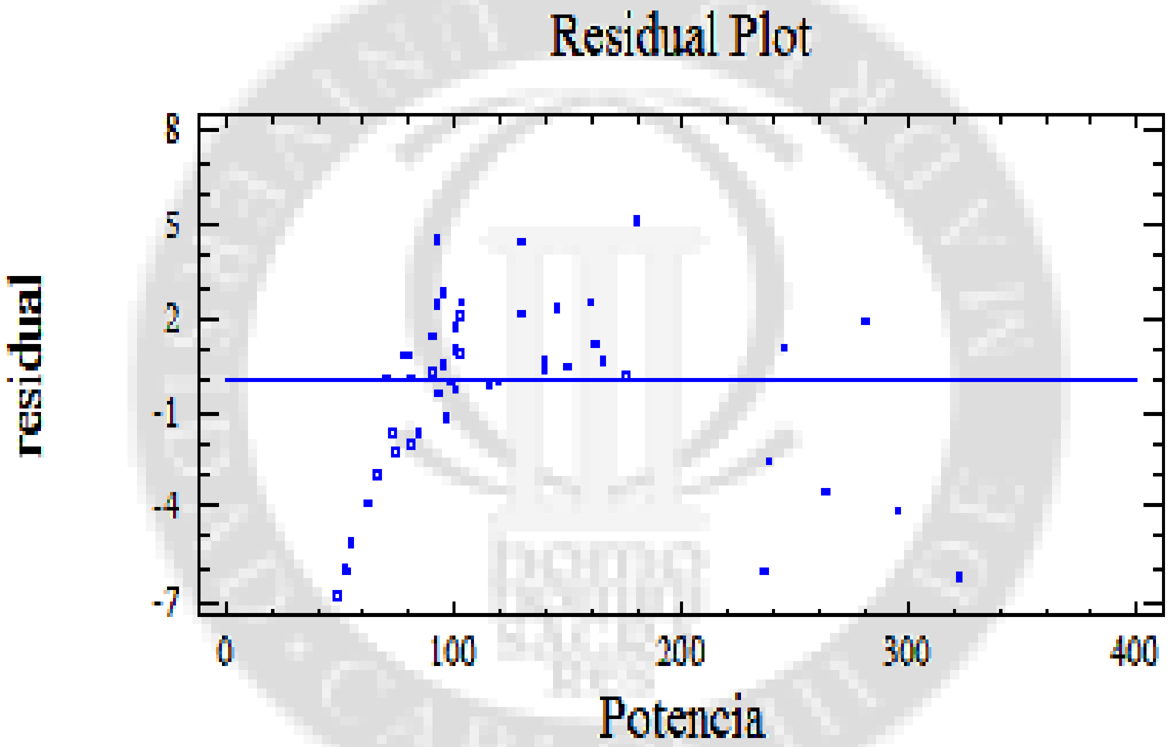
IX. Introduction to Multiple Regression



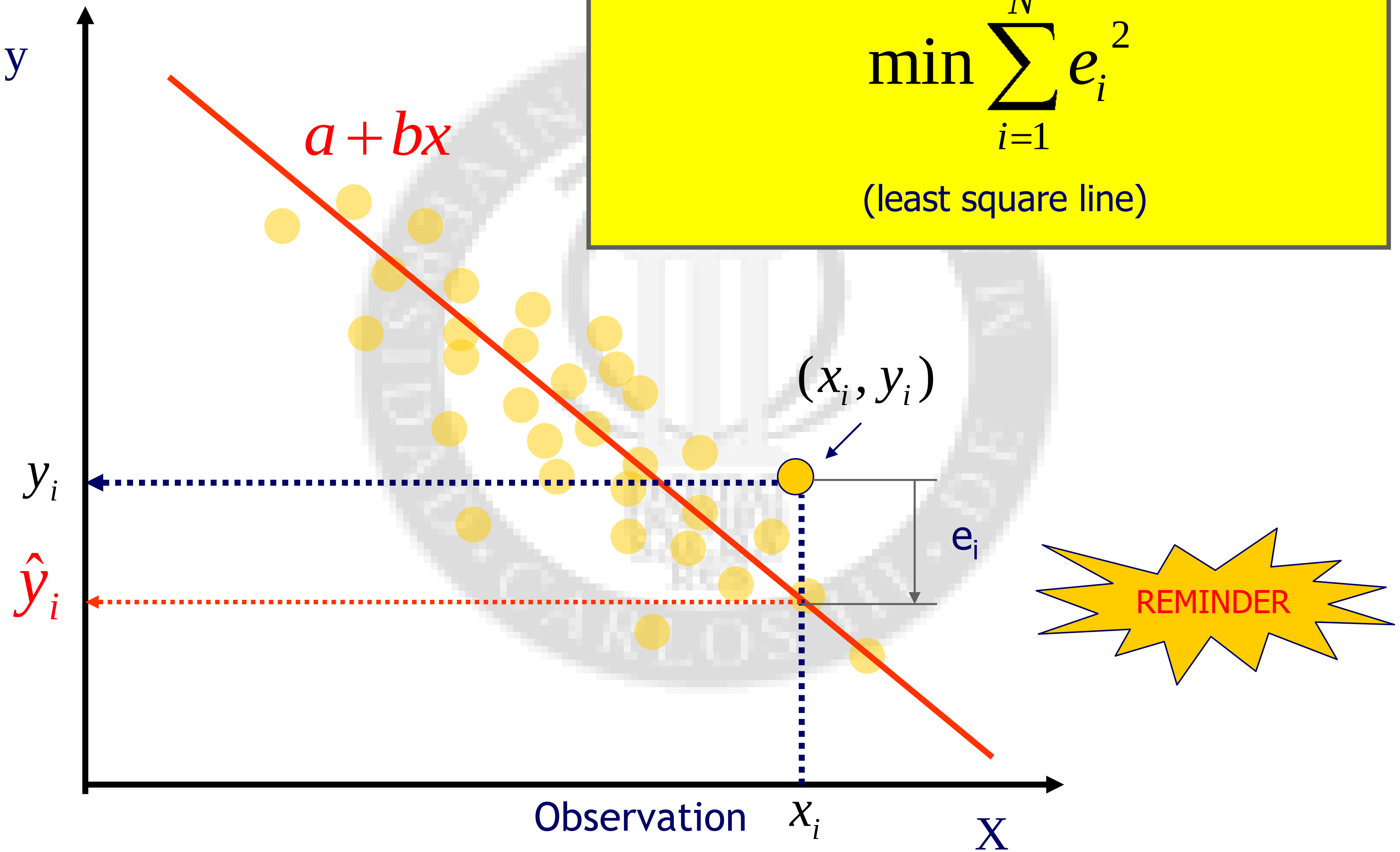
1. Statistical model for Simple Regression.
2. Statistical model for Multiple Regression.
3. Estimation of the Multiple Regression parameters.
4. Inference for Multiple Regression.
5. Test for the Multiple Regression model.
6. Regression with binary variables.

The simple regression line or population regression line

We look for a linear model **Y=a+bX+e** that minimizes the prediction errors:

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

(least square line)

