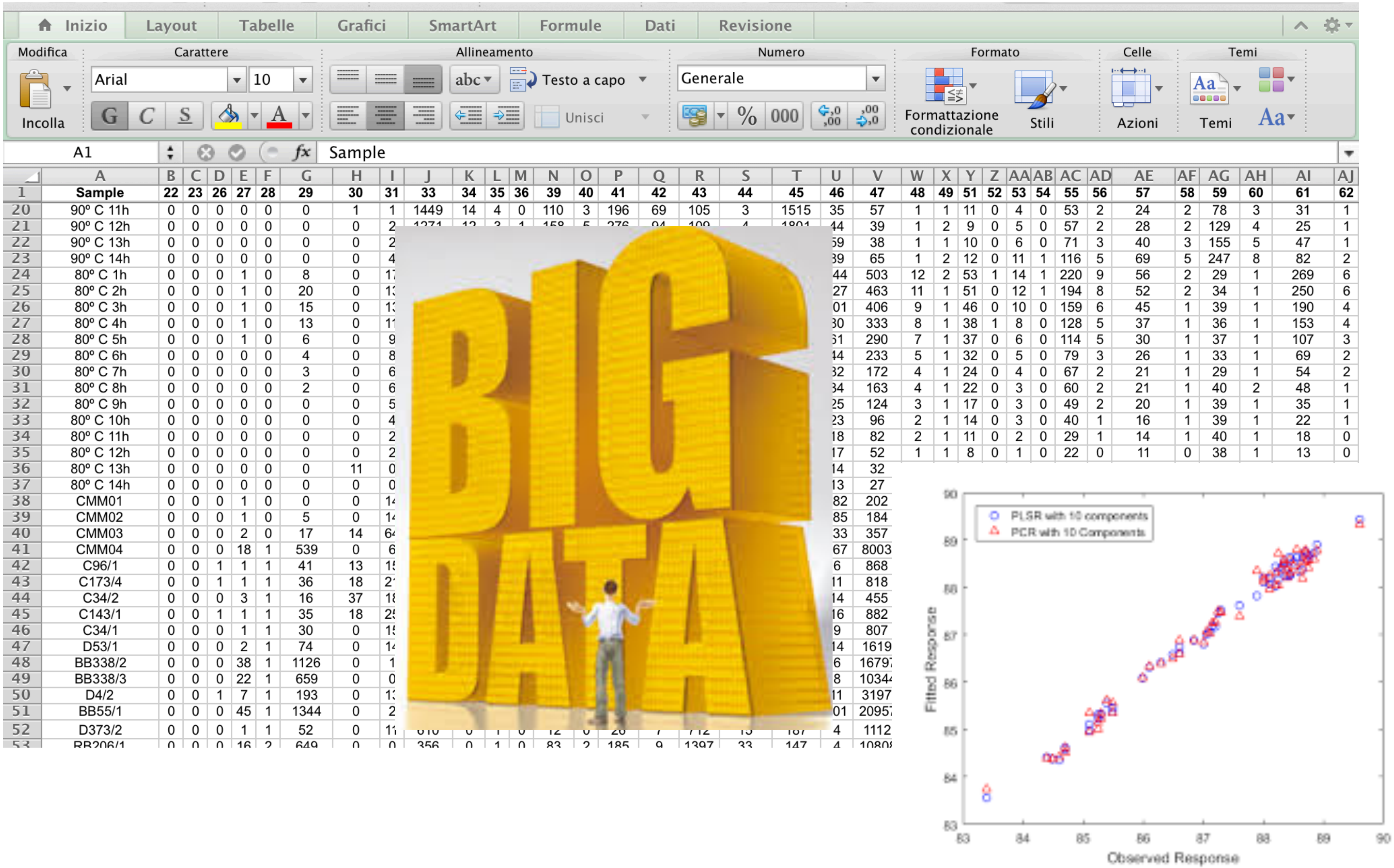
Multivariate regression: MR/PCR/PLS



A1 fx Sample

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1	Sample	22	23	26	27	28	29	30	31	33	34	35	36	39	40	41	42	43	44	45	46	47	48	49	51	52	53	54	55	56	57	58	59	60	61	62
20	90º C 11h	0	0	0	0	0	0	1	1	1449	14	4	0	110	3	196	69	105	3	1515	35	57	1	1	11	0	4	0	53	2	24	2	78	3	31	1
21	90º C 12h	0	0	0	0	0	0	0	2	1271	12	3	1	158	5	276	94	109	4	1801	44	39	1	2	9	0	5	0	57	2	28	2	129	4	25	1
22	90º C 13h	0	0	0	0	0	0	0	2												59	38	1	1	10	0	6	0	71	3	40	3	155	5	47	1
23	90º C 14h	0	0	0	0	0	0	0	4												39	65	1	2	12	0	11	1	116	5	69	5	247	8	82	2
24	80º C 1h	0	0	0	1	0	8	0	17												44	503	12	2	53	1	14	1	220	9	56	2	29	1	269	6
25	80º C 2h	0	0	0	1	0	20	0	13												27	463	11	1	51	0	12	1	194	8	52	2	34	1	250	6
26	80º C 3h	0	0	0	1	0	15	0	13												01	406	9	1	46	0	10	0	159	6	45	1	39	1	190	4
27	80º C 4h	0	0	0	1	0	13	0	11												30	333	8	1	38	1	8	0	128	5	37	1	36	1	153	4
28	80º C 5h	0	0	0	1	0	6	0	9												61	290	7	1	37	0	6	0	114	5	30	1	37	1	107	3
29	80º C 6h	0	0	0	0	0	4	0	8												44	233	5	1	32	0	5	0	79	3	26	1	33	1	69	2
30	80º C 7h	0	0	0	0	0	3	0	6												32	172	4	1	24	0	4	0	67	2	21	1	29	1	54	2
31	80º C 8h	0	0	0	0	0	2	0	6												34	163	4	1	22	0	3	0	60	2	21	1	40	2	48	1
32	80º C 9h	0	0	0	0	0	0	0	5												25	124	3	1	17	0	3	0	49	2	20	1	39	1	35	1
33	80º C 10h	0	0	0	0	0	0	0	4												23	96	2	1	14	0	3	0	40	1	16	1	39	1	22	1
34	80º C 11h	0	0	0	0	0	0	0	2												18	82	2	1	11	0	2	0	29	1	14	1	40	1	18	0
35	80º C 12h	0	0	0	0	0	0	0	2												17	52	1	1	8	0	1	0	22	0	11	0	38	1	13	0
36	80º C 13h	0	0	0	0	0	0	11	0												14	32														
37	80º C 14h	0	0	0	0	0	0	0	0												13	27														
38	CMM01	0	0	0	1	0	0	0	14												82	202														
39	CMM02	0	0	0	1	0	5	0	14												85	184														
40	CMM03	0	0	0	2	0	17	14	64												33	357														
41	CMM04	0	0	0	18	1	539	0	6												67	8003														
42	C96/1	0	0	1	1	1	41	13	15												6	868														
43	C173/4	0	0	1	1	1	36	18	21												11	818														
44	C34/2	0	0	0	3	1	16	37	18												14	455														
45	C143/1	0	0	1	1	1	35	18	25												16	882														
46	C34/1	0	0	0	1	1	30	0	15												9	807														
47	D53/1	0	0	0	2	1	74	0	14												14	1619														
48	BB338/2	0	0	0	38	1	1126	0	1												6	16797														
49	BB338/3	0	0	0	22	1	659	0	0												8	10344														
50	D4/2	0	0	1	7	1	193	0	13												11	3197														
51	BB55/1	0	0	0	45	1	1344	0	2												01	20957														
52	D373/2	0	0	0	1	1	52	0	11	610	0	1	0	12	0	26	7	712	15	187	4	1112														
53	BB206/1	0	0	0	16	2	649	0	0	356	0	1	0	83	2	185	9	1397	33	147	4	10808														

Let’s start from regression …

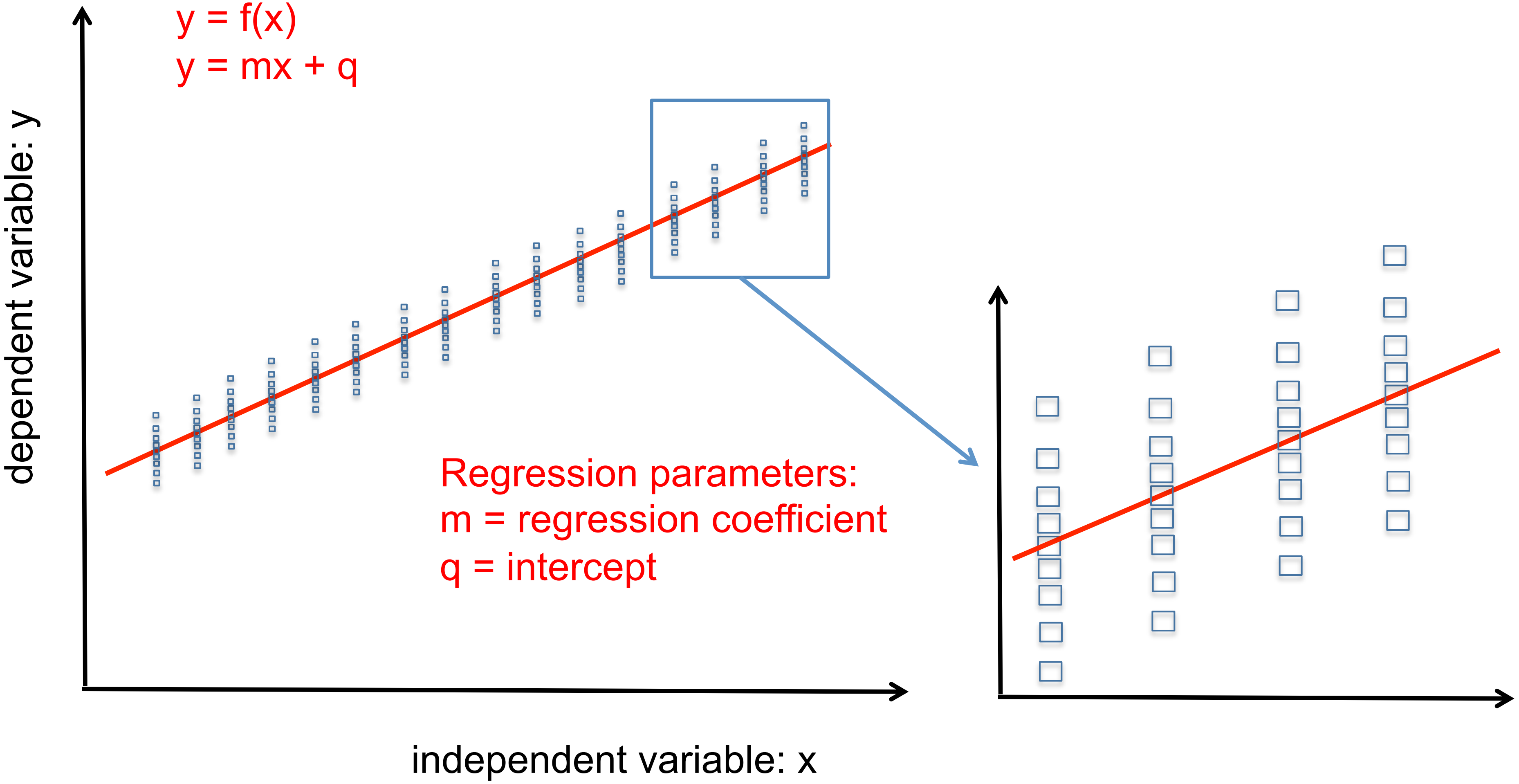
- Regression of a floating value towards a fixed value means that the floating value is progressively approaching the fixed value.

- In statistics ‘regression analysis’ is a statistical process for estimating the relationship among variables.

- In particular: a dependent variable and one or more independent variables (or 'predictors').

Simplest case:

linear regression between two variables



Linear relationship:
$y = f(x)$
$y = mx + q$

Regression parameters:
m = regression coefficient
q = intercept

Regression:

- Linear regression analysis helps to understand how the typical value of the dependent variable (or 'response variable') changes when that of the independent variable (or factor variable) is varied.

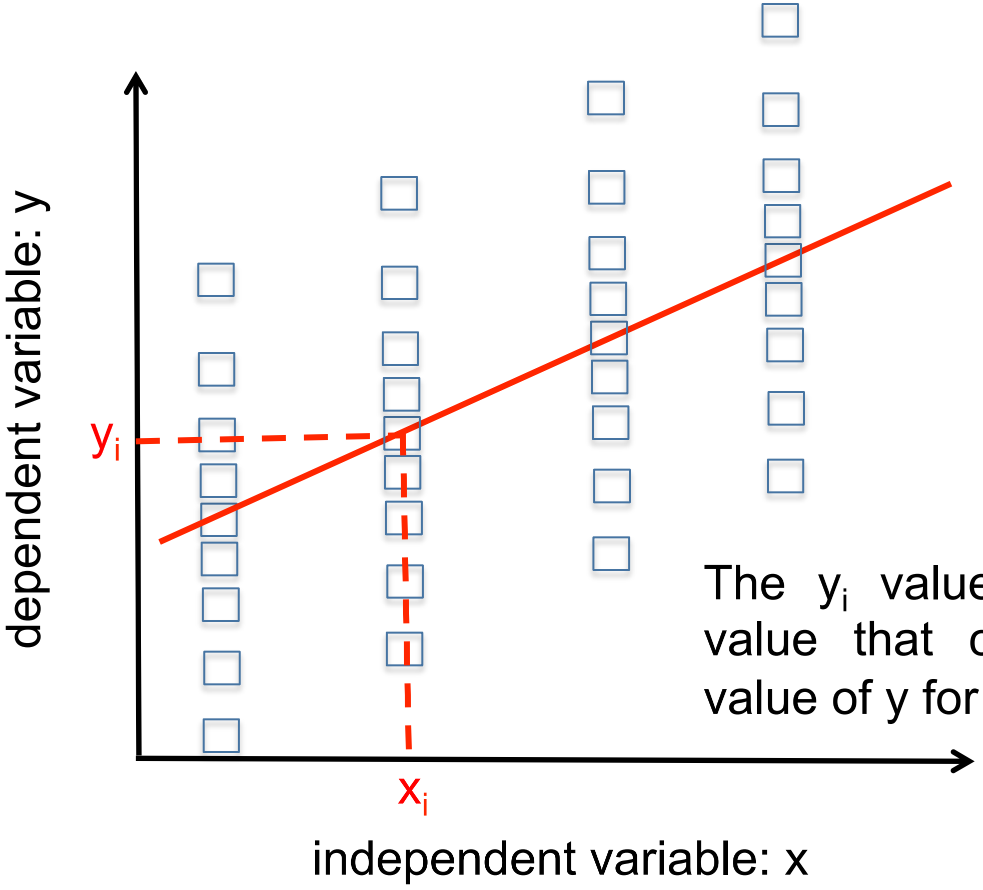
- Typical value = most common value (average value, or mean if the values are normally distributed).

- In practice, regression analysis estimates the average value of the dependent variable when the independent variables is fixed.

- At a given level of x defined as x_i, y values approach a value y_i calculated by the function y = f(x).

Understanding regression

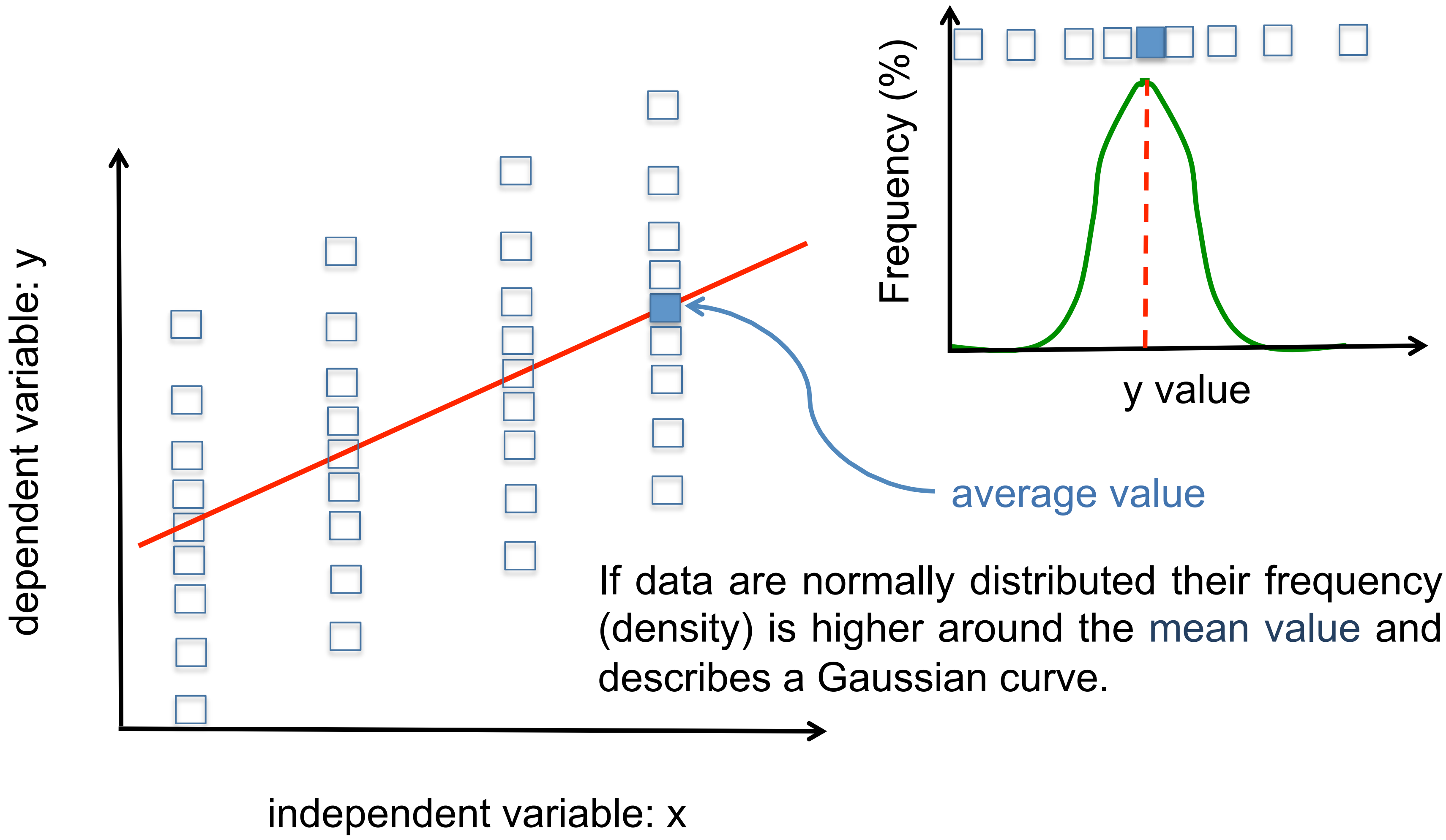
Approach a value does not mean assume a value.
This because any measured y variables has a variability.



