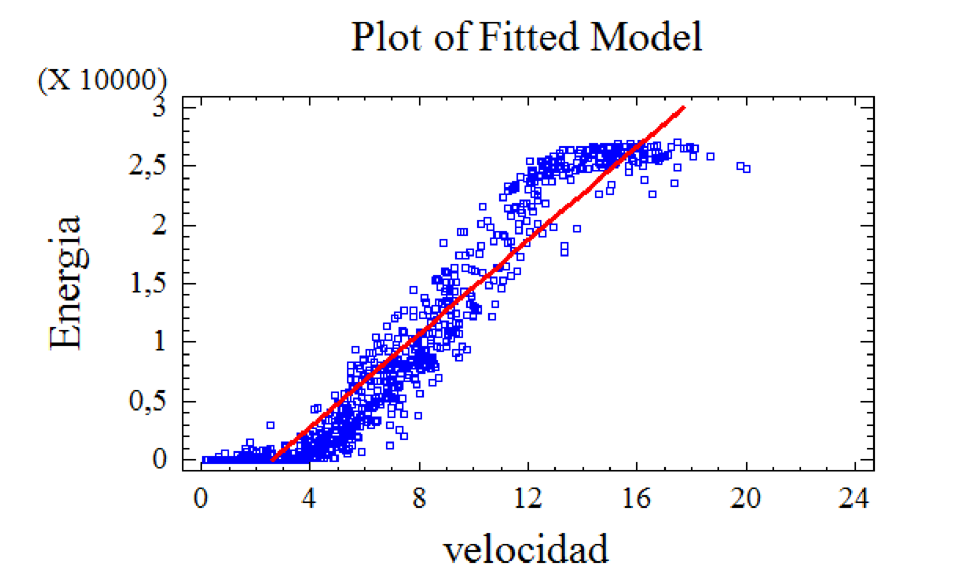
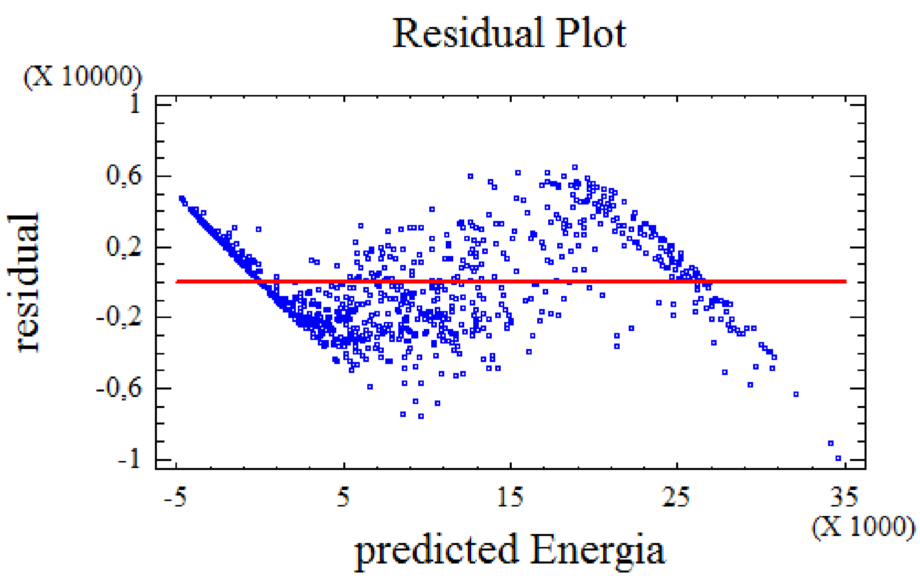
II. BIVARIATE DESCRIPTIVE STATISTICS





Chapter II: Bivariate Descriptive Statistics

1. **Introduction.**
2. **Bivariate Frequency Tables**
3. **Scatterplots**
4. **Measures of linear dependence**
5. **The regression line**

Chapter II: Bivariate Descriptive Statistics

Bivariate tables

We have, for each individual, two variables and to describe them we use a table with a double entrance

Example: for each car we have the number of cylinders and its manufacturing year (file cardata.sf)

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
4	17	12	25	22	28	104
5	1	1	1	0	0	3
6	12	6	2	7	3	30
8	6	10	0	1	0	17
Column Total	36	29	29	30	31	155

2. Bivariate Frequency Tables

Bivariate tables

We have, for each individual, two variables and to describe them we use a table with a double entrance

Example: for each car we have the number of cylinders and its manufacturing year (file cardata.sf)

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
4	17	12	25	22	28	104
5	1	1	1	0	0	3
6	12	6	2	7	3	30
8	6	10	0	1	0	17
Column Total	36	29	29	30	31	155

Each cell contains the **absolute joint frequency**

2. Bivariate Frequency Tables

Bivariate tables

We have, for each individual, two variables and to describe them we use a table with a double entrance

Example: for each car we have the number of cylinders and its manufacturing year (file cardata.sf)

	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
4	17	12	25	22	28	104
5	1	1	1	0	0	3
6	12	6	2	7	3	30
8	6	10	0	1	0	17
Column Total	36	29	29	30	31	155

Univariant: **absolute marginal frequencies**

2. Bivariate Frequency Tables

Bivariate tables

We have, for each individual, two variables and to describe them we use a table with a double entrance

Example: for each car we have the number of cylinders and its manufacturing year (file cardata.sf)

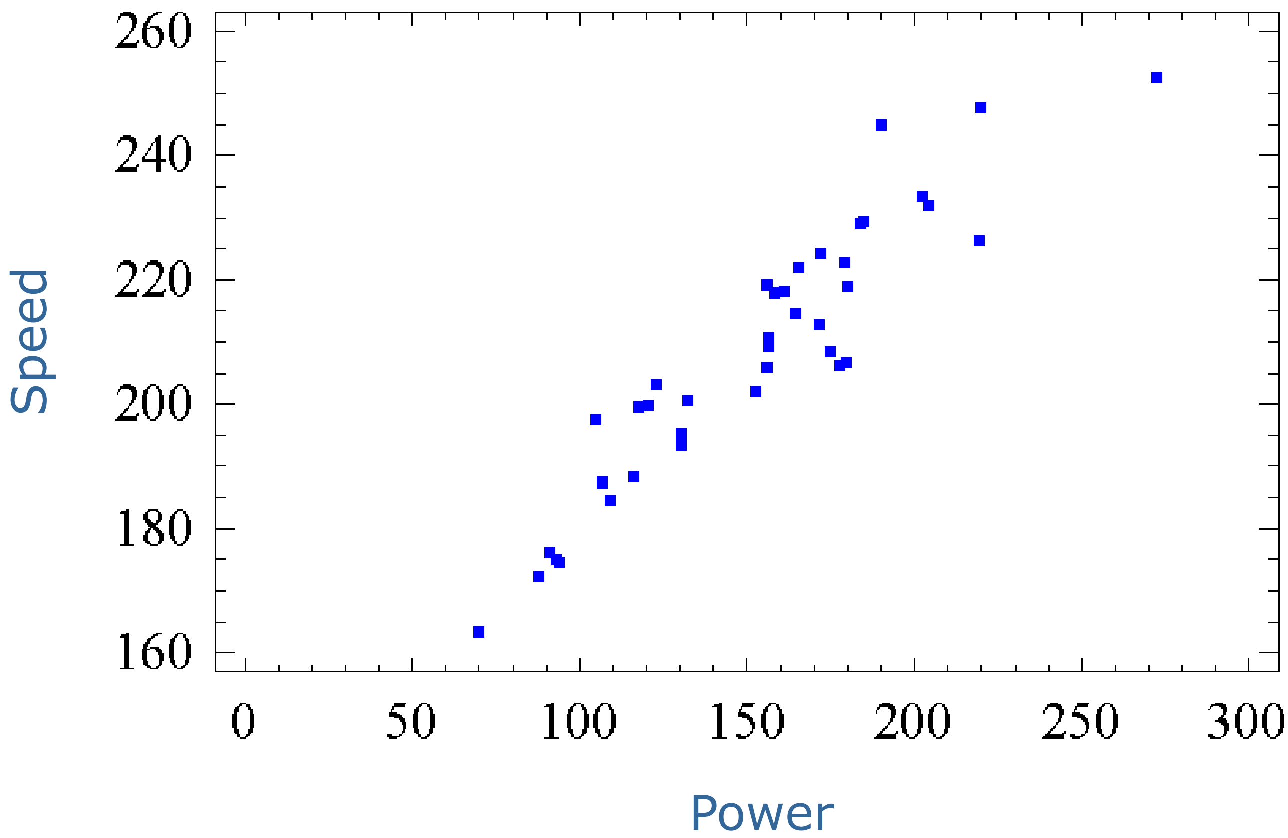
	78	79	80	81	82	Row Total
3	0	0	1	0	0	1
4	17	12	25	22	28	104
5	1	1	1	0	0	3
6	12	6	2	7	3	30
8	6	10	0	1	0	17
Column Total	36	29	29	30	31	155

Each row or column contains the **absolute conditional frequency** (with respect to the value of the row or column)

3. Scatterplot

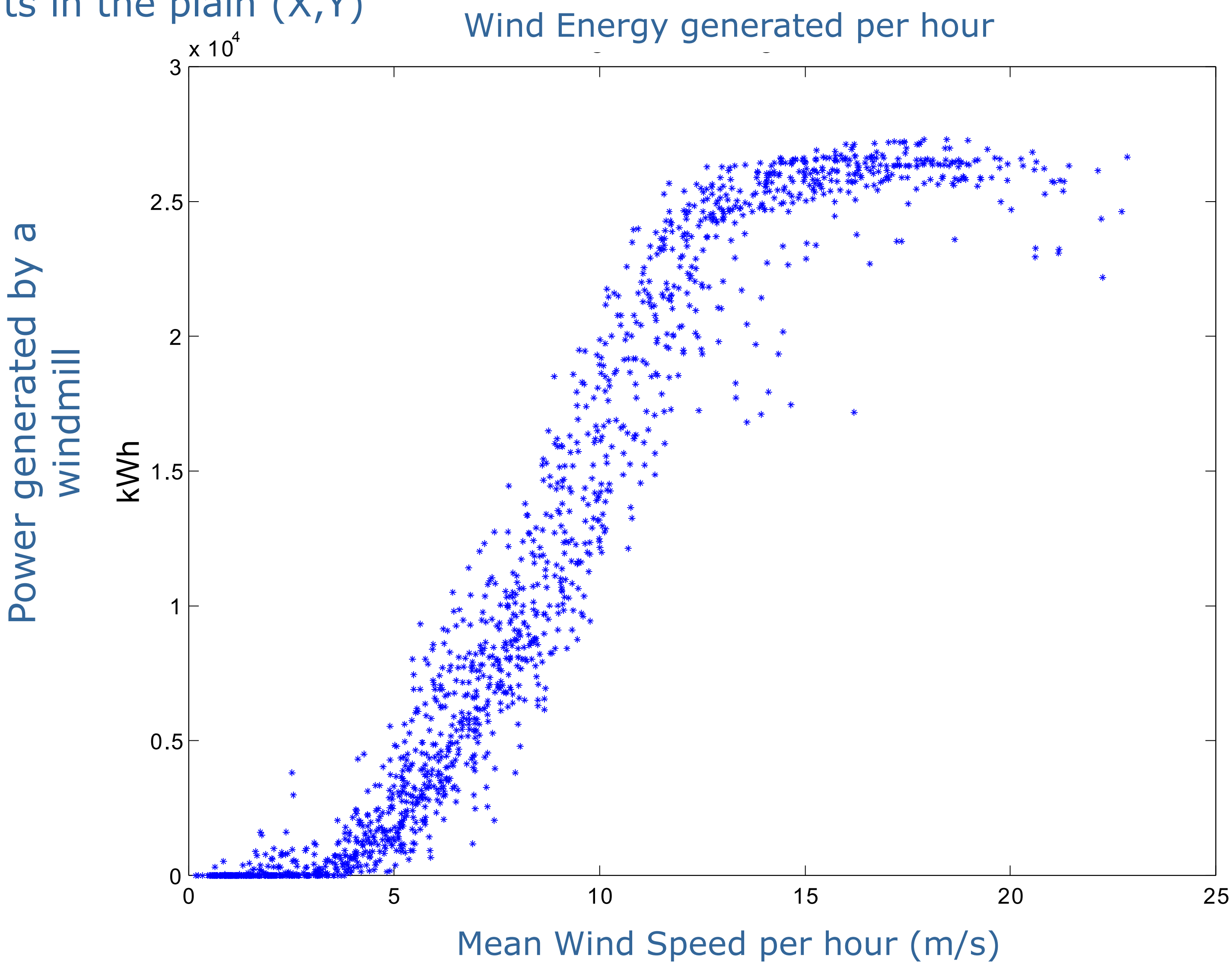
For each individual we have two data, X and Y, and we plot all data as points in the plain (X,Y)