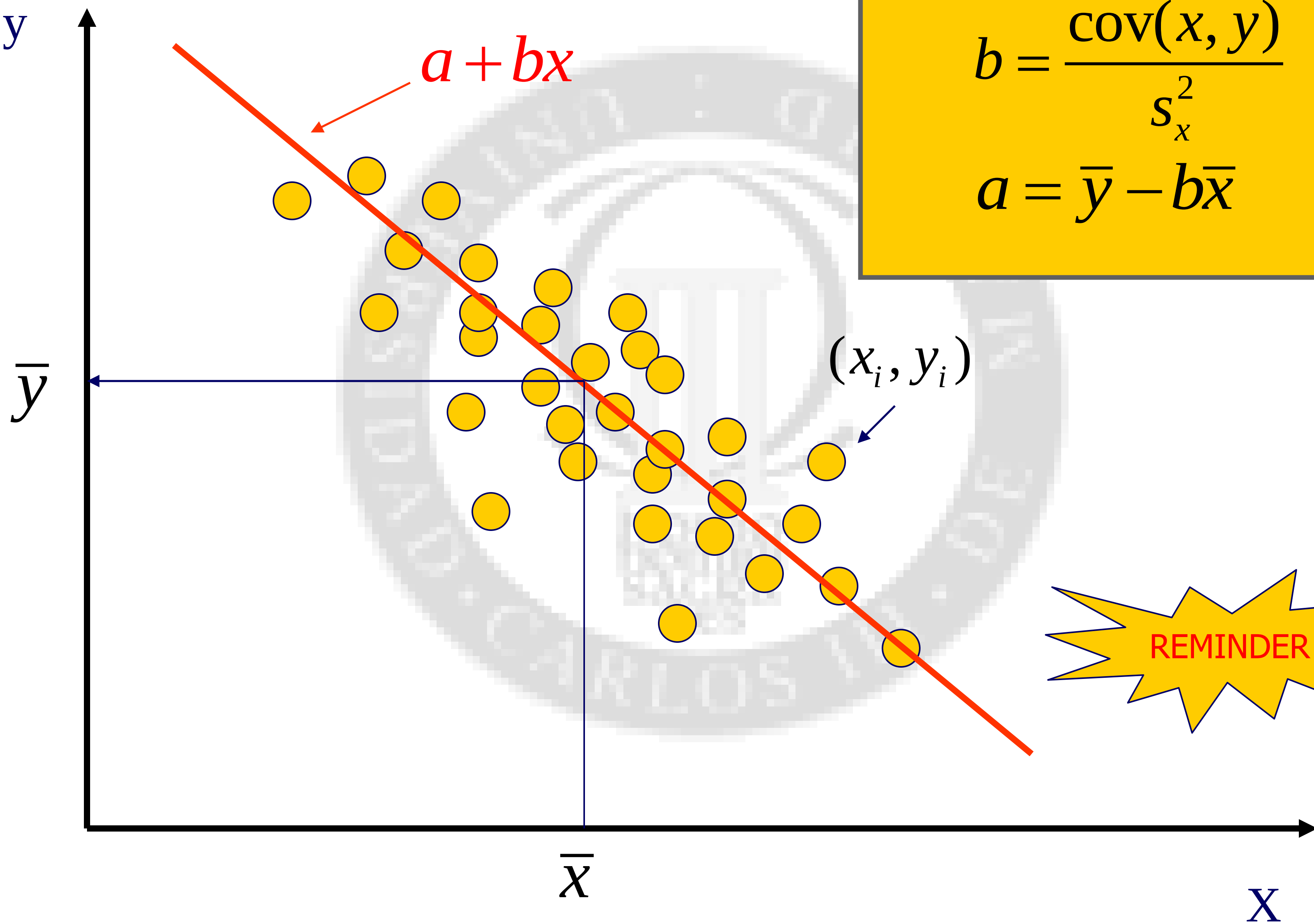


The simple regression line or population regression line

SOLUTION

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

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

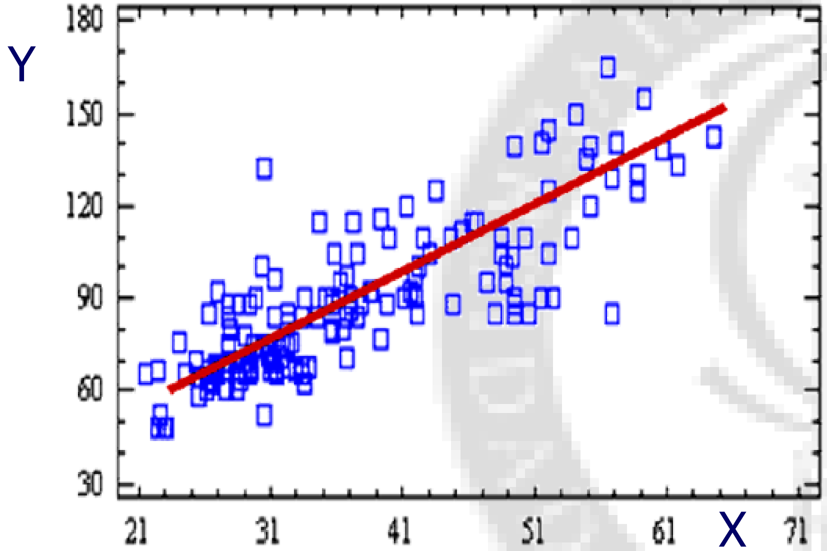


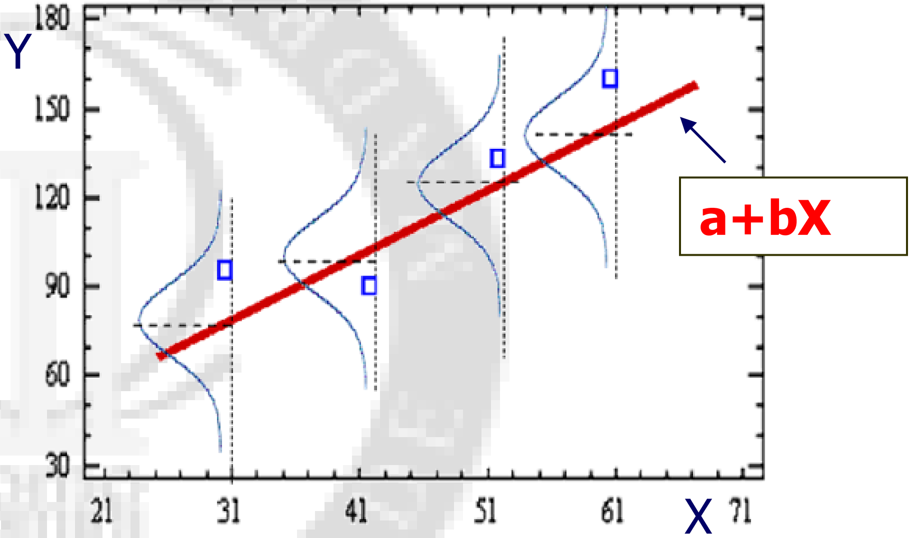
REMINDER

The "Simple Linear Regression Model"

Influence of other factors

$$Y = \underbrace{a + bX}_{\text{fixed}} + \underbrace{e}_{\text{random}} \qquad e_i \sim N(0, \sigma^2)$$





Each point y_i that we observe is interpreted as a realization of normal random variable distributed as $Y_i \sim N(a + bx_i, \sigma^2)$

We assume that the "noise" e is homogeneous along the line: i.e. its variance is constant (homoelasticity assumption)

Example: File *AlumnosIndustriales.sf3*. We want to predict the height of students by knowing their weight

```
-----------------------------------------
Dependent variable: altura
-----------------------------------------
                                       St
Parameter               Estimate
-----------------------------------------
CONSTANT                 138,364        3
peso                    0,535008       0,
-----------------------------------------
```

$$Y = 138{,}4 + 0{,}53X_1 + e.$$

- Individuals who weight 1 kg more are on average 0.53 cm taller
- Individuals who weight 80 kg have an average height of:

$$138 + 0{,}53 \times 80 = 180{,}4 \text{ cm}.$$

1. **Statistical model for Simple Regression.**
2. **Statistical model for Multiple Regression.**
3. **Estimation of the Multiple Regression parameters.**
4. **Inference for Multiple Regression.**
5. **Test for the Multiple Regression model.**
6. **Regression with binary variables.**

2. Statistical model for Multiple Regression

We define now a linear model that explain or predict Y starting from a set of K variables X

"Dependent" variable: $\longrightarrow$ Y

"Independent" or "explicative" variables: $\longrightarrow$ $X = (X_1, X_2, ..., X_K)$

For the i-th observation:
$$\begin{cases} y_i \\ \mathbf{x}_i = (x_{1i}, ..., x_{Ki}) \end{cases}$$

Multiple Regression Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + e_i,$$

The required assumptions can be summarized in the following set:

1. The relation between Y and the explicative variables **X** is linear

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + e_i,$$

2. The error (or residual) **e** is normal distributed with mean 0 and constant variance (homoelasticity assumption)

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + e_i,$$

Influence of unknown variables (if they are many by the CLT the residual would be Normal distributed)

$$E(e_i|\mathbf{x}_i) = 0$$

$$e_i \sim N(0, \sigma^2)$$

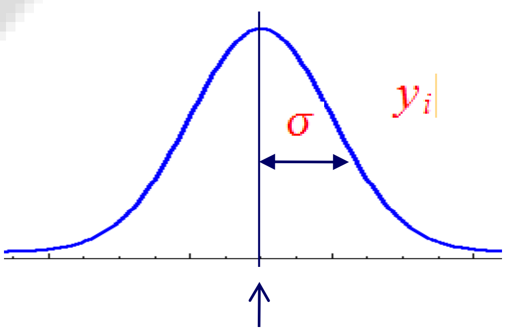
$$E(Y|X = x_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki}$$

Prediction of y_i

$$\text{var}(y_i|X = x_i) = \sigma^2.$$

Conclusion:

$$y_i|\mathbf{x}_i \sim N(\beta_0 + \beta_1 x_{1i} + \cdots \beta_K x_{Ki}; \sigma^2)$$



$$\beta_0 + \beta_1 x_{1i} + \cdots \beta_K x_{Ki}$$

It is useful write the model in matrix form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + e_i,$$

$$\downarrow$$

$$Y = X\boldsymbol{\beta} + \mathbf{e},$$

$$\downarrow$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{K1} \\ 1 & x_{12} & x_{22} & \cdots & x_{K2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{Kn} \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}.$$

Parameters: the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$.
and the variance σ^2

What values should we use?

1. Statistical model for Simple Regression.
2. Statistical model for Multiple Regression.
3. Estimation of the Multiple Regression parameters.
4. Inference for Multiple Regression.
5. Test for the Multiple Regression model.
6. Regression with binary variables.

3. Estimation of the Multiple Regression parameters

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + e_i, \qquad Y = X\boldsymbol{\beta} + \mathbf{e},$$

We do not know the parameters' values.
They are population parameters and as such they are unknown

We can estimate them by using a dataset

$$\left.\begin{array}{l} (y_1, x_{11,\ldots,} x_{K1}); \\ (y_2, x_{12,\ldots,} x_{K2}); \\ \vdots \\ (y_n, x_{1n,\ldots,} x_{Kn}); \end{array}\right\}$$

$Y \quad X$

We look for the values that minimize the residual error (same procedure we used for the simple regression)

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots \beta_K x_{Ki})^2 = \sum_{i=1}^{n} e_i^2.$$

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1} X' Y$$

For K=1 we get the same result as for the simple regression

Example:

File *AlumnosIndustriales.sf3*. We want to predict the height (Y) of students by knowing their weight (X_1) and their shoe size (X_2).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

$$Y = \begin{bmatrix} 180 \\ 161 \\ \vdots \\ 162 \end{bmatrix}; X = \begin{bmatrix} 1 & 72 & 44 \\ 1 & 55 & 39 \\ \vdots & \vdots & \vdots \\ 1 & 49 & 37 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 72 & 55 & \cdots & 49 \\ 44 & 39 & \cdots & 37 \end{bmatrix} \times \begin{bmatrix} 1 & 72 & 44 \\ 1 & 55 & 39 \\ \vdots & \vdots & \vdots \\ 1 & 49 & 37 \end{bmatrix} = \begin{bmatrix} 95 & 6438 & 3839 \\ 6438 & 449382 & 266303 \\ 3839 & 266303 & 160233 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 72 & 55 & \cdots & 49 \\ 44 & 39 & \cdots & 37 \end{bmatrix} \times \begin{bmatrix} 180 \\ 161 \\ \vdots \\ 162 \end{bmatrix} = \begin{bmatrix} 16589 \\ 1131213 \\ 681627 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 3.7756 & 0.0178 & -0.1213 \\ 0.0178 & 0.0002 & -0.0008 \\ -0.1213 & -0.0008 & 0.0043 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} 3.7756 & 0.0178 & -0.1213 \\ 0.0178 & 0.0002 & -0.0008 \\ -0.1213 & -0.0008 & 0.0043 \end{bmatrix} \times \begin{bmatrix} 16589 \\ 1131213 \\ 681627 \end{bmatrix} = \begin{bmatrix} 77.7 \\ 0.13 \\ 2.16 \end{bmatrix}$$

$$Y = 77.7 + 0.13X_1 + 2.16X_2 + e.$$

Example:

File *AlumnosIndustriales.sf3*. We want to predict the height (Y) of students by knowing their weight (X_1) and their shoe size (X_2).

Only weight (simple reg.)

Height=138.4+0.53 Weight + e

If a person weights 80 kg, her/his expected height (mean height of people weighting 80 kg) is

Mean (or predicted) Height=138+0.53x80=180.4cm

As the height depends strongly on the shoe size (that is related to the constitution of the person), the simple regression model gives results very different from the multiple regression one. Indeed the latter considers this relation. If we fix the variable Shoe-Size the influence of the Weight variable is smaller.

Weight and shoe size (multiple reg.)

Height =77.7+0.13 Weight
+2.16 Shoe-Size + e

If a person weights 80 kg, her/his expected height depends on her/his shoe size.

If the shoe size is 37, the expected height (mean height of people weighting 80 kg and with shoe size 37) :

Mean Height =77.7+0.13x80+2.16x37=168.02 cm

If the shoe size is 43, the expected height (mean height of people weighting 80 kg and with shoe size 43) :

Mean Height=77.7+0.13x80+2.16x43=181.98 cm

Simple Regression

$$Y = \alpha_0 + \underline{\alpha_1} X_1 + e;$$

$$Y(X_1 = x) = \alpha_0 + \alpha_1 x + e$$

$$Y(X_1 = x+1) = \alpha_0 + \alpha_1 (x+1) + e$$

$$\Delta Y = Y(X_1 = x+1) - Y(X_1 = x) = \underline{\alpha_1}$$

The coefficient of X1 in a simple regression says how much the variable Y changes (on average) if X1 increased of 1 unit. It measures the (total) influence of X1 on Y.

Multiple Regression

$$Y = \beta_0 + \underline{\beta_1} X_1 + \beta_2 X_2 + e$$

$$Y(X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

$$Y(X_1 = x_1 + 1, X_2 = x_2) = \beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2 + e$$

$$\Delta Y = Y(X_1 = x_1 + 1, X_2 = x_2) - Y(X_1 = x_1, X_2 = x_2) = \underline{\beta_1}.$$

The coefficient of X1 in a simple regression says how much the variable Y changes (on average) if X1 increased of 1 unit with the rest of variable staying fixed. It measures the marginal (differential) influence of X1 on Y when the rest of variable are kept constant.

It lasts to estimate the parameter σ^2

1 – We compute the residuals for any observation

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_K x_{Ki})$$

AlumnosIndustriales.sf3

	altura	peso	zapato	PREDICTED $\hat{y}$	RESIDUALS e
1	180	72	44	181,672	-1,67172
2	161	55	39	168,741	-7,7408
3	180	45	41	171,793	8,20722
4	180	99	44	185,079	-5,0795
5	178	68	41	174,696	3,30431
6	180	64	42	176,348	3,6521
7	182	80	41	176,21	5,78974
8	179	70	41	174,948	4,05188
9	180	80	44	182,681	-2,68143
10	173	55	37	164,427	8,57332
11	177	75	43	179,898	-2,89331
12	182	70	42	[illegible]	[illegible]
13	167	55	38	[illegible]	[illegible]
14	160	50	37	163,796	-3,79561
15	163	55	37	164,427	-1,42668
16	163	50	36	161,639	1,36145
17	185	80	43	180,524	4,47562
18	168	72	40	173,043	-5,04349
19	170	70	41	174,948	-4,94812

Height= $\boxed{\text{77.7+0.13 Weight+2.16 Shoe-Size}}$ +e

It lasts to estimate the parameter σ^2

1 – We compute the residuals for any observation

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_K x_{Ki})$$

2 – We use the following unbiased estimator – Residual Variance –

$$\hat{s}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} \qquad E(\hat{s}_R^2) = \sigma^2$$

Where p = number of beta parameters:

- with the constant term: K+1
- without the constant term: K

Coefficient of Determination R^2 : % measure of the variability of Y explained by the regression (same definition as for the simple regression case)
- The square root of R^2 is also known as Multiple Correlation Coefficient.