The y_i value calculated by regression is a value that closely **approches** the average value of y for that x_i level.

Understanding regression

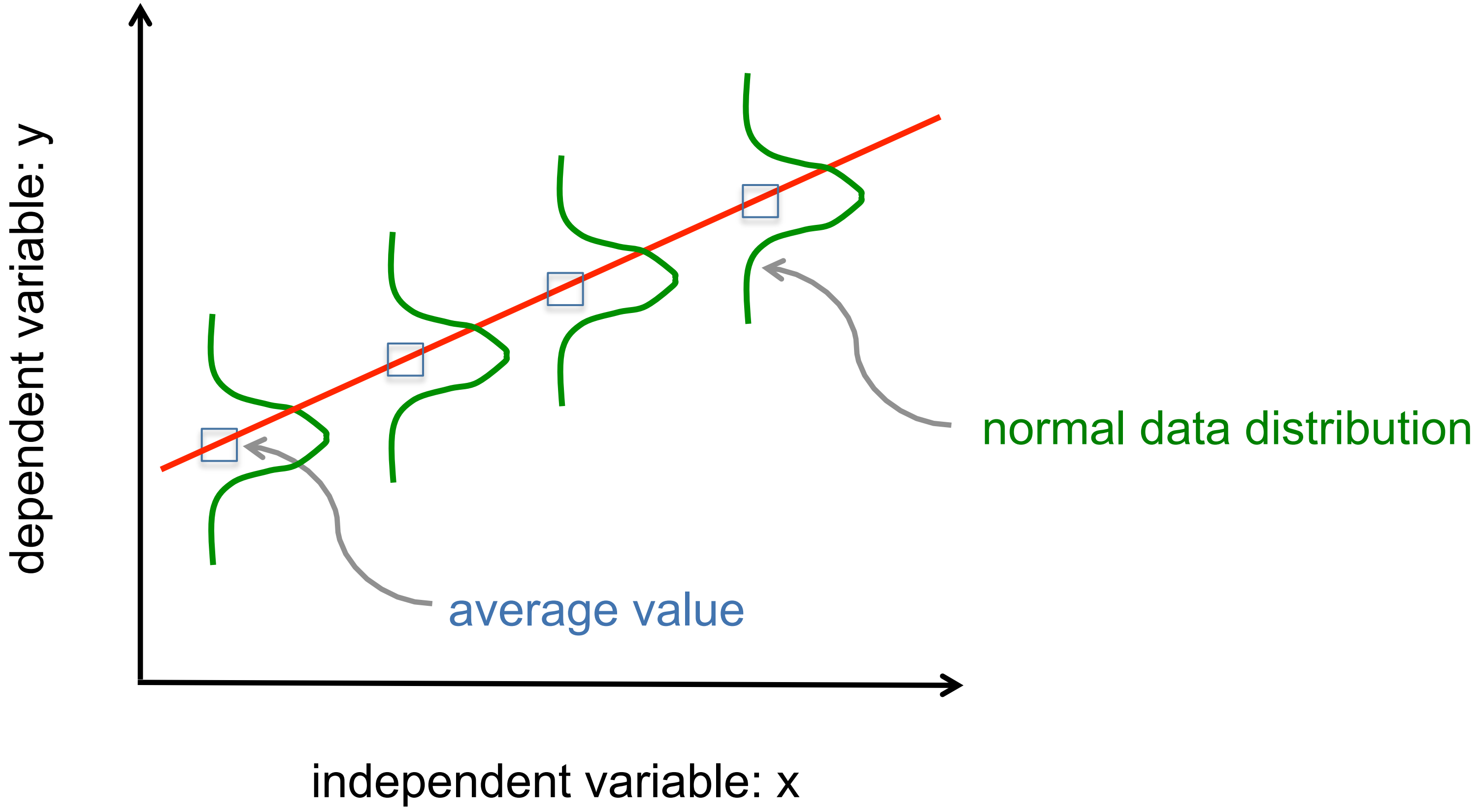
Ideal regression implies y data variability for each x value.



If data are normally distributed their frequency (density) is higher around the mean value and describes a Gaussian curve.

Understanding regression

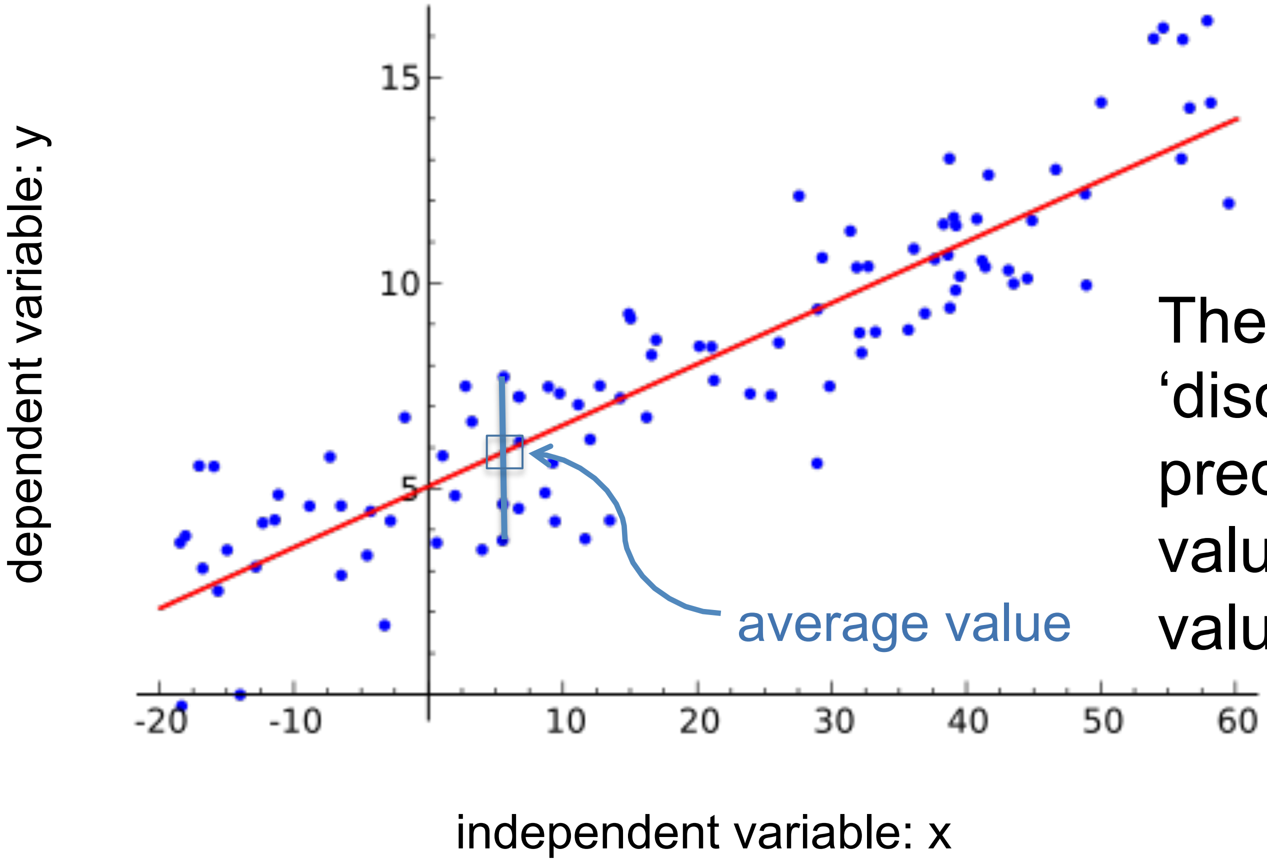
Since regression uses ‘mean values’ as average values, normal (Gaussian) data distribution should be assumed.



Understanding regression

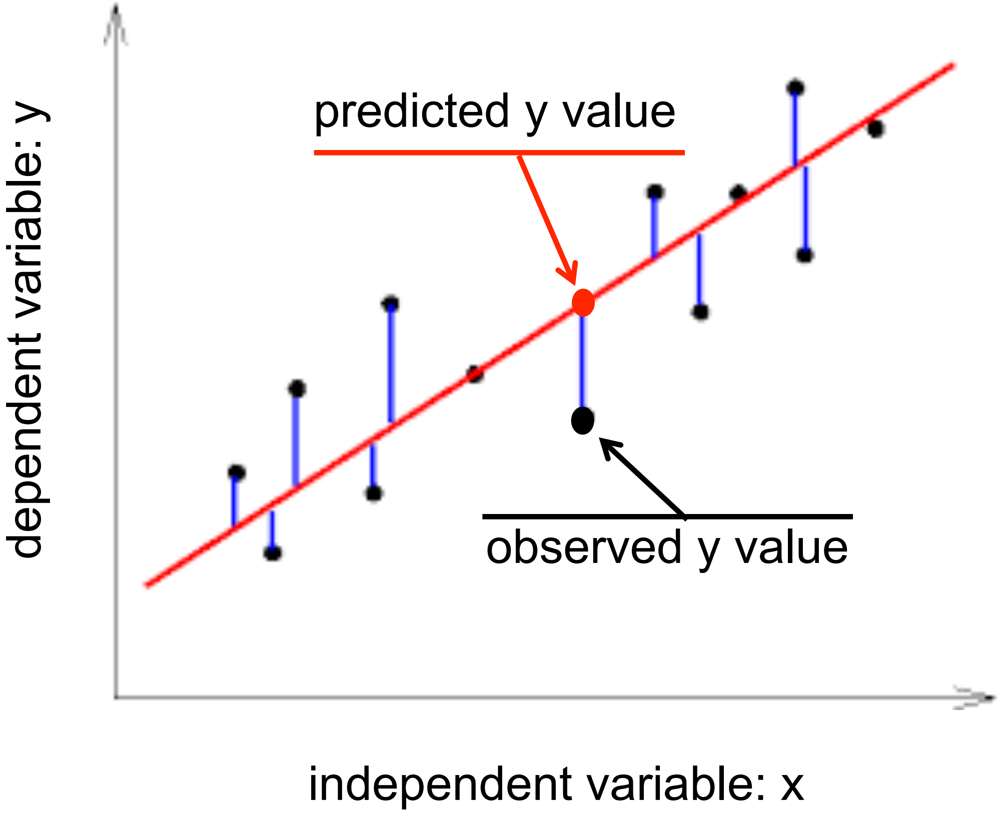
As a statistical tool,
regression has to deal with
data variability or uncertainty.

That's why,
regression has to deal with
'average values'.



There is always a
'discrepancy' between
predicted or expected
values and observed
values.

Understanding regression

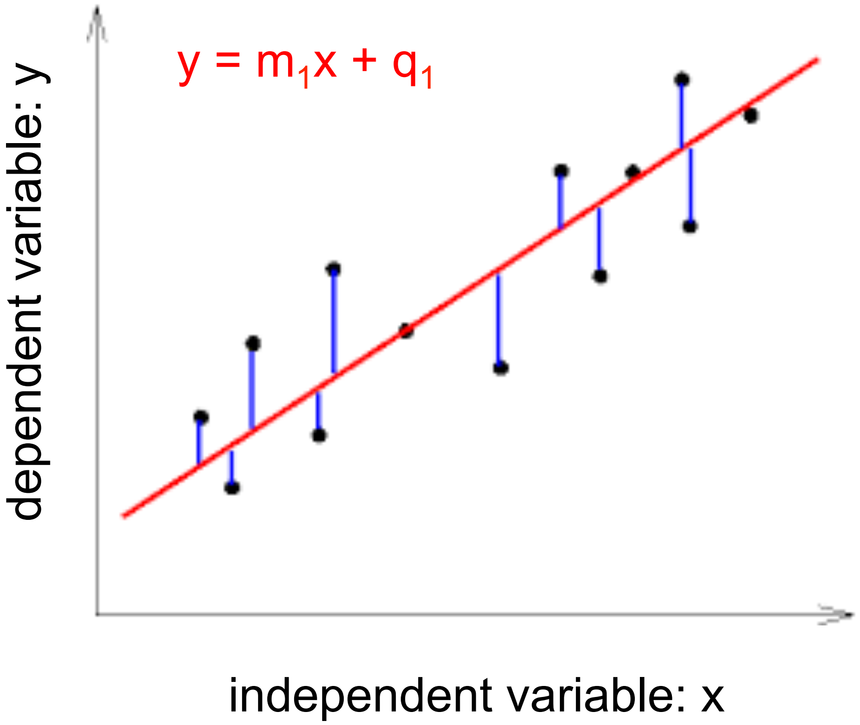


The discrepancies (or differences) between **predicted** and **observed** values are called **residuals**.

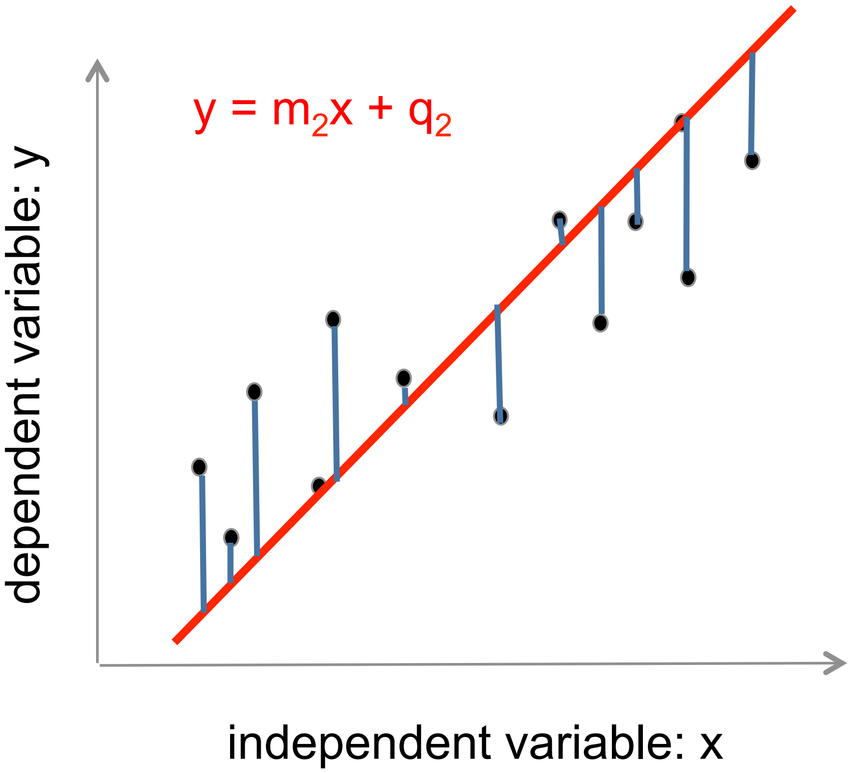
The higher the discrepancy, the lowest is the goodness of the regression.

Understanding regression

Residuals allows to judge the goodness of regression.
As an example: let’s start from the same experimental points and try to draw two different regression lines which depicts data trend at a glance.



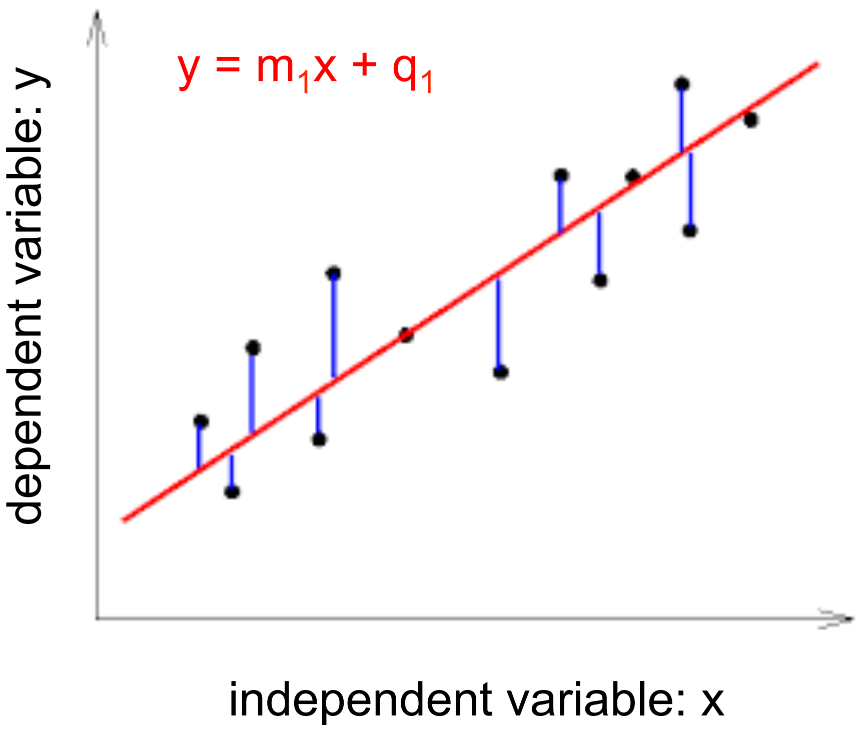
GOOD



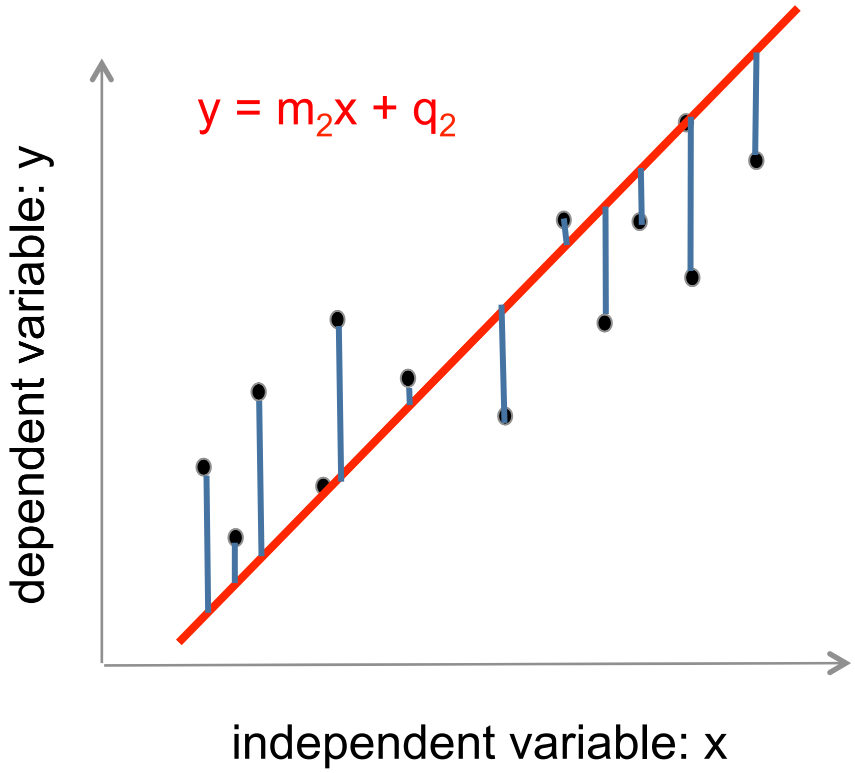
NOT GOOD

Understanding regression

At a first glance, the differences between expected and predicted values (residuals) are higher in the graph on the right side.