Scatter-plot – Speed vs. Power



3. Scatterplot

For each individual we have two data, X and Y, and we plot all data as points in the plain (X,Y)



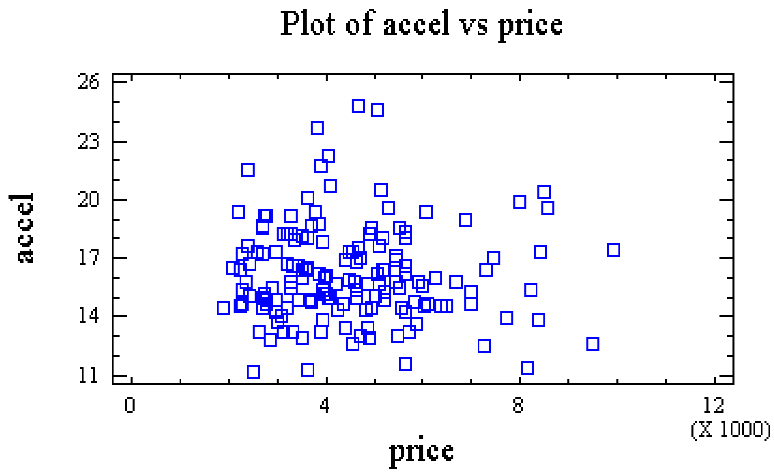
Chapter II: Bivariate Descriptive Statistics

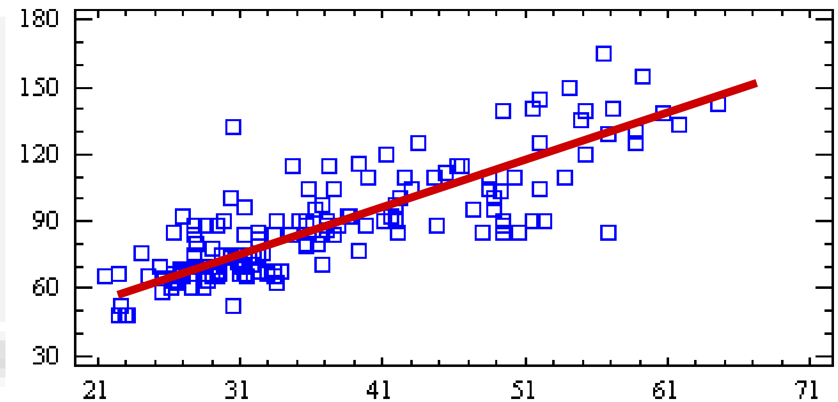
1. Introduction.
2. Bivariate Frequency Tables
3. Scatterplots
4. **Measures of linear dependence**
5. The regression line

3. Measures of linear dependence

Measures of linear dependence

- Covariance coefficient
- Correlation coefficient





Between these variables does not exist a linear relation

Between these variables exist a linear relation

The red line could be a good summary of that relation

For n individuals, we have data of 2 variables

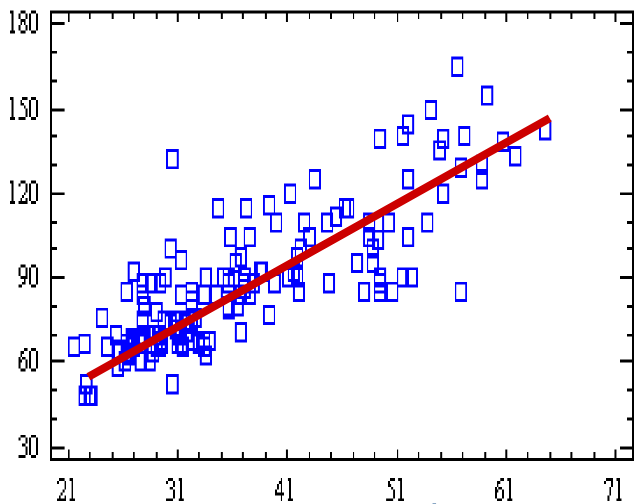
Individuals	**x**	**y**
1	x_1	y_1
2	x_2	y_2
:	:	:
n	x_n	y_n

Covariance

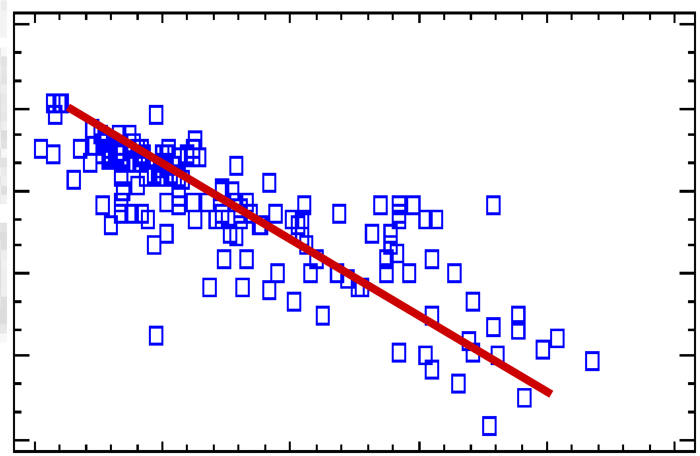
$$\mathrm{cov}(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}.$$

Correlation

$$r = r_{xy} = r(x,y) = \frac{\mathrm{cov}(x,y)}{s_x s_y}.$$



Covariance and positive correlation



Covariance and negative correlation

3. Measures of linear dependence

A usual way to describe this information in a (symmetric) matricial form is by using the

Covariance Matrix

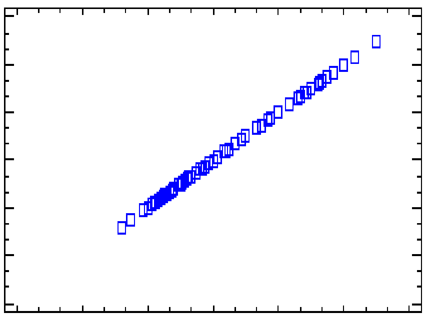
$$M = \begin{bmatrix} s_x^2 & \mathrm{cov}(x,y) \\ \mathrm{cov}(y,x) & s_y^2 \end{bmatrix}$$

Correlation Matrix

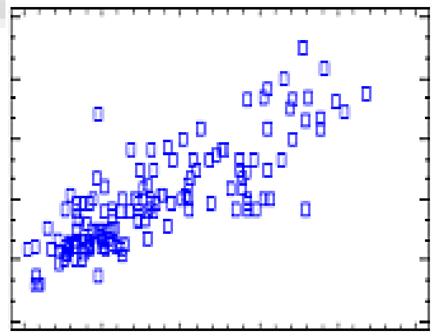
$$R = \begin{bmatrix} 1 & \mathrm{corr}(x,y) \\ \mathrm{corr}(y,x) & 1 \end{bmatrix}$$

$$\mathrm{cov}(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}. \qquad r = r_{xy} = r(x,y) = \frac{\mathrm{cov}(x,y)}{s_x s_y}.$$

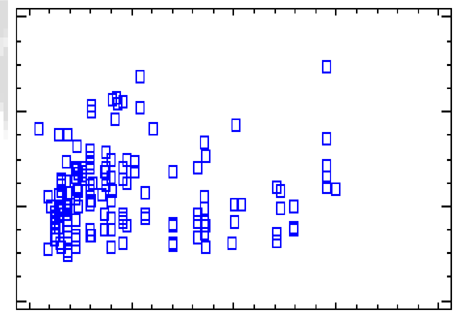
- The covariance depends upon the units in which x and y are measured
- The correlation is dimensionless. IT IS EASIER TO INTERPRET
- The correlation coefficient always is bounded $-1 \leq r \leq 1$