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_K x_{Ki}}_{\text{Part of Y explained by the predicted part estimated by the regression}} + \underbrace{e}_{\text{Part of Y that is not explained by the regression}}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \mathrm{corr}(\hat{y}, y)^2$$

Is it better to have a large R^2?

↓

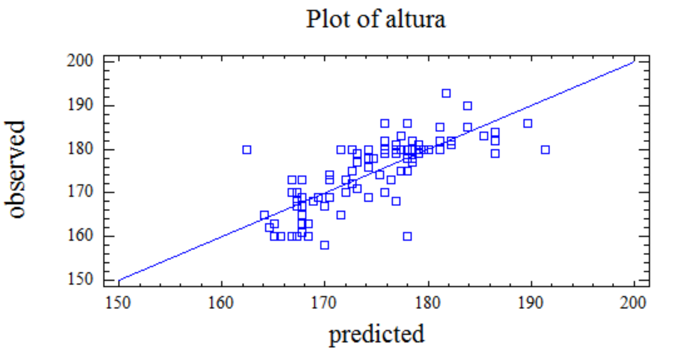
It can be proved that using more variables increases always the value of R^2 even if the included variable do not affect the variable Y, i.e. they are irrelevant

↓

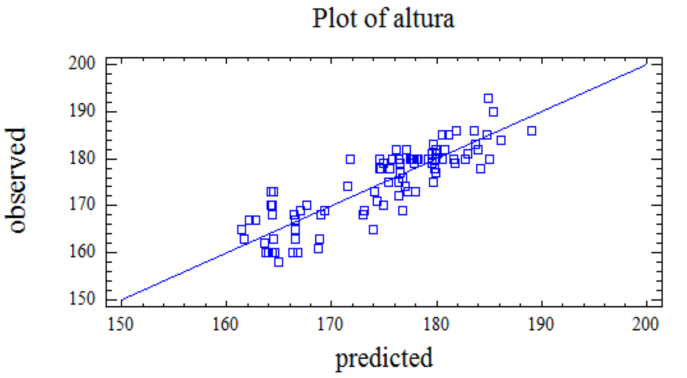
$$\bar{\mathrm{R}}^2 = 1 - \frac{\hat{s}_R^2}{\hat{s}_y^2}$$

Corrected (or Adjusted) Coefficient of Determination. It increases only if we add relevant variables.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: altura
-----------------------------------------------------------------------------
                                          Standard
Parameter                  Estimate          Error
-----------------------------------------------------------------------------
CONSTANT                    138,364        3,18832
peso                       0,535008       0,046357
-----------------------------------------------------------------------------

R-squared = 58,8851 percent
R-squared (adjusted for d.f.) = 58,443 percent
```

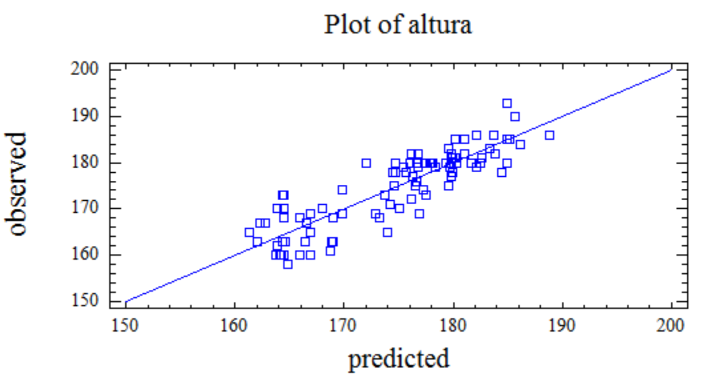



```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: altura
-----------------------------------------------------------------------------
                                          Standard
Parameter                  Estimate          Error
-----------------------------------------------------------------------------
CONSTANT                    77,6738         7,9423
zapato                      2,15706       0,268434
peso                       0,126214      0,0621644
-----------------------------------------------------------------------------

R-squared = 75,8414 percent
R-squared (adjusted for d.f.) = 75,3162 percent
```




```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: altura
-----------------------------------------------------------------------------
                                          Standard
Parameter                  Estimate          Error
-----------------------------------------------------------------------------
CONSTANT                    77,6901         8,0914
hermanos                  -0,245838       0,345273
tiempo                   0,00191048      0,0178027
dinero                 0,0000457647    0,000359008
zapato                        2,171       0,280181
peso                       0,121899      0,0657199
-----------------------------------------------------------------------------

R-squared = 75,9867 percent
R-squared (adjusted for d.f.) = 74,6376 percent
```

Irrelevant variables



Multiple Regression Analysis

How can we know if some variables are or are not relevant?

Should we ask to an expert of the subject or can we deduce it

by looking at the data?

If the variable $\mathbf{x_i}$ do not add anything to the regression model we should have …

$$\beta_i = 0$$

… but we do not observe the values β_i but only their estimations $\hat{\beta}_i$ and in general we have that

$$\hat{\beta}_i \neq 0$$

How can we decide if $\mathbf{x_i}$ is relevant by only looking at $\hat{\beta}_i$?

R-squared = 75,9867 percent

R-squared (adjusted for d.f.) = 74,6376 percent

1. Statistical model for Simple Regression.
2. Statistical model for Multiple Regression.
3. Estimation of the Multiple Regression parameters.
4. Inference for Multiple Regression.
5. Test for the Multiple Regression model.
6. Regression with binary variables.

The numerical values of the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$ are unknown.

We use the estimator $\hat{\beta} = (X'X)^{-1}X'Y$

We apply this estimator to our data and get an estimation

- The estimator is a random variable
- We only observe one sample taken from the population
- What are the properties of this estimator?
- What is its sample distribution?

If the sample size n is large or if $e_i \sim N(0, \sigma^2)$

$$\hat{\beta}_i \sim N(\beta_i, \hat{s}_R^2 q_{ii})$$

here q_{ii} is the i-th element of the diagonal of the matrix $(X'X)^{-1}$

$$\hat{\beta}_i \sim N(\beta_i, \hat{s}_R^2 q_{ii}) \longrightarrow$$ Using this property we can make an hypothesis test to check if a variable is or is not significant

Significant variable = it is relevant to include it in the regression to get information about Y that could not be obtained by the rest of the independent variables

Ideally for a not significant variable: $\beta_i = 0$

Using the fact that $\hat{\beta}_i \sim N(\beta_i, \hat{s}_R^2 q_{ii})$ we can make a hypothesis test :

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

If the p-value is small (<0.05) we reject H_0 and the variable is considered significant (for this p-value)

Example:

File *AlumnosIndustriales.sf3*. We want to predict the height (Y) of students by knowing their weight (peso) and their shoe size (zapato). Should we consider the money they carry (dinero) as well? The sample is made of 95 observations.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: altura
-----------------------------------------------------------------------------
                                         Standard          T
Parameter                Estimate         Error        Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                  77,6132        8,00168        9,69961         0,0000
peso                     0,126793      0,0626927        2,02245         0,0461
zapato                    2,15651       0,269924        7,98934         0,0000
dinero               0,0000419267    0,000355454       0,117953         0,9064
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     4825,54      3      1608,51      95,24       0,0000
Residual                  1536,82     91      16,8882
-----------------------------------------------------------------------------
Total (Corr.)             6362,36     94

R-squared = 75,8451 percent
R-squared (adjusted for d.f.) = 75,0488 percent
```

$$\hat{s}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

The p-value of 'dinero' is very large, that means that this variable is not significant (with significant level 5%) to predict the height of the students. We cannot reject the hypothesis that its associated parameter is 0. We can eliminate this variable and estimate the model again.

If there were more than one not significant variables we would exclude all of them one by one (the significance test of one variable depends on which other variables are included in the regression model).

Example:

File *AlumnosIndustriales.sf3*. We want to predict the height (Y) of students by knowing their weight (peso) and their shoe size (zapato). The sample is made of 95 observations.

Multiple Regression Analysis

Dependent variable: altura

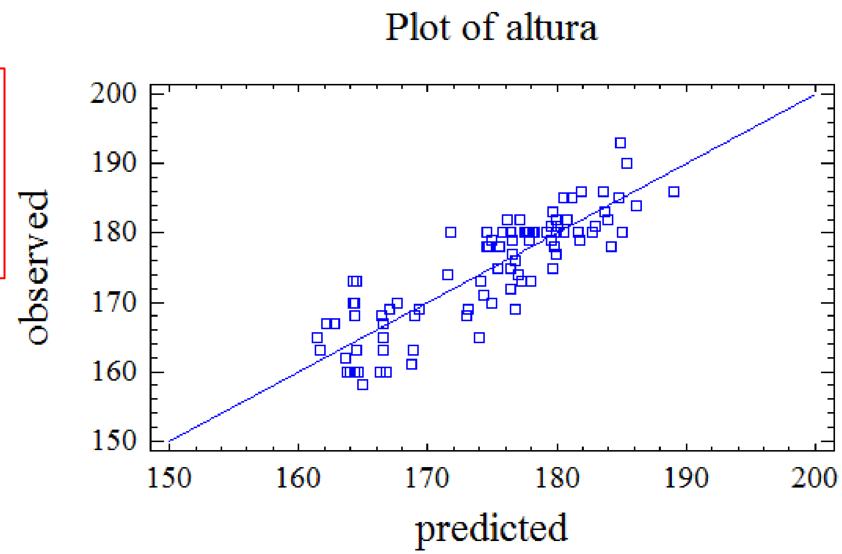
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	77,6738	7,9423	9,77976	0,0000
zapato	2,15706	0,268434	8,0357	0,0000
peso	0,126214	0,0621644	2,03032	0,0452

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	4825,3	2	2412,65	144,41	0,0000
Residual	1537,06	92	16,7071		
Total (Corr.)	6362,36	94			

R-squared = 75,8414 percent
R-squared (adjusted for d.f.) = 75,3162 percent

$$\hat{s}_R^2 = \frac{\sum_{i=1}^n e_i^2}{n-p}$$



Both variables are significant

What is the probability that a person whose shoe size is 40 and whose weight is 60 kg is taller than 185cm?

$$\hat{y} = 171.53$$

$$Y \sim N(171.53; 16.71)$$

P(Y>185)=0.019

1. Statistical model for Simple Regression.
2. Statistical model for Multiple Regression.
3. Estimation of the Multiple Regression parameters.
4. Inference for Multiple Regression.
5. Test for the Multiple Regression model.
6. Regression with binary variables.