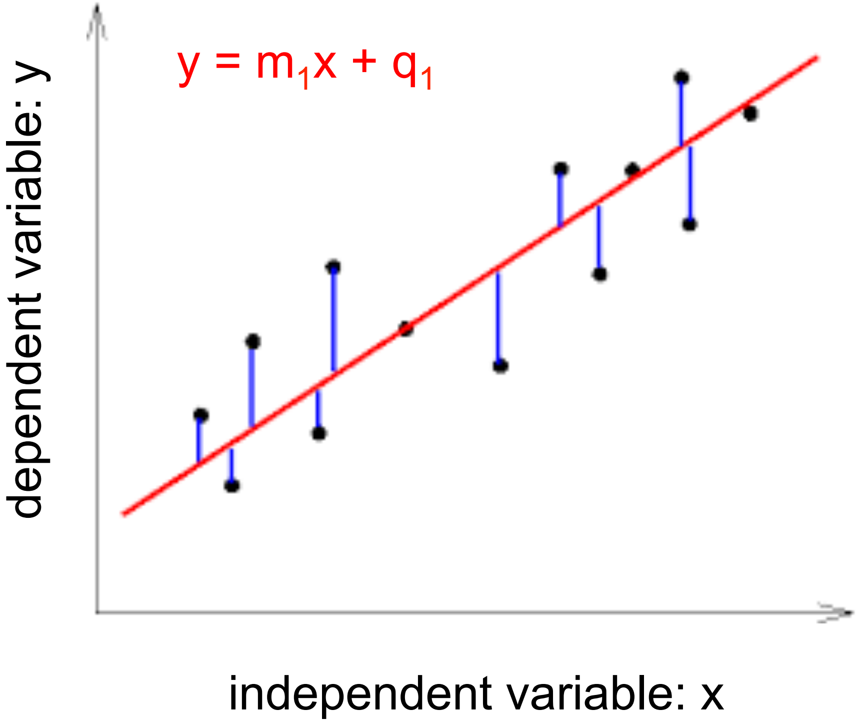
GOOD



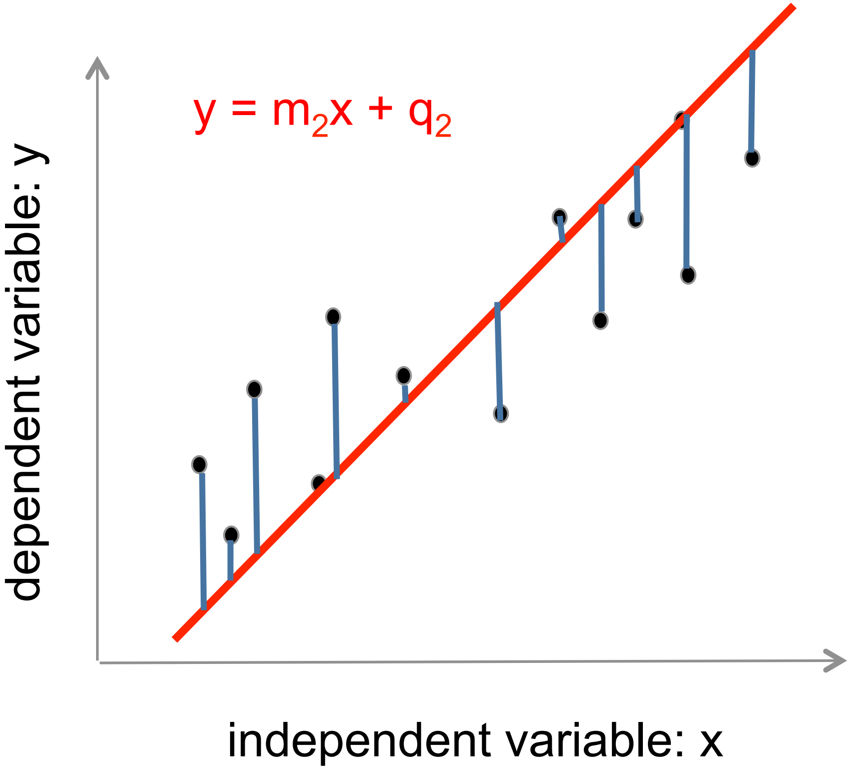
NOT GOOD

Understanding regression

In both graphs residuals could show negative values (below the regression line) or positive values (above the regression line). Negative values mean that predicted values are higher than observed, positive values mean that observed values are higher than predicted.



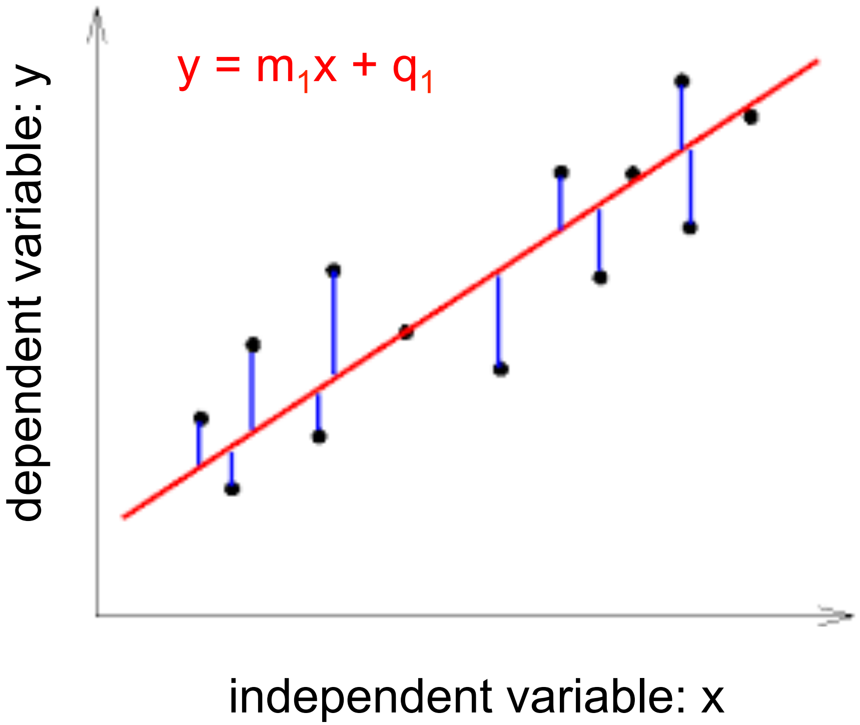
GOOD



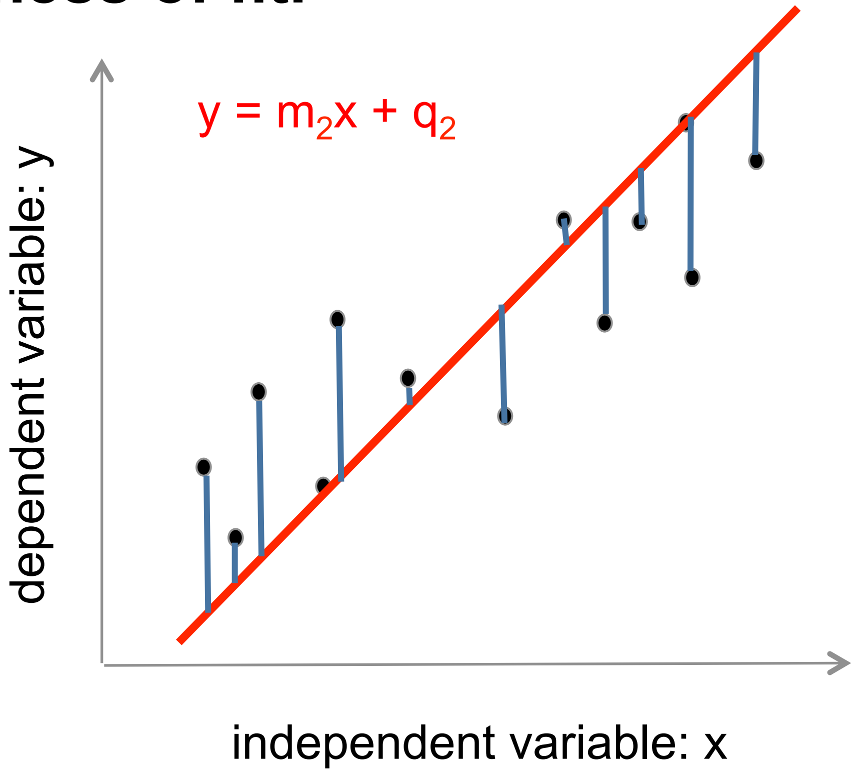
NOT GOOD

Understanding regression

The sum of residuals in both cases is equal to zero since the positive and the negative deviations (or errors) from the regression lines are equal. **The sum of residuals is not a good indicator of goodness of fit.**



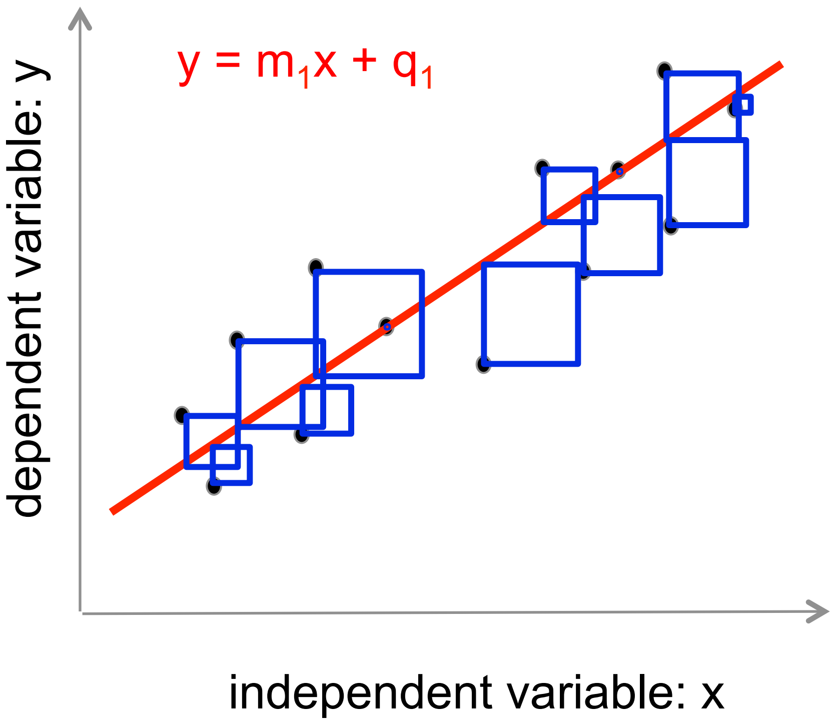
GOOD



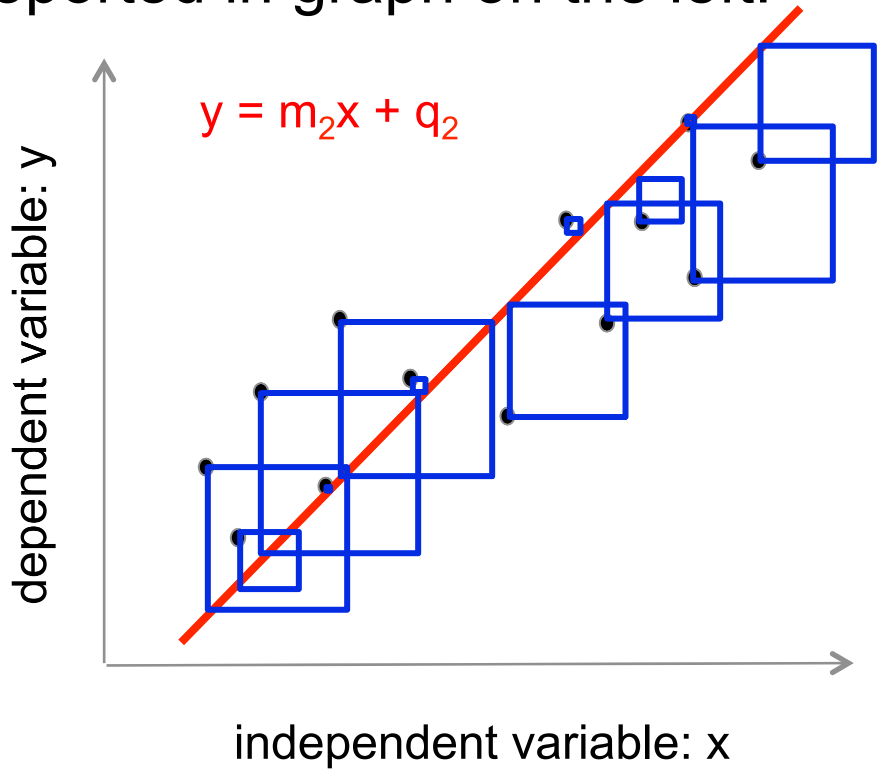
NOT GOOD

Understanding regression

The sum of the squares of residuals (RSS), which is always higher than zero (since squares are always positive), is much higher in the case reported in the graph on the right than in the case reported in graph on the left.



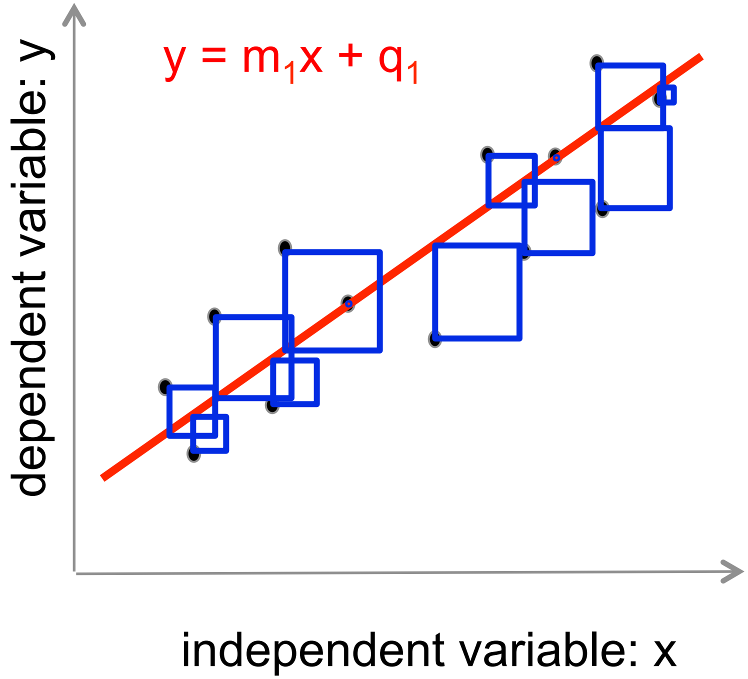
GOOD



NOT GOOD

Understanding regression

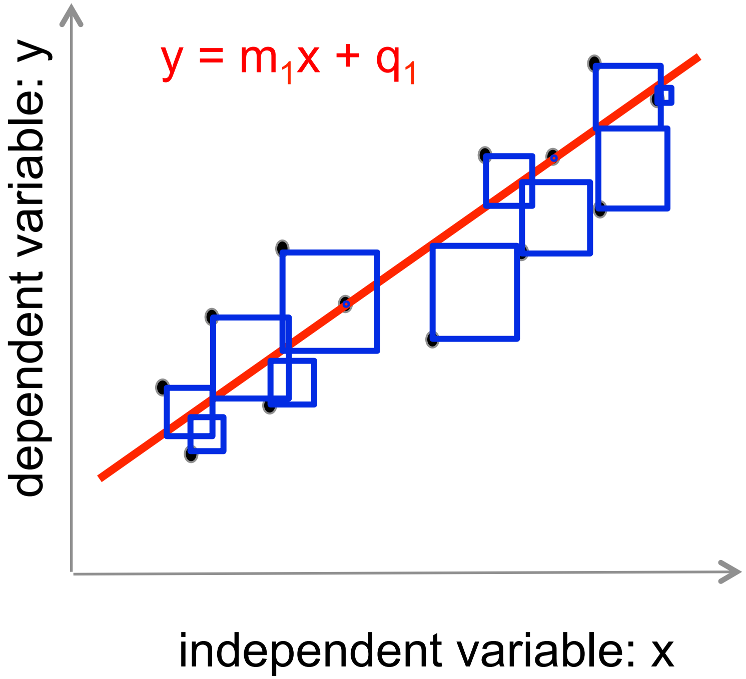
The sum of the squares of residuals (RSS) is generally preferred to the sum of absolute values of residuals, which is also always higher than zero (since absolute values are always positive), because squaring stresses the differences among values.



