Regression Analysis

- ⋆ The **regression analysis** models the random link between two or more variables by a **functional relation**.

- ⋆ Doing a regression analysis means **to relate two or more variables** among them, making the hypothesized functional link esplicit.

- ⋆ One of the variables is the **dependent variable** (or **explanatory variable**), the others are the **independent variables** (or **explanatory variables**).

- ⋆ The regression model, can be interpreted as a cause - effect model. In this case, the independent variables are the causes; the dependent one is the effect.

For example, towards the regression techniques, the analyst can response to questions like:

- Which is the expected variation of the process efficiency if the temperature of the rector is raised, with the same concentration of the catalysis?
- Which is the "formula" that links the height to the weight and age?
- Knowing the engine size of a car, which is the mean error committed while we try to foresee the maximum speed with the best possible formula that links two considered variables?

- ⋆ In the relation among the efficiency of a chemical process, catalysis dose and reactor temperature, the efficiency is the dependent variable. The catalysis dose and the reactor temperature are the independent variables.

- ⋆ In the relation between the maximum speed of a car and the engine size, the maximum speed can be interpreted as the dependent variable. The engine size can be interpreted as the independent variable.

- ⋆ In the relation between the age of a person, his weight and his height, the age can be considered as the independent variable, while the weight and height both as independent and dependent variables.

The steps of a regression analysis can be summerised like follows.

- **Choice of the dependent variable and of the potentially explanatory variables**. Or rather, to determine the variable subject of the analysis and of the potentially explanatory variables. The choice of the explanatory variables is often determined by: the experience of the researcher, the previously done analyses and the available resources.
- **Collection of the matrix of data**. Or rather, creation of the vector $\underline{\boldsymbol{y}}$, that contains the n collected observations of the y response variable, and of the matrix of data $\boldsymbol{X}$, composed by the n records (rows) of the p explanatory variables (columns).
- **Choice of the functional relation and selection of the explanatory variables, estimation and validation of the model**. The realisation of a good model requires not only the knowledge of the undergone statistical techinques, but also a critical judgment and experience on the analysis phenomenon.
- **Use of the model** realised for **esplicative** and / or **predictive** aims.

An exemplification of the vector $\underline{\boldsymbol{y}}$ and of the matrix $\boldsymbol{X}$ is reported in the following table.

Dependent variable ($\underline{\boldsymbol{y}}$)	Independent variables ($\boldsymbol{X}$)					
y_1	1	x_{11}	$\dots$	x_{1j}	$\dots$	x_{1p}
y_2	1	x_{21}	$\dots$	x_{2j}	$\dots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		
y_i	1	x_{i1}	$\dots$	x_{ij}	$\dots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		
y_n	1	x_{n1}	$\dots$	x_{nj}	$\dots$	x_{np}

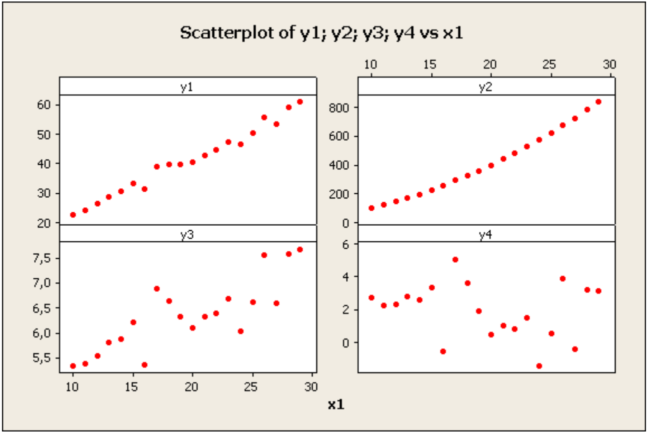
In the **choice, estimation and validation of a model** phase, the following points are crucial:

- The **choice of the variables** is often **limitated** by the available resources and by the insufficient previous knowledge of the phenomenon;
- The **validation of the model** consists in the quantification of the esplicative excellence of the model and of the prediction errors characteristics.

- In mathematical words, the y is called the **response variable** and x_1, x_2, ..., x_p the chosen **explanatory variables**, a generic **regression model** expresses the link between the y and the x like:

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{ip}) + \varepsilon_i.$$

- In other words, the ith observation of the response variable is expressed as the sum of **structural element** ($f(\cdot)$), expressed by a specified function of the chosen explanatory variables, and the sum of a **random element**, ε_i.
- The random element, or error term, does not show something wrong, but it represents a residual element of y that is not taken by the hypothesized relation that depends on not predictable factors.



An example of modeling choices is written in the examples of the plot. Each of the panel represents possible experimental relations among couples of variables.

- ⋆ $y1$ and x show a couple of variables that can be modeled with a **linear trend**. However, in this kind of trend, there is a dispersion of data compared to the central trend.

- ⋆ $y2$ and x show a couple of variables that can be well modeled with **parabolic** model. The functional link seems to be almost perfect.

- ⋆ $y3$ and x show a more uncertain relation between the couple of analysed variables, because the **effect of the residual variability** is really strong. It can be hypothesized both a linear relation, and a not linear relation between the data.

- ⋆ Finally $y4$ and x show an **almost random distribution** of the couples of the noticed observations.

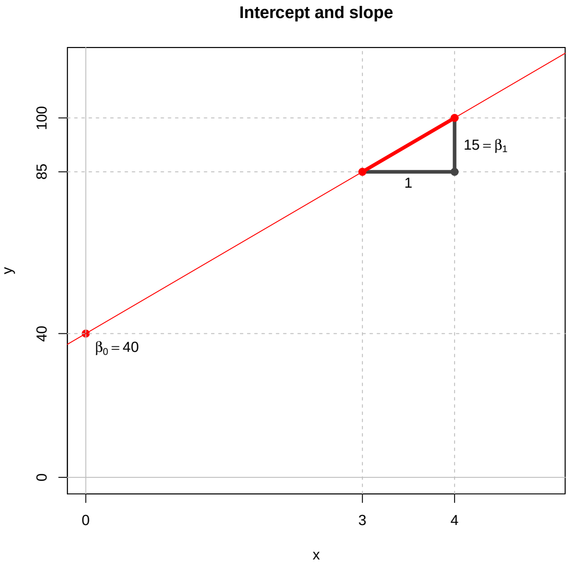
- ⋆ It is a researcher's task that of proving the different hypothesized functional relations and choosing the model that effectively represents the behavior of the phenomenon.