5. Test for the Multiple Regression model.

The built regression model is valid only if the basic assumptions hold. They can be summarized in the following:

1. Linearity $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + e_i,$

2. The error (or residual) **e** is normal distributed with mean 0 and constant variable (homoelasticity assumption)

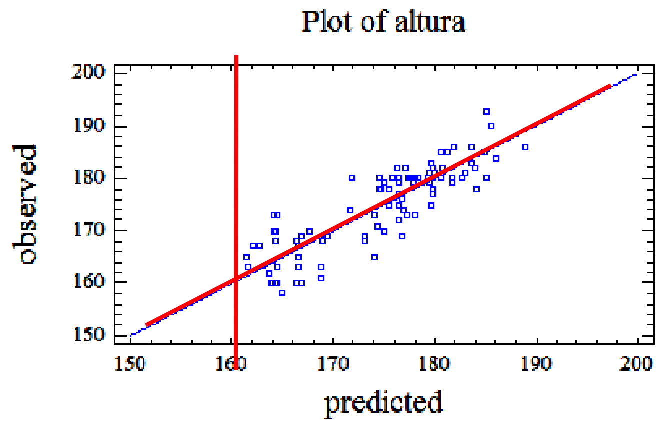
Significance test for the model: it is done by testing that the above hypotheses are valid

We can make it by:

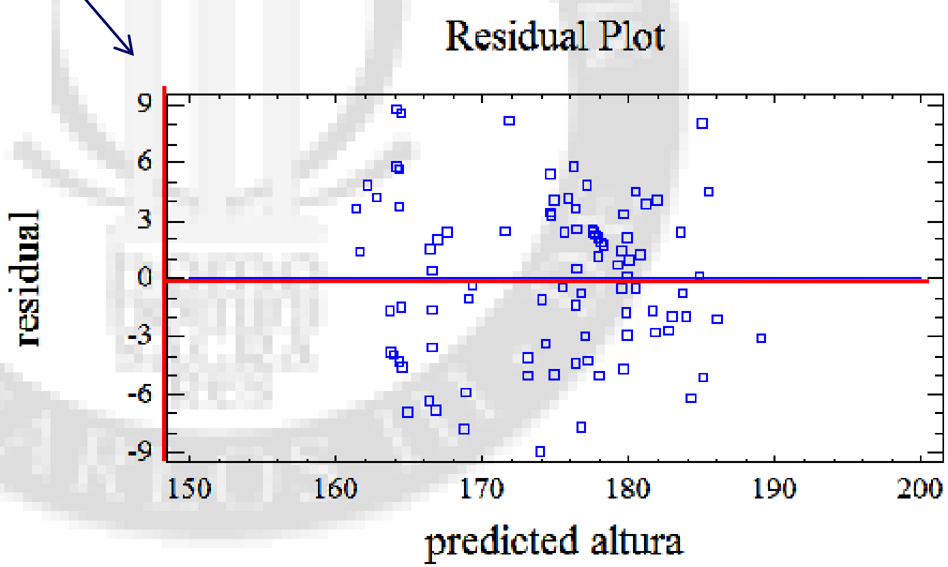
a) Analyzing the residuals vs. predicted values graph
b) Analyzing the graphs residuals vs. the single component X_i
c) Analyzing if the residuals are normal distributed

a) Analyzing the residuals vs. predicted values graph

It is the same as for the case of the simple regression



The residual is the vertical distance of the observed value from the diagonal

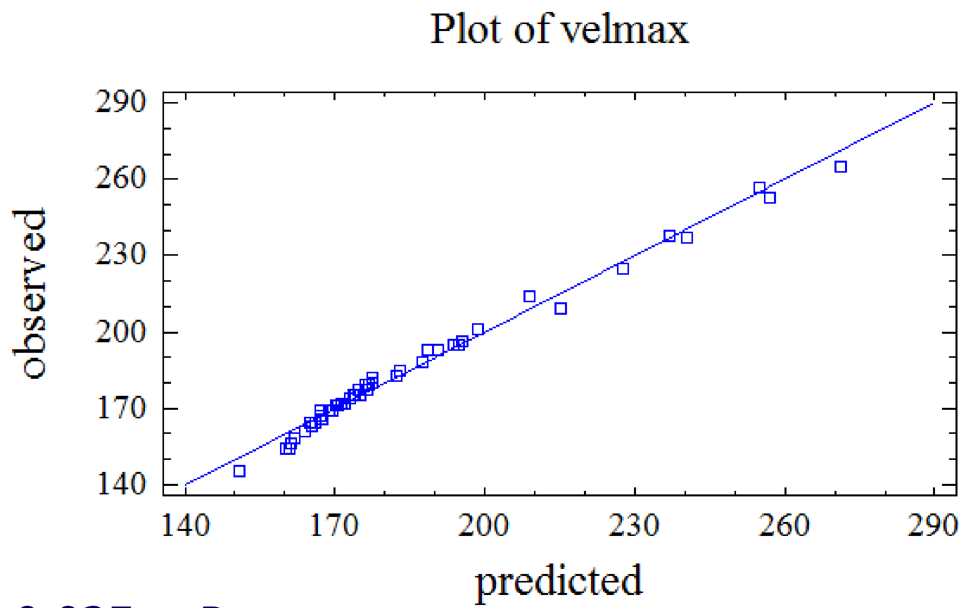


If the model were really linear, the residual would be normal with zero mean and constant variable (homoelasticity assumption). They shouldn't show any especial structure like it is shown in the graph above.

Example:

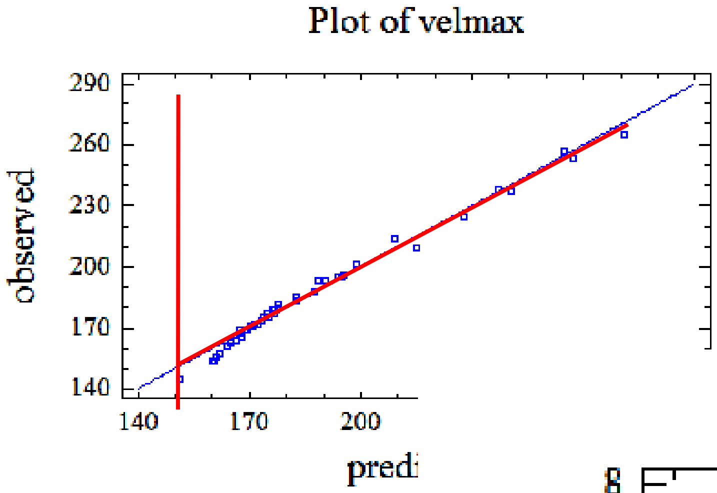
The file *Consumo_coches.sf3* contains data about the maximal speed reached by a sample of cars. What is the relation between the maximum speed (velmax) of a car and its weight (Peso) and power (Potencia)?

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: velmax
-----------------------------------------------------------------------------
                                        Standard          T
Parameter               Estimate         Error        Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 155,465         1,3399        116,027          0,0000
Potencia                0,519647     0,00966429        53,7698          0,0000
Peso                  -0,0252839     0,00148786       -16,9935          0,0000
-----------------------------------------------------------------------------

                           Analysis of Variance
-------------------------------------------------------------
Source                  Sum of Squares     Df  Mean Square
-------------------------------------------------------------
Model                          40555,0      2      20277,5
Residual                       593,746     79      7,51577
-------------------------------------------------------------
Total (Corr.)                  41148,8     81