Linear regression is performed by computer programs that modulate the values of linear regression parameters m and q until a combination m_1, q_1, which minimizes the RSS, is found.

Understanding regression

Since regression minimizes the sum of the squares of residuals (RSS) is also called least square regression.



Statistical computer programs uses different **algorithms** in order to minimize the **RSS**.
An algorithm is a step-by-step set of operations to be performed consecutively.
The most common algorithm is the Marquadt.

Goodness of fit

The residuals sum of squares (RSS) could be used to evaluate the goodness of fit but, generally the **coefficient of determination** or **R^2** is preferred to this purpose.

$$R^2 = 1 - \boxed{\frac{RSS}{TSS}} \longrightarrow \text{Proportion of variance intrinsic to y variable}$$

The total sum of squares (TSS) is the sum of the squares of y values.

R^2 is a number that indicates the proportion of the variance in the dependent variable y that is predictable from the independent variable x.

Variance and covariance

- R^2 account only for y variability or variance.
- In the case of a relationship between x and y both the variables vary.
- In statistics, **covariance** is a measure of the joint variability of the two (x and y) variables.

If the greater values of one variable (x) mainly correspond with the greater values of the other variable (y) (and the same holds for the lesser values), the covariance is positive and the variables tend to show similar behavior.

In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, the covariance is negative and the variables tend to show opposite behavior.

Covariance calculation

- In case x and y have equal probability (equal probability distributions or probability density functions), covariance could be calculated as:

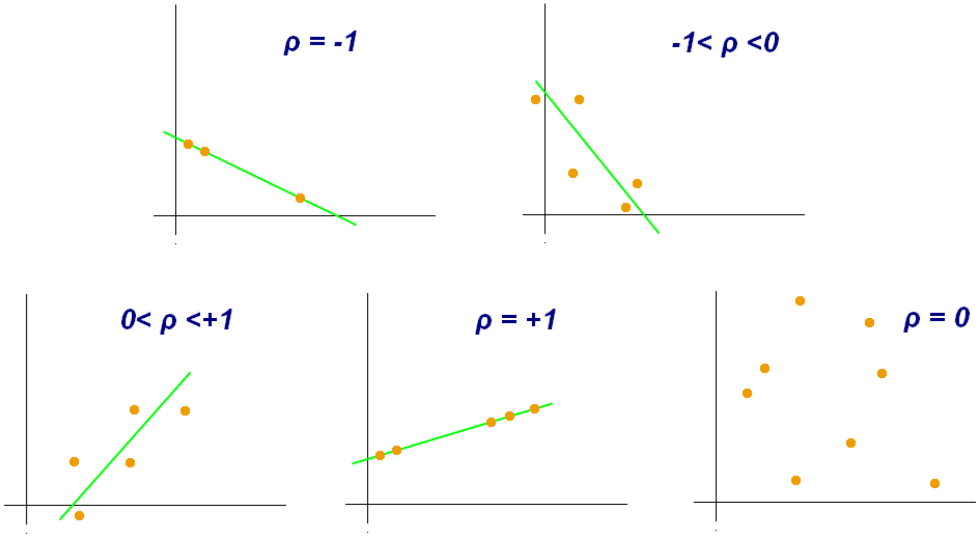
$$\operatorname{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{i=N}(x_i - E(x))(y_i - E(y))$$

Where:
$E(x) = x_{ip}$ = predicted y_i values
$x_i = x_{io}$ = observed y_i values
$E(y) = y_{ip}$ = predicted y_i values
$y_i = y_{io}$ = observed y_i values
N = number of xy observations

Correlation coefficient (*r*)

Pearson's **correlation coefficient (r or ρ)** is the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x \cdot \sigma_y}$$



Is a measure of the linear dependence (correlation) between two variables x and y. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

R^2 and *r*

In case of a linear regression between a finite set of x and y values equally distributed:

$$r = \sqrt{R^2}$$

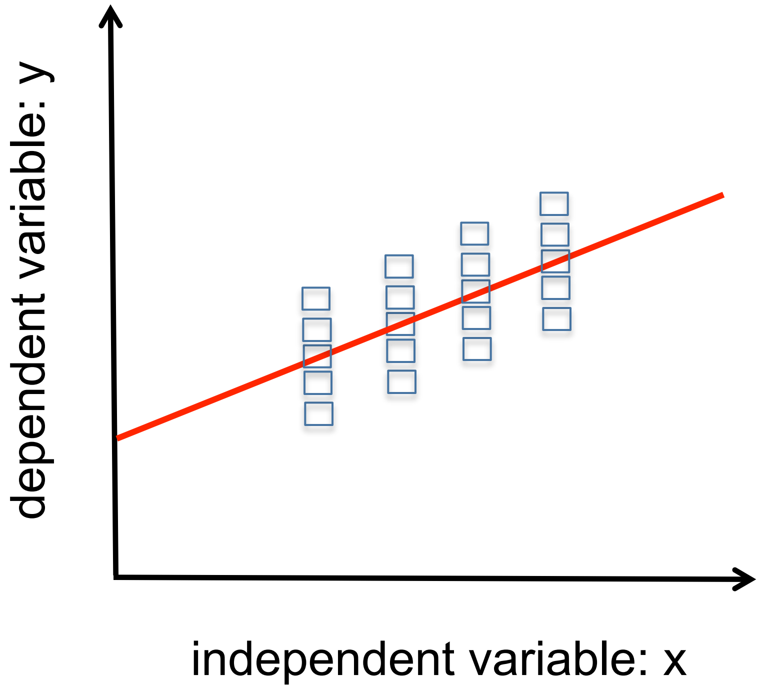
Important!
The two variables have different meanings and are differently calulated, but, in case of a linear regression, they could be easily derived one from the other.

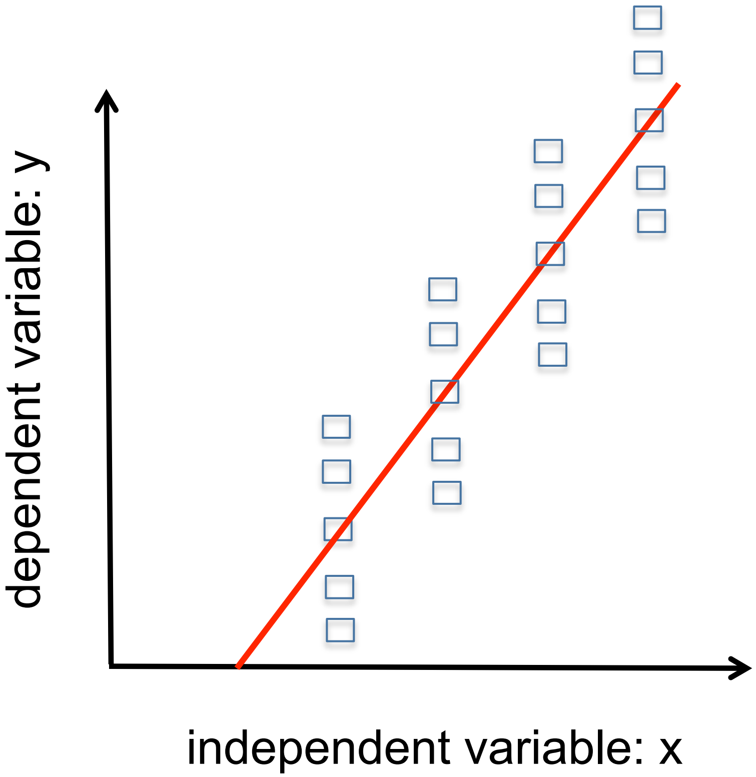
Generally three assumptions should be satisfied for linear regression analysis with two variables:

1. the relationship between y and x should be linear;
2. y should be normally distributed;
3. x and y values should be equally distributed.

Understanding regression

The line on the right has more variability of y values around each mean value (higher RSS) but also higher total y values variability (higher TSS).

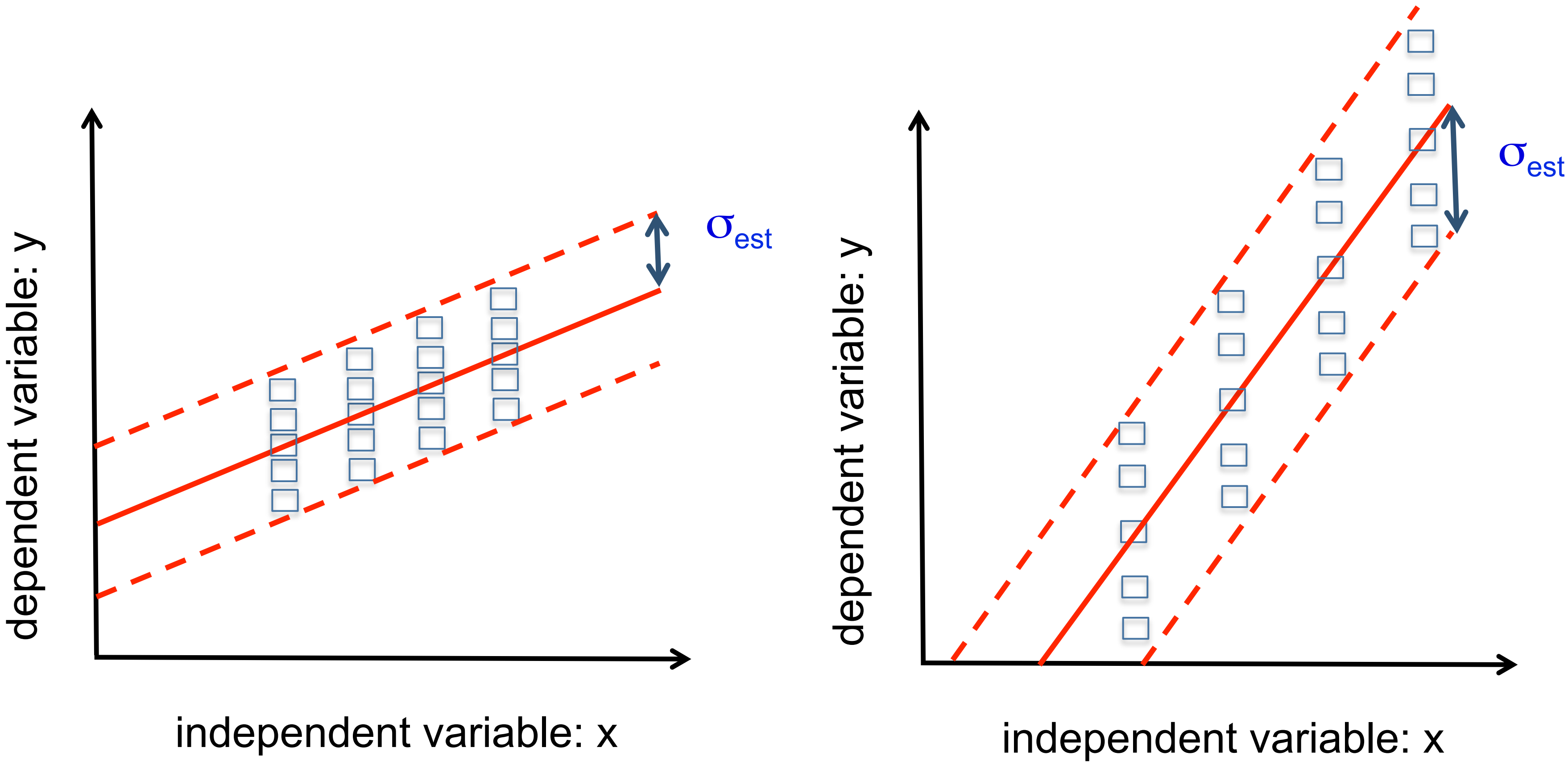




These two regression lines have thus the same R^2 because they have the same RSS/TSS ratio.

Understanding regression

The line on the right has more variability of y values and thus the estimation of y given an x value is less precise.



These two regressions have different errors of estimation (σ_{est}).

Understanding regression

Standard Error of Estimation (σ_{est})

or Root Mean Square Error (RMSE)

or Root Mean Square Deviation (RMSD)

$$RMSE = \sqrt{\frac{RSS}{N}} = \sqrt{\frac{\sum_{i=1}^{i=N} (y_{ip} - y_{io})}{N}}$$

Where:
RSS = residual sum of squares
y_{ip} = predicted y_i values
y_{io} = observed y_i values
N = number of y observations

Understanding regression

RMSE could be expressed as an absolute value or as a percentage (CVRMSE):

$$CVRMSE = \frac{RMSE}{\overline{y}} \cdot 100 = \frac{\sqrt{\frac{\sum_{i=1}^{i=N} (y_{ip} - y_{io})}{N}}}{\overline{y}} \cdot 100$$

Where $\overline{y}$ is the mean value of all y_i data

CVRMSE permits to compare the results of regressions carried out on different set of samples since it is independent from both sample size and the mean value of the dependent variable (y).

Understanding regression

The concept of RMSD or RMSE is generally applied to studies with a big amount of data.

When we have a big amount of data (n > 30) the number of y values is high enough to assume that our data are a population of data.

The mystic number 30 was suggested by an osservation of William Gosset, a statistician and Head Brewer for Guinness, which published several articles under the pseudonymous of Student.

However it should be pointed out that he never said that 30 was a '*magic number*', but, by comparing the correlation coefficients of a n sample with that of a population, he concluded: "*with samples of 30 … the mean value of a correlation coefficient of a sample approaches the real value of the correlation coefficent of a population comparatively rapidly*".

Student (1908). Probable error of a correlation coefficient. *Biometrika*, 6 (2-3): 302–310.

With a limited set of data …

The N number, which compares in the calculation of RMSE and r, is low (< 30).

When the N number is very low, we could not have enough data to carry out a regression.

Why?

With a limited set of data …

A standard deviation (σ) is necessary to describe the variability of an **x variable** in a sample with a limited number of observation.

$$\sigma = SD = \sqrt{\frac{\sum_{i=1}^{i=N}(x_i - \overline{x})}{N-1}}$$

In order to calculate (σ) the mean value $\overline{x}$ should be computed, so a bond is introduced in a system.

The bond is a parameter that is not free to vary; so when we introduce a bond, is like that we limit the variability of x of one degree.

With a limited set data …

In the same way, when we perform a linear regression analysis betwwen an **x variable** and a **y variable**, the m and q parameters should be calculated.

m and q are the bonds in linear regression analysis

The bond is a parameter that is not free to vary; so when we introduce two bonds, is like that we limit the variability of y (the dependent variable) of two degrees.

The degree of freedom in a regression are: N - 2

With a limited set of data …

σ_{est} or Root Mean Square Error (RMSE) for a sample

$$RMSE = \sqrt{\frac{RSS}{(N-2)}} = \sqrt{\frac{\sum\left(y_p - y_o\right)}{(N-2)}}$$

Where:
RSS = residual sum of squares
y_p = predicted y values
y_o = observed y values
N = number of y observations
2 = n° of regression parameters

RMSE could be expressed as CVRMSE as well:

$$CVRMSE = \frac{RMSE}{\overline{y}} \cdot 100 = \frac{\sqrt{\frac{\sum\left(y_p - y_o\right)}{(N-2)}}}{\overline{y}} \cdot 100$$

Regression implies an error

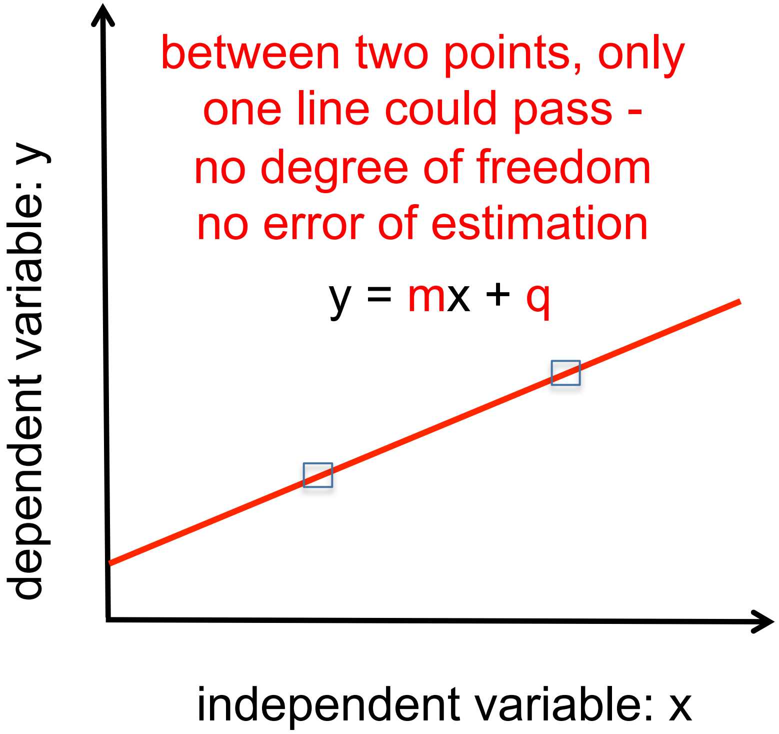
Since regression implies an error of approximation (which could be described by RMSE), the formula of a regression line could be written as:

$y = mx + q + e$

where:
m: regression coefficient
q: intercept
e: error of estimation

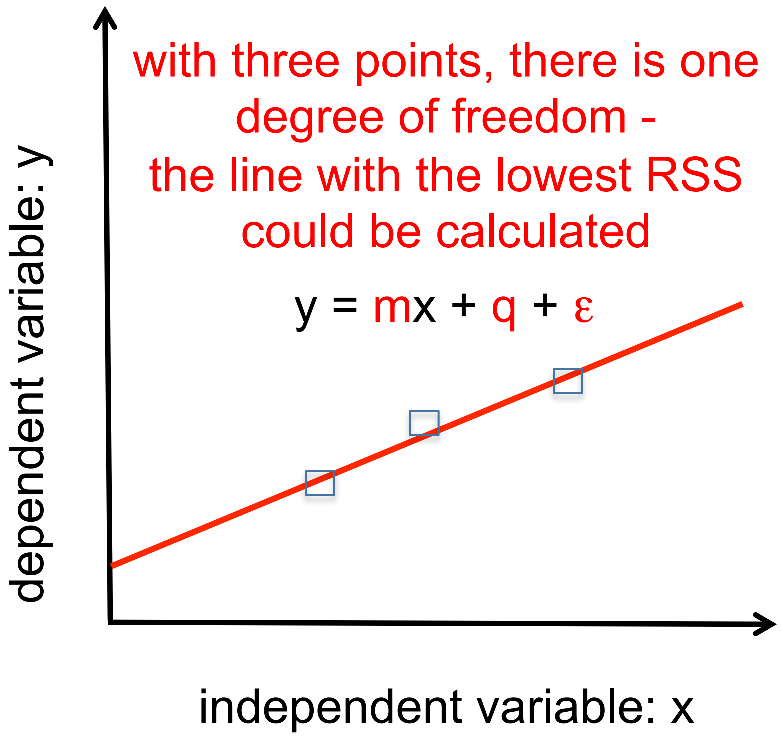
ε and degree of freedom (dof)

Since we have two bonds (one for each regression parameter: m and q) we need at leat three experimental point to perform a linear regression.



between two points, only one line could pass - no degree of freedom no error of estimation

$y = mx + q$



with three points, there is one degree of freedom - the line with the lowest RSS could be calculated

$y = mx + q + \varepsilon$

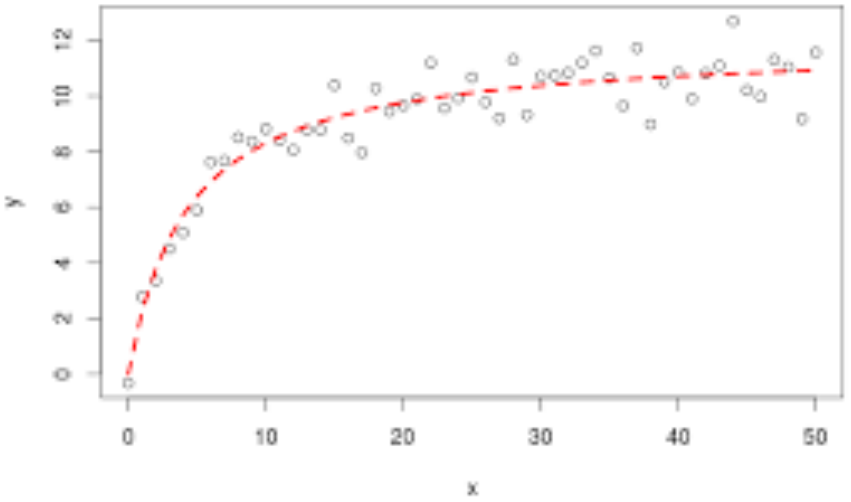
For regression purposes a degree of freedom is needed.

Non linear regression

Modern statistical programs could carry out also non linear regression.

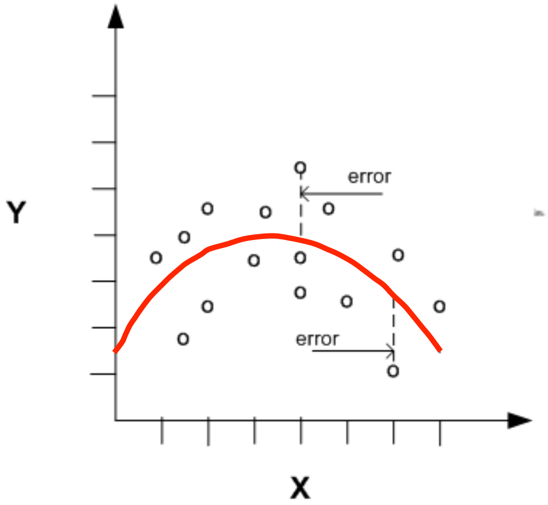
Examples:

asymptotic relationships

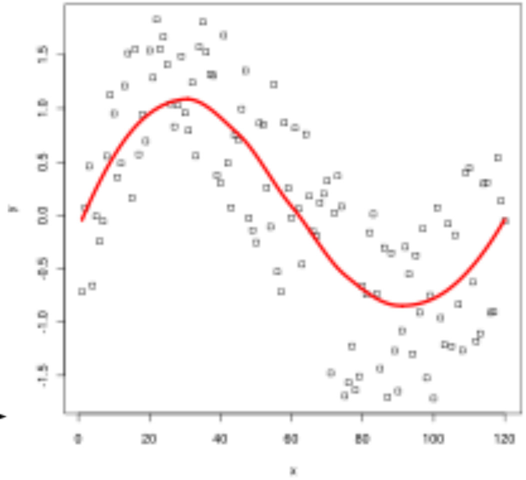


polynomial relationships

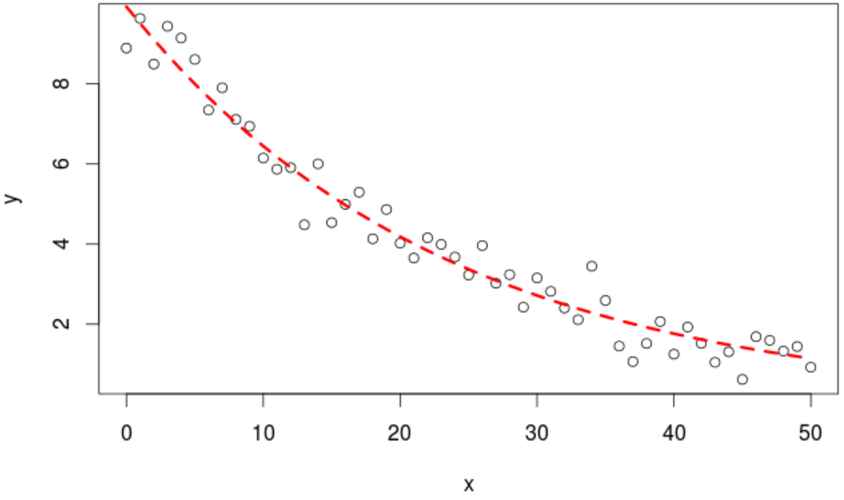
quadratic



cubic



exponential relationship



Non linear models

Non linear models could have more than two parameters.

Quadratic model:

$y = ax^2 + bx + q$ (q being the intercept)

Cubic model:

$y = ax^3 + bx^2 + cx + q$ (q being the intercept)

Asymptotic model:

$y = \frac{x}{a + bx} + q$ (q being the intercept)

Non linear models

Each additional parameter is a new bond and thus an addition data point is required to guarantee a degree of freedom.

Linear regression requires at least three data points.

Quadratic regression requires at leat four data points.

Asymptotic regression requires at least four data points.

Cubic regression requires at leat five data points.

and so on …

Non linear regression

For calculation purposes, non linear regression (NLR) uses the minimization of RSS criterion similarly to linear regression (LR).

RSS and R^2 are calculated in the same way as LR.

The Marquadt algorithm could be used as well for NLR.

In order to calulate RMSE, the number of parameters should be taken into account (parameters = bond).

Attention!!! In non linear regression: $r \neq \sqrt{R^2}$

Multivariate regression (MR)

Multiple regression analysis helps to understand how the typical value of the **dependent variable** (or 'criterion variable') changes when that of **more than one independent variables** (predictors) is varied.

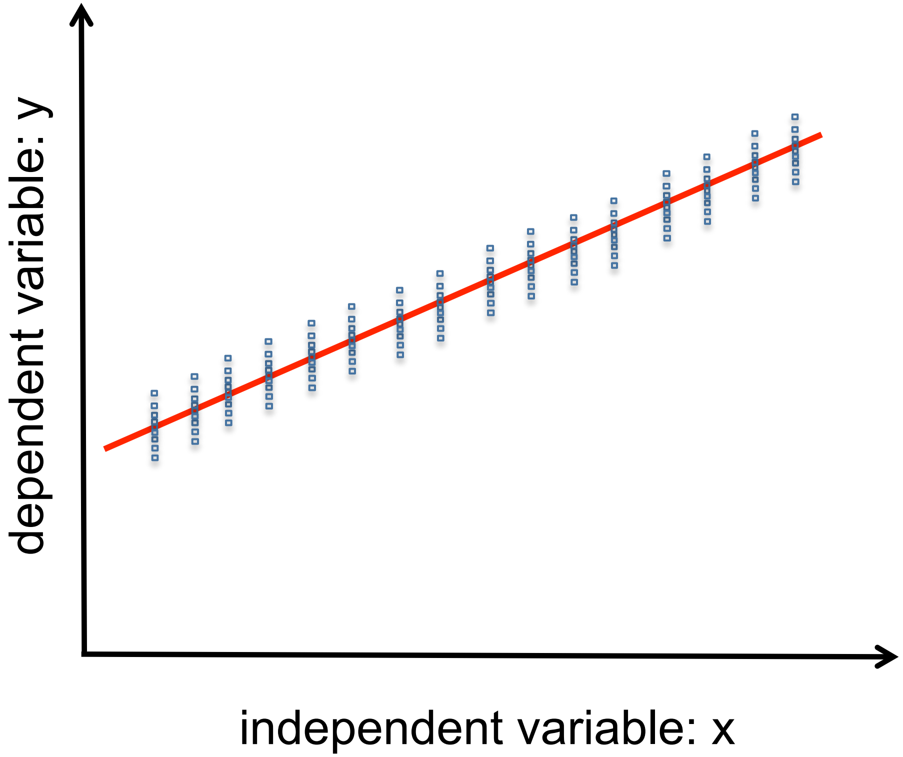
$z = f(x,y)$

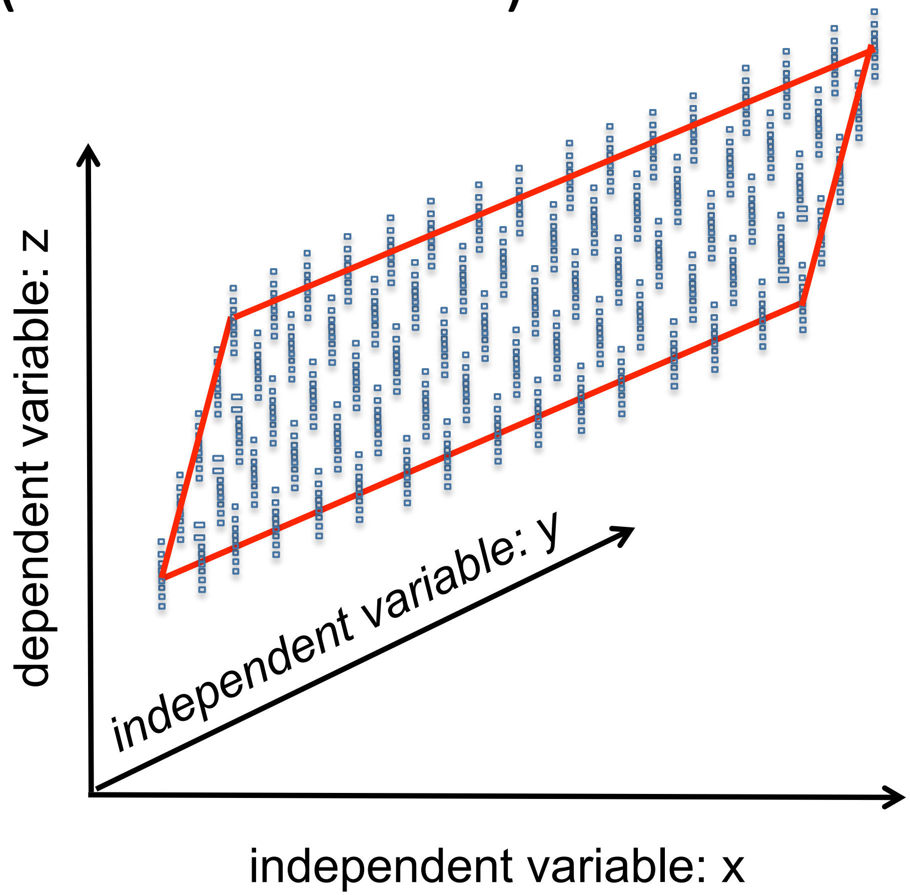
Since it is a regression …

$z = f(x,y) \pm \varepsilon$ where ε is the error of estimation.

Simplest case:

linear regression between a dependent variable and two independent variables (factor variables)



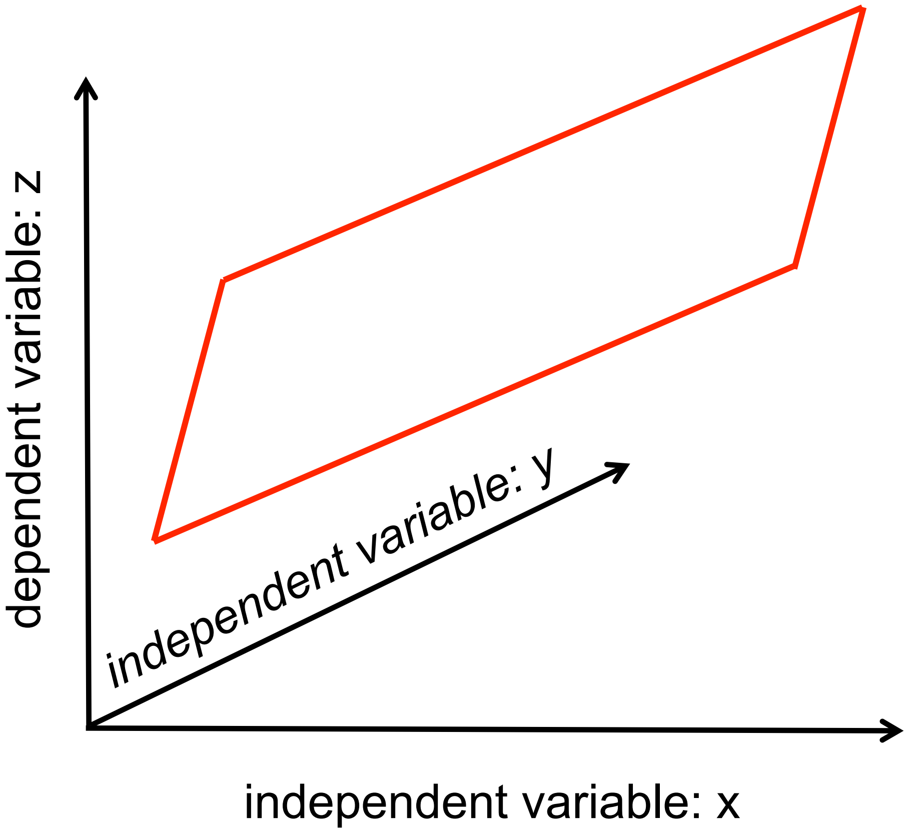


$y = mx + q$

$z = m_1x + m_2y + q$

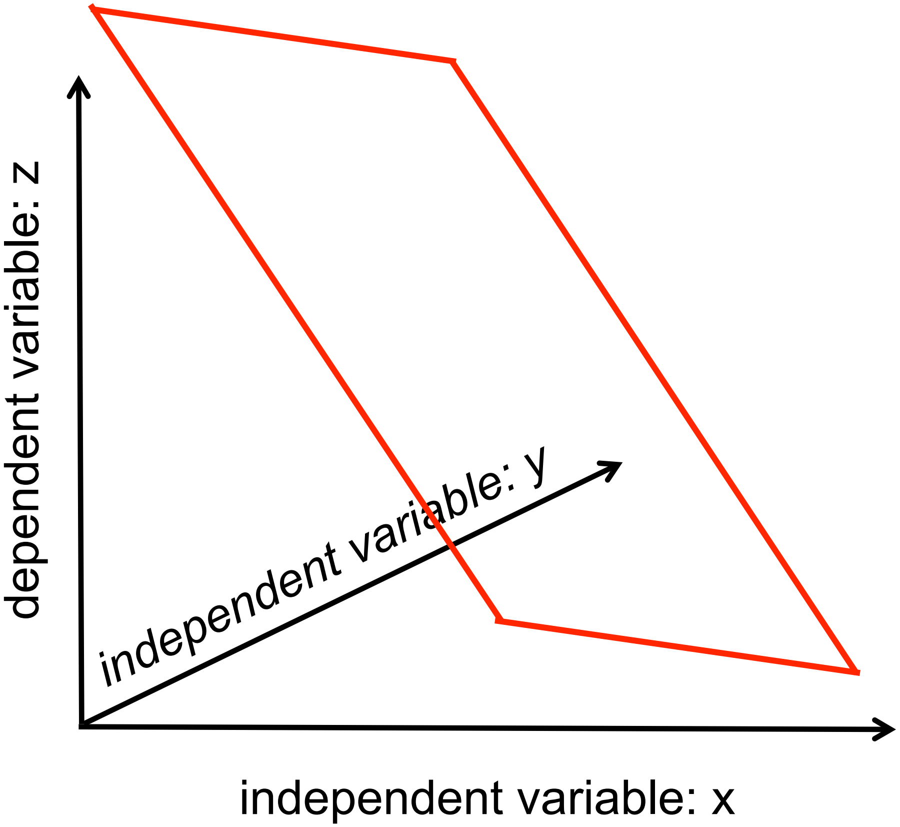
Simplest case:

linear regression between a dependent variable and two independent variables (factor variables)



$z = m_1x + m_2y + q$

$m_1 > 0 ; m_2 > 0 ; q > 0$



$z = m_1x + m_2y + q$

$m_1 < 0 ; m_2 < 0 ; q > 0$

Multivariate regression (MR)

The regression model is a polynomial model that has **additive properties** and could have one or more independent variables.

$$y = m_1x_1 + m_2x_2 + m_3x_3 + m_4x_4 + \dots m_ix_i + q \pm \varepsilon$$

where:

x_i is an independent variable

m_i is the regression coefficient of x_i $(-\infty < m < \infty)$

q is the intercept $(-\infty < q < \infty)$

ε is the error of estimation.

red letters correspond to regression parameter

Multiple linear regression model

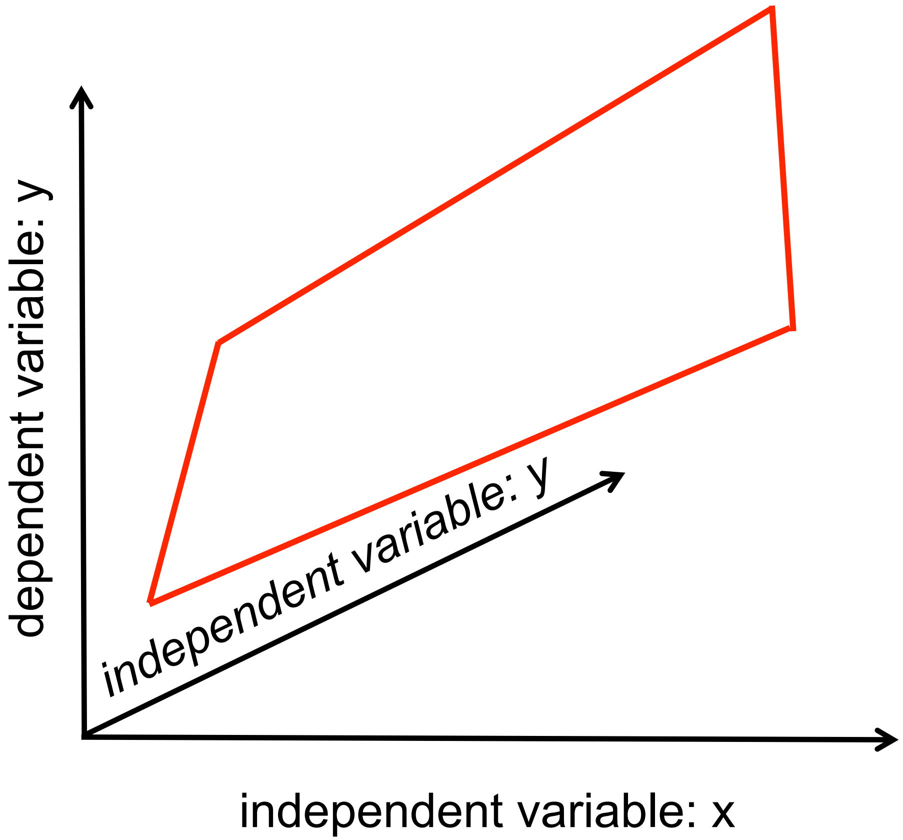
Multiple linear regression (MLR) models have more than two parameters.