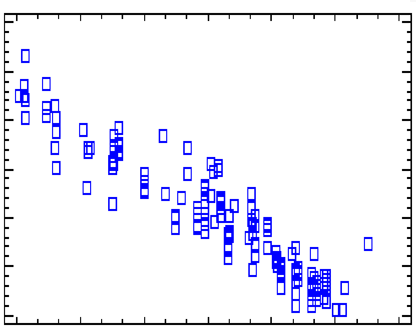
r=1



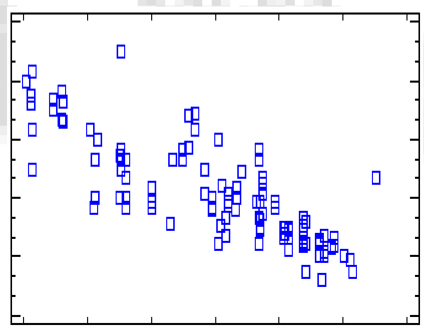
r=0.8



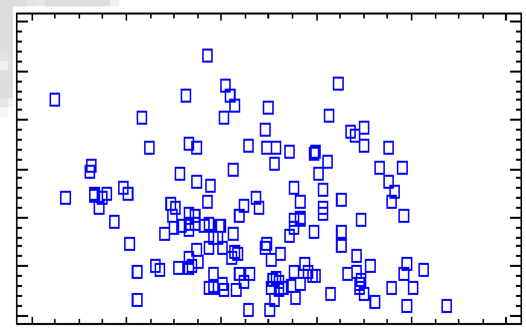
r=0.06



r=-0.94



r=-0.83

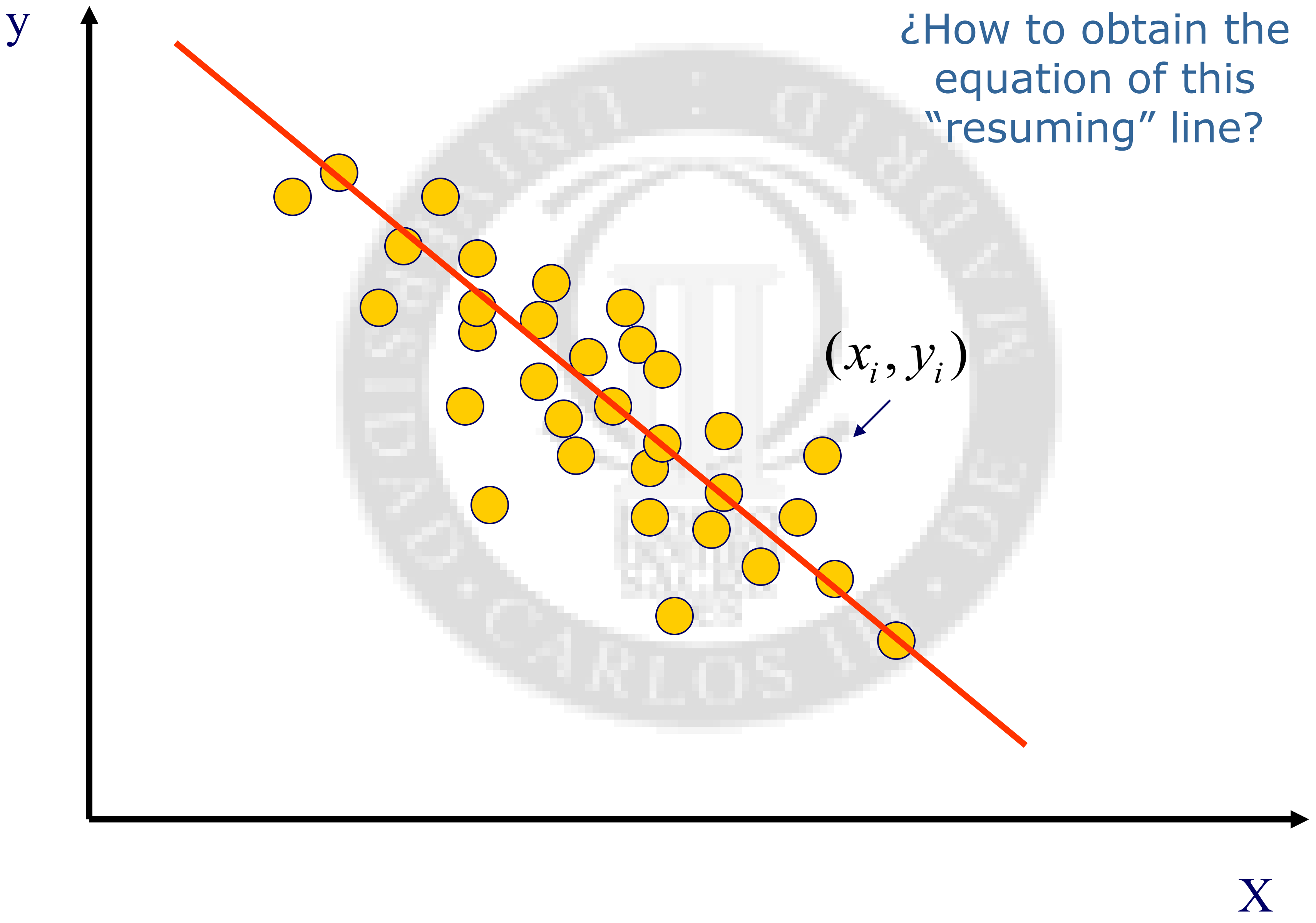


r=-0.08

Chapter II: Bivariate Descriptive Statistics

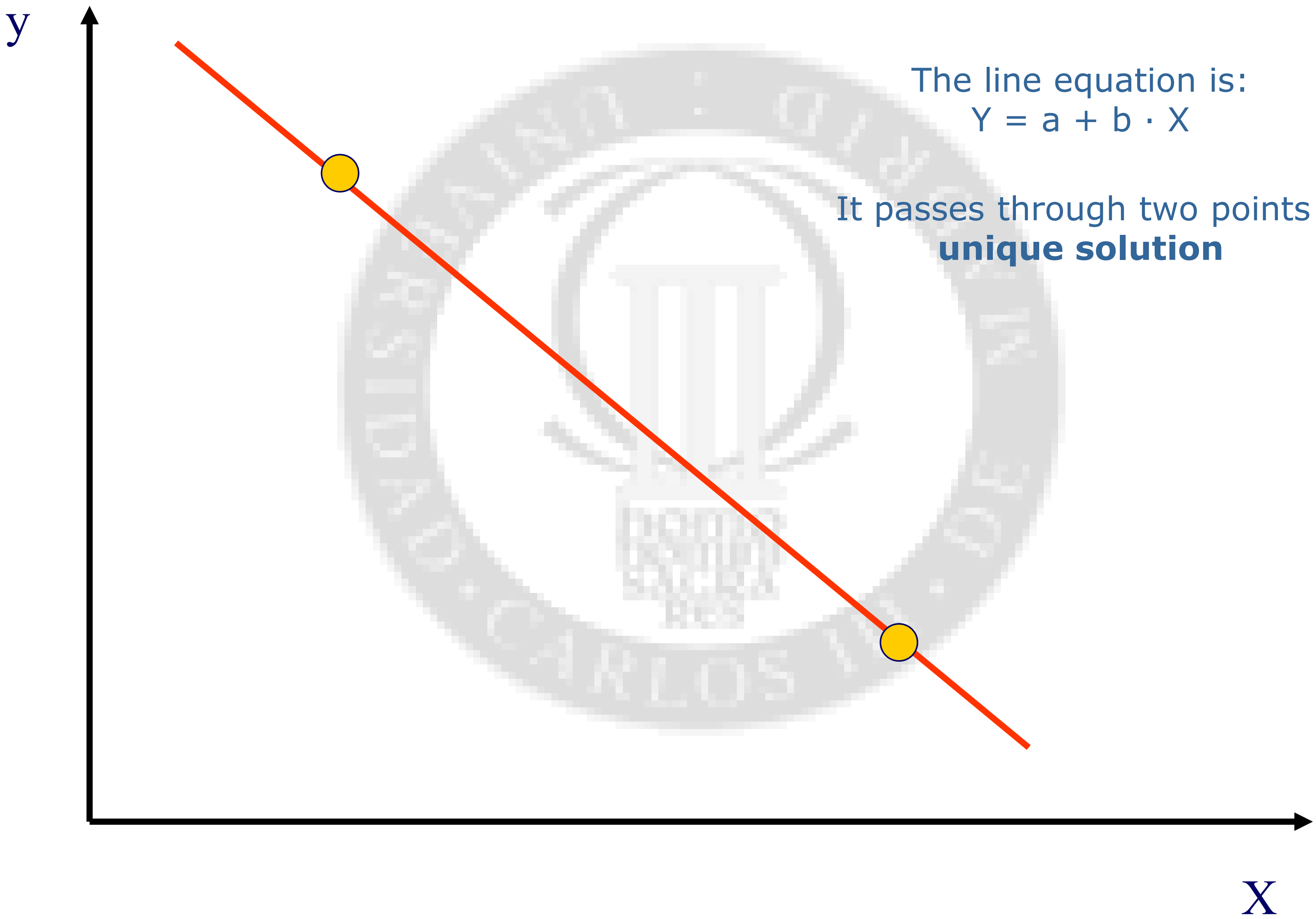
1. Introduction.
2. Bivariate Frequency Tables
3. Scatterplots
4. Measures of linear dependence
5. **The regression line**

4. The regression line



¿How to obtain the equation of this "resuming" line?

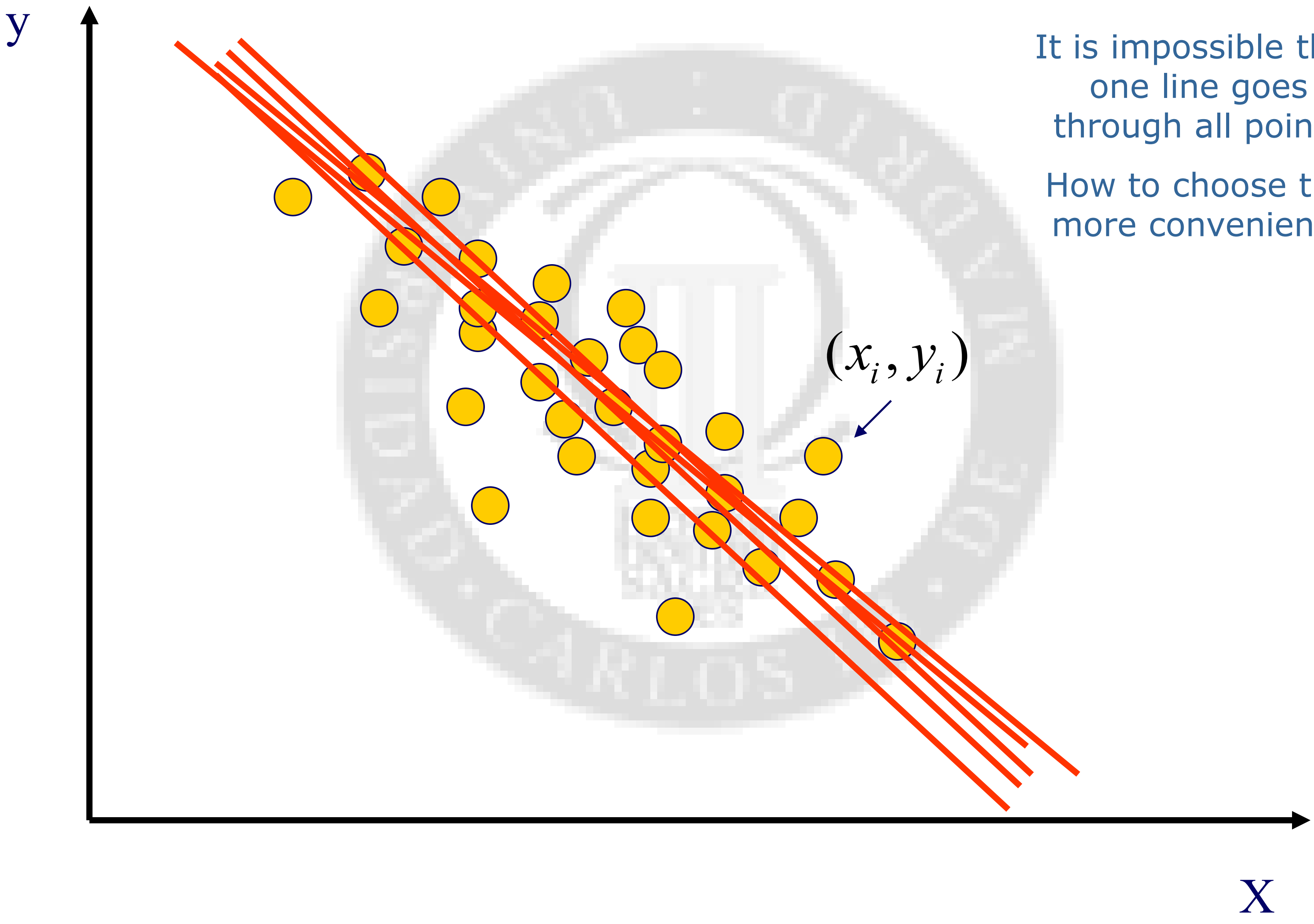
4. The regression line



The line equation is:
$Y = a + b \cdot X$

It passes through two points:
unique solution

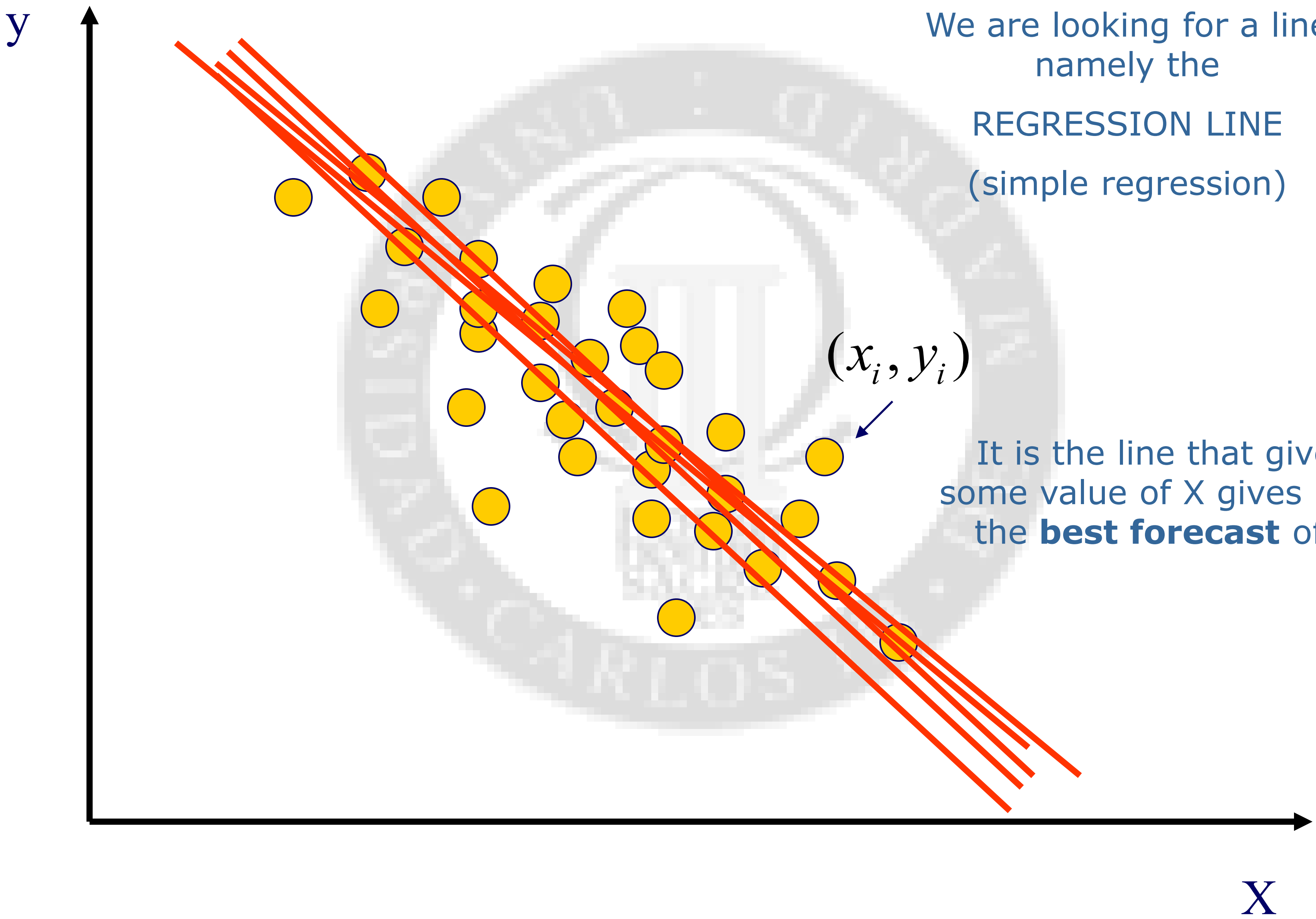
4. The regression line



It is impossible that one line goes through all points

How to choose the more convenient?

