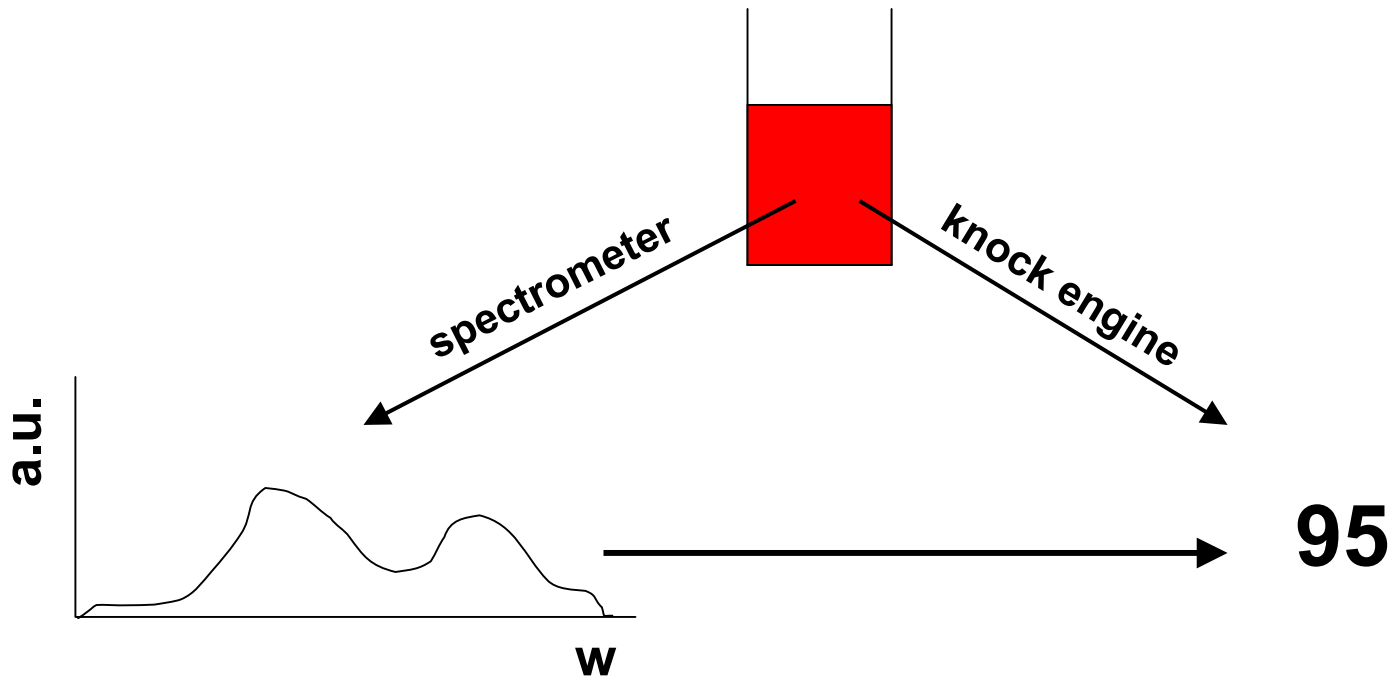
Regression Methods

Process Analysis & Chemometrics

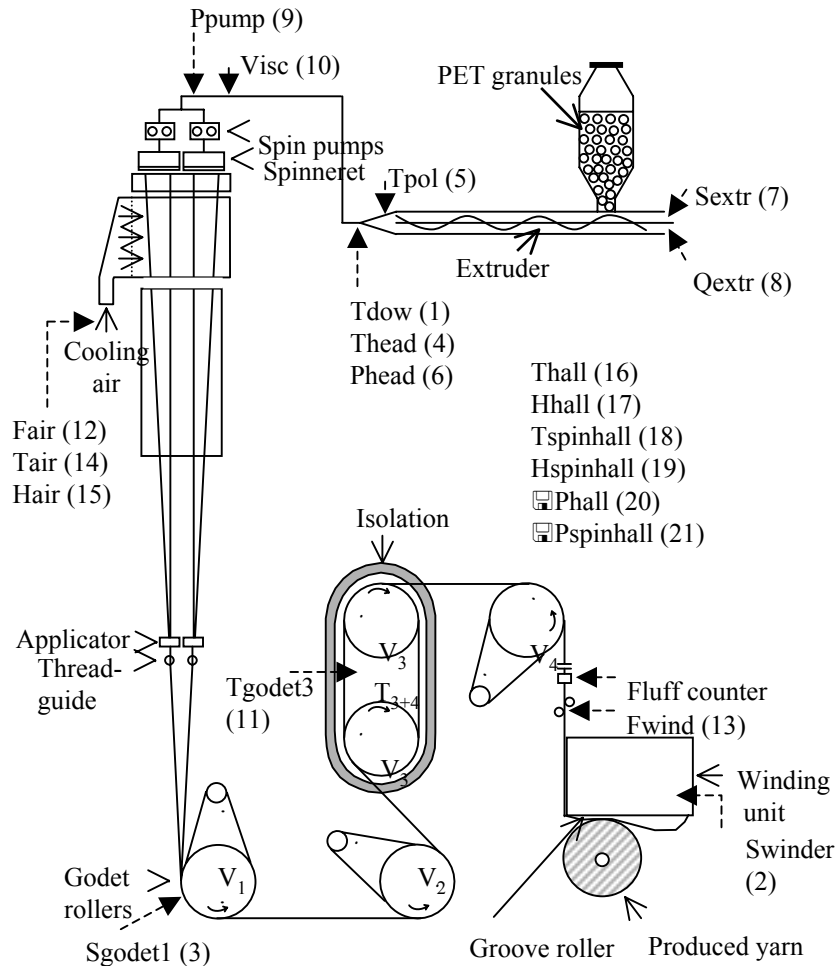


Regression: Example (1)

Predict octane number of gasoline from NIR spectrum:



Regression: Example (2)



Predict yarn quality:

Use process measurements to predict yarn breaking strength

Regression Problem

Predictors X :

- absorbances at different wavelengths
- temperature and other process variables

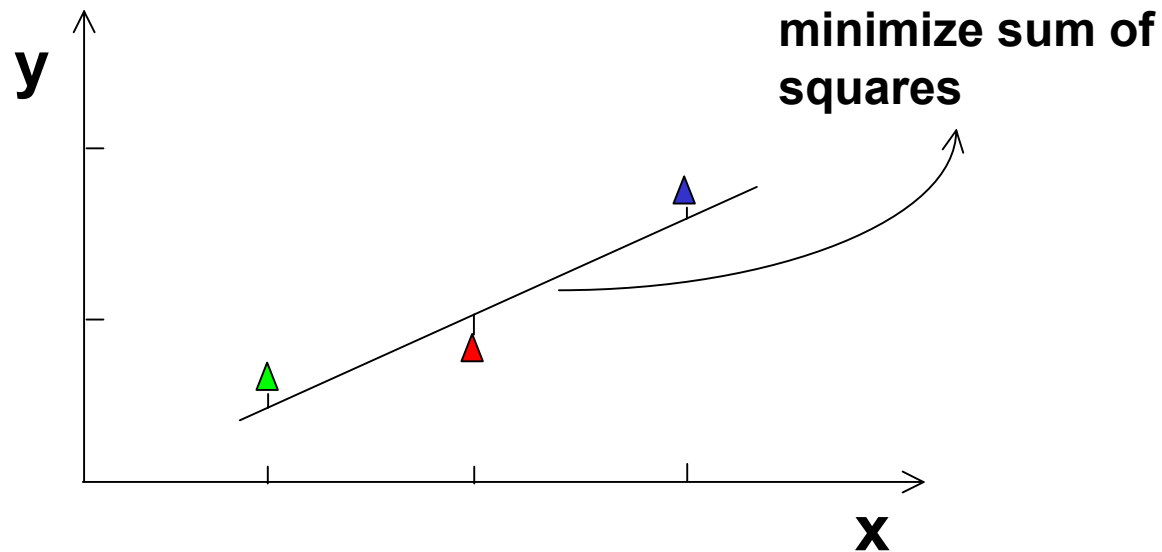
Responses (predictands) y :

- octane number
- yarn quality

Predictors X are used to predict response y

Univariate Regression: Idea (1)

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \begin{matrix} \blacktriangle \\ \blacktriangle \\ \blacktriangle \end{matrix} \longrightarrow \mathbf{y} = \begin{bmatrix} 0.7 \\ 0.9 \\ 1.4 \end{bmatrix} \begin{matrix} \blacktriangle \\ \blacktriangle \\ \blacktriangle \end{matrix}$$

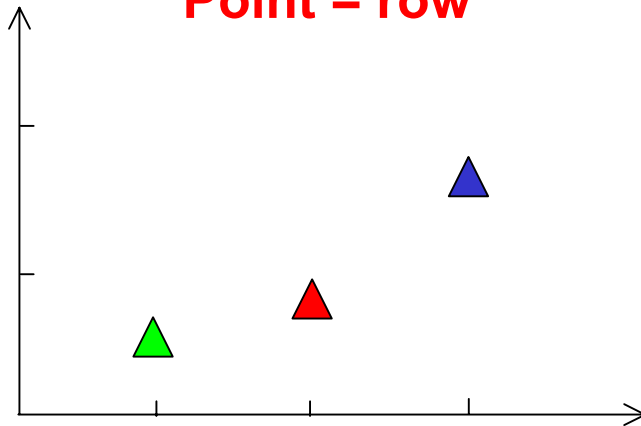


Intermezzo: plotting matrices

	●	●	
	1	0.7	▲
	2	0.9	▲
	3	1.4	▲

Row-space:

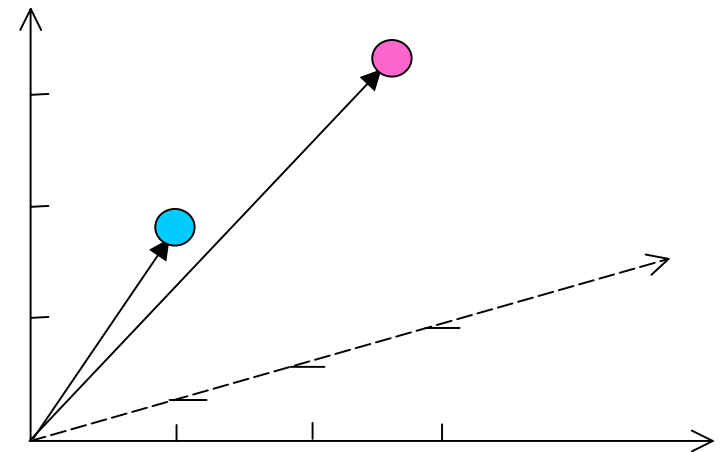
Point = row



Relationships between objects:
(dis)similarities

Column-space:

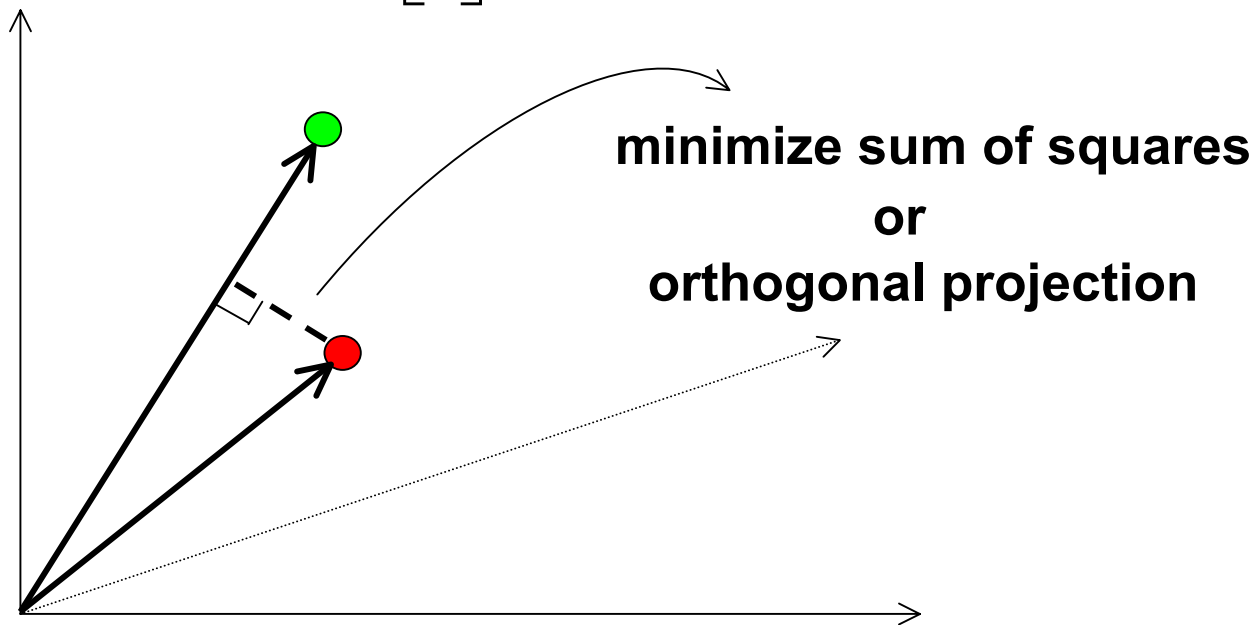
Point = column



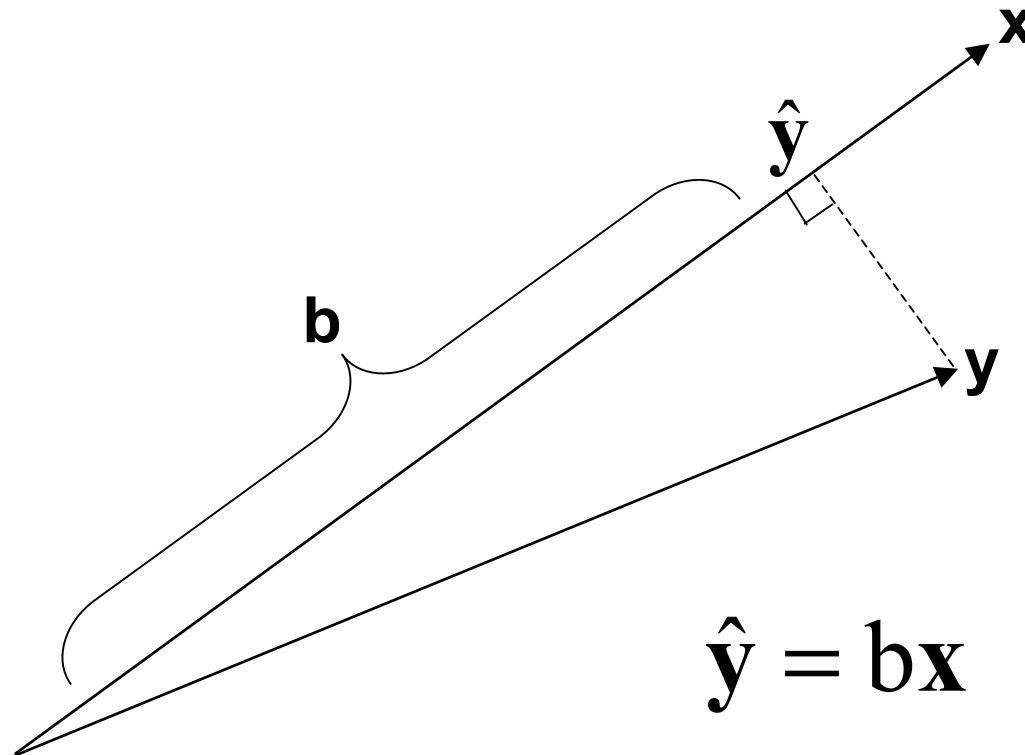
Relationships between variables:
correlation, variance

Univariate Regression: Idea (2)

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \longrightarrow \mathbf{y} = \begin{bmatrix} 0.7 \\ 0.9 \\ 1.4 \end{bmatrix}$$



Univariate Regression: Geometry



Univariate Regression: Model

Assumption:

Linear model between x and y

Sample or object i:

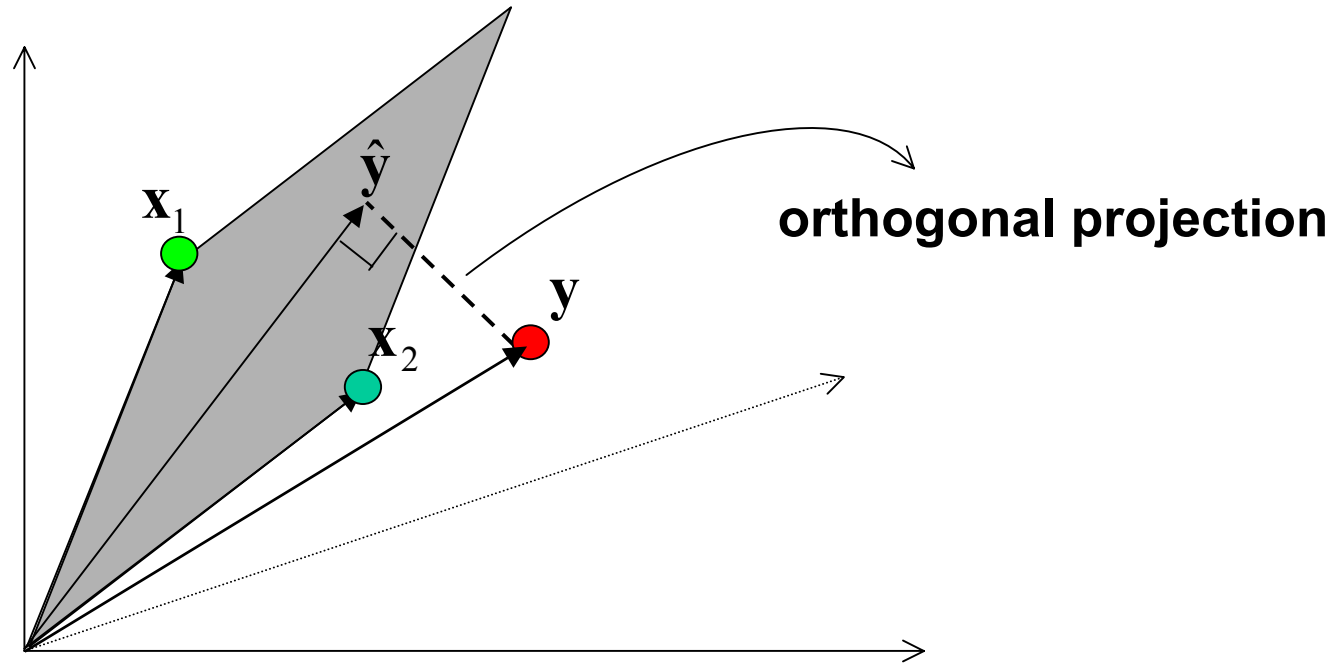
response y_i and predictor x_i

Model:

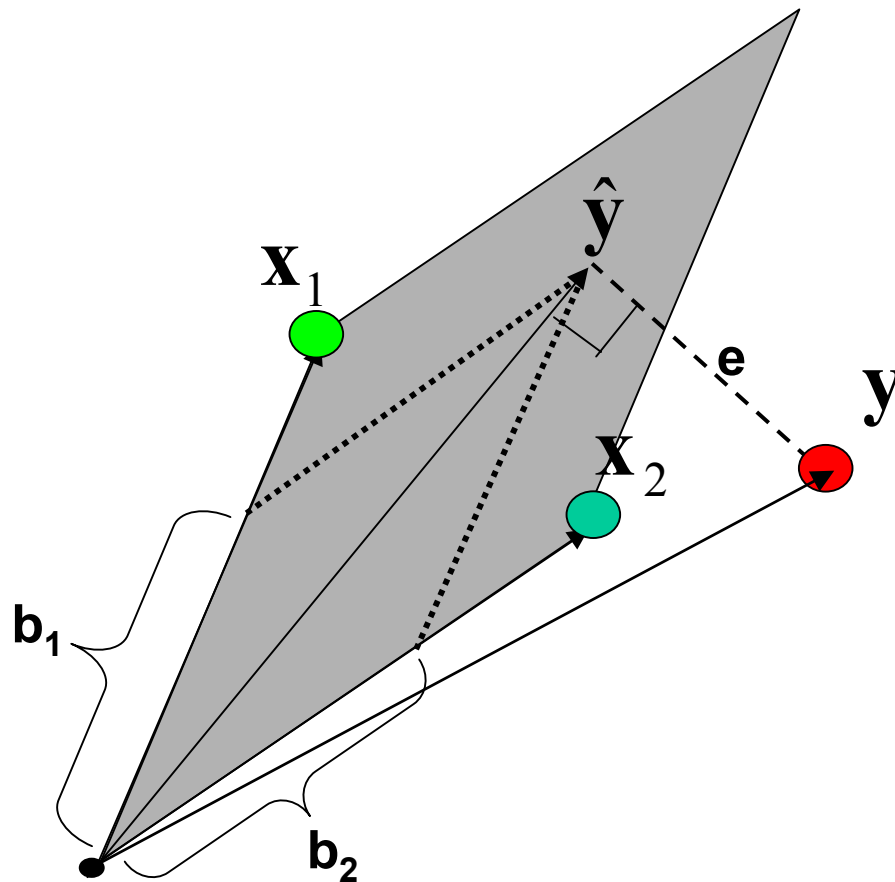
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Multiple Regression: Idea (1)

$$\mathbf{x}_1 = \begin{bmatrix} \bullet \\ \bullet \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} \bullet \\ \bullet \end{bmatrix} \longrightarrow \mathbf{y} = \begin{bmatrix} \bullet \\ \bullet \end{bmatrix}$$



Multiple Regression: Idea (2)



Model:

$$y = b_1 x_1 + b_2 x_2 + e$$

$$\hat{y} = b_1 x_1 + b_2 x_2$$

Multiple Regression: Model (1)

Assumption:

Linear model between X and y

Sample or object i :

response y_i and predictors $x_{i1}, \dots, x_{ij}, \dots, x_{iJ}$

Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_J x_{iJ} + \varepsilon_i$$

Multiple Regression: Model (2)

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_J x_{iJ} + \varepsilon_i$$

y_i : response of *i*th object

x_{ij} : *j*th predictor for *i*th object

β_0 : intercept or offset

β_j : *j*th regression coefficient

ε_i : error of *i*th object

Multiple Regression: Model (3)

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_J x_{iJ} + \varepsilon_i$$



$$y_1 = \beta_0 + \beta_1 x_{11} + \dots + \beta_J x_{1J} + \varepsilon_1$$

⋮

⋮

⋮

$$y_I = \beta_0 + \beta_1 x_{I1} + \dots + \beta_J x_{IJ} + \varepsilon_I$$



$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1 + \dots + \beta_J \mathbf{x}_J + \boldsymbol{\varepsilon}$$

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

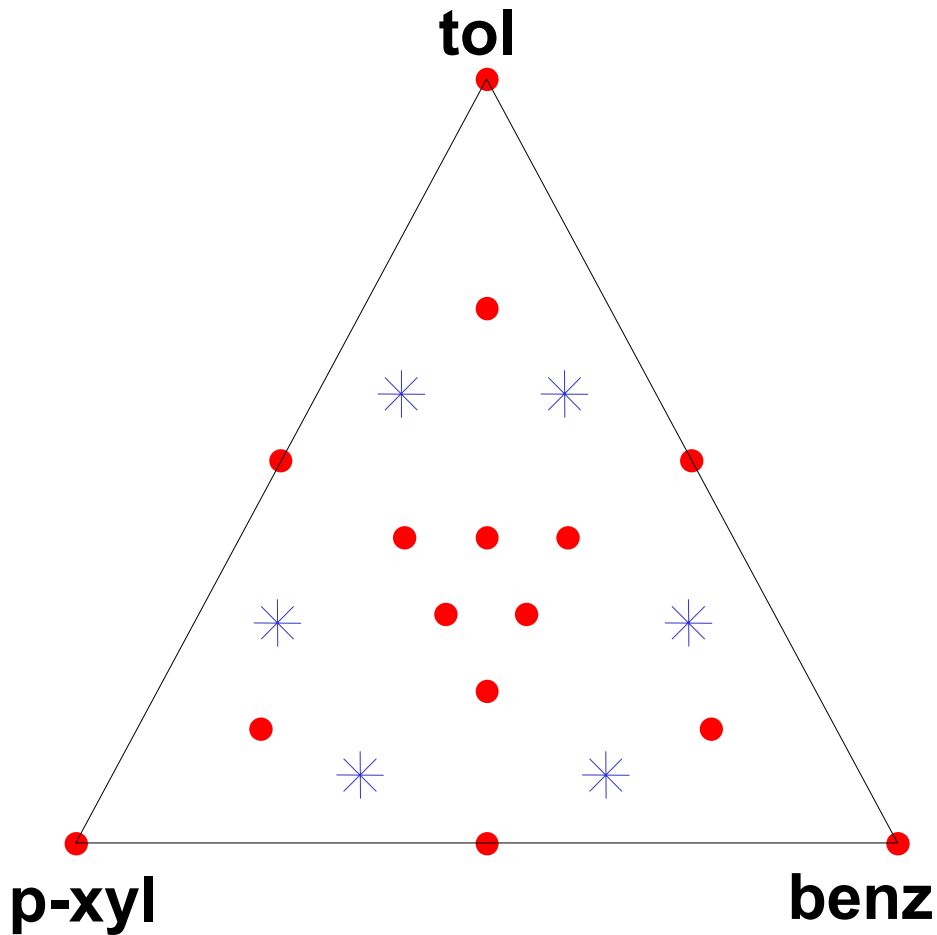
Multiple Regression: Model (4)

$$y = \beta_0 \mathbf{1} + \beta_1 \mathbf{x}_1 + \dots + \beta_J \mathbf{x}_J + \boldsymbol{\varepsilon}$$

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

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ y_I \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \cdot \\ \cdot \\ \beta_J \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \varepsilon_I \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdot & \cdot & \cdot & x_{J1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{I1} & \cdot & \cdot & \cdot & x_{IJ} \end{bmatrix}$$

Mult. Regression: Example (1)