R-squared = 98,5571 percent
R-squared (adjusted for d.f.) = 98,5205 percent
```

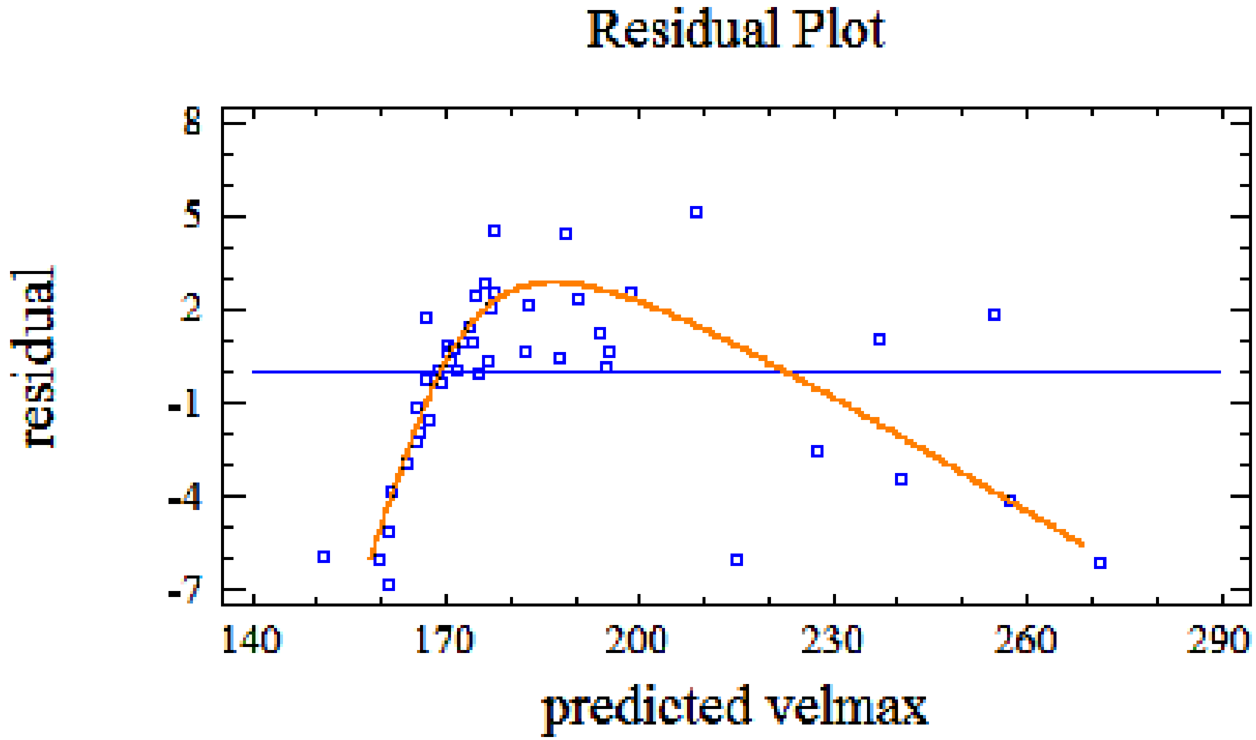


$$velmax = 155.5 + 0.52 \times Potencia - 0.025 \times Peso + e$$

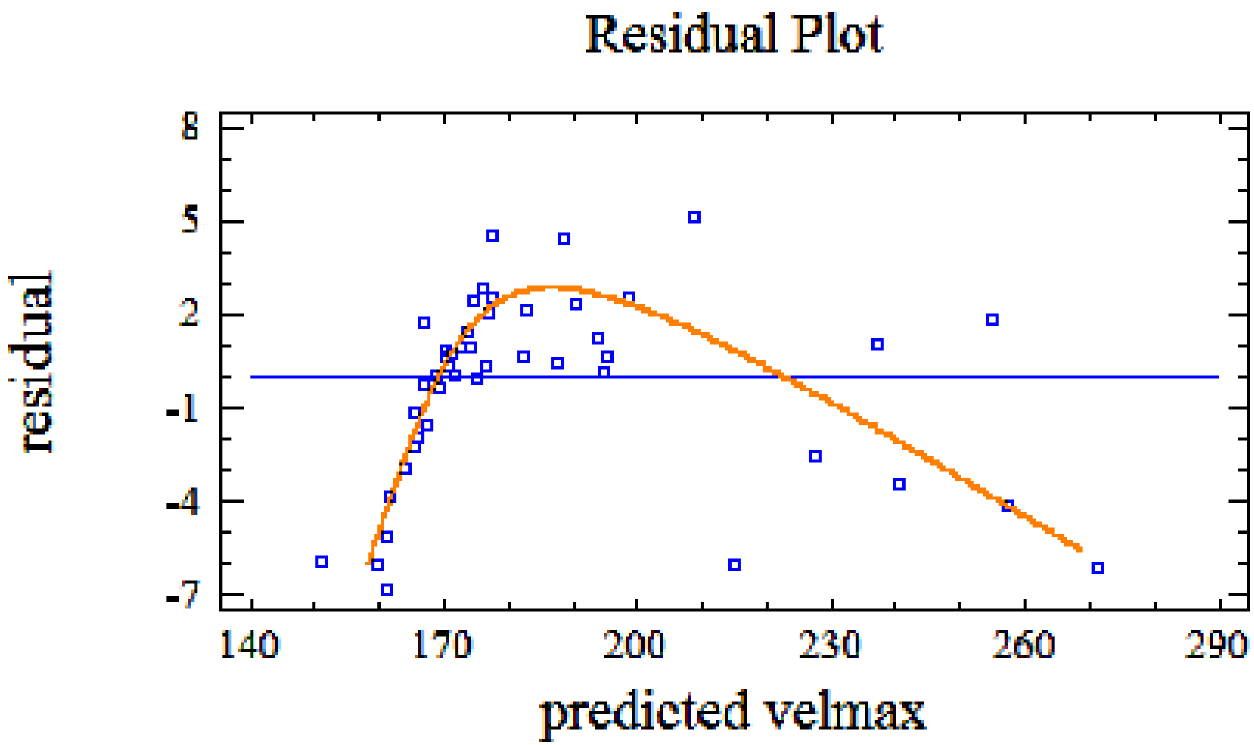
a) Analyzing the residuals vs. predicted values graph



(file *Consumo_coches.sf3*: maximal speed as function of the car weight and power)

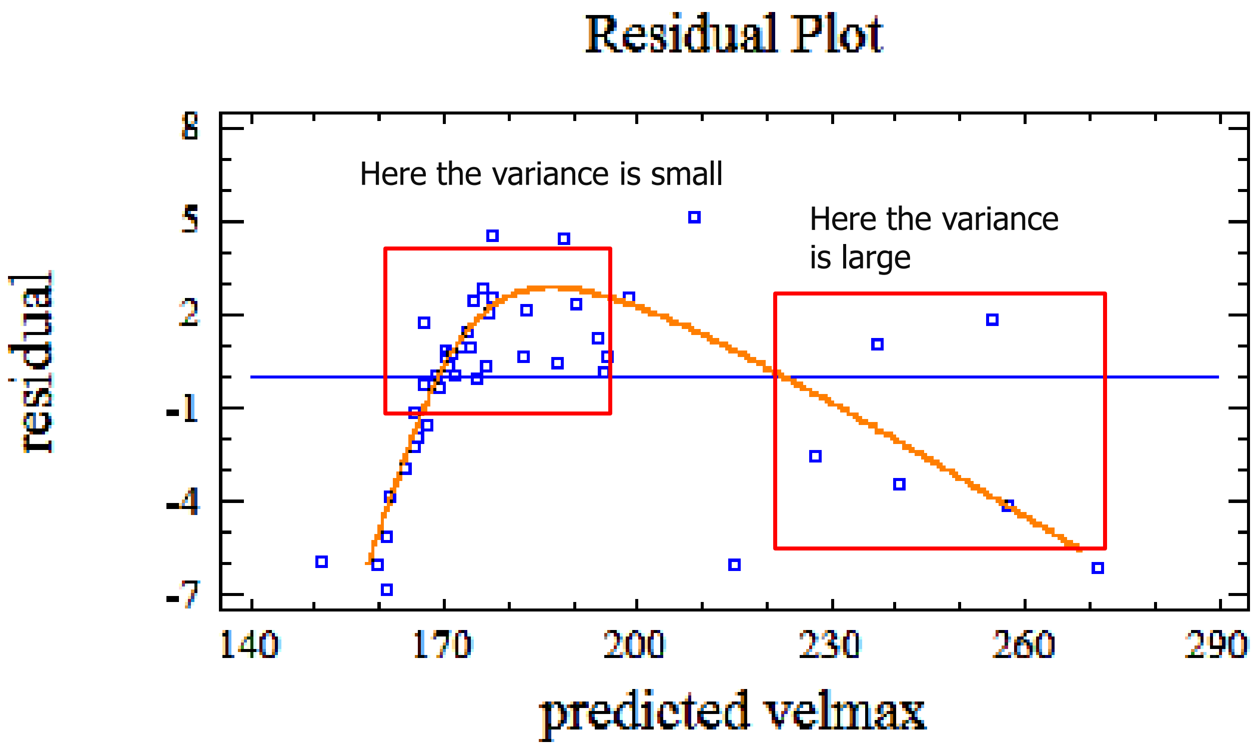


In this example, the residuals show an evident structure. Therefore the regression model is not valid.



- The structure is not linear: the relation between Y and X is not linear

a) Analyzing the residuals vs. predicted values graph



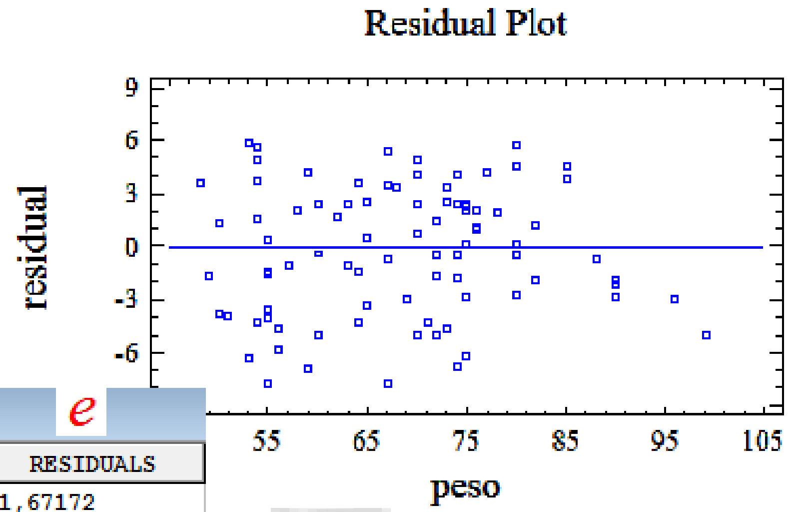
- The residuals has non-constant variance

$$\text{var}(e_i) \neq \sigma^2$$

$$\text{var}(e_i) = \sigma_i^2$$

b) Analyzing the graphs residuals vs. the single component X_i

This graph allows to detail the analysis to the individual independent variable. Also in this case if the model were correct we should observe no structure

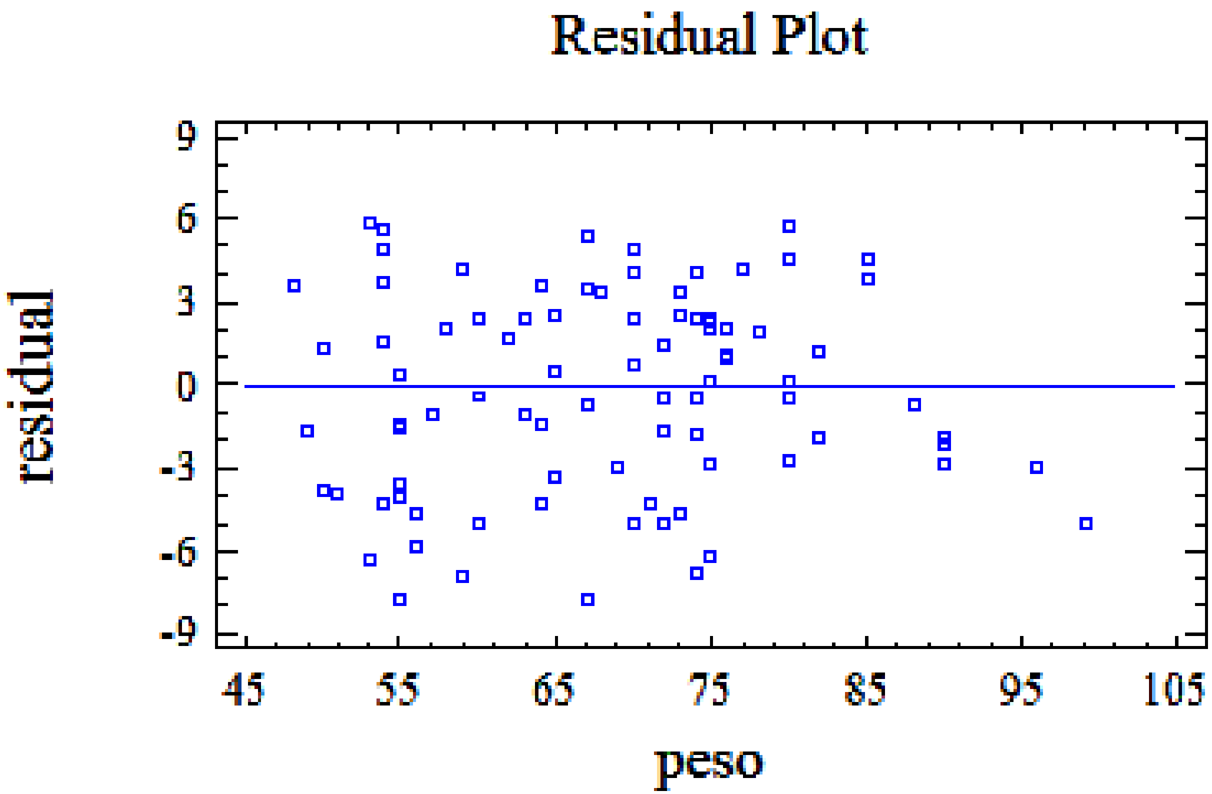


AlumnosIndustriales.sf3

	altura	peso	zapato	PREDICTED ($\hat{y}$)	RESIDUALS (e)
1	180	72	44	181,672	-1,67172
2	161	55	39	168,741	-7,7408
3	180	45	41	171,793	8,20722
4	180	99	44	185,079	-5,0795
5	178	68	41	174,696	3,30431
6	180	64	42	176,348	3,6521
7	182	80	41	176,21	5,78974
8	179	70	41	174,948	4,05188
9	180	80	44	182,681	-2,68143
10	173	55	37	164,427	8,57332
11	177	75	43	179,8[illegible]3	-2,89331
12	182	70	42	[illegible]	[illegible]
13	167	55	38	[illegible]	[illegible]
14	160	50	37	163,796	-3,79561
15	163	55	37	164,427	-1,42668
16	163	50	36	161,639	1,36145
17	185	80	43	180,524	4,47562
18	168	72	40	173,043	-5,04349
19	170	70	41	174,948	-4,94812

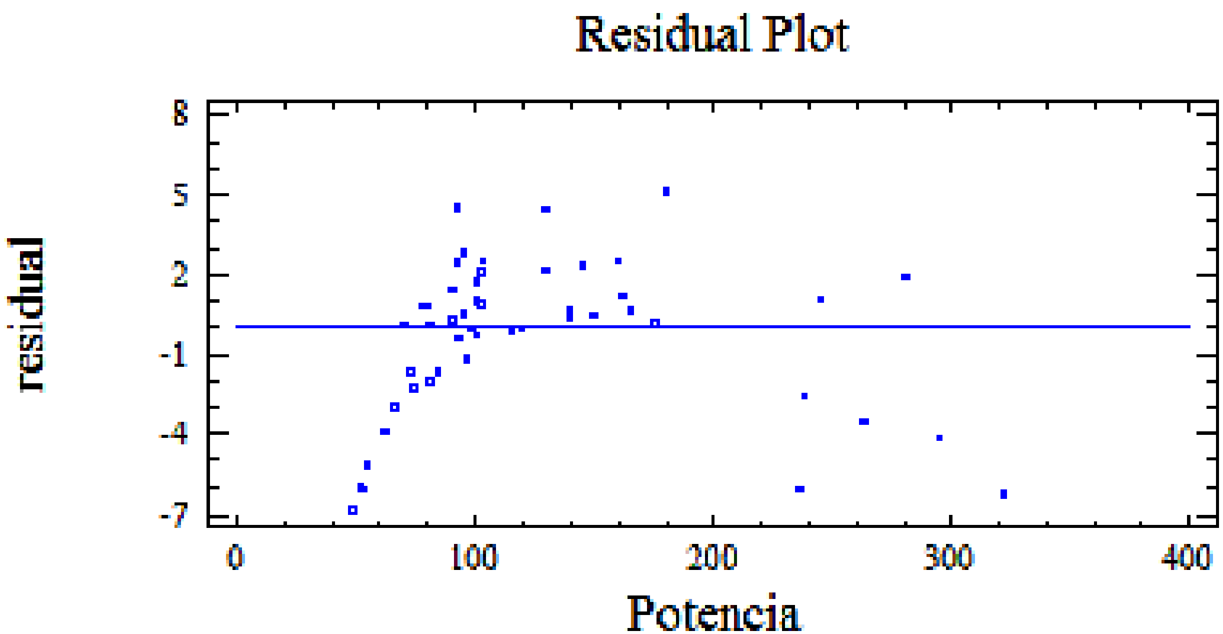
$$\text{altura} = \boxed{77.7 + 0.13 \times \text{Peso} + 2.16 \times \text{zapato}} + e$$

b) Analyzing the graphs residuals vs. the single component X_i



This graphs does not show any problem

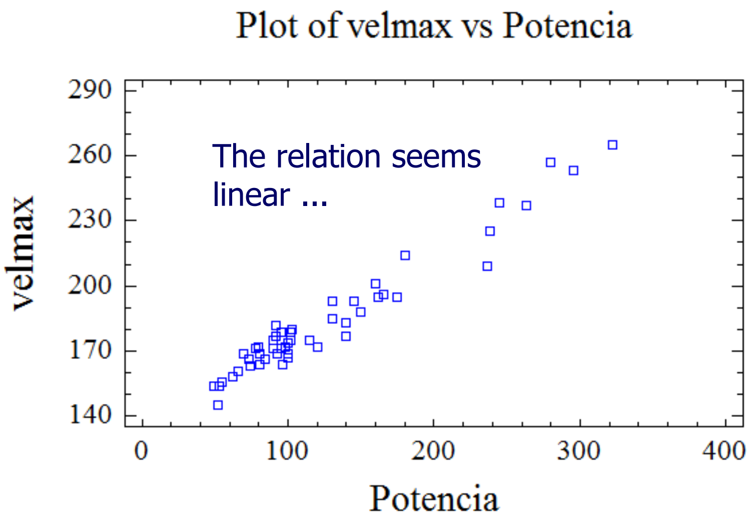
Focusing again on example of the maximal car speeds:

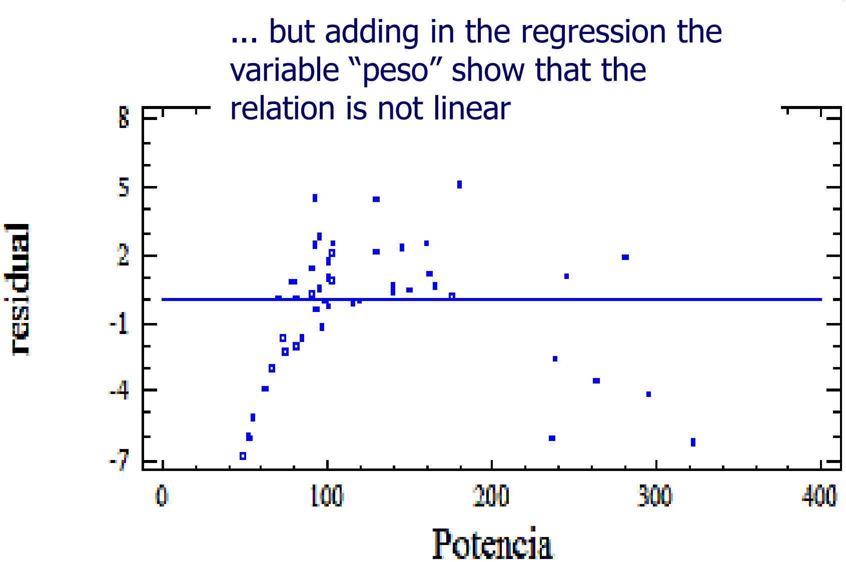


This variable is problematic: we need to look for a transformation of the type X^c and estimate again the model. How should we choose the power exponent c?

-- c>1 or c<1? -- $\longrightarrow$ We look at the Component Effect Graphs