4. The regression line



We are looking for a line namely the

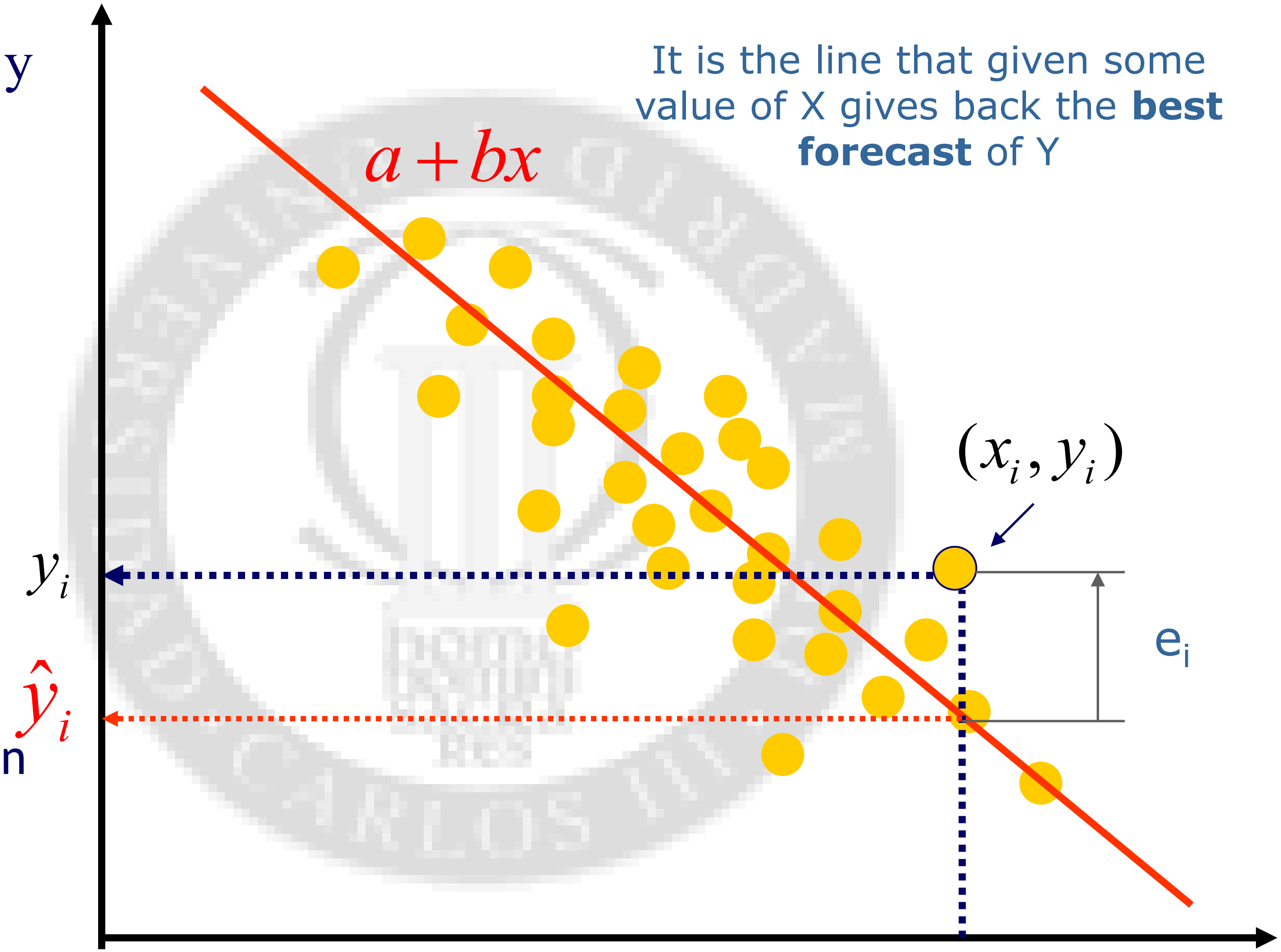
REGRESSION LINE

(simple regression)

It is the line that given some value of X gives back the **best forecast** of Y

4. The regression line

It is the line that given some value of X gives back the **best forecast** of Y



Actual value

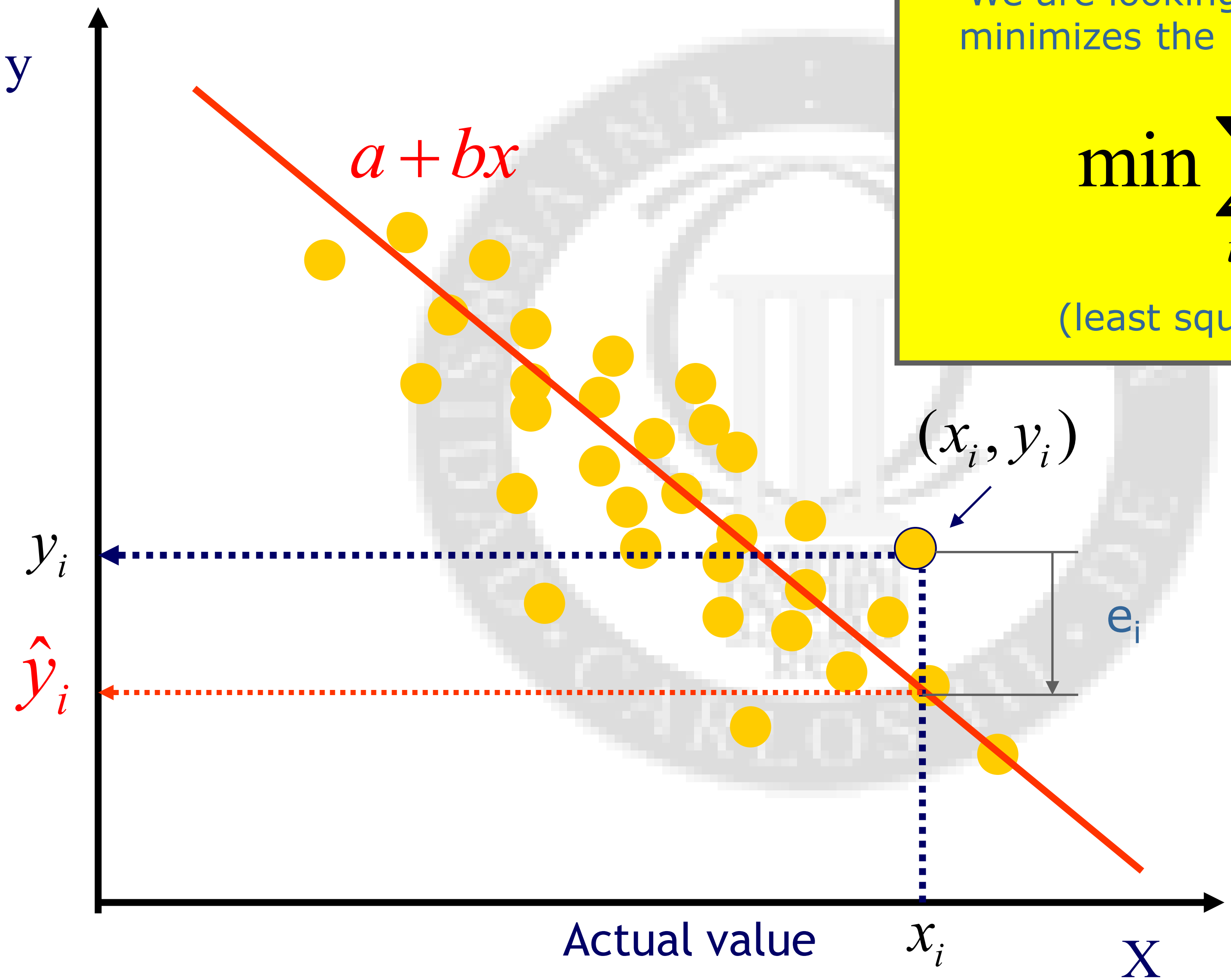
Predicted value by the regression line

Actual value x_i

X

Prediction errors: $e_i = y_i - \hat{y}_i$

4. The regression line



We are looking for the line that minimizes the prediction errors:

$$\min \sum_{i=1}^{N} e_i^2$$

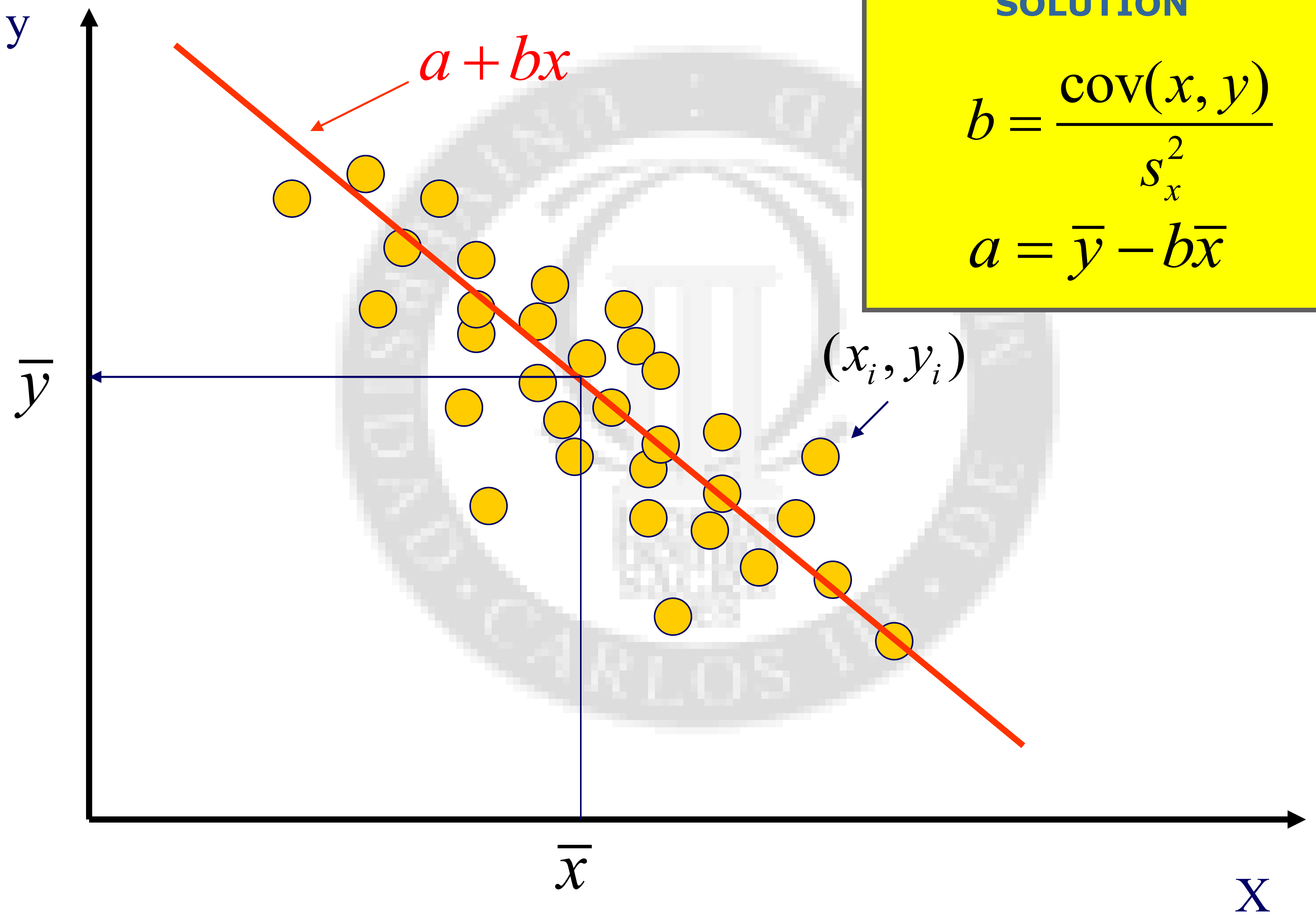
(least squares line)