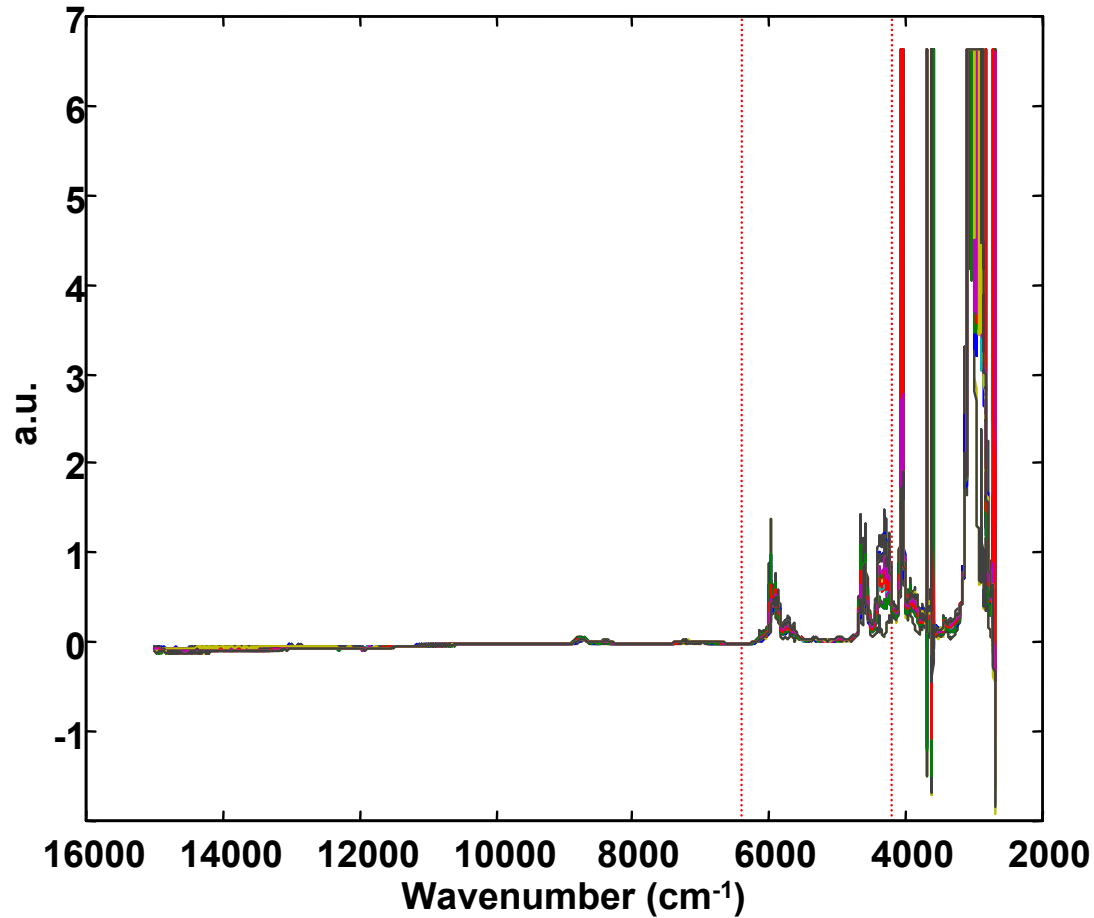


Predict toluene in presence of others using NIR

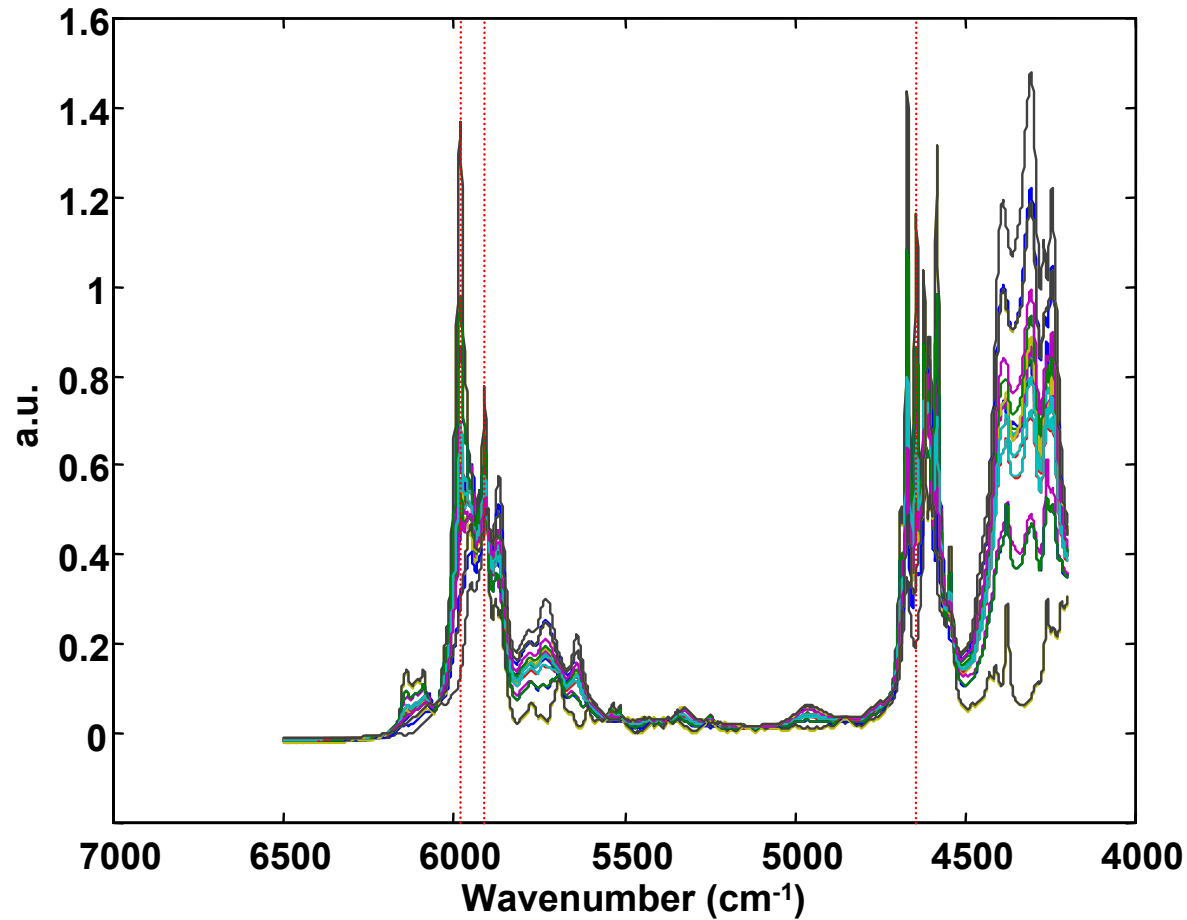
**Training set:
21 samples;
some in duplo**

**Test set:
6 samples**

Mult. Regression: Example (2)

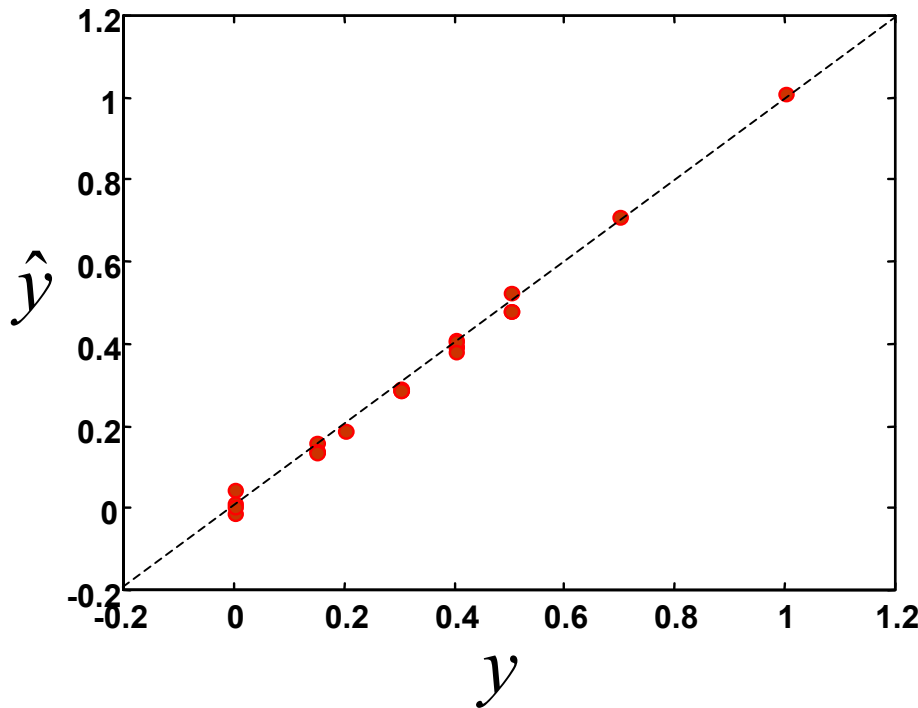


Mult. Regression: Example (3)



Mult. Regression: Example (4)

$$y_{tol} = 1.77 - 3.90x_{5984} - 5.20x_{5911} + 6.42x_{4643} + e$$



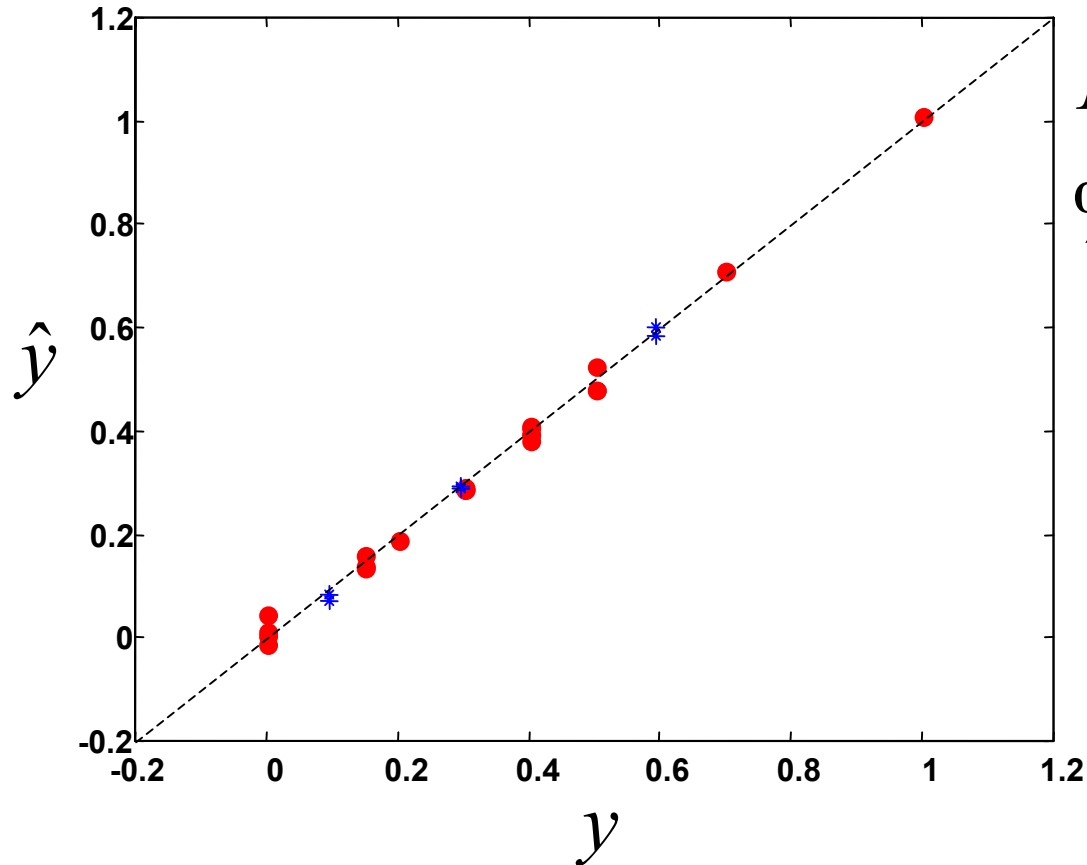
$$R_{MLR}^2 = 0.998$$

$$RMSEC = 0.0175$$

$$\%RMSEC = 5.43\%$$

Mult. Regression: Example (5)