We look for a transformation of the type X^c that improves linearity.
How should we choose the power exponent c? -- c>1 or c<1? --





In the case of simple regression, the graph XY would have been helpful to take a decision about the exponent c.

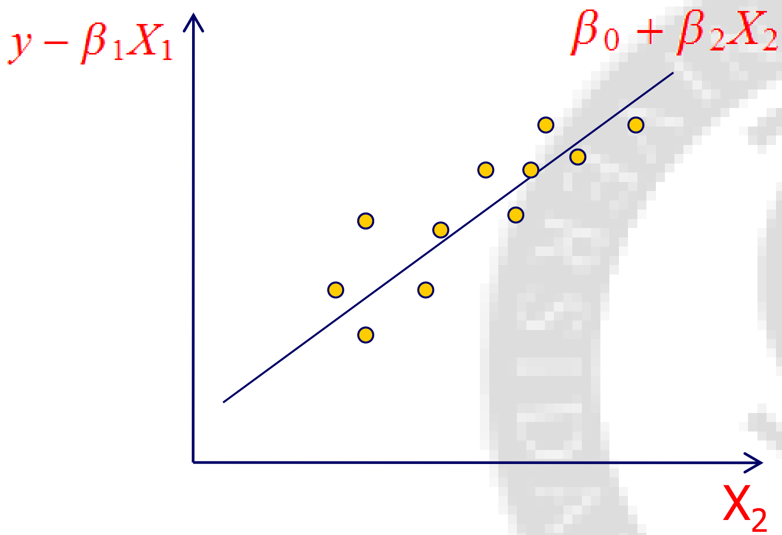
In the Multiple Regression it is not anymore so useful since the relation we want to analyze is still the one between Y and X_i but now taking into account also the relation between Y and the rest of independent variables.

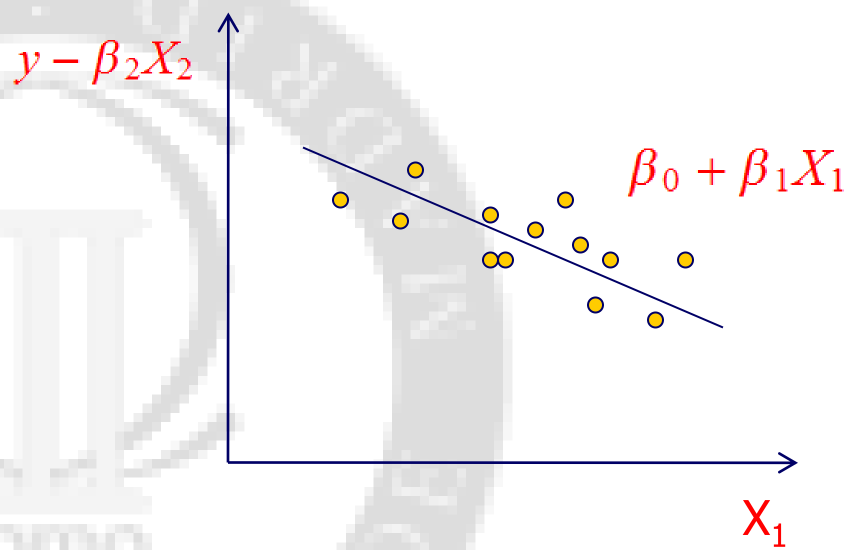
This means that we have to take out form Y the amount that can be explained by the information carried by the other variables and then plot the remaining part versus X_i.

For example considering a multiple regression with two variables,

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

we would like to show the following two graphs:





This kind of graphs is called Component Effect Graphs
One way to build them is using the following expression:

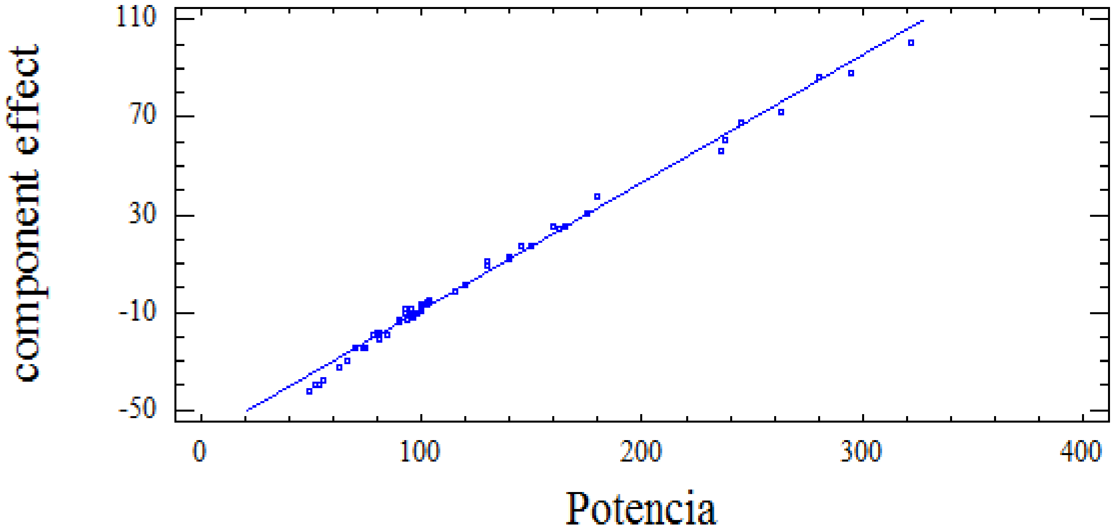
$$y - \beta_1 X_1 = \beta_0 + \beta_2 X_2 + e$$

and therefore plotting

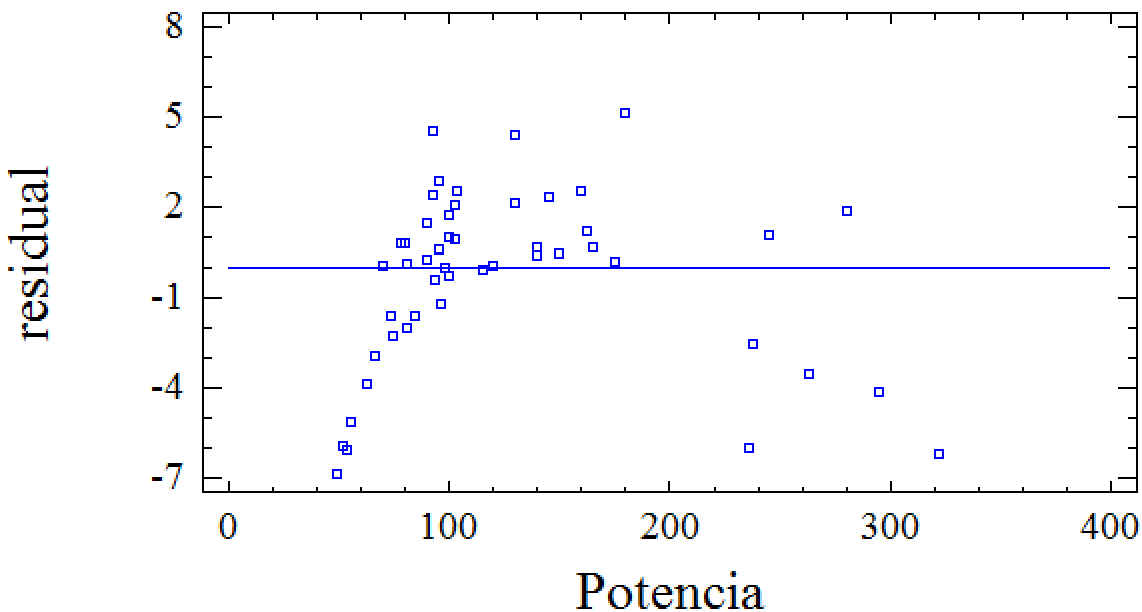
(Statgraphics plots the in this way)

$$\begin{cases} e + \hat{\beta}_2 (X_2 - \bar{X}_2) & \text{versus } X_2 \\ e + \hat{\beta}_1 (X_1 - \bar{X}_1) & \text{Versus } X_1 \end{cases}$$

Component+Residual Plot for velmax

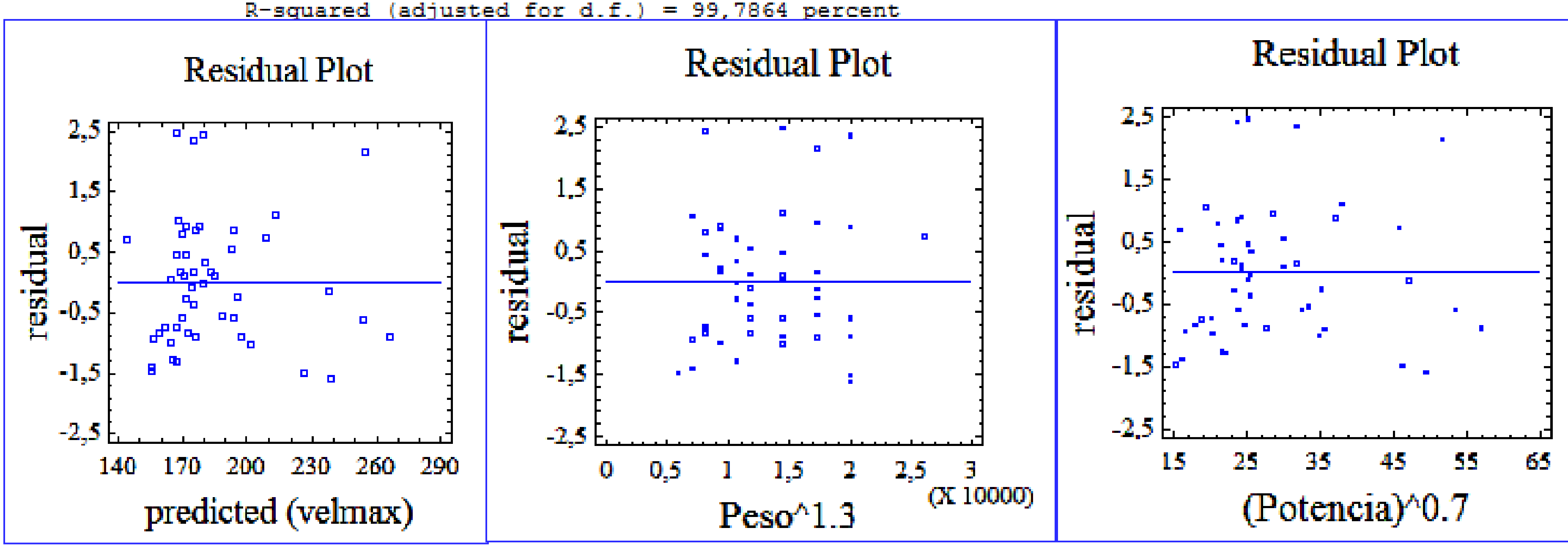


Residual Plot



Looking at these graphs we can appreciate that the whished transformation is of kind Potencia^{c} with $c<1$

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: velmax
-----------------------------------------------------------------------------
                                           Standard          T
Parameter               Estimate            Error        Statistic      P-Value
-----------------------------------------------------------------------------
CONSTANT                 117,107         0,380632         307,665       0,0000
Potencia^(0.7)           3,62663        0,0254188         142,675       0,0000
Peso^(1.3)           -0,00287637      0,000052611        -54,6724       0,0000
-----------------------------------------------------------------------------


R-squared = 99,7917 percent
R-squared (adjusted for d.f.) = 99,7864 percent
```