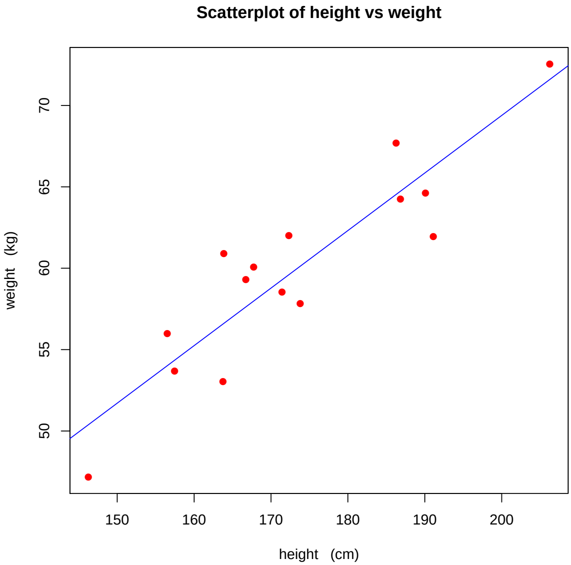
- If we want to study the relation between weight and height, it is possible to hypothesize that the weight is influenced by the height. In a regression model, the height is the independent variable and the weight is the dependent one.
- If y is the dependent variable and x the independent one, it is possible to write the relation like:

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

- Each observation y_i is characterised by a **deterministic element** (or structural element), $f(x) = \beta_0 + \beta_1 \cdot x_i$, and by a **random element**, ε_i.
- The presence of the random element, ε_i, makes the y_i random values. In other words, it is not possible to know exactly the value of y_i, if we only know the value of x_i.

- As in the previous case, when the deterministic element is represented by a straight line, we talk about a **simple linear regression**.
- In the case of a simple linear regression:
  - β_0 is the **intercept**, and it shows the y value when x is equal to 0;
  - β_1 is the **slope** (or gradient), and it shows how much y changes when x increases in one unity.
- The plot in the following slide shows the geometrical meaning of the intercept and the slope of a line.
- The line represented in the plot has an equation $y = 40 + 15 \cdot x$.


Intercept and slope
y
100
85
40
0
x
0
3
4
$15 = \beta_1$
1
$\beta_0 = 40$


Scatterplot of height vs weight
weight (kg)
70
65
60
55
50
150
160
170
180
190
200
height (cm)

- ⋆ The plot in the previous slide shows the weights and the corresponding height of $n = 15$ people.
- ⋆ The line represents the linear regression estimated between weight and height.
- ⋆ The estimated regression line is the line that "better crosses the points".
- ⋆ It is possible to write the estimated regression line's equation like $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively represent the estimation of β_0 and β_1.
- ⋆ The differences among the y_i observed values and the y_i estimated values through the regression line $(\hat{y}_i)$, are called **residuals**:

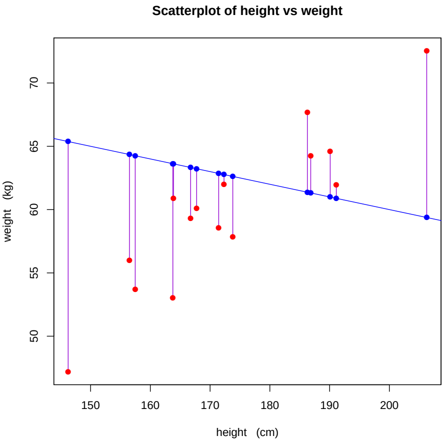
$$e_i = y_i - \hat{y}_i$$

- ⋆ How is it possible to obtain the estimations of β_0 and β_1, or rather those values for which the line “better crosses the points”?

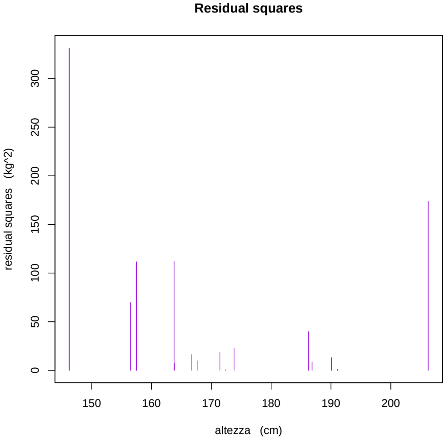
- ⋆ One of the most used criteria for the parameters’ estimation ($\hat{\beta}_0$ and $\hat{\beta}_1$) of the regression model is the **ordinary least squares criterion**.

- ⋆ The ordinary least squares criterion finds the parameters of the curve that **minimize the sum of squared residuals**.

- ⋆ The next slides show three possible lines.

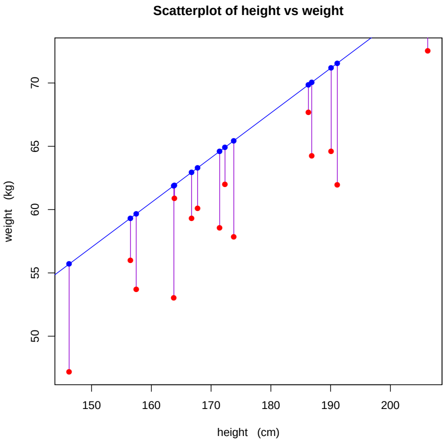


The purple segments represent the residuals, or rather the differences between the true value of the weight (red points) and the estimated values of the lines (blue points).

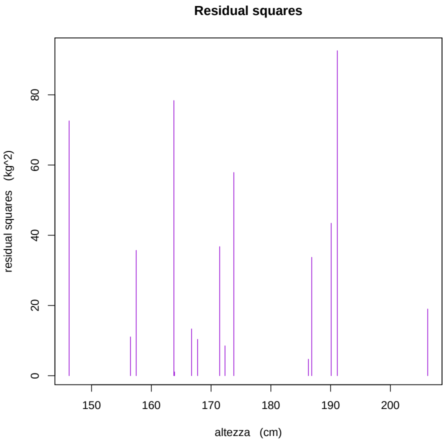


The purple segments represent the squared residuals.

The sum of squared residuals is 936.22.