Each additional parameter is a new bond and thus an addition data point is required to guarantee a degree of freedom.

The simplest model: $z = ax + by + q$, with three parameters, requires at least four data point for calculation purposes.

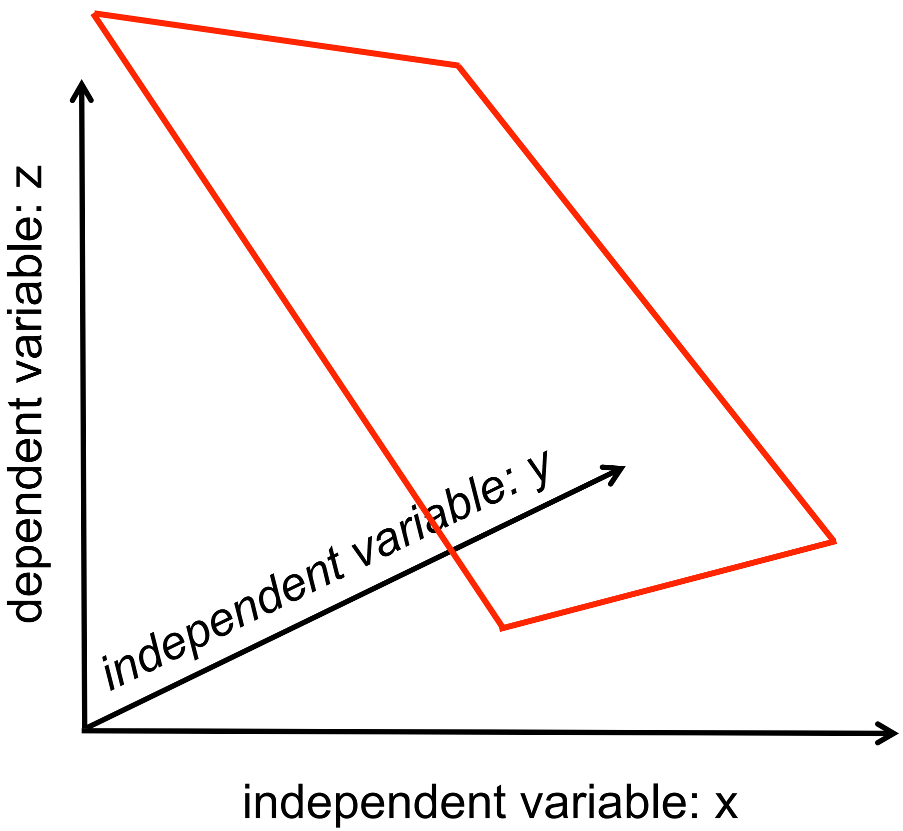
Combined effects:

MLR could consider combined effect between independent/factor variables (es. xy)



$z = m_1x + m_2y + m_3xy + q$

$m_1 > 0$; $m_2 > 0$; m3 > 0; $q > 0$



$z = m_1x + m_2y\ m_3xy + q$

$m_1 < 0$; $m_2 < 0$; $m_3 < 0$; $q > 0$

MLR combined effects

When m_3 is > 0

Synergistic effect (the combined effect is higher than the sum of the effects of the two variables)

$m_1m_2 > m_1 + m_2$

When m_3 is < 0

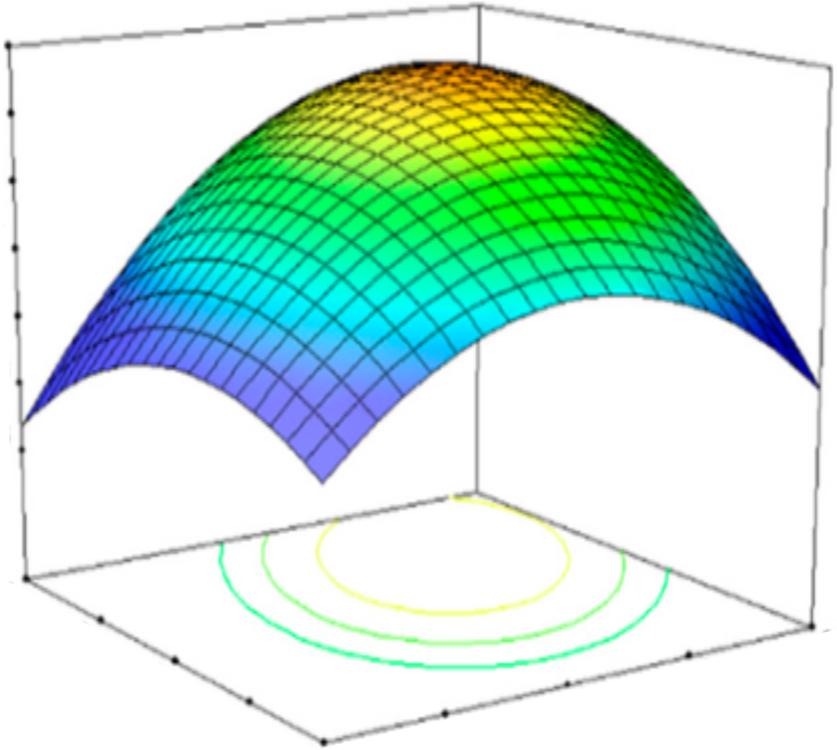
Antagonistic effect (the combined effect is lower than the sum of the effects of the two variables)

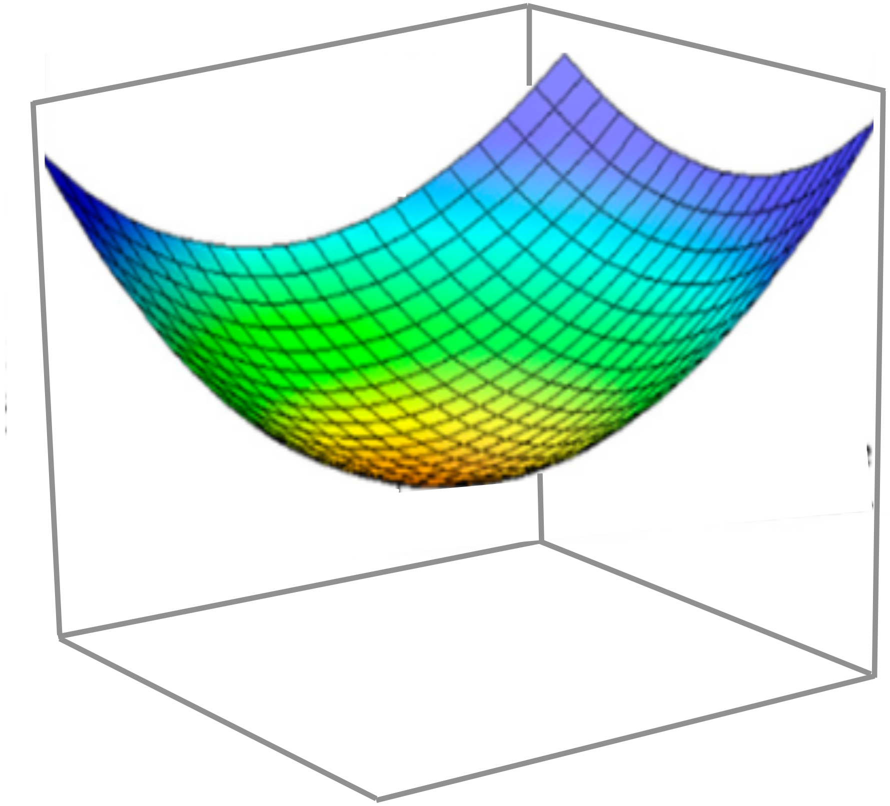
$m_1m_2 < m_1 + m_2$

Multiple regression:

could consider non linear effects

quadratic effects





$z = m_1x^2 + m_2x + m_3y^2 + m_4y + q$

$m_1 < 0$; $m_2 < 0$; $m_3 < 0$; $m_4 < 0$; $q > 0$

$z = m_1x^2 + m_2x + m_3y^2 + m_4y + q$

$m_1 > 0$; $m_2 < 0$; $m_3 > 0$; $m_4 < 0$; $q > 0$

Multivariate regression (MR)

A second degree regression model between one dependent variable (z) and two independent/factor variables (x,y).

$$z = m_1x^2 + m_2x + m_3y^2 + m_4y + m_5xy + q \pm \varepsilon$$

where:

x and y are an independent variable

m_i is the regression coefficient of a variable ($-\infty < m < \infty$)

q is the intercept ($-\infty < q < \infty$)

ε is the error of estimation.

red letters correspond to regression parameters

Multivariate linear regression (MLR)

In case of combined and quadratic effects, MLR could be carried out by using both original variables (e.g. x and y) and calculated variables (x^2, y^2, xy)

This is a x and y data linearization procedure that allows to fit z data that are not linearly correlated to x and y, by using a linear model.

Limits of multivariate regression (MLR)

MLR could be imply the use of many factors.

Multivariate linear regression requires that independent/factor variables are not correlated among them.

When the number of factors gets too large (e.g. greater than the number of observations), it is possible to have data overfitting.

Partial MLR for factor reduction

- **Partial multiple regression analysis** could be carried out in order to reduce the number of x variables.

- In Partial MLR each variable is inserted in the model using a stepwise method (one by one/step by step) and its regression coefficient is calculated by setting all the other variables constant.

- This tecnique does not take into account the effect of other variables in the model and could cause a loss of information.

Stepwise analysis

- **Forward stepwise analysis**: variables are inserted one by one in the model starting from the x variable that is most significantly correlated to y (in terms of *r* and *p*) and its regression coefficient is calculated by setting all the other variables constant, then another variable is inserted. When the first x variable non significantly correlated to y is encountered, it is removed from the model. This removal process is repeated for all the other uncorrelated variables.

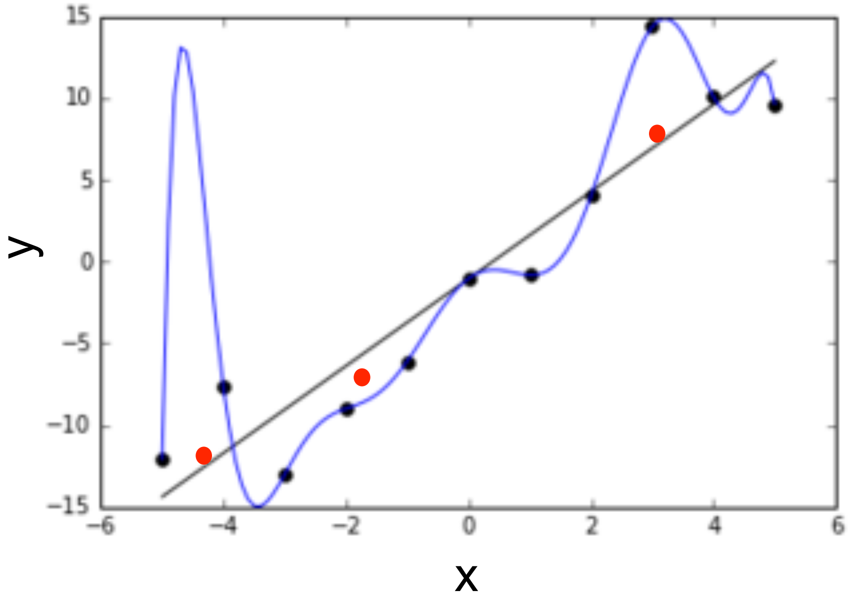
- **Backward stepwise analysis**: all variables are inserted in the model, then the x variables that are less correlated to y (in terms of *r* and *p*) are removed from the system one by one and the regression coefficients of other variables are recalculated each time by partial regression. When the first significantly correlated variable is identified by the analysis, the procedure of removal is stopped.

Overfitting due to many factors

The blue curve which is polynomial model, with n factorial variables, fits data perfectly ($R^2 = 1$).

When new data (red dots) are added, the model fails to predict them.

A simple linear model with only one factor works better in prediction.



Validation

Validation with an external data set (**full cross validation**) could be used to test/avoid overfitting.

Another possibility is to carry out **leave-one-out validation** (LOOV) .

LOOV validation, uses all but one sample to calculate a MLR model and different MLR models are calculated by leaving each sample out from the data set.

Given n data, n models are obtained. The one that better predicts the leaved out sample is the best.

Correlation among factors (MLR)

Another limit of MLR is correlation among factors

Correlation coefficient *r* between x and y variables (or among all x_i independent variables) should be tested.

If the probability value (*p*) associated to *r* for a given N number of observations is significant, MR could not be performed.

Significance of correlation among factors

Could be determined by computer program.

Could be found in statistical tables which reports *r* and *p* values as a function of N.

In the latter case, care should be taken in correcting N by taking into account the number of bonds.

Correlation and collinearity

When there are a lot of independent variables (factor variables) or combinations among them (quadratic or combined effects) is more probable to find significant correlations among some of the factor variables.

In this case it is likely to have **collinearity** of variables (e.g. factor x increase or decrease lienarly with the increase of factor y)

MLR limits

- The overfitting of MLR models could be avoided by stepwise analysis and validation.

- Correlation among variables, when present, is a limit that could not be overcome without resorting to latent variables extraction.

Significance of correlation and PCR

When factors are many and highly collinear among them:

- It is possible to carry out a PCA analysis and extract new latent variables (PCs) that are not correlated among them by construction.
- Then, it is possible to carry out a MLR or a MNLR by using PCs instead of the original variables.

This procedure is called **Principal Component Regression** (PCR).

Principal Component Regression

PCR permits to carry out MLR or MNLR avoiding the problem of independent variables/factor correlation.

PCs explain only the maximum variance within a given variables data set and do not consider the variation of an average dependent/response variable (e.g. y variable) value with the variation of the average values of independent/factor variables (e.g. n x variables).

For this reason PCs could not be the best latent variables to take in consideration in order to carry out a multiple regression analysis using latent variables (or latent structures).

Multiple regression by PLS

The acronym PLS stands for:

Partial Least Square

but nowadays the best definition is retained:

Projection on Latent Structure

since the regression is not performed by using the original x_i variables but using a number of new variables called components calculated from a linear combination of the original variables.

Extraction of components

PCA calculates components that maximize the explained variance of the data matrix.

PLS calculates components (latent structures) by seeking **directions** in a n dimensional space defined by a set of x variables.

A direction correspond to new factor described by a vectors) that are associated to an high variation of the response y variables.

‘Components’ or ‘Factors’?

Some texts (or statistical computer programs) define the components calculated by PLS as **factors** but, in some cases, this definitions could be misleading, since the original independent variables (x variables) are also called **factorial variables** or **factors**.

Hereby, in these slides, they will be defined components in order to avoid misleading.

Why ‘latent structures’?

Components are **latent structures**; that is variables whose existence is inferred (deducted) from the existing relationship among observed items (factors and response variables).

Since the latent structure are calculated by observing a relationship (linear relationship) among variables, PLS is a regression method by definition; for this reason many researchers call it ‘PLS’ and not ‘PLS regression’.

Extraction of components

Similarly to PCA, PLS calculates a number of components (latent structures) that are linear combinations of the original variables.

The number of components could range from i (where i is the number of the original factor variables) to 1.

n components explain the 100% of y variability.