Test set predictions:



$$RMSEP = 0.0157$$

$$\%RMSEP = 4.70\%$$

Mult. Regression: Example (6)

Predicting toluene in unknown samples:

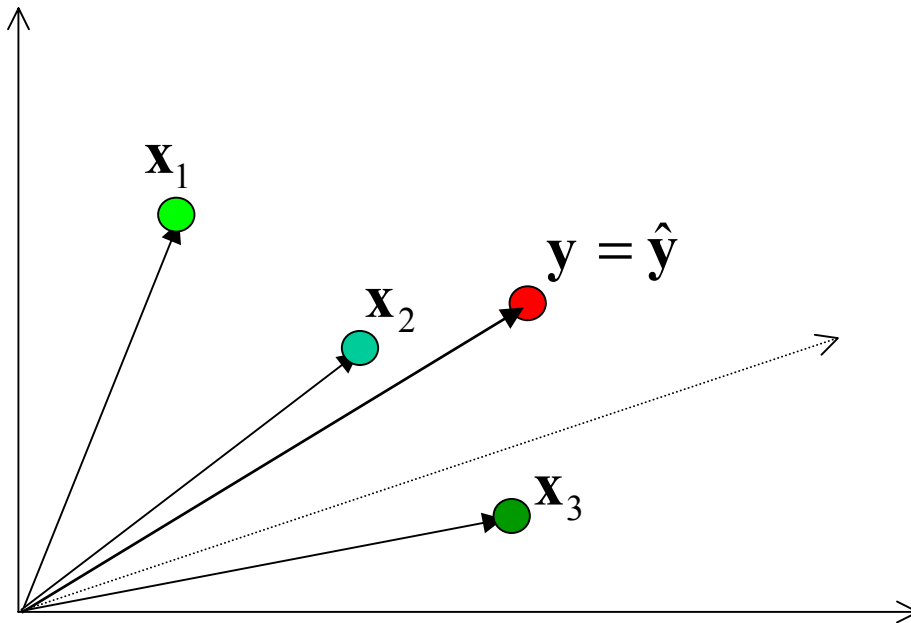
Predicted : **True:**

0.1379	0.1600	} OK!
0.6939	0.7100	
0.0352	0.0400	
0.6065	0.6300	
-0.0065	0.1000	→ Tert.butylbenz.
0.6390	0.4400	→ Meth. cyclohex.

No error message!!

Mult. Regression: Problems (1)

$$\mathbf{x}_1 = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix} \quad \longrightarrow \quad \mathbf{y} = \begin{bmatrix} \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

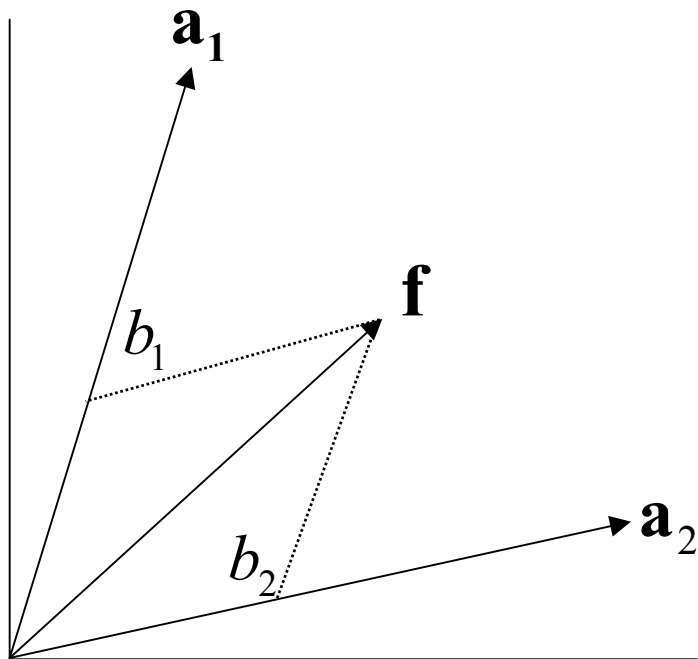


Perfect fit!

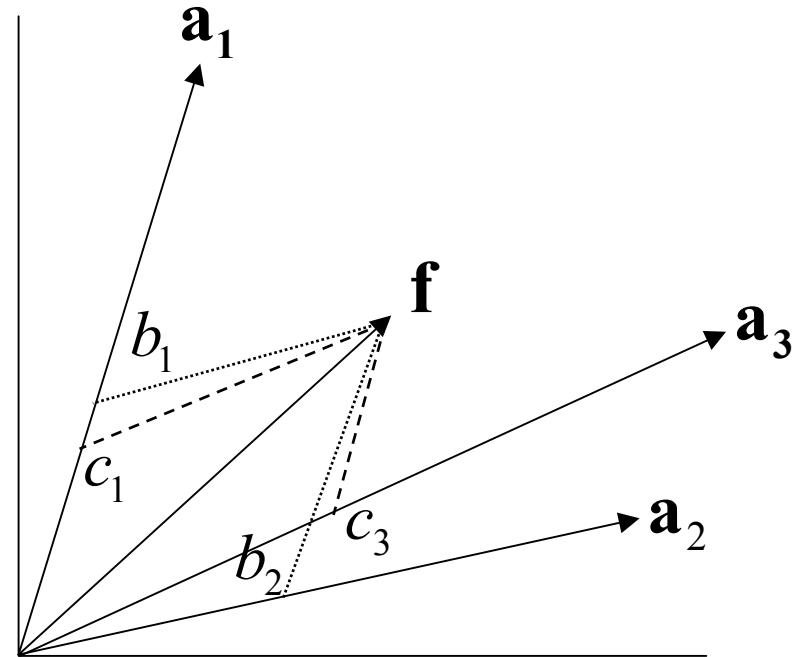
No residual

Realistic?

Intermezzo: bases

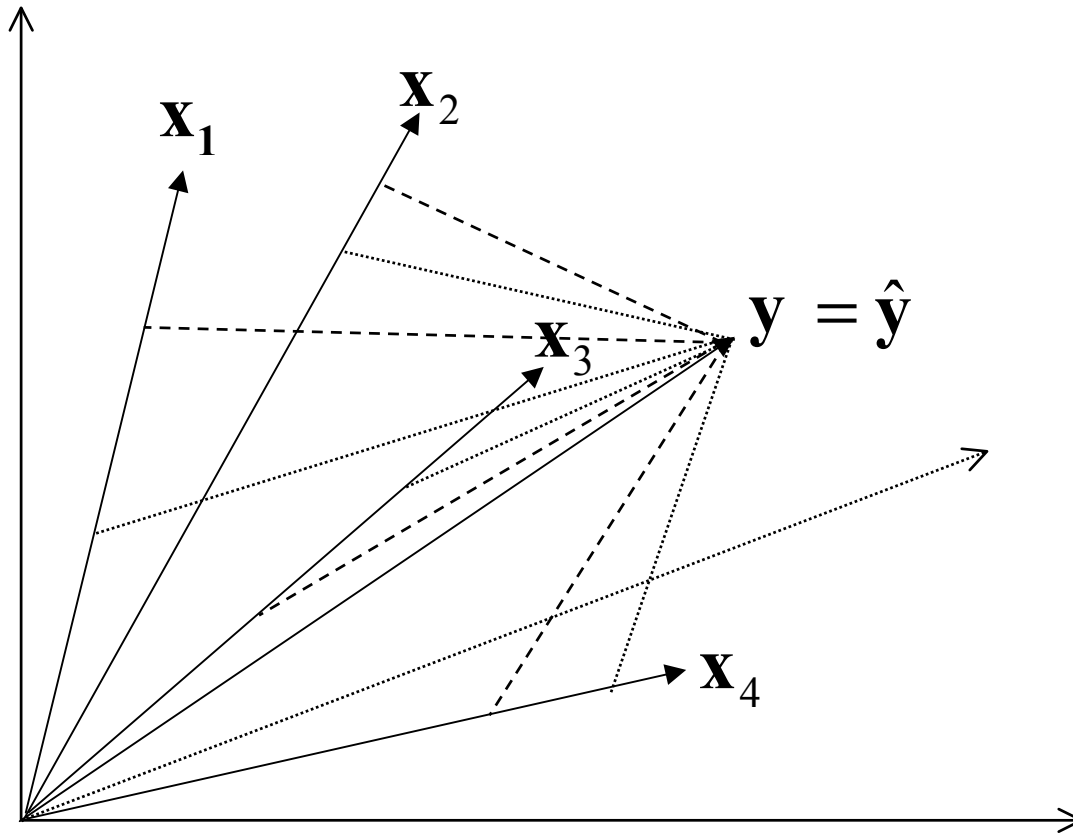


Unique representation



No unique representation

Mult. Regression: Problems (2)