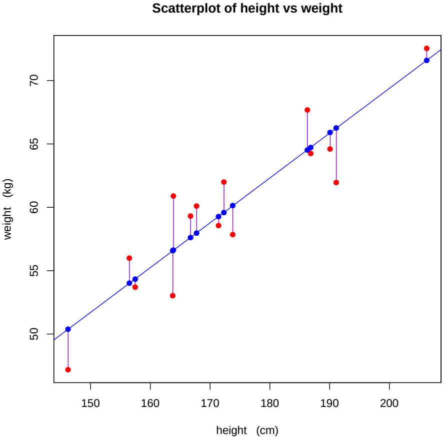


The purple segments represent the residuals, or rather the differences between the true value of the weight (red points) and the estimated values of the lines (blue points).

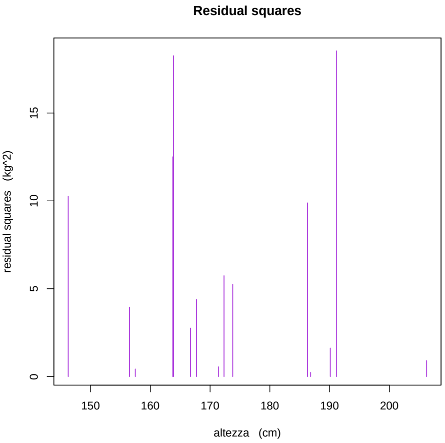


The purple segments represent the squared residuals.

The sum of squared residuals is 519.13.



The purple segments represent the residuals, or rather the diffferences between the true value of the weight (red points) and the estimated values of the lines (blue points).



The purple segments represent the squared residuals.

The sum of squared residuals is 95.37.

- ⋆ From the previous plots it is possible to notice how the first two lines have a greater sum of squared residuals. On the contrary, the third one has a lower sum of squared residuals.
- ⋆ As a matter of fact, the first line has been calculated using random values for the intercept and the slope, that certainly do not “better cross the points”.
- ⋆ In the second line it has been used the estimated value with the least squares method for the slope, and a random value for the intercept.
- ⋆ The third line is the regression one, estimated with the least squares method.

- ⋆ After having shown, in an intuitive way, what is the least squares method, we will briefly show its corresponding formula.
- ⋆ In mathematical words, we look for the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ by which we have

$$(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} \left\{ \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 \cdot x_i)\right]^2 \right\}$$

- ⋆ After appropriate mathematical calculations, it is possible to find the formula that allows to calculate $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- With the data we used in the plot and we wrote in the table of the following slide, we will find that $\hat{\beta}_0 = -1.32$ and $\hat{\beta}_1 = 0.35$.
- The regression line represented in the graphic, is the line whose equation is $y = -1.32 + 0.35 \cdot x$.
- As we said before, from the mathematical point of view, the **intercept** ($\hat{\beta}_0$) represents the y value when x is equal to zero. The intercept does not often have a practical meaning, mainly when there are phisical measurements. In the example, it has no sense to think that the height can be equal to zero and then the weight can be negative.
- The **slope** ($\hat{\beta}_1$) represents the (mean) growth of the y when an unit of x increases. In the example, it is possible to state that, when the height increases one centimeter, the weight increases 0.35 kilograms.

i	weight (y)	height (x)
1	60.08	167.75
2	57.83	173.80
3	59.29	166.74
4	61.94	191.12
5	64.24	186.84
6	72.55	206.25
7	58.54	171.44
8	67.68	186.28
9	53.69	157.47
10	47.18	146.25
11	60.89	163.88
12	53.04	163.77
13	62.00	172.33
14	64.61	190.10
15	56.00	156.51
mean	59.97	173.37

- It has been considered the easiest case in which the dependent variable is the linear function of one independent variable so far.

- This regression model is called **simple linear regression model**.

- The model often hypothesizes that the dependent variable is the linear function of more independent variables:
$$y_i = f(\underline{\boldsymbol{x_i}}; \underline{\boldsymbol{\beta}}) + \varepsilon_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \cdots + \beta_p \cdot x_{ip} + \varepsilon_i$$
where p represents the number of independent variables.

- In this case the model is called **multiple linear regression model**.

- ⋆ Also in this case, for the estimation of the parameters β_j, $j = 0, 1, \ldots, p$, the ordinary least squares method is used.

- ⋆ In mathematical words, the formulation of the estimation problem is similar to the previous one:

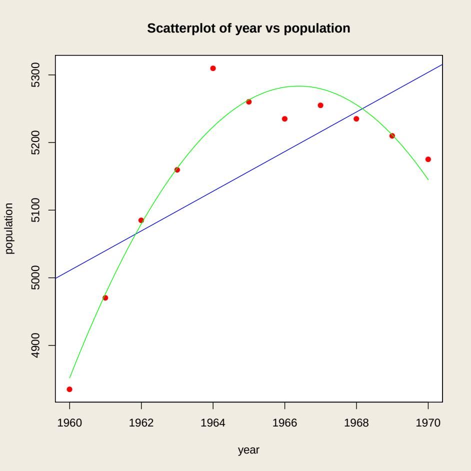
$$\underline{\hat{\boldsymbol{\beta}}} = \min_{\underline{\boldsymbol{\beta}}} \left\{ \sum_{i=1}^{n} \left[y_i - f\left(\underline{\boldsymbol{x_i}}; \underline{\boldsymbol{\beta}} \right) \right]^2 \right\}$$

- ⋆ Similarly to the simple linear regression case, the **coefficients** β_j indicate the awaited variance of the response variable for an unitary increase of the relative explanatory variable x_j, when the value of the other variables is constant.

⋆ In the previous slides it has been used the matrix notation too:

- the vector $\underline{\boldsymbol{y}}$ represents the vector, with length n, of the dependent variables;
- the vector $\underline{\boldsymbol{\beta}}$ represents the vector, with length $p+1$ (it is important to consider β_0), of the parameters that have to be estimated;
- the matrix $\boldsymbol{X}$ represents the matrix of the independent variables.
- the matrix $\boldsymbol{X}$ has sizes n x $(p+1)$ and the first column contains all the values equal to 1. This column allows to calculate the value of the constant (or **intercept**).
- The vector $\underline{\boldsymbol{x_i}}$ represents the ith row of the matrix $\boldsymbol{X}$.

As we have already seen in the slide about the functional relation, the relation between the x and the y can be not linear, as shown in the following graphic.