4. The regression line

SOLUTION

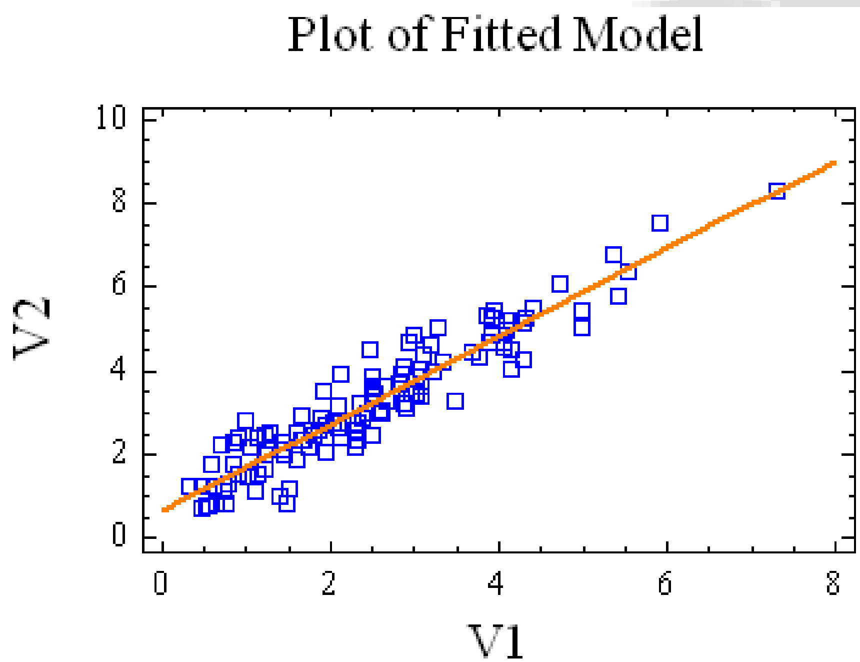
$$b = \frac{\operatorname{cov}(x, y)}{s_x^2}$$

$$a = \bar{y} - b\bar{x}$$



Example

The variable V1 is the wind speed registered in location 1, while the variable V2 is the speed registered at the same time in location 2. There are a total of 115 pairs of measures.



Loc.1:
mean: 2.51
variance: 1.91

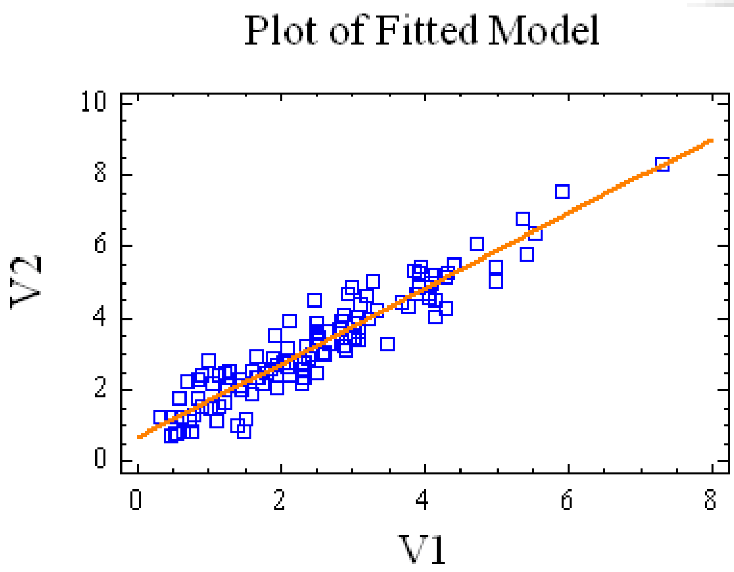
Loc.2:
mean: 3.28
variance: 2.36

cov (V1,V2)=1.995

In location 1 it is going to install a computer system to telemeasure the wind speed, but not for location 2. We want to calculate the regression line which allows to predict the speed in location 2 knowing the speed in location 1.

Example

The variable V1 is the wind speed registered in location 1, while the variable V2 is the speed registered at the same time in location 2. There are a total of 115 pairs of measures.



Loc.1:	Loc.2:
mean: 2.51	mean: 3.28
variance: 1.91	variance: 2.36

cov (V1,V2)=1.995

b=cov(x,y)/var(x)=1.995/1.91=1.045

$$a = \overline{y} - b\overline{x} = 3.28 - 1.045 \times 2.51 = 0.657$$

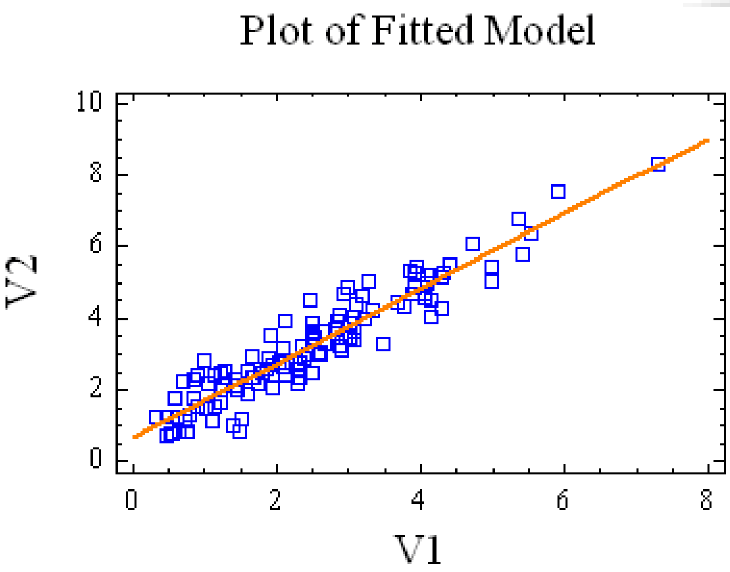
$$\hat{V}_2 = 0.657 + 1.045 \times V_1$$

If, for example, in Location 1 we measure a speed wind of 5 m/s, the wind speed prediction for Location 2 is

0.657+1.045x5=5.88 m/s

Example

The variable V1 is the wind speed registered in location 1, while the variable V2 is the speed registered at the same time in location 2. There are a total of 115 pairs of measures.



$$b = cov(x,y)/var(x) = 1.995/1.91 = 1.045$$

$$a = \bar{y} - b\bar{x} = 3.28 - 1.045 \times 2.51 = 0.657$$

$$\hat{V}_2 = 0.657 + 1.045 \times V_1$$

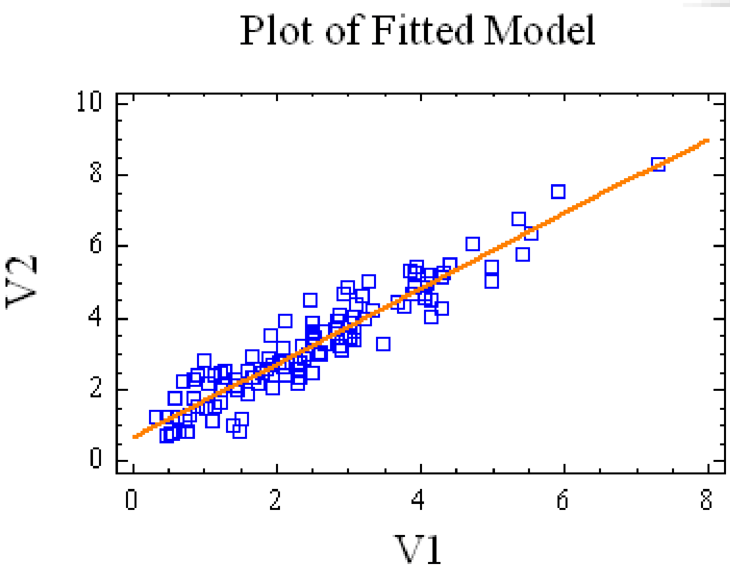
y=a+bx

Interpretation of b: if x increases of one unit y increases of b units

If in location 1 the wind speed increases of 1 m/s, in location 2 it is predicted to increase of 1.045 m/s

Example

The variable V1 is the wind speed registered in location 1, while the variable V2 is the speed registered at the same time in location 2. There are a total of 115 pairs of measures.



b=cov(x,y)/var(x)=1.995/1.91=1.045

$$a = \bar{y} - b\bar{x} = 3.28 - 1.045 \times 2.51 = 0.657$$

$$\hat{V}_2 = 0.657 + 1.045 \times V_1$$

y=a+bx
Interpretation of a: if the value of x is 0, y has value a

If in location 1 there is no wind, in location 2 there would be a wind speed of 0.657 m/s, that is a small value

4. The regression line

Evaluation of the regression

The regression to predict y starting from the value x will be good if:

1. The relation between x and y is linear
2. The linear relation is strong (big correlation)

Correlation and
Square of correlation coefficient R^2

Graphs

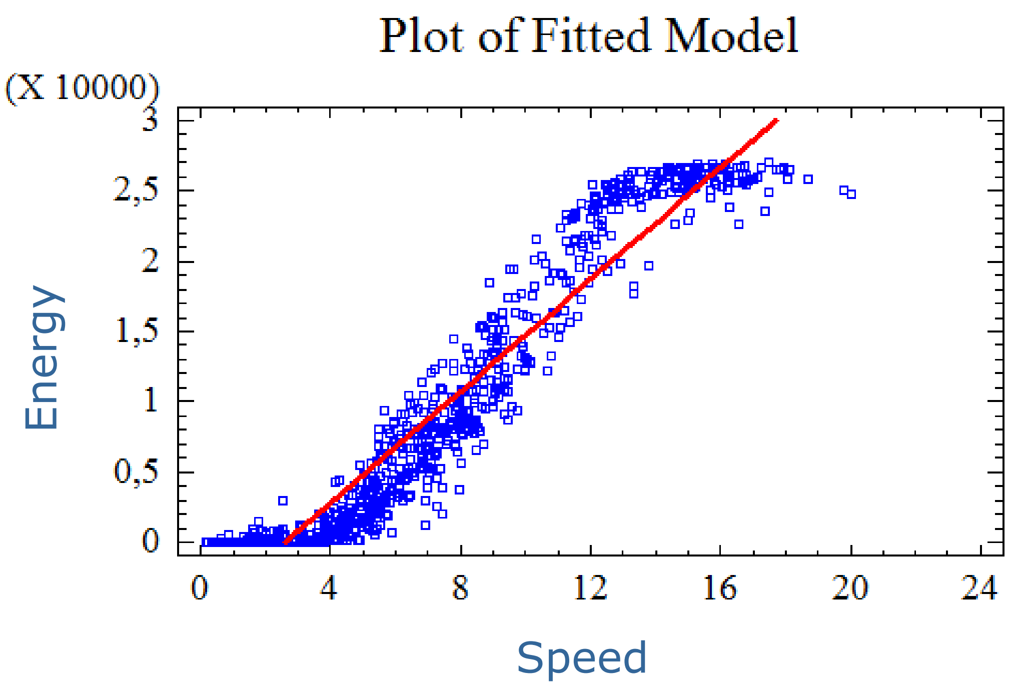
Evaluation of the regression

The regresion to predict y starting from the value x will be good if:

1. The relation between x and y is linear

1.1 Scatterplot
1.2 Graph of predictions vs. observation
1.3 Graph of residuals vs. predicted values

1.1 Scatterplot



Data: parqueeolico.sf3

This simple scatterplot shows that there is no linear realation.
The regression line gives bad predictions

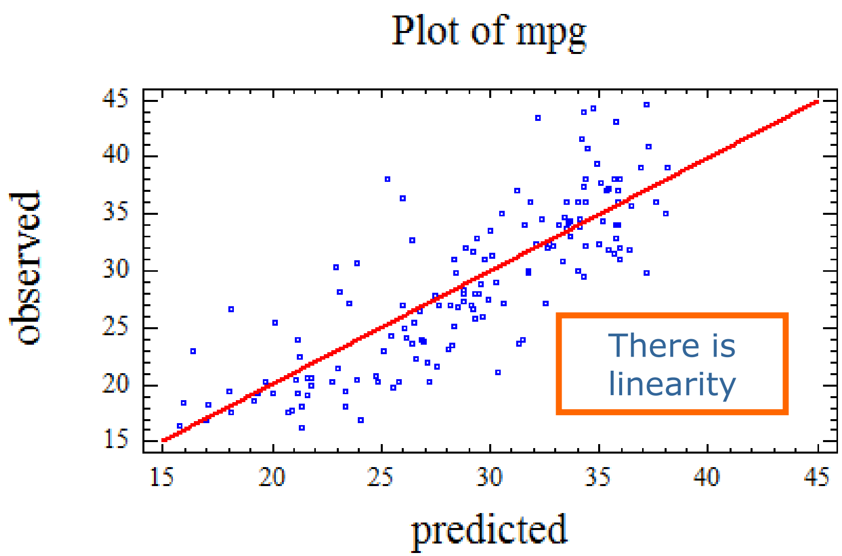
Evaluation of the regression

The regresion to predict y starting from the value x will be good if:

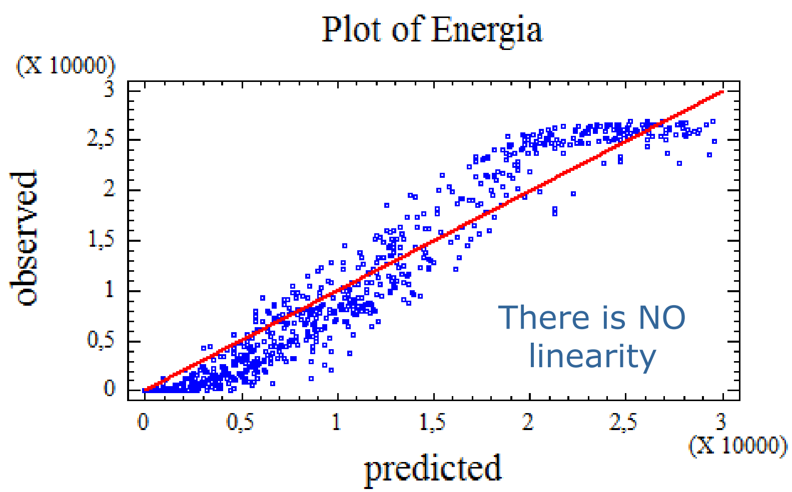
1. The relation between x and y is linear

1.1 Scatterplot
1.2 Graph of predictions vs. observation
1.3 Graph of residuals vs. predicted values

1.2 Graph of predictions vs observation



Cardata.sf: we want to express mpg (miles per gallon) as function of the weight (weight)

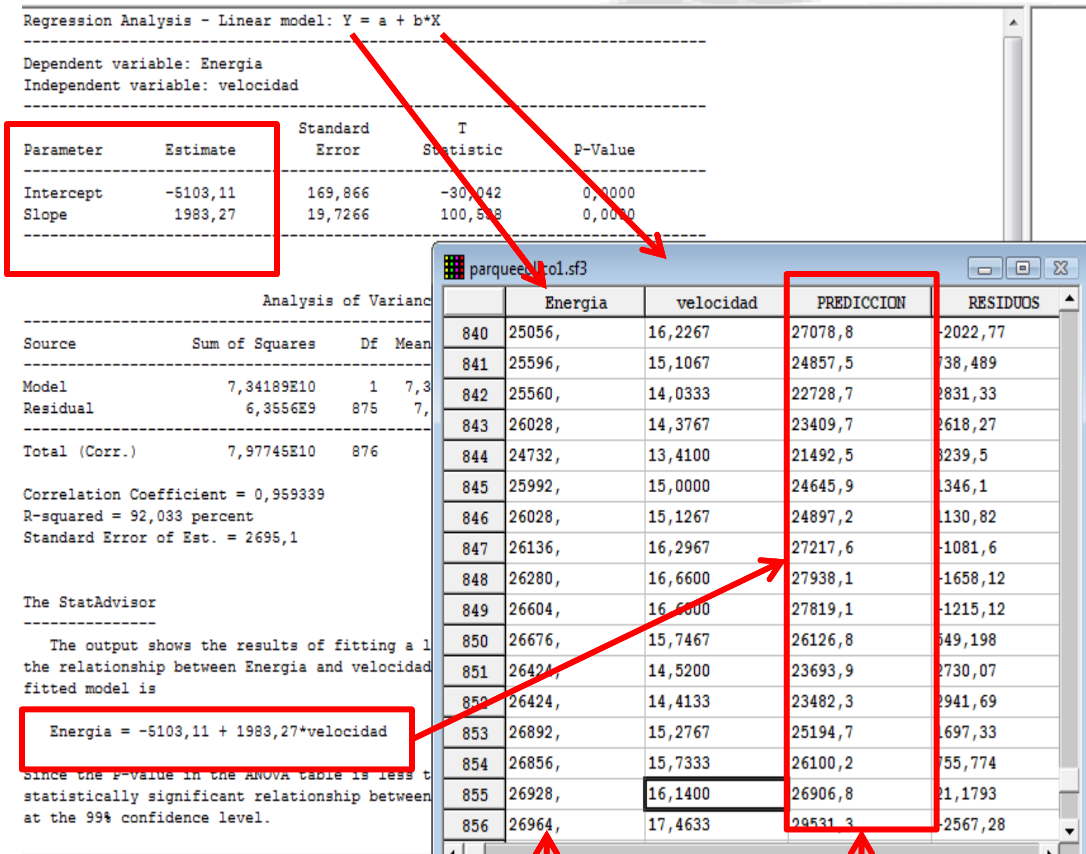


Parqueeolico.sf3: we want to express the generated energy as function of the wind speed

Evaluation of the regression

1.3 Graph of residuals vs. predicted values

It is the most important graph representation to evaluate a regression



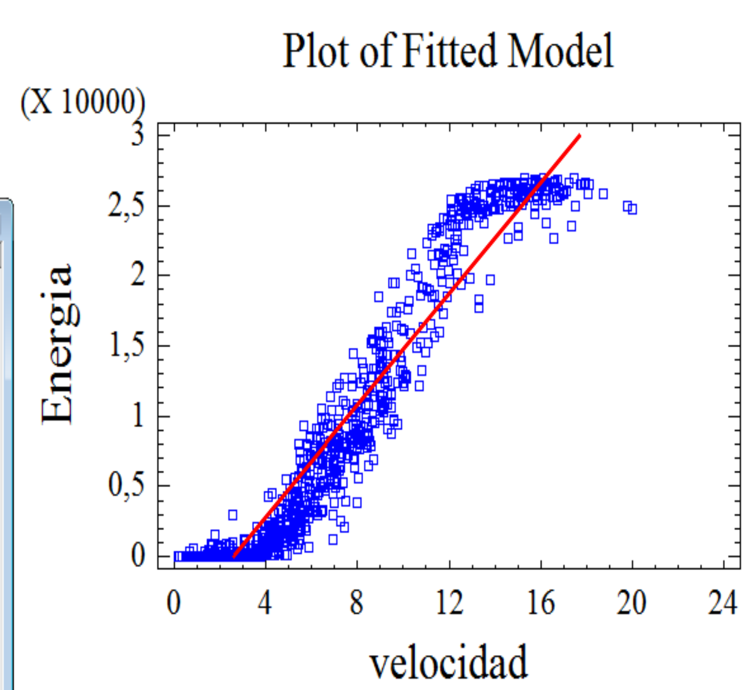
```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Energia
Independent variable: velocidad
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept       -5103,11       169,866       -30,042          0,0000
Slope            1983,27       19,7266       100,538          0,0000
-----------------------------------------------------------------------------

                         Analysis of Varianc
-----------------------------------------
Source             Sum of Squares     Df  Mean
-----------------------------------------
Model                  7,34189E10      1   7,3
Residual                6,3556E9     875   7,
-----------------------------------------
Total (Corr.)          7,97745E10    876

Correlation Coefficient = 0,959339
R-squared = 92,033 percent
Standard Error of Est. = 2695,1


The StatAdvisor
---------------
   The output shows the results of fitting a l
the relationship between Energia and velocidad
fitted model is

   Energia = -5103,11 + 1983,27*velocidad

Since the P-value in the ANOVA table is less t
statistically significant relationship between
at the 99% confidence level.
```

parqueeolico1.sf3

	Energia	velocidad	PREDICCION	RESIDUOS
840	25056,	16,2267	27078,8	-2022,77
841	25596,	15,1067	24857,5	738,489
842	25560,	14,0333	22728,7	2831,33
843	26028,	14,3767	23409,7	2618,27
844	24732,	13,4100	21492,5	3239,5
845	25992,	15,0000	24645,9	1346,1
846	26028,	15,1267	24897,2	1130,82
847	26136,	16,2967	27217,6	-1081,6
848	26280,	16,6600	27938,1	-1658,12
849	26604,	16,6000	27819,1	-1215,12
850	26676,	15,7467	26126,8	549,198
851	26424,	14,5200	23693,9	2730,07
852	26424,	14,4133	23482,3	2941,69
853	26892,	15,2767	25194,7	1697,33
854	26856,	15,7333	26100,2	755,774
855	26928,	16,1400	26906,8	21,1793
856	26964,	17,4633	29531,3	-2567,28

Observados values

Predicted values



Evaluation of the regression

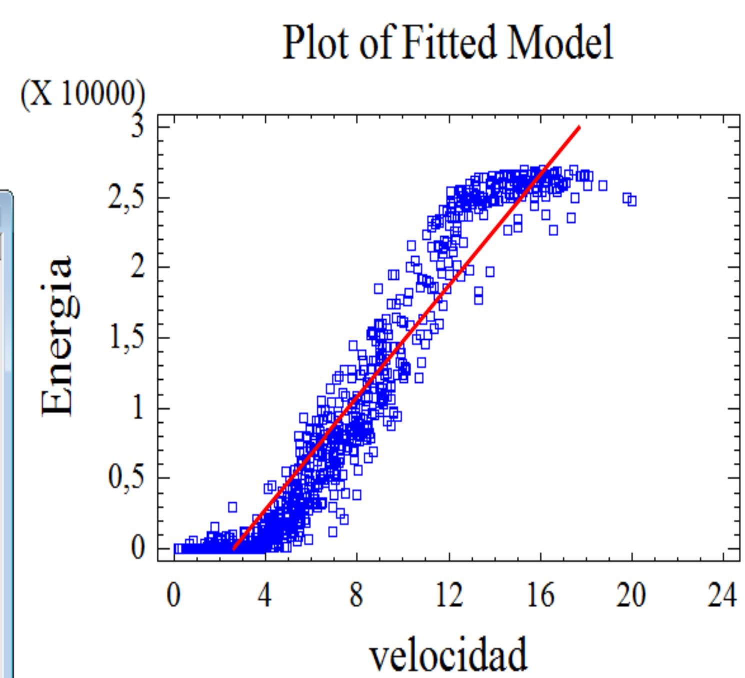
1.3 Graph of residuals vs. predicted values

It is the most important graph representation to evaluate a regression

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Energia
Independent variable: velocidad
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept       -5103,11       169,866        -30,042         0,0000
Slope            1983,27       19,7266        100,538         0,0000
-----------------------------------------------------------------------------


                              Analysis of Varianc
-----------------------------------------------
Source            Sum of Squares     Df  Mean
-----------------------------------------------
Model                7,34189E10       1   7,3
Residual              6,3556E9      875    7,
-----------------------------------------------
Total (Corr.)        7,97745E10     876

Correlation Coefficient = 0,959339
R-squared = 92,033 percent
Standard Error of Est. = 2695,1


The StatAdvisor
---------------
   The output shows the results of fitting a l
the relationship between Energia and velocidad
fitted model is