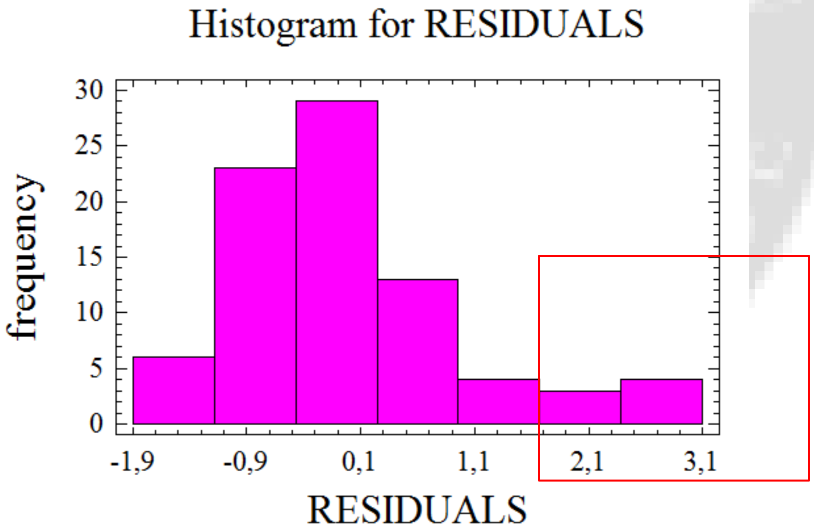


This new model improves the linearity

c) Analyzing if the residuals are normal distributed

- Normality is important to compute probabilities about predicted values as these computations assume normality.
- If n is large, the estimations and the tests are valid (if we can assume linearity) even if data are not themselves normal distributed

It is therefore sufficient to plot the histogram and verify that the data distribution looks unimodal and almost normal shaped



This asymmetry can be due to a not well solved linearity or to atypical values

1. Statistical model for Simple Regression.
2. Statistical model for Multiple Regression.
3. Estimation of the Multiple Regression parameters.
4. Inference for Multiple Regression.
5. Test for the Multiple Regression model.
6. Regression with binary variables.

The binary or dichotomous variable is a variable that only takes two values. We assume that those two values are 1 and 0

This variable can be used to define the presence/absence of some attribute or the membership/not-membership to a group

This variable is quantitative and doing regression it is treated as the rest of the variables.

Example:

The file *AlumnosIndustriales.sf3* contains the variable sex (sexo): 1 for male and 0 for female.
Is it relevant to predict the height (altura)?

AlumnosIndustriales.sf3

	nacimiento	altura	peso	zapato	sexo	dinero
1	1	180	72	44	1	1100
2	1	161	55	39	0	287
3	1	180	45	41	1	2000
4	1	180	99	44	1	25
5	1	178	68	41	1	3225
6	1	180	64	42	1	1300
7	2	182	80	41	1	4000
8	3	179	70	41	1	75
9	3	180	80	44	1	115
10	3	173	55	37	0	350
11	4	177	75	43	1	50
12	4	182	70	42	1	2000
13	4	167	55	38	0	500
14	4	160	50	37	0	1600
15	4	163	55	37	0	55
16	5	163	50	36	0	1000

$$\text{altura} = \beta_0 + \beta_1 \text{ sexo} + e$$

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: altura
-----------------------------------------------------------------------------
                                          Standard          T
Parameter                 Estimate           Error      Statistic       P-Value
-----------------------------------------------------------------------------
CONSTANT                   165,313        0,856112        193,097        0,0000
sexo                       14,0367         1,05129        13,3519        0,0000
-----------------------------------------------------------------------------
```

$$\text{altura} = 165.313 + 14.0367 \text{ sexo} + e$$

The "usual" interpretation of the regression is :
If the variable sexo increases of one unit, the average height increases of 14 cm

Being sexo a binary variable, the coefficient measures the difference between the mean height of the individuals with value 1 and the one of the individuals with value 0

We can separate the model into two parts: one for each value of the binary variable

altura = 165.313 + 14 sexo + e

When sexo=0:

E(altura|female)=165.313+14.0367 x0=165.313 cm

When sexo=1:

E(altura|male)=165.313+14.0367 x1= 179.3497 cm

For each "group" the model estimate the mean value of the dependent variable

The result is exactly equal to compute the sample means of each separate group (0 and 1)…

Summary Statistics for altura

	sexo=0	sexo=1
Count	32	63
Average	165,313	179,349
Variance	19,6411	25,36
Standard deviation	4,43183	5,03587
Minimum	158,0	165,0
Maximum	174,0	193,0
Range	16,0	28,0
Skewness	0,23093	-0,293724
Kurtosis	-0,978498	0,946925

We can separate the model into two parts: one for each value of the binary variable

altura = 165.313 + 14 sexo + e

When sexo=0:

E(altura|female) = 165.313 + 14.0367 x 0=165.313cm

When sexo=1:

E(altura|male) = 165.313 + 14.0367 x 1 = 179.3497cm

For each "group" the model estimate the mean value of the dependent variable

The result is exactly equal to compute the sample means of each separate group (0 and 1)…

... with the advantage that the p-value tells us if the difference is significant

$$\text{altura} = \beta_0 + \beta_1 \text{ sexo} + e$$

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: altura
-----------------------------------------------------------------------------
                                          Standard          T
Parameter               Estimate           Error        Statistic       P-Value
-----------------------------------------------------------------------------
CONSTANT                 165,313         0,856112        193,097         0,0000
sexo                     14,0367          1,05129        13,3519         0,0000
-----------------------------------------------------------------------------
```

$\mu_{\text{female}} = \beta_0$

$\mu_{\text{male}} = \beta_0 + \beta_1$

$\mu_{\text{male}} = \mu_{\text{female}} \Rightarrow \beta_1 = 0$

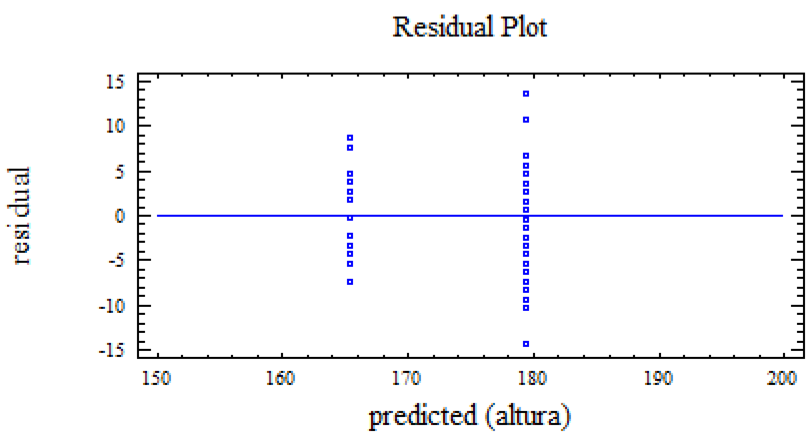
$H_0 : \mu_{\text{male}} = \mu_{\text{female}}$ $\quad\Rightarrow\quad$ $H_0 : \beta_1 = 0$

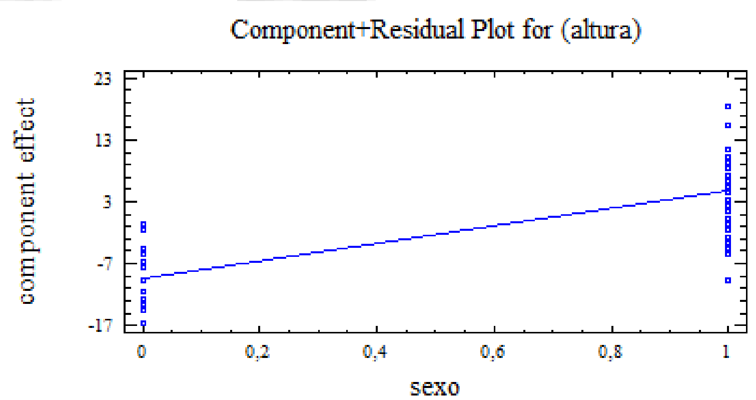
$H_1 : \mu_{\text{male}} \neq \mu_{\text{female}}$ $\qquad\qquad$ $H_1 : \beta_1 \neq 0$

The "predicted values" are the means of each group and therefore there are only two possible values

AlumnosIndustriales.sf3

	nacimiento	altura	sexo	PREDICTED
1	1	180	1	179,349
2	1	161	0	165,313
3	1	180	1	179,349
4	1	180	1	179,349
5	1	178	1	179,349
6	1	180	1	179,349
7	2	182	1	179,349
8	3	179	1	179,349
9	3	180	1	179,349
10	3	173	0	165,313
11	4	177	1	179,349
12	4	182	1	179,349
13	4	167	0	165,313
14	4	160	0	165,313
15	4	163	0	165,313
16	5	163	0	165,313





Example:

The file *AlumnosIndustriales.sf3* contains the variable sex (sexo): 1 for male and 0 for female. The mean heights for male students is higher than the one for female students
What if we compare the heights (altura) of male and female students with same weight (peso)?