- The plot shows the population of a town in the 1960s.
- As we can see, the estimated linear regression, represented by the blue line, does not catch the relation between the x and the y.
- The parabola, the green one, seems to be better in catching this relation.
- In similar cases, it is possible to use a **quadratic regression** analysis, whose model can be formulated like

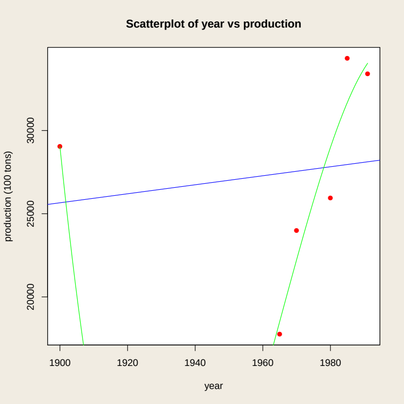
$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i1}^2 + \varepsilon_i$$

- This model, even though it is not linear in the variable, is **linear in the parameters**.
- As a matter of fact, if we consider that it is possible that $x_{i2} = x_{i1}^2$, this model can be considered as a particular case of a multivariate linear regression model.

- ⋆ Sometimes the polynomial can be a third degree one (**cubic regression**), or an higher degree one (**polinomial regression** of r degree).
- ⋆ Anyway, even though the model improves when the parameters increase, it is important to remember that a **parsimonious model**, or rather one with few parameters, is easiest to interpret.
- ⋆ Checking of the parameters' significance and using criteria such as the R^2 can help to choose the best model. These methods are described later.
- ⋆ Obviously, a polinomial model can include more variables, for example:

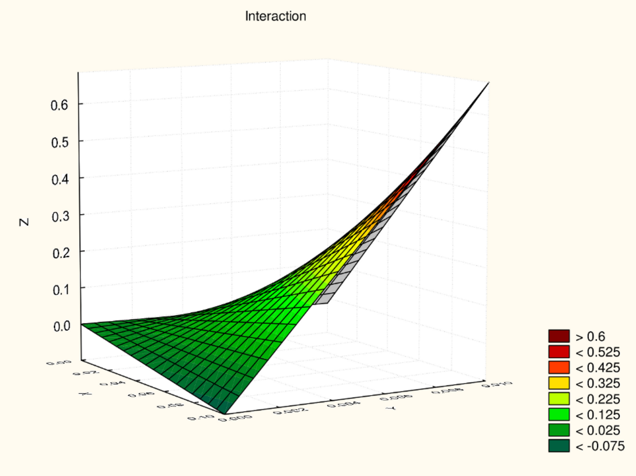
$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i1}^2 + \beta_3 \cdot x_{i1}^3 + \beta_4 \cdot x_{i2} + \varepsilon_i$$

The plot shows how the relation between year and citrus fruits' production in Italy can be described by a cubic regression model. The green curve represents the curve $y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 \cdot x_{i1}^2 + \hat{\beta}_3 \cdot x_{i1}^3$.



The example is taken by A. Quarteroni, F. Saleri, Calcolo Scientifico: Esercizi E Problemi Risolti Con Matlab E Octave, Springer

- ⋆ In some occasions, the relations that link the independent variables to the dependent one can be not "released" among them, but they can interact.
- ⋆ The synergic effect of the presence of two influent factors, can be "depressive" or "explosive" on the response variable.
- ⋆ For example, it is known that assuming alcohol together with drugs produces a synergic effect on the lucidity and the reactivity of a person. The synergic effect reduces the ability better than the sum of the effects.
- ⋆ In statistical and "regression analysis" words, this kind of effect, "depressive" or "explosive", is called **interaction effect** (or simply "interaction").



A plot example of an interaction effect between two continuous variables (x and y) on the dependent variable (z).

- It is possible that it is necessary to express the interaction effects among independent variables but we do not have specific information on the functional shape the relation has. In this case "product" operators are used.

- For example, going back to the reaction times in function of the quantity of drug and alcohol assumed by a person, in order to express the joint relation of the explanatory variables (called x_1 e x_2) on the response variable(y), it can be used the following:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i1} \cdot x_{i2} + \varepsilon_i$$

where i, $i = 1, \ldots, n$, represents the ith sample observation.

- ⋆ The model can be elaborated and estimated by the creation of a third variable given by the row-by-row product of x_1 and x_2 values ($x_{i3} = x_{i1} \cdot x_{i2}$) and then adding it to the model itself:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + \varepsilon_i$$

- ⋆ At this point, the model can be estimated like a normal model of multiple regression.

- Through the regression it is possible to analyse the qualitative effects on the response variable too.

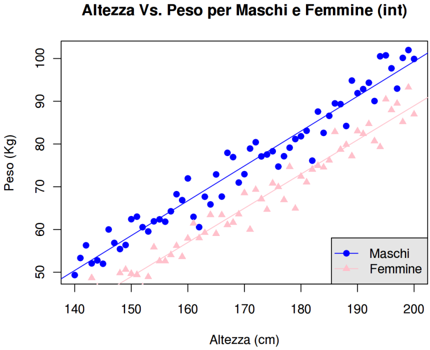
- This is usually done using the **dummy variables**.

- The dummy variables (we will indicate them with the letter Z) are (independent) variables that can assume only two values (0 o 1) and can be useful to indicate presence/lack of qualitative characteristics. The dummy variables conventionally:
  - assume value **0** when the interest characteristic is **absent**;
  - assume value **1** when the interest characteristic is **present**.

- The dummy variables are used also when a characteristic can assume just two different modalities, or rather where there are **dichotomous variables**. Some examples of these variables can be: male/female, minor/adult, employed/unemployed...

  In this case the choice to put a modality equal to 1 and the other equal to 0, or vice versa, is arbitrary and does not influence the results.

- ⋆ It is supposed that we want to analyse the relation between height and weight of a series of people, distinguishing between males and females.
- ⋆ It is hypothesized that the functional relation is locally linear and equal both for men and women, unless an intercept gap:



- ⋆ It is possible to express in mathematical words this relation as a model like the following one:

$$y_i = \beta_0 + \beta_D \cdot Z_i + \beta_1 \cdot x_{i1} + \varepsilon_i$$

- ⋆ Z_i is the dummy independent variable “female” that will have:
  - value 0 if the ith subject is male;
  - value 1 if the ith subject is female.

- ⋆ β_D is the coefficient associated to the dummy variable that identifies the “intercept gap” between males and females.