Original data matrix for PLS

y
y_1
y_2
y_3
y_4
y_5
y_6
y_7
y_8
y_9
…
y_i

x1	x2	x3	x4	x5	…	xn
$x1_1$	$x2_1$	$x3_1$	$x4_1$	$x5_1$	…	xn_1
$x1_2$	$x2_2$	$x3_2$	$x4_2$	$x5_2$	…	xn_2
$x1_3$	$x2_3$	$x3_3$	$x4_3$	$x5_3$	…	xn_3
$x1_4$	$x2_4$	$x3_4$	$x4_4$	$x5_4$	…	xn_4
$x1_5$	$x2_5$	$x3_5$	$x4_5$	$x5_5$	…	xn_5
$x1_6$	$x2_6$	$x3_6$	$x4_6$	$x5_6$	…	xn_6
$x1_7$	$x2_7$	$x3_7$	$x4_7$	$x5_7$	…	xn_7
$x1_8$	$x2_8$	$x3_8$	$x4_8$	$x5_8$	…	xn_8
$x1_9$	$x2_9$	$x3_9$	$x4_9$	$x5_9$	…	xn_9
…	…	…	…	…	…	…
$x1_i$	$x2_i$	$x3_i$	$x4_i$	$x5_i$	…	xn_i

where:
y is the dependent (response) variable
xn are the (independent) factor variables
i is the number of observations

Data matrix after PLS extraction

y
y_1
y_2
y_3
y_4
y_5
y_6
y_7
y_8
y_9
…
y_n

C1	C2	C3	C4	C5	…	**Cn**
$C1_1$	$C2_1$	$C3_1$	$C4_1$	$C5_1$	…	Cn_1
$C1_2$	$C2_2$	$C3_2$	$C4_2$	$C5_2$	…	Cn_2
$C1_3$	$C2_3$	$C3_3$	$C4_3$	$C5_3$	…	Cn_3
$C1_4$	$C2_4$	$C3_4$	$C4_4$	$C5_4$	…	Cn_4
$C1_5$	$C2_5$	$C3_5$	$C4_5$	$C5_5$	…	Cn_5
$C1_6$	$C2_6$	$C3_6$	$C4_6$	$C5_6$	…	Cn_6
$C1_7$	$C2_7$	$C3_7$	$C4_7$	$C5_7$	…	Cn_7
$C1_8$	$C2_8$	$C3_8$	$C4_8$	$C5_8$	…	Cn_8
$C1_9$	$C2_9$	$C3_9$	$C4_9$	$C5_9$	…	Cn_9
…	…	…	…	…	…	…
$C1_i$	$C2_i$	$C3_i$	$C4_i$	$C5_i$	…	Cn_i

where:
y is the dependent (response) variable
Cn are the components calculated by PLS analysis
i is the number of observations

PLS model

The final PLS model is:

$$y = m_1C_1 + m_2C_2 + m_3C_3 + \ldots + m_nC_n + q + \varepsilon$$

where:

C is a calculated component (latent structure)

m_n is the regression coefficient of x_i $(-\infty < m < \infty)$

q is the intercept $(-\infty < q < \infty)$

ε is the error of estimation

red letters correspond to regression parameter

PLS model

Since components are linear combinations of the original factor variables (xn), statistical programs could easily recalculate the model using xn as variables.

$$y = m_1x_1 + m_2x_2 + m_3x_3 + \ldots + m_ix_n + q + \varepsilon$$

where:

x is the original factor variable

m_i is the regression coefficient of x_i $(-\infty < m < \infty)$

q is the intercept $(-\infty < q < \infty)$

ε is the error of estimation

red letters correspond to regression parameter

Extraction of components

n components explain the 100% of y variability.

By taking into account the components that account for the maximum y variability, PLS could be used to reduce the system dimensionality.

If the number of components is too large (for example greater than the number of observation) overfitting could occur also in PLS regression analysis.

Regression technique

Even though it is carried out with calculated components (latent structures) instead of original variables, PLS uses the least square method for regression purposes.

By considering only the components that account for the maximum y variability, PLS could be used to reduce the system dimensionality.

If the number of components is too large (for example greater than the number of observation) overfitting could occur also in PLS regression analysis.

PLS and ‘partial’ least square

- PLS method uses a ‘**partial** regression analysis’, which means that each variable (component in this case) is inserted to the model using a stepwise method (step by step) and its regression coefficient is calculated by setting all the other variables constant.

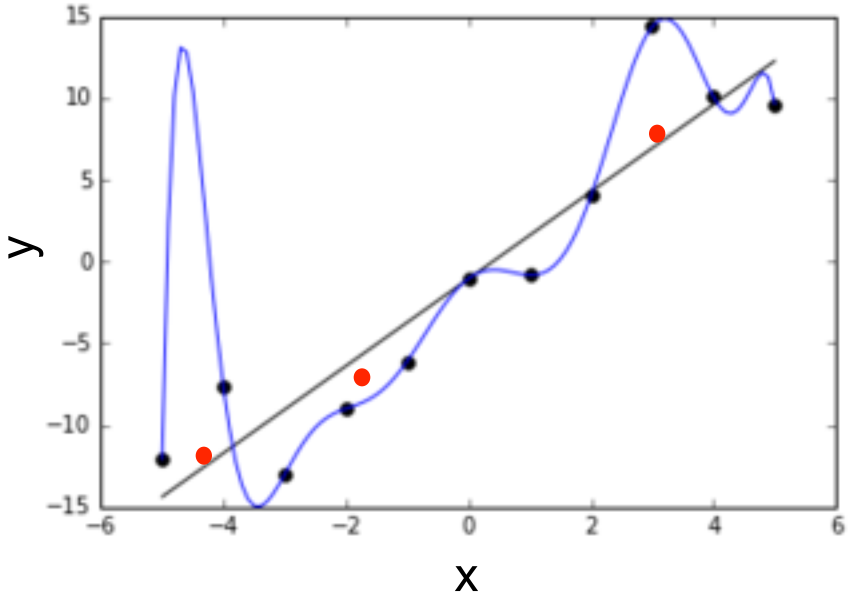
- This tecnique does not take into account the effect of other variables (components) in the model but the use of latent structures offers the advantage that the variables are not correlated among them, thus their effect is independent from the variable to be inserted.

Overfitting in PLS

The blue curve which is a PLS model, with n factorial variables (LS), fits data perfectly ($R^2 = 1$).

When new data (red dots) are added, the model fails to predict them.

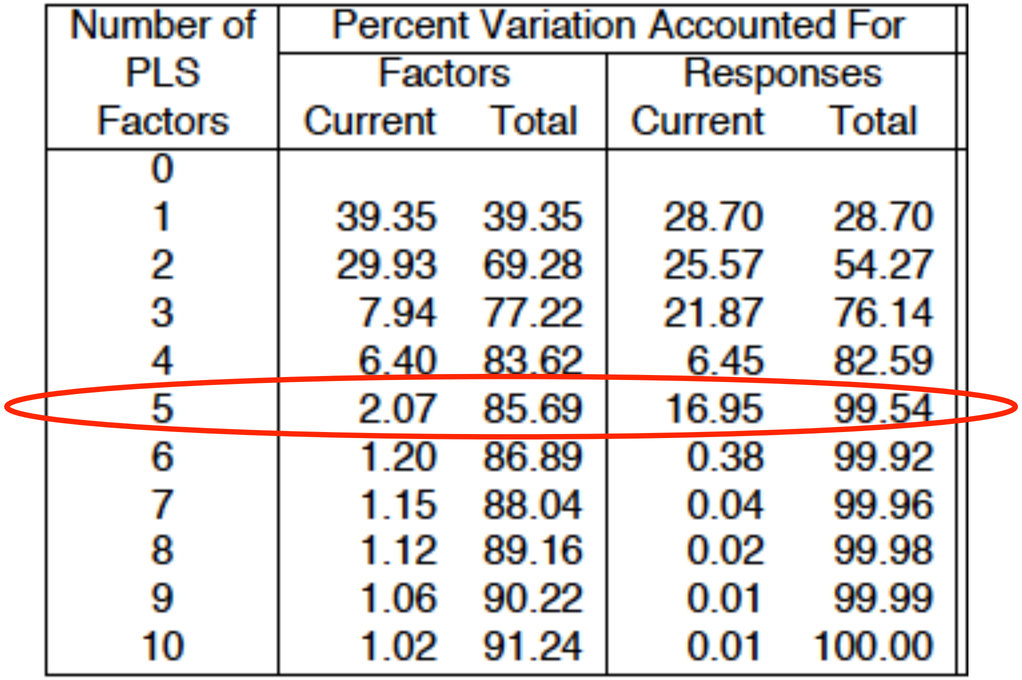
A simple PLS model with only one factor works better in prediction.



Avoiding overfitting

- Identification of the number of components which significantly increase the percent variation of the response variable (y) of the PLS model.

5 new components (Factors) calculated from 10 original factor variables could account for the 99.54% of the response variable variation.



Number of PLS Factors	Percent Variation Accounted For			
	Factors		Responses	
	Current	Total	Current	Total
0				
1	39.35	39.35	28.70	28.70
2	29.93	69.28	25.57	54.27
3	7.94	77.22	21.87	76.14
4	6.40	83.62	6.45	82.59
5	2.07	85.69	16.95	99.54
6	1.20	86.89	0.38	99.92
7	1.15	88.04	0.04	99.96
8	1.12	89.16	0.02	99.98
9	1.06	90.22	0.01	99.99
10	1.02	91.24	0.01	100.00

Validation of PLS

Validation with an external data set (full cross validation) could be used to avoid overfitting.

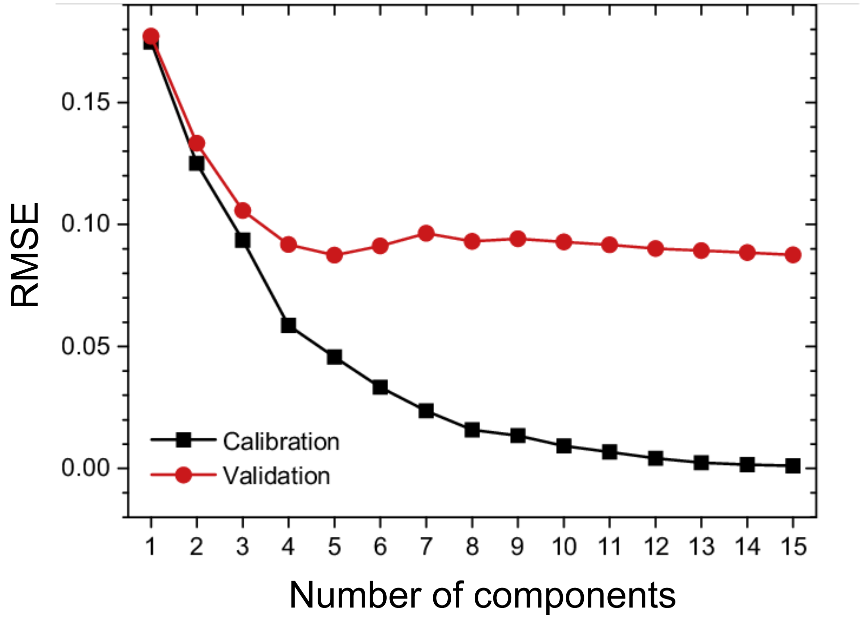
Another possibility is to carry out leave-one-out (LOOV) validation.

LOOV validation, uses all but one sample to calculate a MLR model and different MLR models are calculated by leaving each sample out from the data set.

Given n data, n models are obtained. The one that better predicts the leaved out sample is the best.

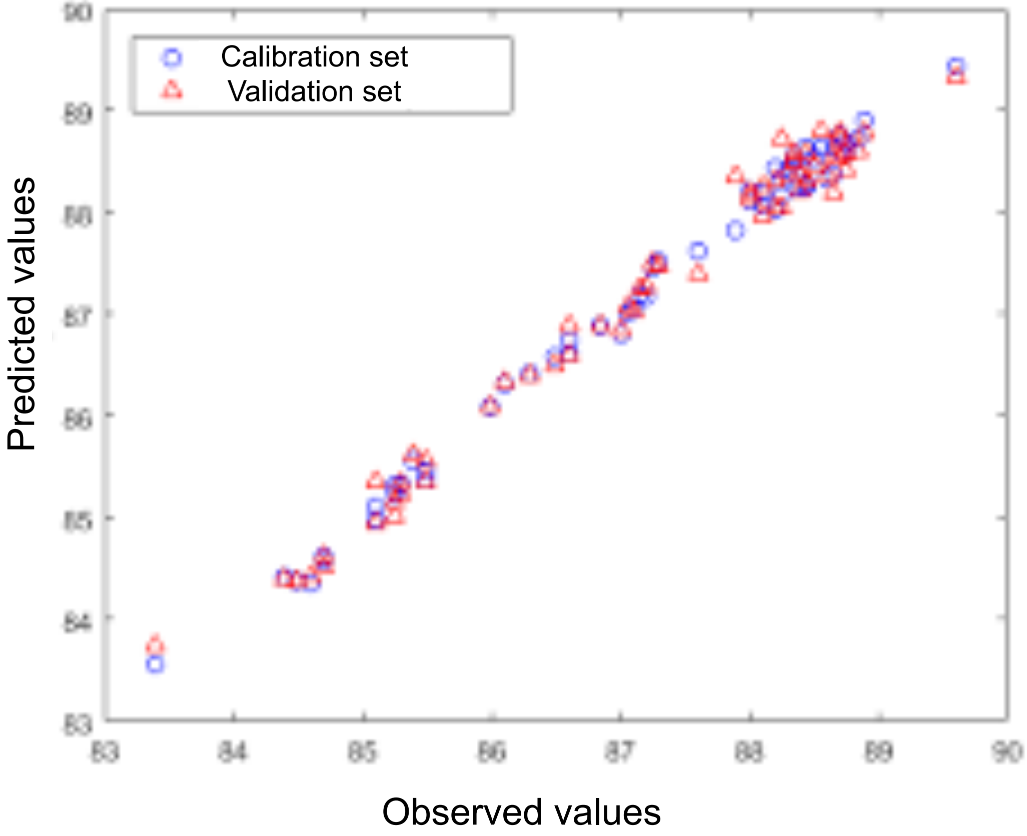
Avoiding overfitting by validation

Identification of the number of components which significantly decrease the RMSE of calibration of the model (RMSEC) and the RMSE of prediction or validation (RMSEV) of the PLS model.



In the case of the results reported on the graph, 4 components, out of 15 calculated components, are enough to minimize RMSEV below 10% and to reduce the RMSEC to a value that is lower than that of RMSEV (about 5%).

Visualization of PLS results



An observed vs predicted data plot for calibration and validation sets could offer an opportunity to visualize the goodness of fit of the regression model.

Orthogonal PLS (O-PLS)

Components extracted by PLS are orthogonal among them for construction.

However since PLS maximizes the dependent variable (y) variance at the variation of the factor variables (xn), it is possible to have some systematic variation in the response variable that is unrelated, or orthogonal, to the factors variables.

O-PLS is a data pre-treatment used to avoid y systematic variation unrelated to x variables.

It is very usefull in spectrometric data processing.

PLS and PLS2

- PLS regression could be performed both using one single y variable and a set of y variables.
- Just in the case that y variables are more than one (y1, y2, y3, … yn) PLS is defined as PLS2.
- PLS maximizes the dependent variables (yn) variance at the variation of the factor variables (xn).

Original data matrix for PLS2

y1	y2	y3	**…**	**yi**
$y1_1$	$y2_1$	$y3_1$	…	yn_1
$y1_2$	$y2_2$	$y3_2$	…	yn_2
$y1_3$	$y2_3$	$y3_3$	…	yn_3
…	…	…	…	…
$y1_i$	$y2_i$	$y3_i$	…	yn_i

x1	x2	x3	**…**	**xn**
$x1_1$	$x2_1$	$x3_1$	…	xn_1
$x1_2$	$x2_2$	$x3_2$	…	xn_2
$x1_3$	$x2_3$	$x3_3$	…	xn_3
…	…	…	…	…
$x1_i$	$x2_i$	$x3_i$	…	xn_i

where:
yn are the dependent (response) variables
xn are the (independent) factor variables
i is the number of observations

Data matrix after PLS2 extraction

y1	y2	y3	**…**	**yi**
$y1_1$	$y2_1$	$y3_1$	…	yn_1
$y1_2$	$y2_2$	$y3_2$	…	yn_2
$y1_3$	$y2_3$	$y3_3$	…	yn_3
…	…	…	…	…
$y1_i$	$y2_i$	$y3_i$	…	yn_i

C1	C2	C3	**…**	**Cn**
$C1_1$	$C2_1$	$C3_1$	…	Cn_1
$C1_2$	$C2_2$	$C3_2$	…	Cn_2
$C1_3$	$C2_3$	$C3_3$	…	Cn_3
…	…	…	…	…
$C1_i$	$C2_i$	$C3_i$	…	Cn_i

where:
yn are the dependent (response) variables
Cn are the components (or factors) extracted by PLS analysis
i is the number of observations

PLS Discriminant Analysis (PLS-DA)

As previously discussed, PLS is a regression analysis.

However, since PLS2 maximizes the dependent variable (yn) variance at the variation of the factor variables (xn), it is possible to use PLS2 as a discriminant analysis by using y variables as classification variables and by applying 0-1 binomial values to each yn variable.

When a sample does not belong to a y class, its values for that class is 0; whilst when a sample belongs to a y class, its value for that class is 1.

PLS-DA maximizes the variation of yn at variation of xn.

Data matrix for PLS-DA

y1	y2	y3
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

x1	x2	x3	…	xn
$x1_1$	$x2_1$	$x3_1$	…	xn_1
$x1_2$	$x2_2$	$x3_2$	…	xn_2
$x1_3$	$x2_3$	$x3_3$	…	xn_3
$x1_4$	$x2_4$	$x3_4$	…	xn_4
$x1_5$	$x2_5$	$x3_5$	…	xn_5
$x1_6$	$x2_6$	$x3_6$	…	xi_6
$x1_7$	$x2_7$	$x3_7$	…	xn_7
$x1_8$	$x2_8$	$x3_8$	…	xn_8
$x1_9$	$x2_9$	$x3_9$	…	xn_9
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
$x1_i$	$x2_i$	$x3_i$	…	xn_1

This example is for three grouping variable but in PLS-DA there is no limit for grouping variables.