   Energia = -5103,11 + 1983,27*velocidad

Since the P-value in the ANOVA table is less t
statistically significant relationship between
at the 99% confidence level.
```

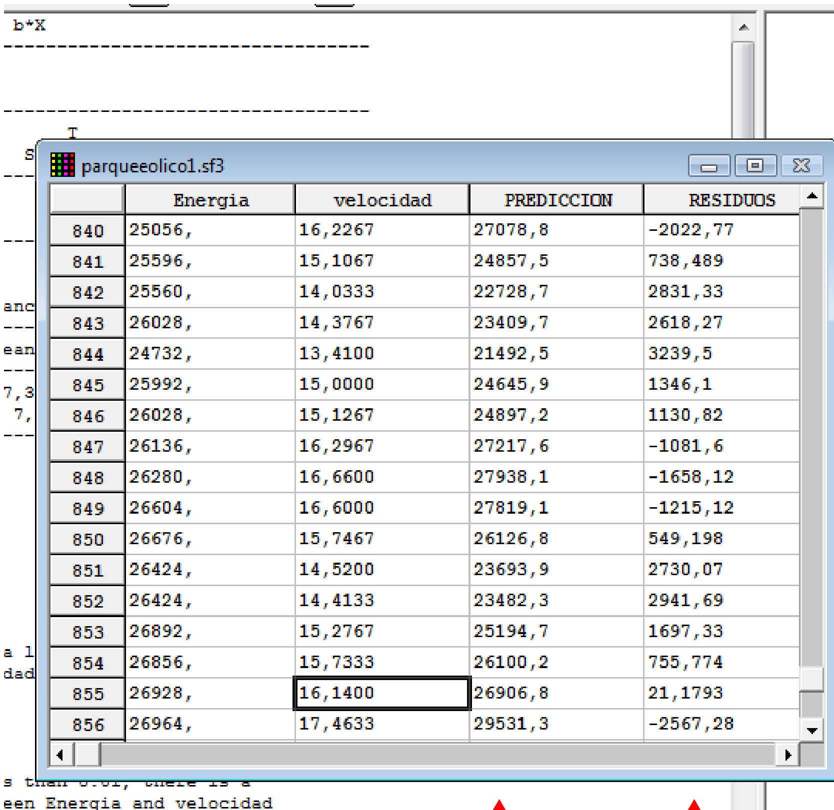
parqueeolico1.sf3

	Energia	velocidad	PREDICCION	RESIDUOS
840	25056,	16,2267	27078,8	-2022,77
841	25596,	15,1067	24857,5	738,489
842	25560,	14,0333	22728,7	2831,33
843	26028,	14,3767	23409,7	2618,27
844	24732,	13,4100	21492,5	3239,5
845	25992,	15,0000	24645,9	1346,1
846	26028,	15,1267	24897,2	1130,82
847	26136,	16,2967	27217,6	-1081,6
848	26280,	16,6600	27938,1	-1658,12
849	26604,	16,6000	27819,1	-1215,12
850	26676,	15,7467	26126,8	549,198
851	26424,	14,5200	23693,9	2730,07
852	26424,	14,4133	23482,3	2941,69
853	26892,	15,2767	25194,7	1697,33
854	26856,	15,7333	26100,2	755,774
855	26928,	16,1400	26906,8	21,1793
856	26964,	17,4633	29531,3	-2567,28



Predicted values

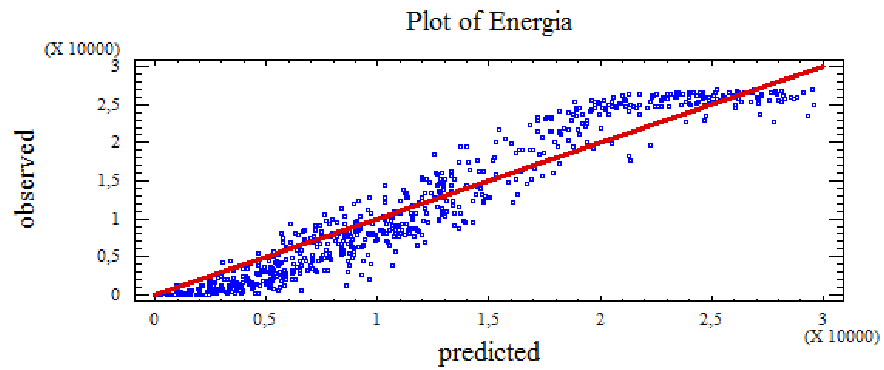
Energia-PREDICCION

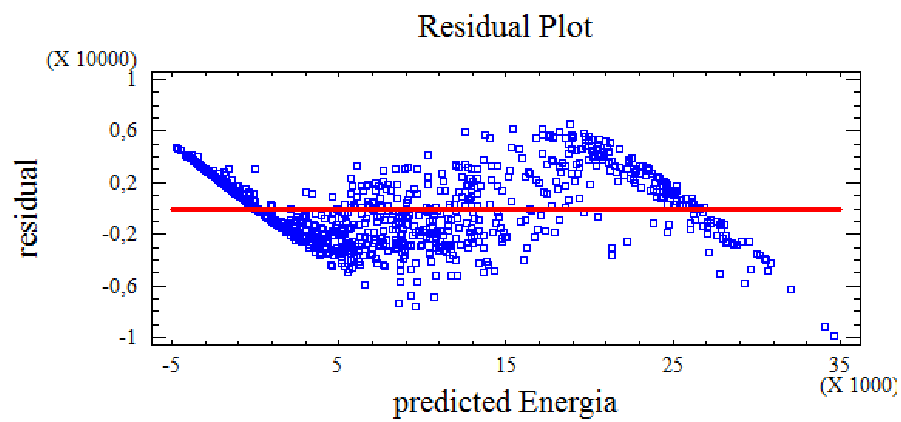


parqueeolico1.sf3

	Energia	velocidad	PREDICCION	RESIDUOS
840	25056,	16,2267	27078,8	-2022,77
841	25596,	15,1067	24857,5	738,489
842	25560,	14,0333	22728,7	2831,33
843	26028,	14,3767	23409,7	2618,27
844	24732,	13,4100	21492,5	3239,5
845	25992,	15,0000	24645,9	1346,1
846	26028,	15,1267	24897,2	1130,82
847	26136,	16,2967	27217,6	-1081,6
848	26280,	16,6600	27938,1	-1658,12
849	26604,	16,6000	27819,1	-1215,12
850	26676,	15,7467	26126,8	549,198
851	26424,	14,5200	23693,9	2730,07
852	26424,	14,4133	23482,3	2941,69
853	26892,	15,2767	25194,7	1697,33
854	26856,	15,7333	26100,2	755,774
855	26928,	16,1400	26906,8	21,1793
856	26964,	17,4633	29531,3	-2567,28

een Energia and velocidad





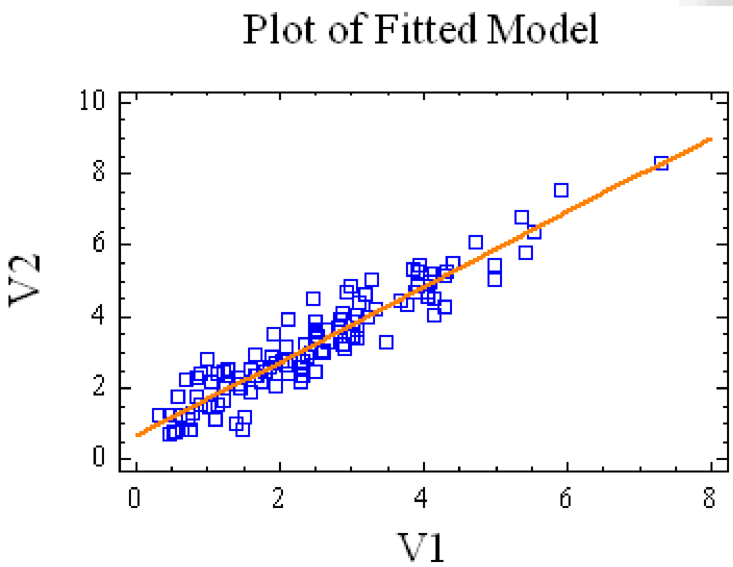
Predicted values

Energia-PREDICCION

The absence of linearity is clear.
In this case the regression line is not an useful information for prediction.

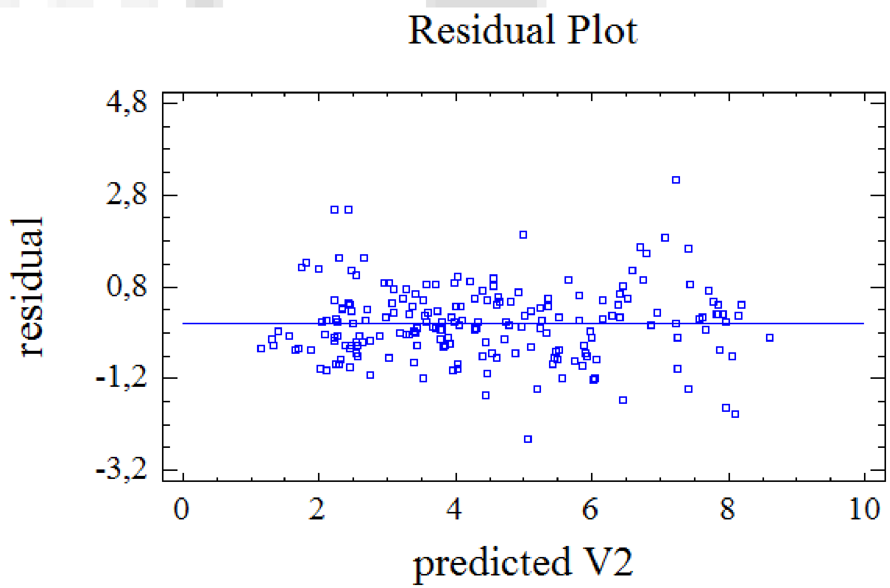
Example

The variable V1 is the wind speed registered in location 1, while the variable V2 is the speed registered at the same time in location 2. There are a total of 115 pairs of measures.

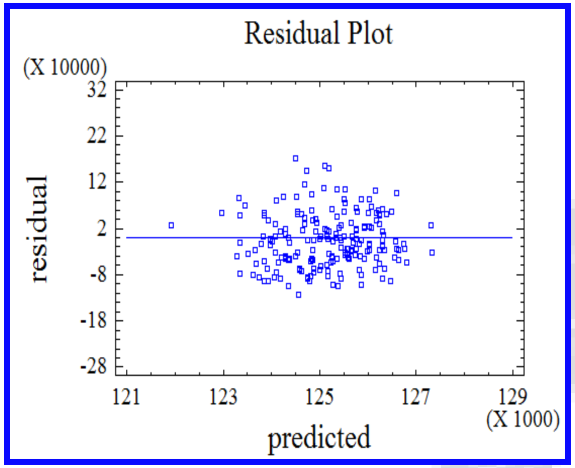


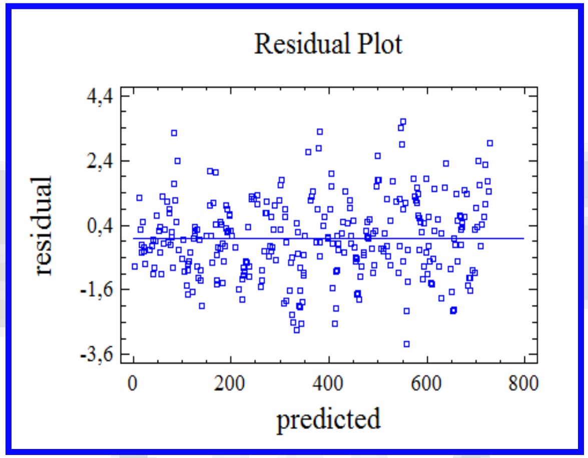