- ⋆ If the subject is a male, as $Z_i = 0$, the model will be like:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \varepsilon_i = \beta_{0M} + \beta_1 \cdot x_{i1} + \varepsilon_i$$

- ⋆ If the subject is a female, as $Z_i = 1$, the model will be like:

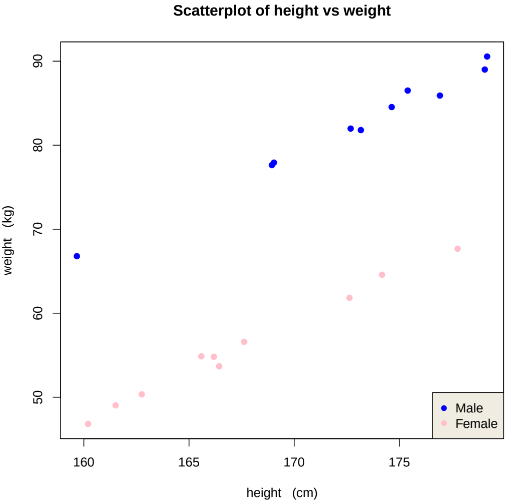
$$y_i = (\beta_0 + \beta_D) + \beta_1 \cdot x_{i1} + \varepsilon_i = \beta_{0F} + \beta_1 \cdot x_{i1} + \varepsilon_i$$

- ⋆ As it can be noticed, the two models are almost identical, except for intercepts (β_{0M} and β_{0F}) that will be different for males and females.

Example. To study the linear relation between weight and height, according to the gender.

Men		Women	
Height	Weight	Height	Weight
167.8050	76.04820	162.2464	49.02631
174.5437	84.51962	159.5950	45.25246
175.9317	86.13436	164.6700	53.92118
175.3069	83.19130	174.5236	65.15938
177.6868	88.64845	162.6091	50.31901
170.0375	78.77422	162.9598	50.66751
173.8890	84.85968	165.0699	52.52929
180.0417	91.75516	162.4670	50.16048
170.7094	78.73333	165.4449	56.30775
174.7580	82.41510	165.6076	55.06725

The following plot shows the data.



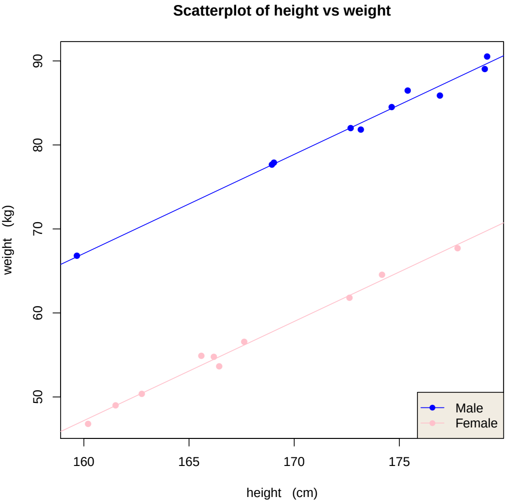
The estimated model can be expressed like:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i + \hat{\beta}_2 \cdot Z_i$$

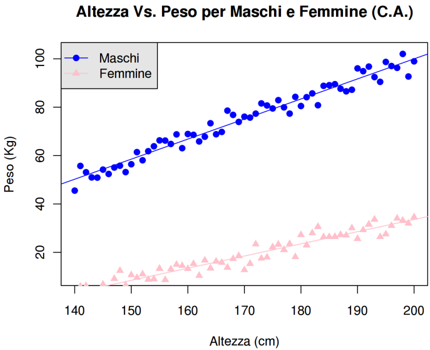
where $\hat{y}_i$ is the weight estimated value by the regression of the ith subject, $\hat{\beta}_0$ and the intercept for the women, $\hat{\beta}_1$ is the coefficient associated to the height (x_i) and $\hat{\beta}_2$ is the estimation of the difference between men and women weight.

	estimate	Std. Error	t value	Pr($>$ |t|)
Intercept	-141.53018	4.82671	-29.32	5.35e-16
x	1.17950	0.02879	40.97	$<$ 2e-16
Z	19.88756	0.35212	56.48	$<$ 2e-16

The plot shows the data and the regression lines that have just been estimated.



- ⋆ If we want to analyse, in this case too, the relation between height and weight of a series of subjects, distiguishing between males and females.
- ⋆ If we hypothesize that the functional relation is locally linear and equal both for males and females, unless a slope gap:



- ⋆ It is possible to express this relation in mathematical words with a model like the one that follows:

$$y_i = \beta_0 + \beta_D \cdot Z_i \cdot x_{i1} + \beta_1 \cdot x_{i1} + \varepsilon_i$$

- ⋆ Z_i is the dummy independent variable "female" that will have:
  - value 0 if the ith subject is male;
  - value 1 if the ith subject is female.

- ⋆ β_D is the coefficient associated to the dummy variable that identifies the "slope gap" between males and females.

- If the subject is a male, as $Z_i = 0$, the model is like:

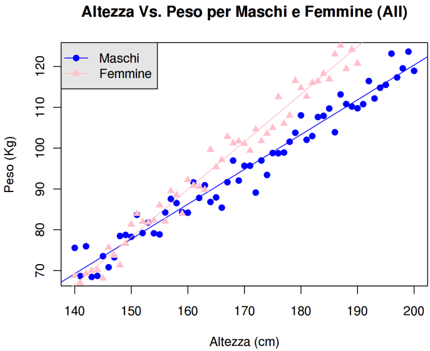
$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \varepsilon_i = \beta_0 + \beta_{1M} \cdot x_{i1} + \varepsilon_i$$

- If the subject is a female, as $Z_i = 1$, the model is like:

$$y_i = \beta_0 + (\beta_D + \beta_1) \cdot x_{i1} + \varepsilon_i = \beta_0 + \beta_{1F} \cdot x_{i1} + \varepsilon_i$$

- As we can notice, the two models are almost identical, except for slopes (β_{1M} and β_{1F}), that will be different for males and females.

- ⋆ If we want to analyse the relation between height and weight of a series of subject, distinguishing between males and females.
- ⋆ If it is hypothesized that the functional relation is locally linear and equal to males and females, unless a slope and an intercept gaps:



- It is possible to express this relation in mathematical words with a model like the one that follows:

$$y_i = \beta_0 + \beta_{0D} \cdot Z_i + \beta_{1D} \cdot Z_i \cdot x_{i1} + \beta_1 \cdot x_{i1} + \varepsilon_i$$

- Z_i is the independent dummy variable “female” that will have:
  - value 0 if the ith subject is a male;
  - value 1 if the ith subject is a female.