Data matrix after PLS-DA extraction

y1	y2	y3
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

C1	C2	C3	…	Cn
$C1_1$	$C2_1$	$C3_1$	…	Cn_1
$C1_2$	$C2_2$	$C3_2$	…	Cn_2
$C1_3$	$C2_3$	$C3_3$	…	Cn_3
$C1_4$	$C2_4$	$C3_4$	…	Cn_4
$C1_5$	$C2_5$	$C3_5$	…	Cn_5
$C1_6$	$C2_6$	$C3_6$	…	Ci_6
$C1_7$	$C2_7$	$C3_7$	…	Cn_7
$C1_8$	$C2_8$	$C3_8$	…	Cn_8
$C1_9$	$C2_9$	$C3_9$	…	Cn_9
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
$C1_i$	$C2_i$	$C3_i$	…	Cn_1

This example is for three grouping variable but in PLS-DA there is no limit for grouping variables.

Orthogonal PLS-DA (O-PLS-DA)

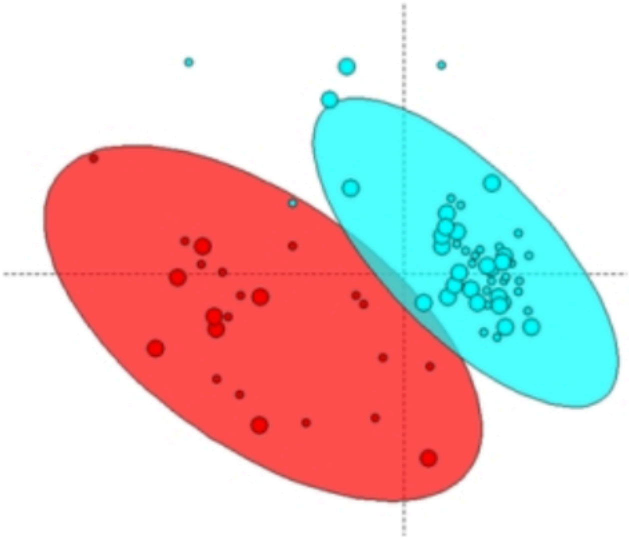
O-PLS is a data pre-treatment used to avoid y systematic variation unrelated to x variables.

It is very useful in spectrometric data processing.

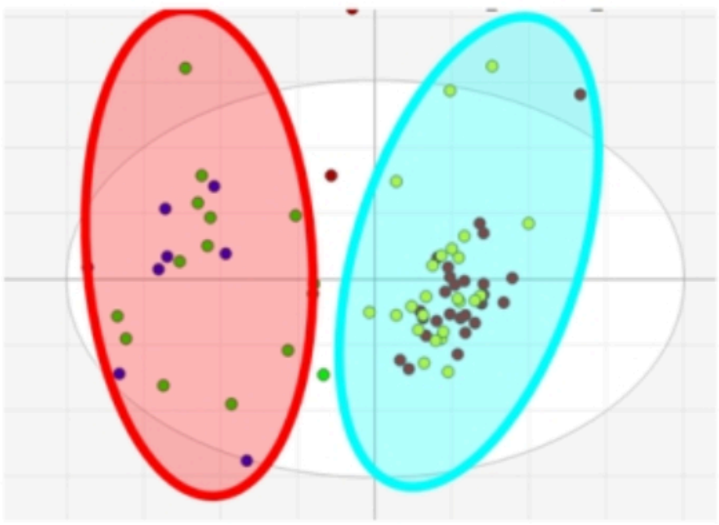
O-PLS data pre-treatment could be used also for PLS-DA and this improves the classification power of PLS-DA.

O-PLS pretreatment for classification

PLS-DA



O-PLS-DA



Exercise

- Carry out PLS-DA on the data set used for PCA and LDA analysis

- Individuate the three most important components for classes discrimination

- Visualize data using 3D graph

Original table

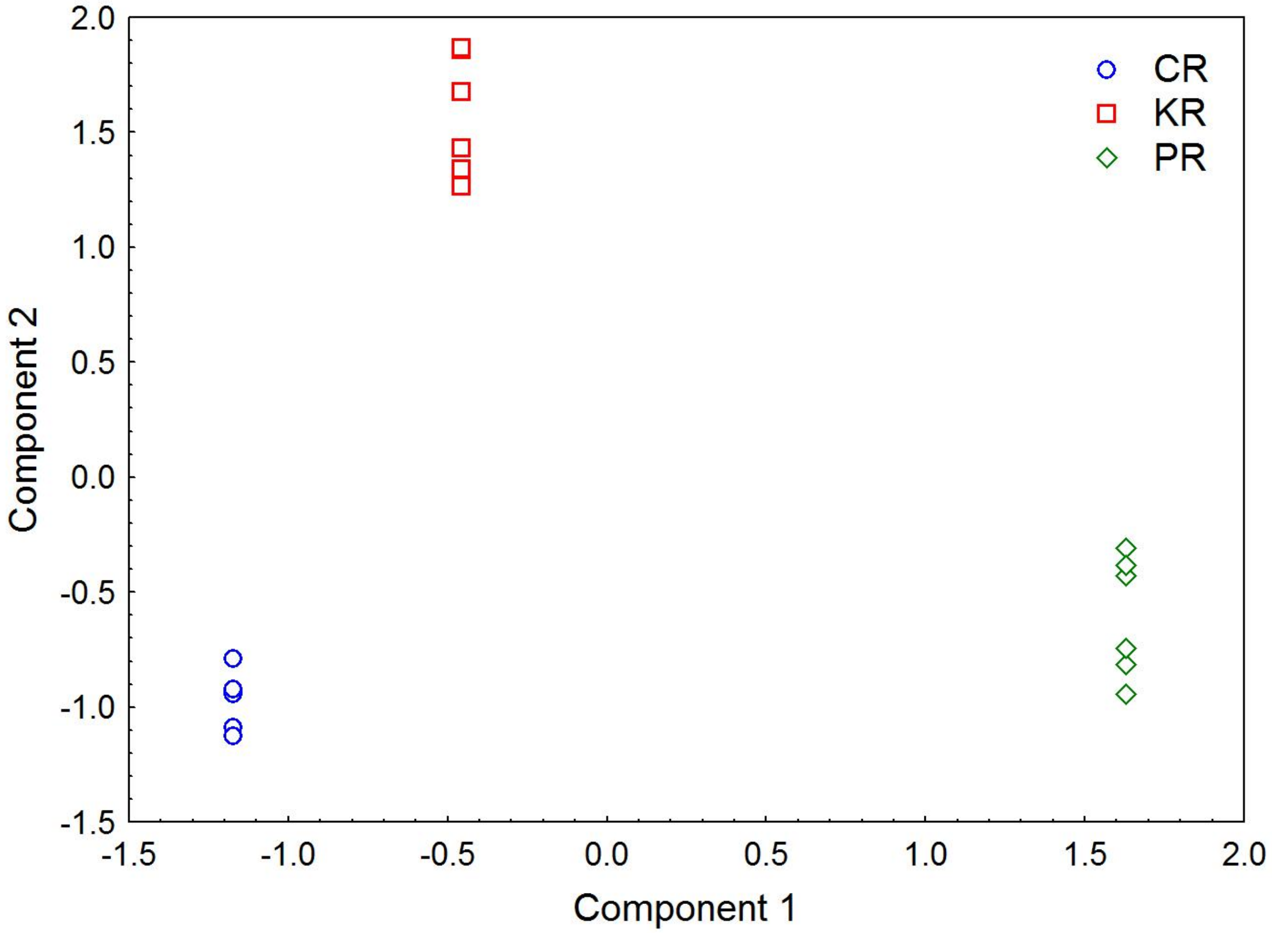
Group	Age	V1	V2	V3	V4	V5	V6	…	**Vn**
CR	2	1.05	26.65	3.90	27.19	2.37	1.48	…	1.01
CR	15	0.64	7.23	4.76	35.98	3.01	1.29	…	1.51
CR	30	1.10	6.26	4.90	24.39	4.03	1.53	…	1.34
CR	60	0.89	2.80	1.87	18.21	3.85	6.98	…	1.05
CR	90	0.73	3.99	0.00	20.28	2.07	7.01	…	0.57
CR	180	1.18	1.93	2.03	13.66	4.56	6.55	…	1.15
KR	2	0.18	1.95	3.73	21.31	3.69	4.02	…	0.52
KR	15	0.54	0.04	5.05	16.89	2.29	2.95	…	1.28
KR	30	0.33	2.31	5.39	29.44	3.72	4.23	…	0.81
KR	60	0.43	1.40	9.49	10.38	1.35	2.92	…	0.64
KR	90	0.57	1.18	9.53	9.30	0.84	5.42	…	0.9
KR	180	0.43	1.88	6.65	14.61	2.03	2.75	…	1.03
PR	2	0.35	2.69	11.07	10.21	0.46	2.84	…	0.77
PR	15	0.45	4.65	8.77	11.77	0.33	4.03	…	0.82
PR	30	2.74	0.86	11.35	7.47	0.93	1.27	…	1.18
PR	60	0.87	0.60	16.12	5.22	0.28	3.95	…	1.05
PR	90	0.43	0.68	13.66	6.72	0.35	2.11	…	1.3
PR	180	0.31	1.68	11.77	8.17	0.43	1.47	…	1.76

Data matrix for PLS-DA

CR	KR	PR
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1

V1	V2	V3	V4	V5	V6	…	Vn
1.05	26.65	3.90	27.19	2.37	1.48	…	1.01
0.64	7.23	4.76	35.98	3.01	1.29	…	1.51
1.10	6.26	4.90	24.39	4.03	1.53	…	1.34
0.89	2.80	1.87	18.21	3.85	6.98	…	1.05
0.73	3.99	0.00	20.28	2.07	7.01	…	0.57
1.18	1.93	2.03	13.66	4.56	6.55	…	1.15
0.18	1.95	3.73	21.31	3.69	4.02	…	0.52
0.54	0.04	5.05	16.89	2.29	2.95	…	1.28
0.33	2.31	5.39	29.44	3.72	4.23	…	0.81
0.43	1.40	9.49	10.38	1.35	2.92	…	0.64
0.57	1.18	9.53	9.30	0.84	5.42	…	0.9
0.43	1.88	6.65	14.61	2.03	2.75	…	1.03
0.35	2.69	11.07	10.21	0.46	2.84	…	0.77
0.45	4.65	8.77	11.77	0.33	4.03	…	0.82
2.74	0.86	11.35	7.47	0.93	1.27	…	1.18
0.87	0.60	16.12	5.22	0.28	3.95	…	1.05
0.43	0.68	13.66	6.72	0.35	2.11	…	1.3
0.31	1.68	11.77	8.17	0.43	1.47	…	1.76

Results

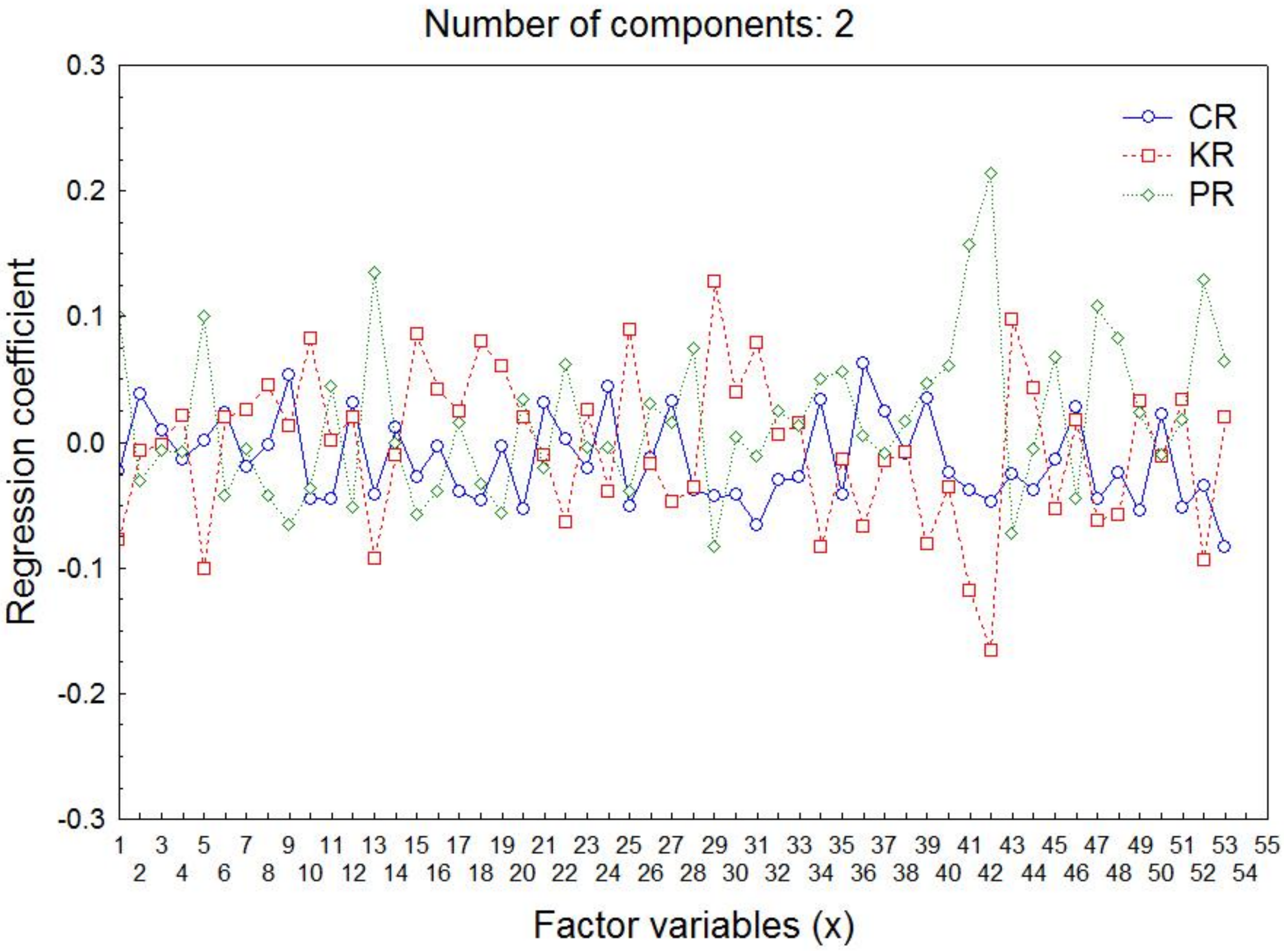


The plot of y scores along the first two components permitted to discriminate among classes.

Variables

Compound	IUPAC name	ID	PC1 loading	PC2 loading
acetone	propan-2-one	1	-	-
ethyl acetate	ethyl acetate	2	-	-
2-butanone	butan-2-one	3	-0.82	0.42
ethyl alcohol	ethanol	4	0.88	0.13
diacetyl	butane-2,3-dione	5	0.83	-0.23
2-pentanone	pentan-2-one	6	0.11	-0.75
1-ethanone	ethan-1-one	7	0.71	0.59
2-butanol	butan-2-ol	8	-0.87	0.30
3-methyl-(2 o 3)-heptanol	3-methylheptan-(2 o 3)-ol	9	-	-
thiophene	thiophene	10	0.80	0.15
1-propyl alcohol	propan-1-ol	11	-	-
ethyl butyrate	ethyl butanoate	12	-	-
methyl butyrate	methyl butanoate	13	-	-
2-hexanone	hexan-2-one	14	-	-
5-methyl-2-hexanone	5-methylhexan-2-one	15	-	-
hexanal	hexanal	16	0.73	-0.05
isobutyl alcohol	2-methylpropan-1-ol	17	-	-
3-methyl-2-butanol	3-methylbutan-2-ol	18	-	-
2-pentanol	pentan-2-ol	19	-	-
butyl alcohol	butan-1-ol	20	-	-
2-heptanone	heptan-2-one	21	0.03	-0.92
heptanal	heptanal	22	0.92	0.28
isoamyl alcohol	3-methylbutan-1-ol	23	-	-
ethyl hexanoate	ethyl hexanoate	24	-	-
2-methyl hexanoate	2-methyl hexanoate	25	-	-
1-pentanol	pentan-1-ol	26	0.93	0.21
2-octanone	octan-2-one	27	-0.16	-0.94
acetoin	3-hydroxybutan-2-one	28	0.87	0.08
octanal	octanal	29	-	-
1-heptanol	heptan-1-ol	30	-0.75	-0.03
isobutyl hexanoate	2-methylpropyl hexanoate	31	-0.70	0.42
hexanol	hexan-1-ol	32	0.76	0.53
2-methyl-3-pentanol	2-methylpentan-3-ol	33	-	-
2-nonanone	nonan-2-one	34	-0.26	-0.88
nonanal	nonanal	35	-	-
ethyl heptanoate	ethyl heptanoate	36	-	-
ethyl octanoate	ethyl octanoate	37	-	-
acetic acid	acetic acid	38	-0.87	0.09
8-nonen-2-one	non-8-en-2-one	39	-0.11	-0.82
propionic acid	propanoic acid	40	-0.84	0.27
2-nonenale	non-2-enal	41	-	-
benzaldehyde	benzaldehyde	42	-	-
2-undecanone	undecan-2-one	43	-	-
butyric acid	butanoic acid	44	-0.92	0.26
isovaleric acid	3-methylbutanoic acid	45	-0.79	0.22
2-thiopheneethanol	2-thiophen-2-yl ethanol	46	-	-
phenylacetaldehyde	2-phenylacetaldehyde	47	-	-
2-thiopheneacetic acid	2-thiophen-2-yl acetic acid	48	-	-
hexanoic acid	hexanoic acid	49	-0.86	0.28
phenethyl alcohol	2-phenylethanol	50	0.78	0.20
octanoic acid	octanoic acid	51	-	-
nonanoic acid	nonanoic acid	52	-	-
decanoic acid	decanoic acid	53	-0.06	0.75

Variables for classification



The regression coefficient between x and y could permit to individuate the most important variables for samples classification out of the 55 initial variables.

Results

PLS-DA was carried out on the data set.

2 components permitted to discriminate samples to classes and to individuate the most important variables for sample classification.

The results were well presented by plotting the y scores of samples on the plane defined by C1 and C2.