Perfect fit

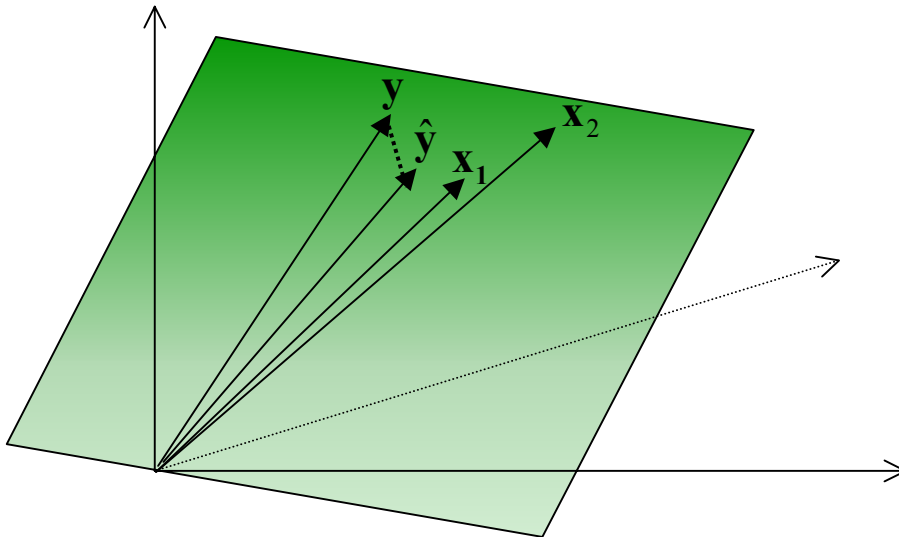
$$\text{-----} = \mathbf{b}$$
$$\text{- - - - -} = \tilde{\mathbf{b}}$$

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}\tilde{\mathbf{b}}$$

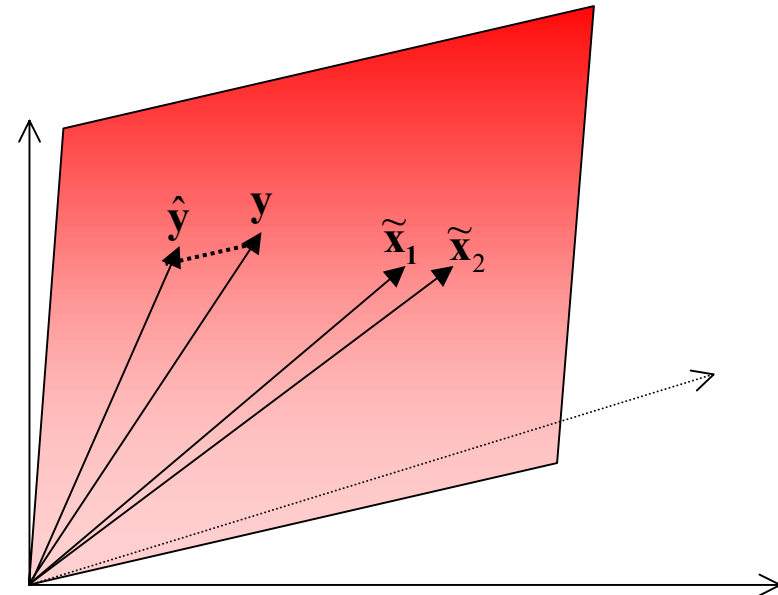
Regression coefficients not unique!

Mult. Regression: Problems (3)

First:



Second:



Very unstable predictions!!

Multicollinearity

Mult. Regression: Problems (4)

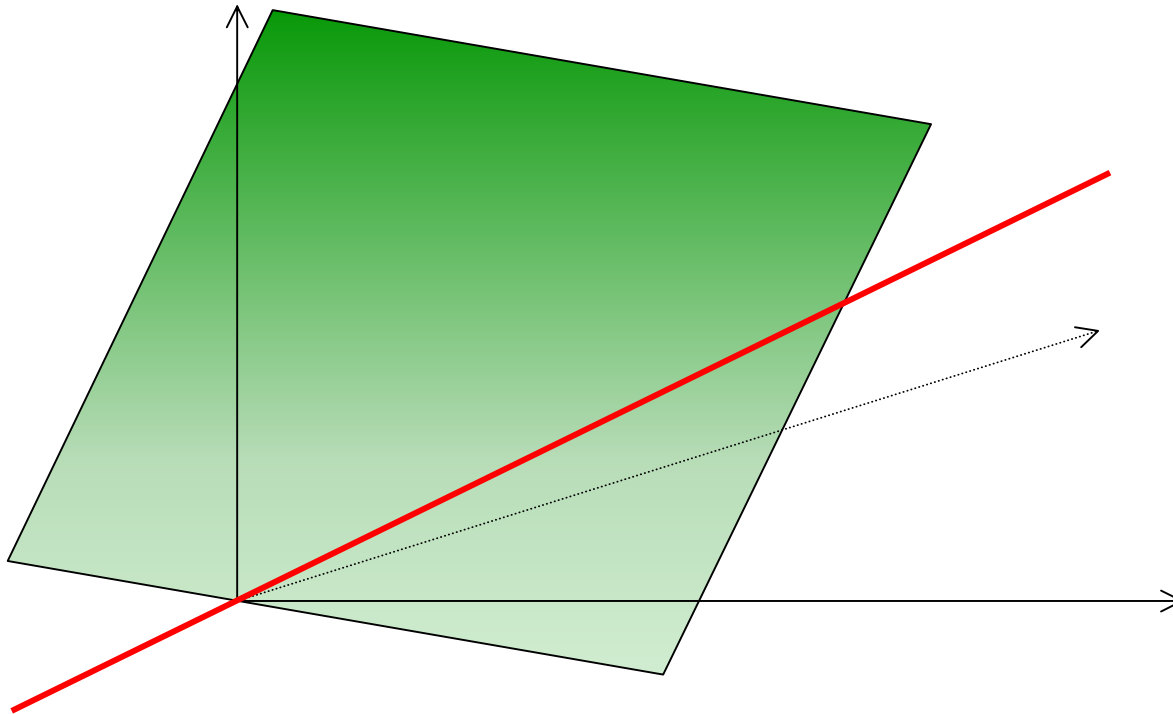
Large number of variables:

- **collinearity**
- **perfect fit**
- **nonunique regr.coeff.**

Nonrepresentative samples:

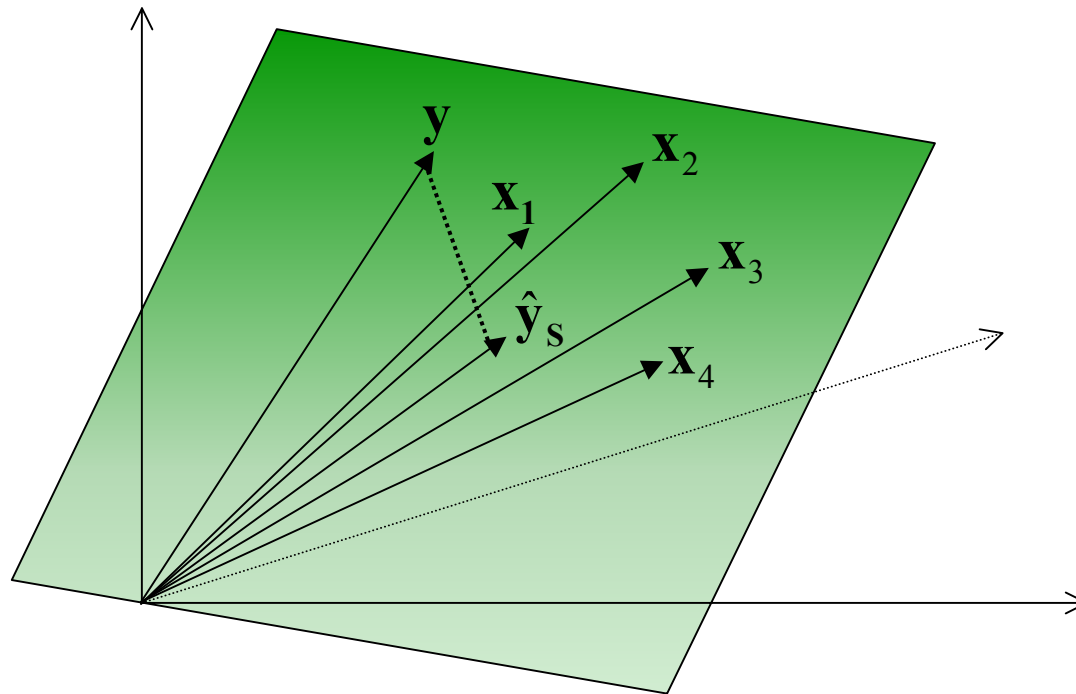
- **large errors**
- **no diagnostics**

Intermezzo: subspaces



- **Line through origin**
- **Plane through origin**

Subspace Regression: Idea (1)