- β_{0D} is the coefficient associated to the dummy variable that identifies the “intercept gap” between males and females.
- β_{1D} is the coefficient associated to the dummy variable that identifies the “slope gap” between males and females.

- If the subject is a male, as $Z_i = 0$, the model will be:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \varepsilon_i = \beta_{0M} + \beta_{1M} \cdot x_{i1} + \varepsilon_i$$

- If the subject is female, as $Z_i = 1$, the model will be:

$$y_i = (\beta_0 + \beta_{0D}) + (\beta_{1D} + \beta_1) \cdot x_{i1} + \varepsilon_i = \beta_{0F} + \beta_{1F} \cdot x_{i1} + \varepsilon_i$$

- As we can noticed, the two model will be different because of both the intercepts (β_{0M} e β_{0F}) and the slopes (β_{1M} e β_{1F}), that will be different for males and females.

- ⋆ The use of the dummy variables is really flexible. It allows to build even complex models.

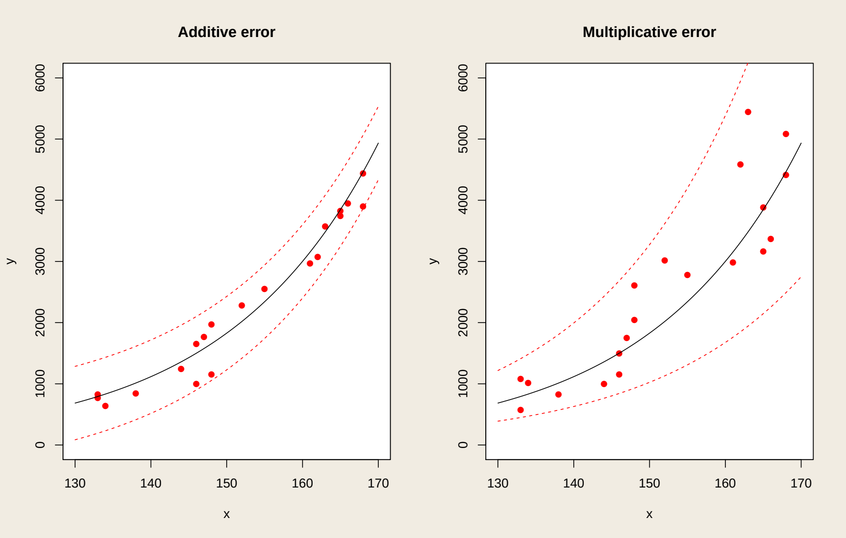
- ⋆ It has to be noticed, for example, that the use of dummy can be extended to the case of qualitative variables that can have more than two modalities (for example: income brackets, colors, used materials etc.).

- ⋆ If we want to analyse the relation between weight and height of a series of subjects, distinguishing them according to the continent they come from, it is possible to consider five different dummy variables. Each of these variables is equal to 1 if the subject comes from that continent and 0 if he does not.

- ⋆ In the practice, the construction of the dummy variables is delegated to the statistical analysis' softwares that make the work easier.
- ⋆ It has to be noticed that, when we study the results given by the software, the dummy variables that have been used will be one less than the categorical variables modalities, e.g. than the number of continents.
- ⋆ In the case of the differences between males and females we saw in the previous examples, the dummy regarded just the "female" and the corresponding coefficients indicated the gaps compared to the"male".
- ⋆ Similarly, in the case of the continents, the dummy only regards four continents and the corresponding coefficients will indicate the gaps compared to the fifth continents, considered the reference one.

- Sometimes it can happen that the relation between the dependent variable and the independent variables is not linear even in the parameters.
- Sometimes it is possible that a not linear model in the parameters can be transformed in a linear one with appropriate transformations.
- An exhaustive report of the different situations where the not linear model can be linearised is not possible in this course.
- We will following give an example with two models: one that cannot be linearised and another that, on the contrary, can be linearised.

- It is supposed that the relation is expressed by functions like $y_i = \beta_0 \cdot e^{\beta_1 \cdot x_i} + \varepsilon_i$ or $y_i = \beta_0 \cdot e^{\beta_1 \cdot x_i} \cdot \varepsilon_i$.
- The two models distinguish one from the other by the relationship, respectively additive and multiplicative, of the error term.
- The following slide shows two plots, one with the additive error term, where the confidence bands are equidistant to the regression curve, and the other one with the multiplicative error term, where the confidence bands detach themselves when x increases.


Additive error
Multiplicative error
y
x
0
1000
2000
3000
4000
5000
6000
130
140
150
160
170

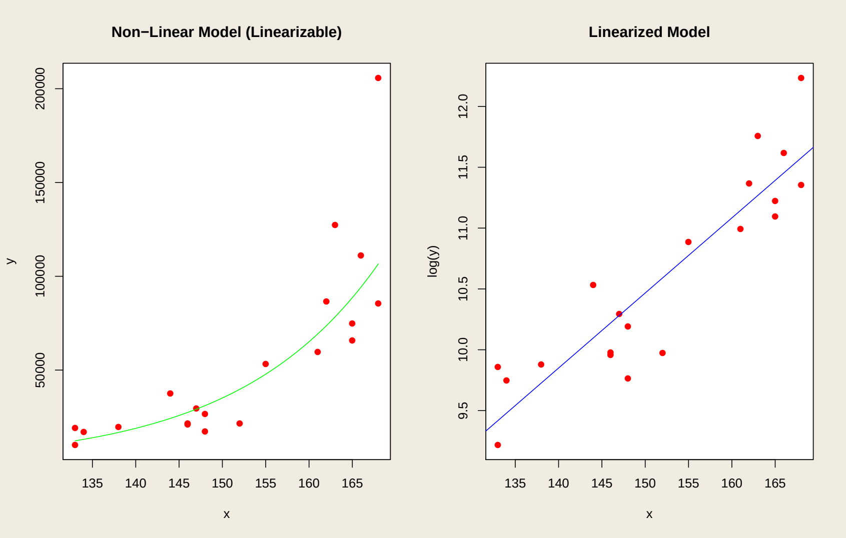
- ⋆ The first case requires the use of not linear techniques in order to find the regression curve.
- ⋆ The second case, thanks to the logarithm characteristics, allows to lead to a linear model.

$$y_i = \beta_0 \cdot e^{\beta_1 \cdot x_i} \cdot \varepsilon_i$$

$$\ln(y_i) = \ln(\beta_0 \cdot e^{\beta_1 \cdot x_i} \cdot \varepsilon_i)$$

$$y_i^* = \beta_0^* + \beta_1 \cdot x_i + \varepsilon_i^*$$

- ⋆ The parameters estimations are: $\hat{y}_i = e^{\hat{y_i}^*}$, $\hat{\beta}_0 = e^{\hat{\beta_0}^*}$ and $\hat{\varepsilon}_i = e^{\hat{\varepsilon_i}^*}$.
- ⋆ It has to be noticed that $\hat{y_i}^*$ is a not distorted estimator of y_i^*, but $\hat{y}_i$ it is not a not distorted estimator of y_i.
- ⋆ Furthermore, as for the hypothesis of the linear model ε_i^* has to have a normal distribution, then ε_i has to have a lognormal distribution.