$$\hat{V}_2 = 0.657 + 1.045 \times V_1$$

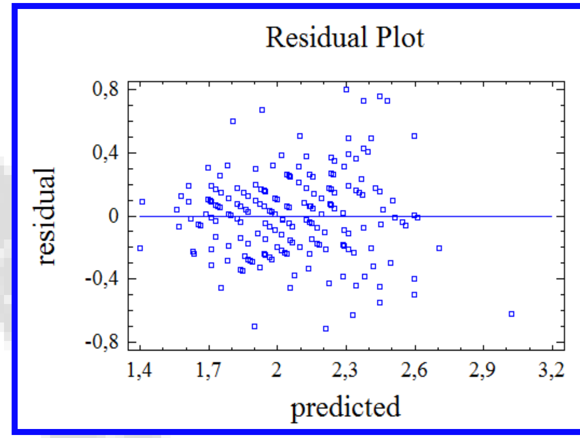
These residuals do not show any clear structure. This means that the linear model is aduquate



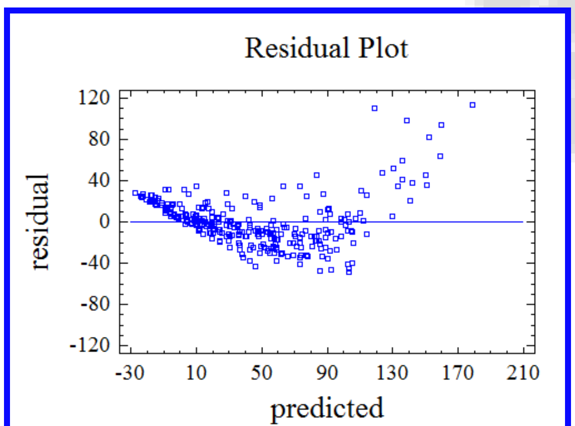
These residual graphs DO be acceptable

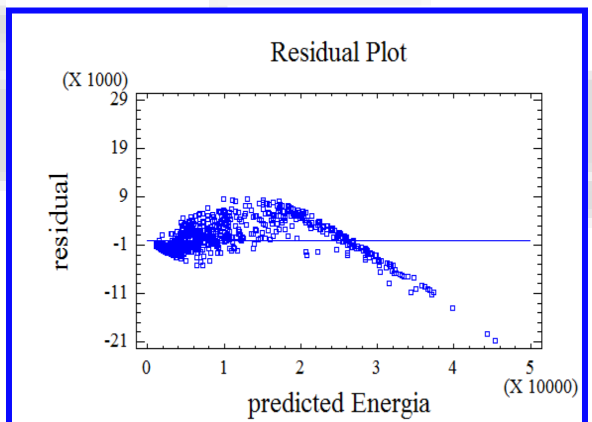


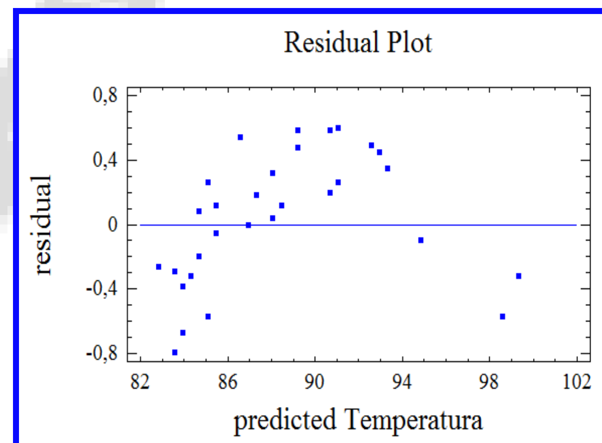




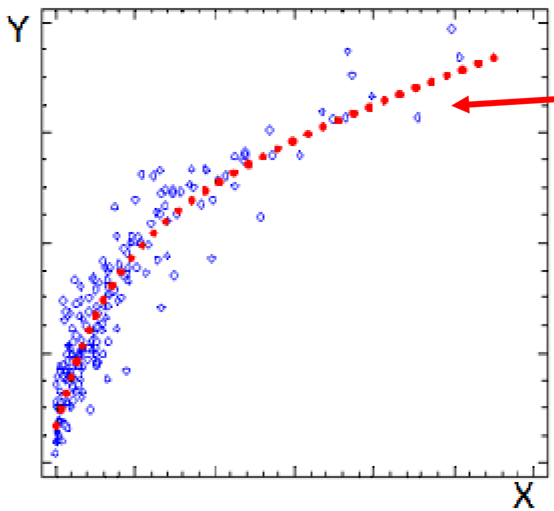
These residual graphs DO NOT be acceptable





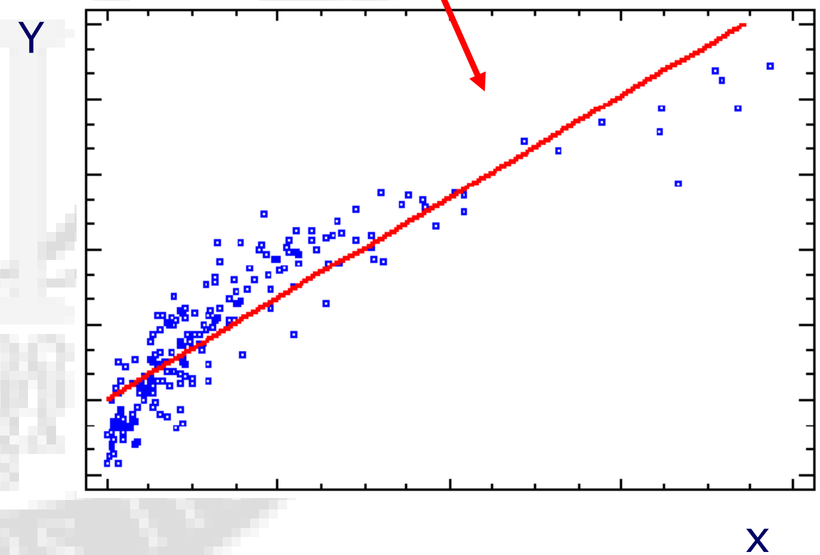


Some types of no linearity can be corrected by transforming the variable

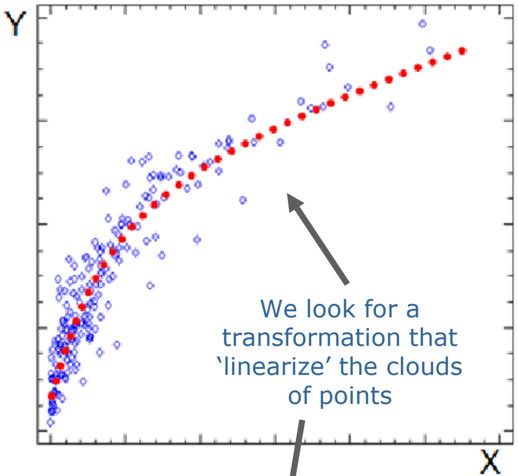


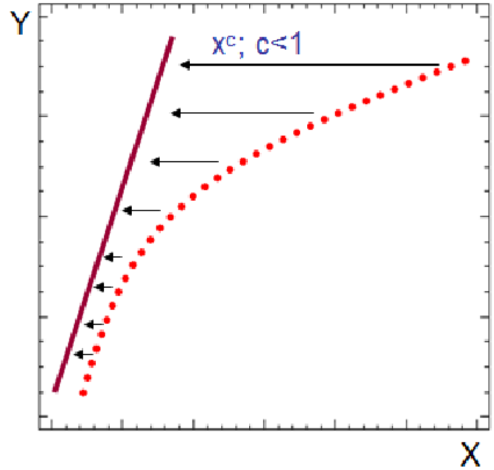
This curve is the one we would like to use to resume the relation…

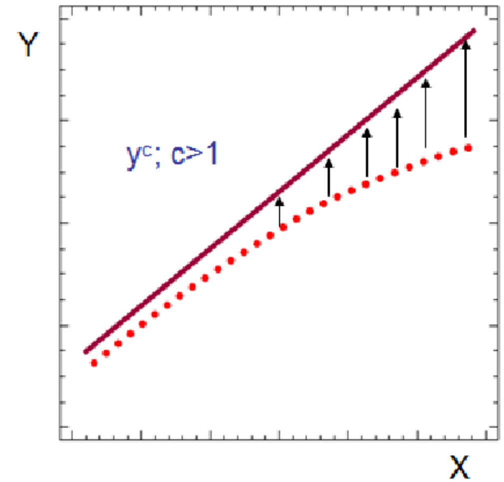
… but the simple regression tecnique will give only this type of solutions



We look for an additional variables
$y^*=f(y)$ and $x^*=g(x)$
such that between them there is a linear relation







$y=a+bx^c$

$y^c=a+bx$

If c<1, the biggest values are reduced the most.
In this case applied to X we will 'linearize' the curve.

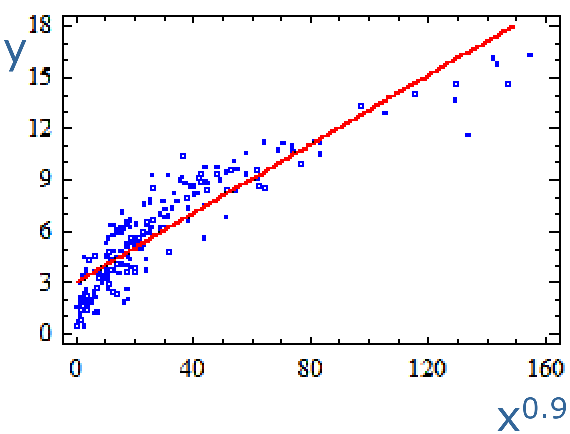
If c>1, we get the opposite effect: the smallest values are expanded the most.
In this case applied to Y we will 'linearize' the curve as well.

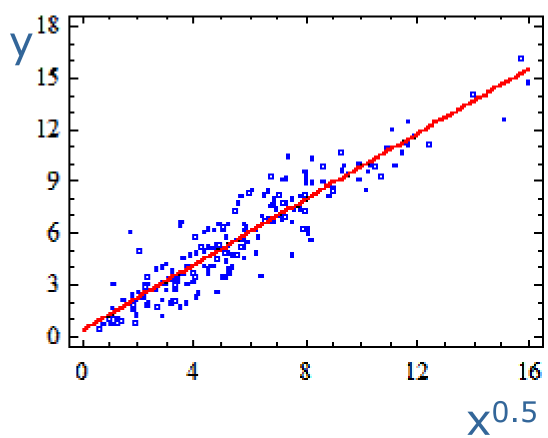
$y=a+bx^c$

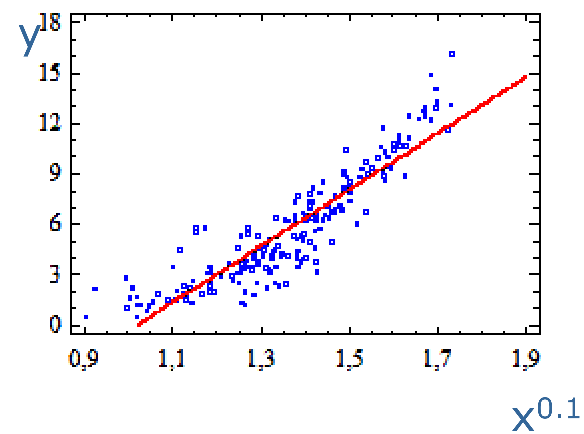
c=0.9 **Insufficient**

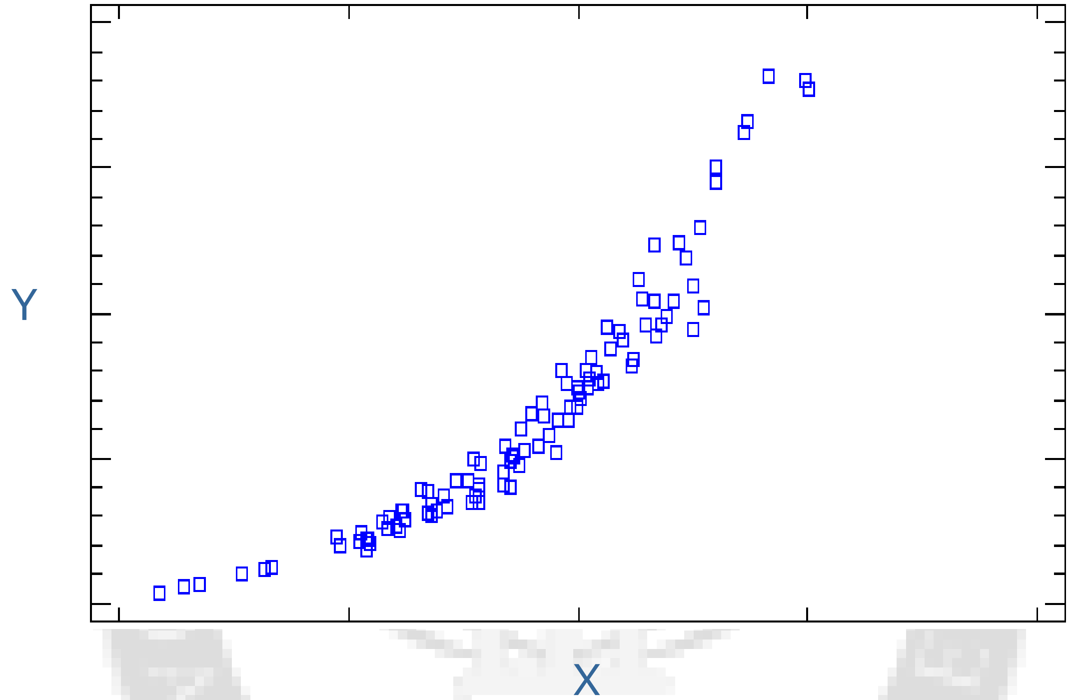
c=0.5 **Perfect!**

c=0.1 **Too much**









Assuming positive values. How the cloud of point will change if we use the following transormations:

- y^2
- x^2
- $y^{0.5}$
- $\log(y)$

Evaluation of the regression

The regression to predict y starting from the value x will be good if:

1. The relation between x and y is linear ← Graphs
2. The linear relation is strong (big correlation)

Correlation and Square of correlation coefficient R^2

$$\boxed{R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.}$$

- Between 0 y 1
- $R^2 = \mathrm{corr}(x, y)^2$
- The square of correlation (or coefficient of determination) tells us which proportion of the dispersion of the dependent variable y is used to compute the regression line