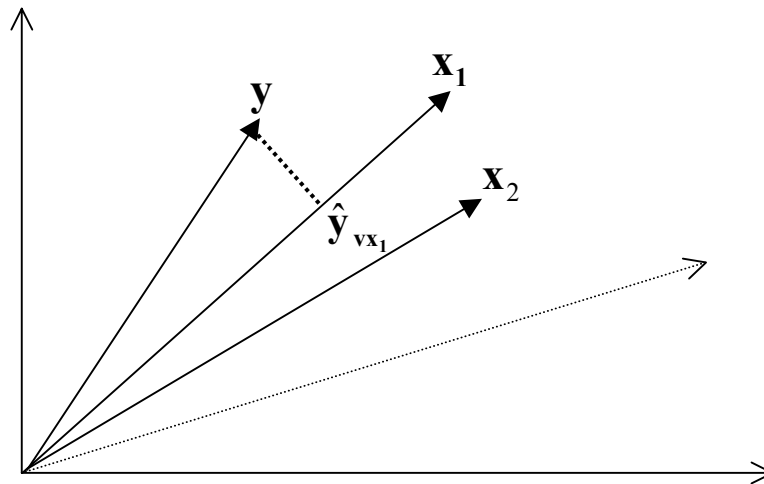


Define suitable subspace

Project y on that subspace

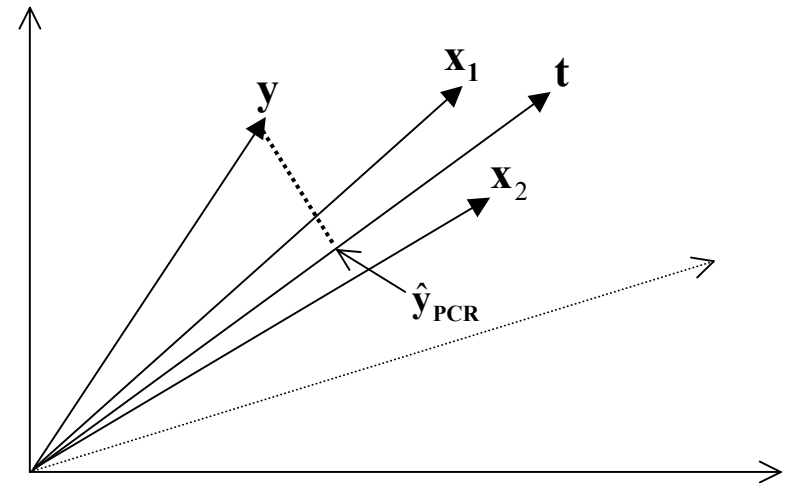
Subspace Regression: Idea (2)

Select x_1



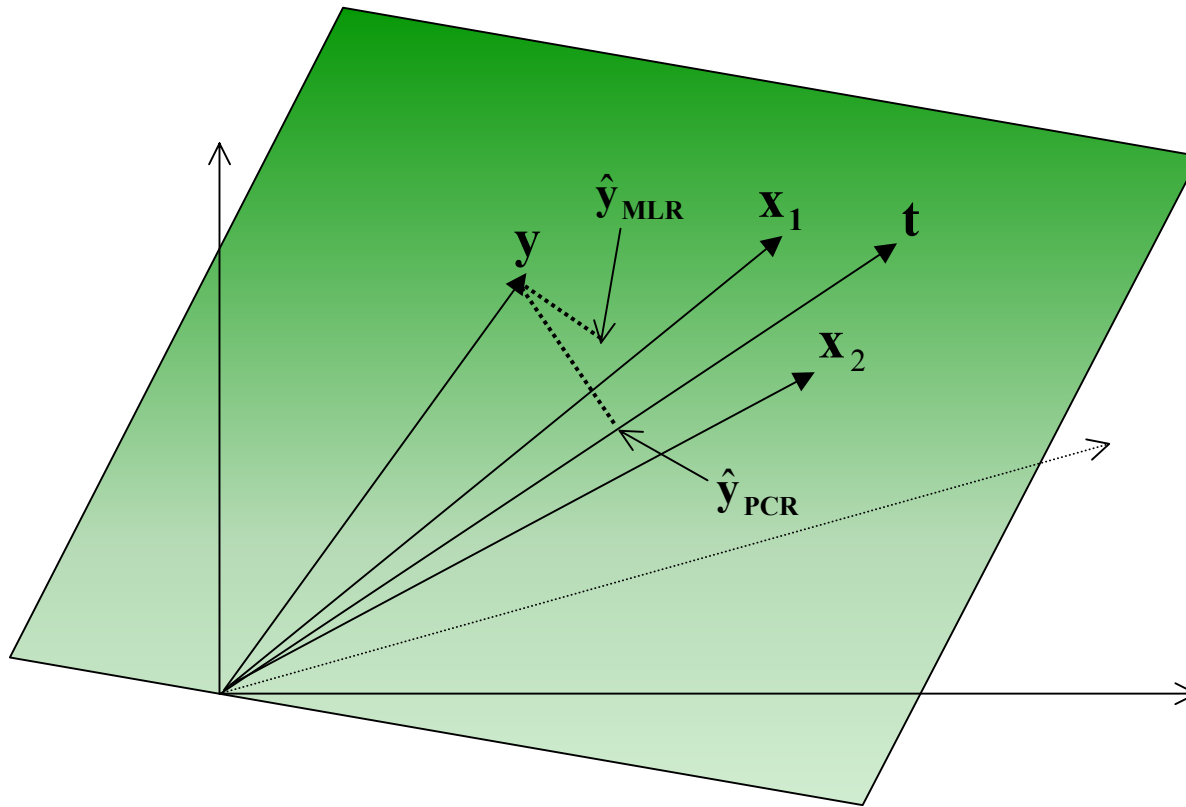
Variable Selection

Select PC1 (= t)



Principal Component Regression (PCR)

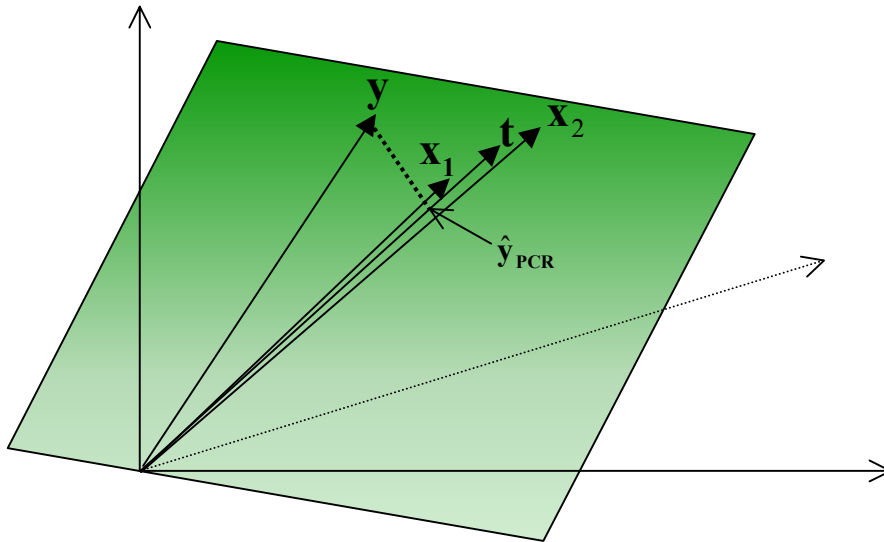
Subspace Regression: PCR (1)



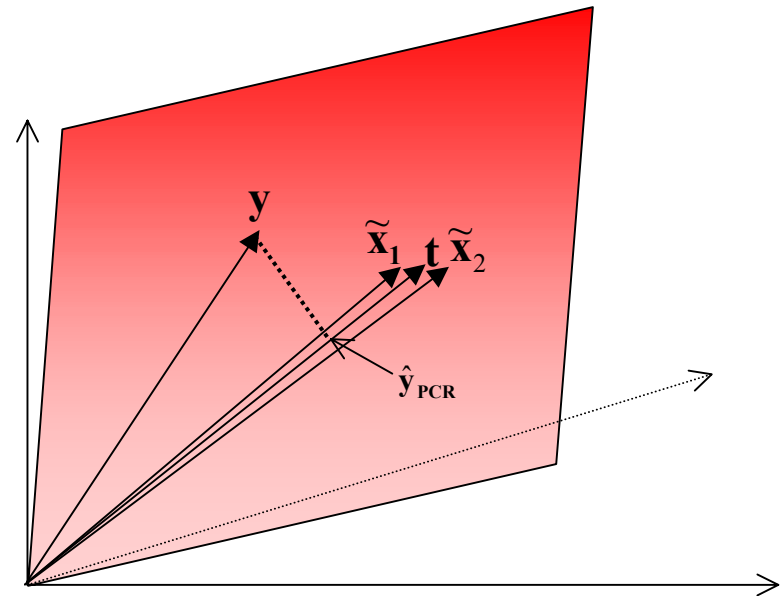
\hat{y}_{MLR} closer to y than \hat{y}_{PCR} \implies **MLR fits better than PCR**

Subspace Regression: PCR (2)

First:



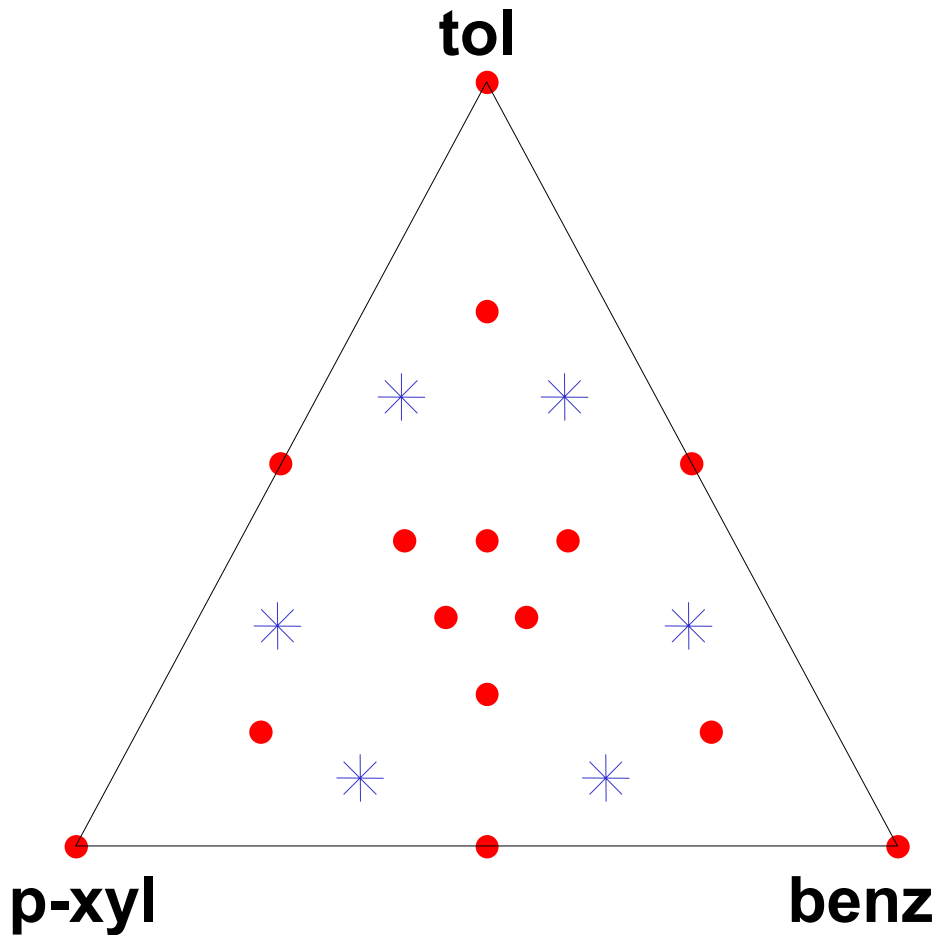
Second:



PCR gives stable predictions!!

t is more stable than x_1, x_2

PCR: Example (1)



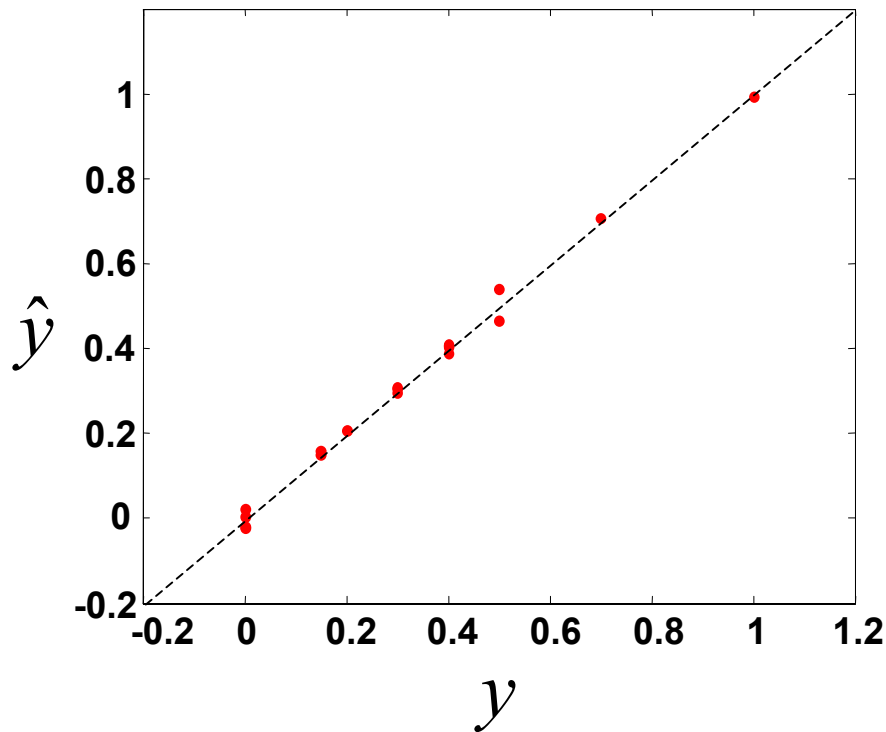
Predict toluene in presence of others using NIR

**Training set:
21 samples;
some in duplo**

**Test set:
6 samples**

PCR: Example (2)

$$y = \mathbf{T}b + e = \mathbf{X}\mathbf{P}b + e = \mathbf{X}b_{\text{PCR}} + e$$



$$R_{\text{PCR}}^2 = 0.998$$

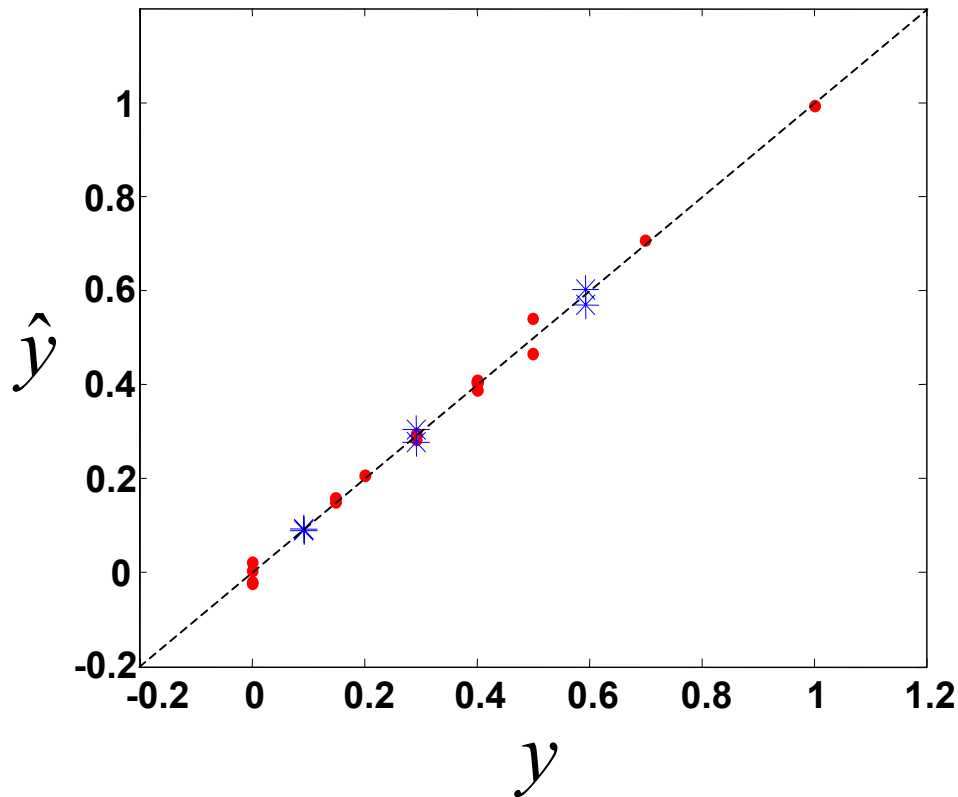
$$RMSEC = 0.0184$$

$$\%RMSEC = 5.72\%$$

2 PC's

PCR: Example (3)

Test set predictions:



$$RMSEP = 0.0129$$

$$\%RMSEP = 4.02\%$$

PCR: Example (4)

Predicting toluene in unknown samples:

Predicted : **True:**

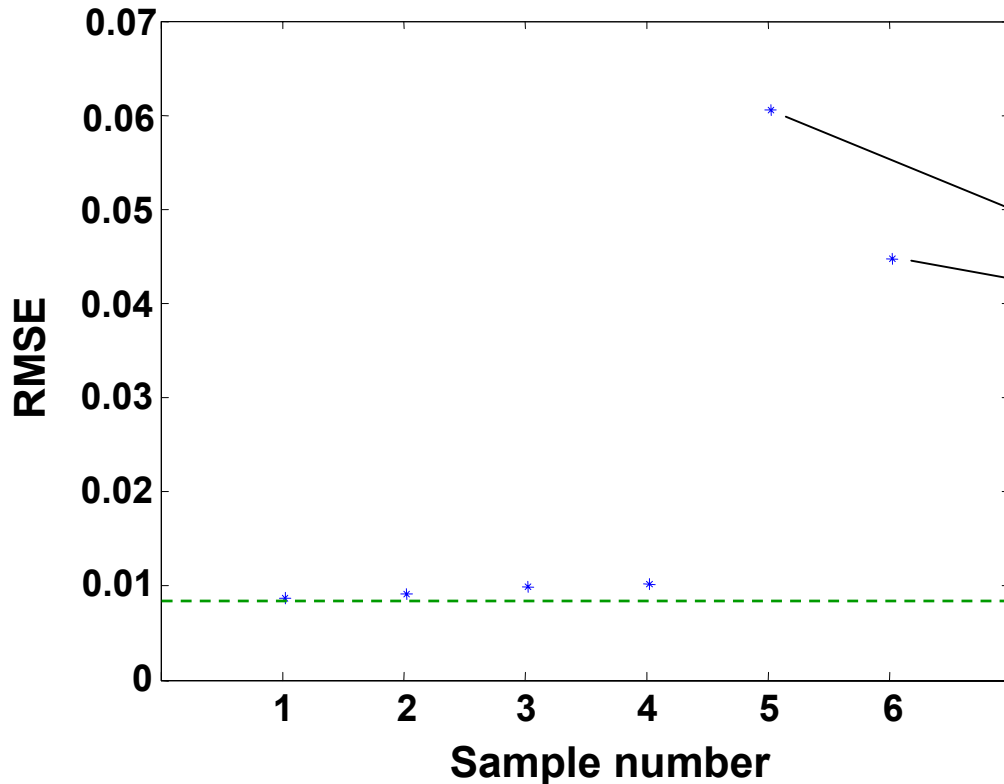
0.1597	0.1600	} OK!
0.6948	0.7100	
0.0254	0.0400	
0.5998	0.6300	
0.0463	0.1000	→ Tert.butylbenz.
0.4608	0.4400	→ Meth. cyclohex.

Error message??

PCR: Example (5)

$$X=TP'+E$$

RMSEX: 0.0028



Warning!!

3xRMSEX