Non−Linear Model (Linearizable)
y
200000
150000
100000
50000
135
140
145
150
155
160
165
x
Linearized Model
log(y)
12.0
11.5
11.0
10.5
10.0
9.5
135
140
145
150
155
160
165
x

- The **main hypothesis** of the linear regression is that the structural link between the dependent variable, the explanatory variables' parameters and the error term is, as it has been expressed in the previous formula, linear.

- Futher **hypothesis** on the behavior of the **error** $\underline{\varepsilon}$ are added to the main hypothesis. Or rather:

  1. $E(\varepsilon_i) = 0 \qquad (i = 1, \ldots, n)$
  2. $\boldsymbol{X}$ not stochastic
  3. $\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \qquad (i = 1, \ldots, n)$
  4. $Corr(\varepsilon_i, \varepsilon_{i-j}) = 0 \qquad (i = 1, \ldots, n; 0 < j < i)$

- ⋆ The first hypothesis, $\boldsymbol{E(\varepsilon_i) = 0}$, indicates the **not systematicity of the errors**. To accept this hypothesis means to affirm that the model does not systematically "make mistakes" in excess or in fault compared to the central trend of the phenomenon.

- ⋆ The second hypothesis affirms that **the matrix $\boldsymbol{X}$ is not stochastic**, or rather that the independent variables are "firmly" determined by the experimenter. The independent variables are not affected by gathering error; if this is present, it has to be considered unimportant.

- ⋆ The third hypothesis, $\varepsilon_i \sim \ \boldsymbol{N(0,\, \sigma_\varepsilon^2)}$, beyond restating the first one, affirms that the **errors distribution has to be gaussian**, and **the errors variability has to be constant**. This means that the dispersion of the prediction error is not influenced, in a systematical way, by the value of the response variable or by the value of the explanatory variables.

  To accept that the error term follows a **normal distribution**, with null mean and constant variance, allows:

  - the development of hypothesis tests on the parameters (t test);
  - the development of hypothesis tests on the whole model;
  - a criterion for the identification of the outlier values, or rather of those observation that present an excessive gap compared to the values that have been anticipated by the model.

- The fourth hypothesis, $Corr(\varepsilon_i, \, \varepsilon_{i-j}) = 0$, affirms that **the prediction errors are not correlated in sequence**. To accept this hypothesis means being able to affirm that the time order of the observations collection does not affect the behavior of the response variable.
- The four previous statements represent the hypothesis that undergo the gaussian linear model.
- If the empirical data do not confirm these hypothesis, it is possible that the model is bad specified in the hypothesized functional relation and / or in the missing consideration of an important explanatory variable.
- The **error normality test** is one of the main point of the model's validation phase. The **graphical analysis** of the residuals and the **Anderson-Darling test** are commonly used to check this hypothesis.

- In the previous slides we showed how it is possible to calculate the coefficients estimations $\underline{\boldsymbol{\beta}}$ towards the ordinary least squares method. The values $\underline{\hat{\boldsymbol{\beta}}}$ represent the **estimates** of $\underline{\boldsymbol{\beta}}$.

- Each observations is given by the sum of a **structural componenent** (or deterministic one) and by an **aleatory element** ε_i. As consequence, the sample of values $\underline{\boldsymbol{y}}$ that has been collected, with the same values of the matrix $\boldsymbol{X}$, represents one, among the infinite samples, of the realizations of the random ε_i.

- The vector $\underline{\hat{\boldsymbol{\beta}}}$ results then a **random variable** with a specific distribution, which can be calculated in the linear regression.

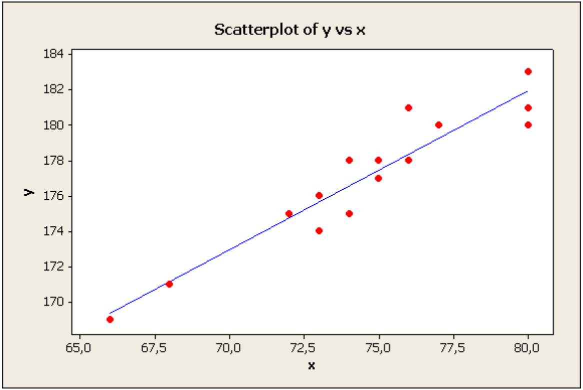
- ⋆ If the five statements of the gaussian linear model are valid, it can be demonstrated that:
  - $E(\underline{\hat{\boldsymbol{\beta}}}) = \underline{\boldsymbol{\beta}}$, or rather that $\underline{\hat{\boldsymbol{\beta}}}$ is one **not distorted estimator** of the parameters vector $\underline{\boldsymbol{\beta}}$;
  - $Var(\hat{\beta}_j) = \sigma_\varepsilon^2 \cdot g_j(\boldsymbol{X})$, where σ_ε^2 is the error variance ε $g_j(\cdot)$ is a function (that can be calculated, deterministic and different for each j) of the data $\boldsymbol{X}$ and, in particular, of the number n of sample observations: $g_j(\boldsymbol{X})$ tends to decrease when n increases.
- ⋆ It can be demonstrated that $\underline{\hat{\boldsymbol{\beta}}}$ is an **efficient estimator**, or rather with the smallest variance among the estimators of the same parameter, and it is with mean null error.
- ⋆ Furthermore, if the distributive normality of $\underline{\varepsilon}$ is valid, it is possible to demonstrate that also the single estimated parameters $(\hat{\beta}_j)$ follow a normal distribution.

- ⋆ If it wants to be estimated the variance estimations ($Var(\hat{\beta}_j)$), the result has to be an estimation $\hat{\sigma}_\varepsilon^2$ of σ_ε^2.