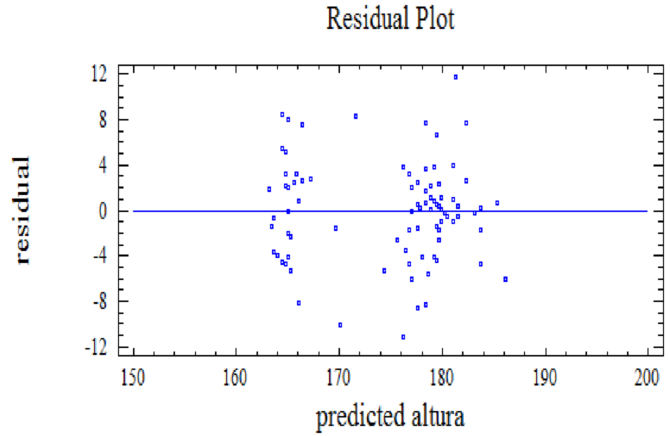
$$\text{altura} = \beta_0 + \beta_1 \text{ sexo} + \beta_2 \text{ peso} + e$$

Multiple Regression Analysis

Dependent variable: altura

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	150,306	3,12145	48,1528	0,0000
peso	0,267968	0,0540419	4,95853	0,0000
sexo	9,28133	1,34214	6,91531	0,0000

Between a male and a female students of same weight, the male student is on average 9.28 cm taller



Here is an example with more than just one group:

Example:

We want to compare the behavior of three hard disk with the aim to find the one with highest speed. To test them we save a file whose size is 200 MB in each of them and record the time of the this task. We repeat this experiment a given number of times and the results are contained in the file *Discosduros.sf3*. What is the quickest hard disk?

Discosduros.sf3

	Tiempo	Disco
1	38750	1
2	39812	1
3	38453	1
4	38203	1
5	37609	2
6	38609	2
7	37344	2
8	38328	2
9	37015	2
10	38000	2
11	37675	3
12	38631	3
13	39566	3
14	38377	3
15	39268	3
16	38020	3
17	38985	3
18	37708	3
19	38753	3
20	39786	3
21	38392	3

We create 3 binary variables: each of them denotes if the data belong to one of the three hard disks

$$D1= \begin{cases} 1, \text{ if it is the HD 1} \\ 0, \text{ if it is NOT the HD 1} \end{cases}$$

$$D2= \begin{cases} 1, \text{ if it is the HD 2} \\ 0, \text{ if it is NOT the HD 2} \end{cases}$$

$$D3= \begin{cases} 1, \text{ if it is the HD 3} \\ 0, \text{ if it is NOT the HD 3} \end{cases}$$

Here is an example with more than just one group:

Example:

We want to compare the behavior of three hard disk with the aim to find the one with highest speed. To test them we save a file whose size is 200 MB in each of them and record the time of the this task. We repeat this experiment a given number of times and the results are contained in the file *Discosduros.sf3*. What is the quickest hard disk?

Discosduros.sf3

	Tiempo	Disco	D1	D2	D3
1	38750	1	1	0	0
2	39812	1	1	0	0
3	38453	1	1	0	0
4	38203	1	1	0	0
5	37609	2	0	1	0
6	38609	2	0	1	0
7	37344	2	0	1	0
8	38328	2	0	1	0
9	37015	2	0	1	0
10	38000	2	0	1	0
11	37675	3	0	0	1
12	38631	3	0	0	1
13	39566	3	0	0	1
14	38377	3	0	0	1
15	39268	3	0	0	1
16	38020	3	0	0	1
17	38985	3	0	0	1
18	37708	3	0	0	1
19	38753	3	0	0	1
20	39786	3	0	0	1

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + e$$

$$Y = X\beta + e$$

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

The first column is just the sum of the other three ones

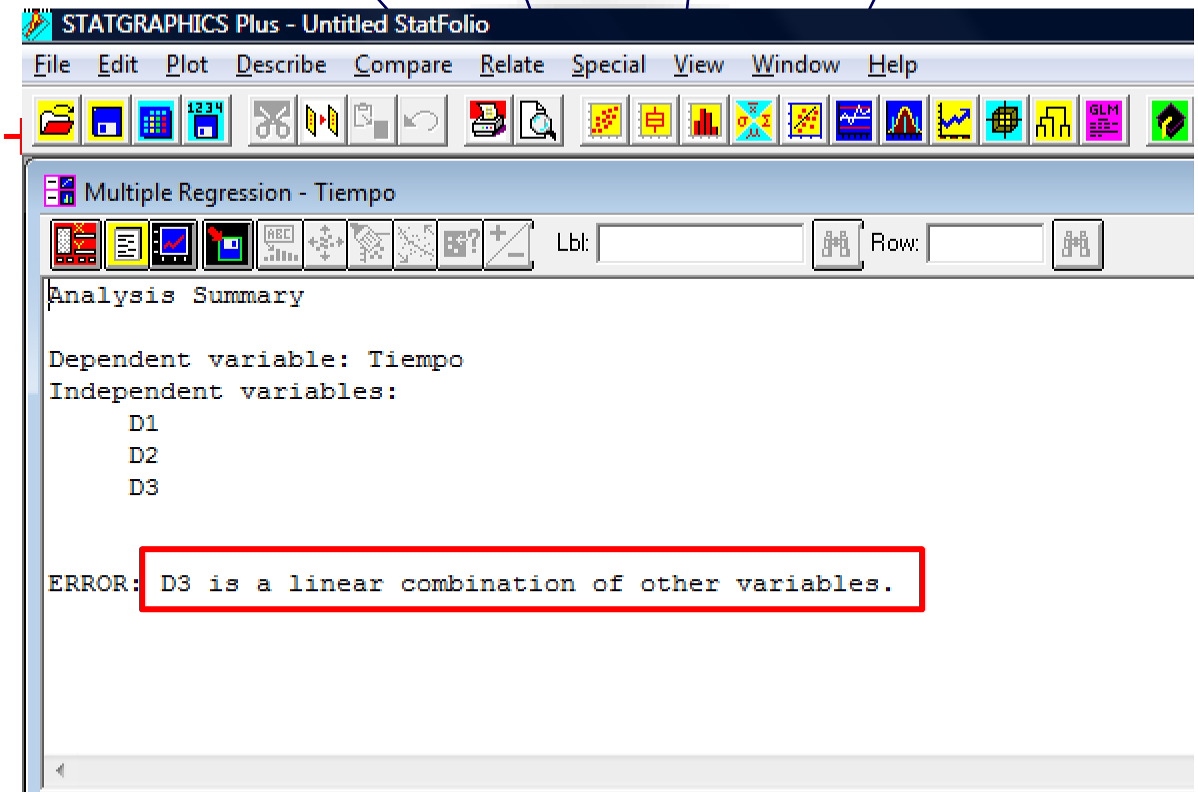
(X'X) is not invertible

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Therefore it is not possible to estimate the parameters

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + e$$

$Y = X\beta +$



The first column is just the sum of the other three ones → $(X'X)$ is not invertible

$$\hat{\beta} = (X'X)^{-1}X'Y$$ ← Therefore it is not possible to estimate the parameters

If we have G groups we have to make the model for only G-1 of them

$$Y = \beta_0 + \beta_1 D_1 + \dots + \beta_{G-1} D_{G-1} + e$$

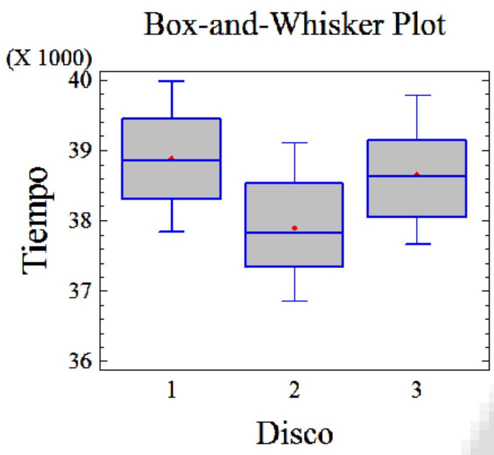
$E[Y \mid \text{group } G] = \beta_0$ ← The constant term is the mean of the excluded group

$E[Y \mid \text{group } g] = \beta_0 + \beta_g;\ g=1, \dots, G-1$ ← The g-th parameter is the difference between the mean of the selected group and the mean of the excluded one

Is the mean of the g-th group different from the mean of the excluded group G?

$$H_0 : \beta_g = 0$$

$$H_1 : \beta_g \neq 0$$



The best thing to do is to start by excluding the group with highest or lowest mean.

Multiple Regression Analysis

Dependent variable: Tiempo

$$Y = \beta_0 + \beta_1 D_1 + \beta_3 D_3 + e$$

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	37896,3	75,572	501,46	0,0000
D1	978,018	107,235	9,12029	0,0000
D3	747,922	118,785	6,29644	0,0000

The 2 is significantly better

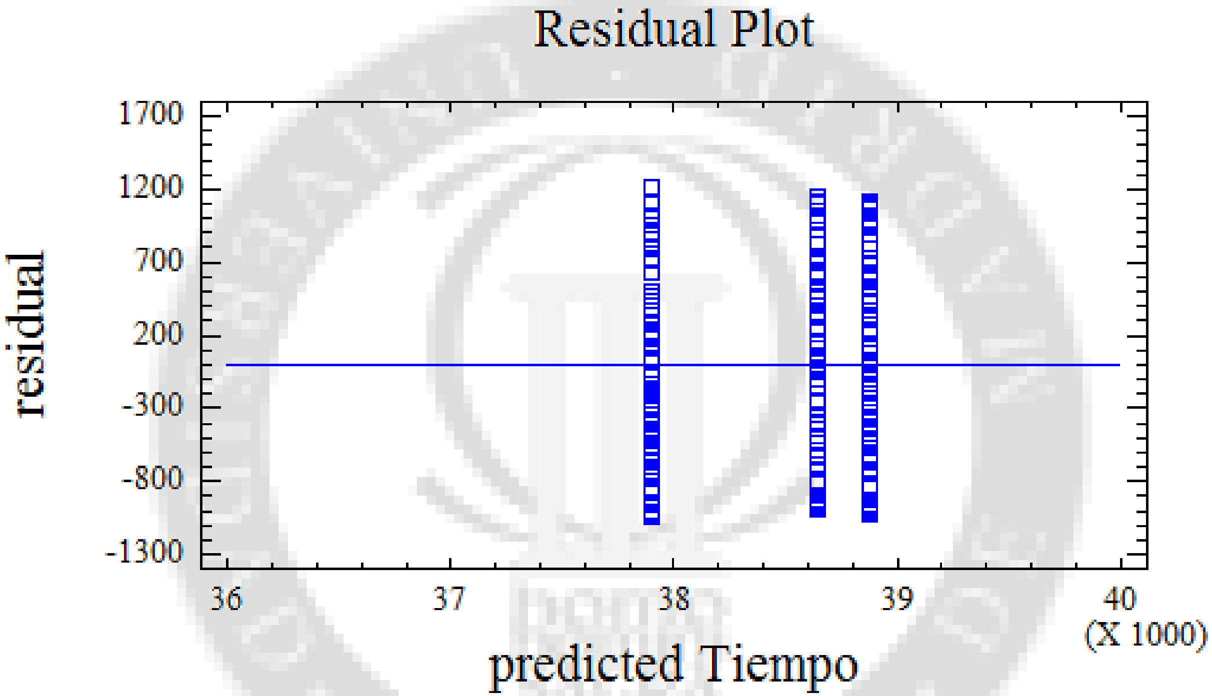
Multiple Regression Analysis

Dependent variable: Tiempo

$$Y = \alpha_0 + \alpha_2 D_2 + \alpha_3 D_3 + e$$

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	38874,4	76,0809	510,96	0,0000
D2	-978,018	107,235	-9,12029	0,0000
D3	-230,096	119,109	-1,93181	0,0548

There is not significant difference between the hard disks 1 and 3



Could you explain the obtained graph looks like this?