- ⋆ The term $\hat{\sigma}_\varepsilon^2$ represents the **estimated standard deviation** of the prediction error. The standard deviation of the prediction error is the residual variability for the dependent variable $\underline{\boldsymbol{y}}$, once that the variability due to the “predictable” or deterministic element $f(\cdot)$ of the model has been removed. $\hat{\sigma}_\varepsilon^2$ is calculable as the square of:

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{\sum_{i=1}^{n} \left[y_i - f\left(\underline{\boldsymbol{x_i}}; \underline{\boldsymbol{\beta}}\right)\right]^2}{n-p-1}}$$

- ⋆ In the slides about the simple linear regression it has been presented an example of parameters estimation with the ordinary least squares method.
- ⋆ Another example of estimation is presented afterwards. The plot shows the relation between a variable x and a variable y.



- The relation between y and x has been modeled with a simple linear regression.
- The estimated **coefficients** of the regression line, reported above, are $y = 109.95 + 0.90 \cdot x$
- The coefficient 0.90 is the **slope**. It indicated the (mean) increase of the dependent variable if x increases one unity.
- The value 109.95 represents the **intercept** of the line. It indicates which is the mean value of y if x is (hypothetically) equal to 0. The pratic interpretation of the intercept depends on the sense the null value in all the explanatory variables has on the application taken into consideration.
- The value of $\hat{\sigma}^2_\varepsilon$ estimated for this model is equal to 1.244. This value represents the estimation of the **residual variance** of the y, after the link between the y and the x.

- ⋆ The **excellence of a regression model** as regards its adaptation to the data depends on the variability quota of the dependent variable that the model is able to explain.
- ⋆ If the intercept is included in the regression function, the **total deviance** of the dependent variable (**SST**) can be decomposed in the sum of two elements, **SSR** and **SSE**:

$$\mathsf{SST} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$\mathsf{SST} = \mathsf{SSR} + \mathsf{SSE}$$

$$\mathsf{SSR} = \sum_{i=1}^{n} \left(y_i - f(\underline{\boldsymbol{x_i}}; \underline{\boldsymbol{\beta}})\right)^2$$

$$\mathsf{SSE} = \sum_{i=1}^{n} \left(f(\underline{\boldsymbol{x_i}}; \underline{\boldsymbol{\beta}}) - \overline{y}\right)^2$$

- **SST**: *Sum of Squares of Total*
- **SSE**: *Sum of Squares Explained*
- **SSR**: *Sum of Squares of Residuals*

- In other words, the whole variability of the phenomenon we are studying (SST) can be divided into two elements: the “explained” element of the regression (SSE) and the “residual” element (SSR).

- Towards the sum of squares of the differences between the expected values and the mean, **SSE** measures the variabilty quota that the model catched.
- Towards the sum of squares of the differences between the expected values and the observed ones, **SSR** measures the variability side of the y that the regression model did not catch.

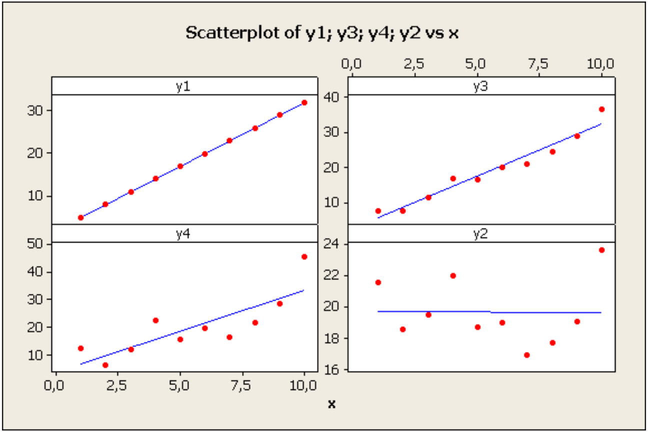
- Thence, a whole excellence index of the model is given by the weight of SSE on SST. The greater is the quota of SSE on SST, the greater is the percentage of variability of the y explained by the regression model.

- The adaptation index that sums up the relation between the elements of the variance is the **coefficient of determination**:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- ⋆ R^2 (R-squared) is an index that is always included between 0 and 1, included;
- ⋆ R^2 is, at the maximum, equal to one when all the variability is explained by the regression function (or rather, when the regression function “crosses” all the points);
- ⋆ R^2 is, at the minimum, equal to zero when the regression function is not more useful than the sample mean in order to describe the phenomenon;
- ⋆ R^2 assumes intermediate values in function of the “adaptation level” (or rather of the ability to explain the phenomenon) of the regression function.
- ⋆ The R^2 values is decreasing with the number of explanatory variables introduced in the regression function. This means that it is possible to obtain a R^2 value really near to 1 simply adding explanatory variables even though they are not really “linked” with the analysed phenomenon.

- ⋆ The value of the coefficient of determination does not depend on the "direction" of the estimated relation between independent variables and the dependent variable. Or rather, the value of the coefficient of determinaton does not depend on the "sign" of the estimated parameters but by the residual variability degree for the y that results from leaving the "captured" variability from the estimated model.
- ⋆ To raise the adaptation index does not have to overlook the analytic parsimony and the simplicity of the interpretation. In general words, it is preferable to have values of the R^2 coefficient slightly lesser if the model that results is more easily interpretable;
- ⋆ In general, a R^2 value can be defined as good if it has been compared to the R^2 value of alternative models for the same phenomenon.



These plots represent linear relations between the y and the x with decreasing R^2 values. The panel in the top left-hand corner shows a model with $R^2 = 1$. The panel in the low right-hand corner shows the model with $R^2 = 0$.

- As the R^2 value is also influenced by the presence of explanatory variables, even though these are not effectively linked to the analysed phenomenon, it has been considered the possibility to introduce a variation of the index, the **adjusted $\boldsymbol{R^2}$**:

$$\overline{R}^2 = 1 - \frac{(n-1)}{(n-p-1)}\frac{SSR}{SST} = 1 - \frac{(n-1)}{(n-p-1)}(1-R^2)$$

- The adjusted R^2 has the following **characteristics**:
  - The adjusted R^2 is lesser or equal to 1, but can be also lesser than 0;
  - The **correction factor** $-\frac{(n-1)}{(n-p-1)}$ "reduces" the R^2 value of an element that depends on the **number of explanatory variables** that have been used. The most are the explanatory variables, the most the adjusted R^2 will be lesser than the R^2.

- With the formula we presented, it is possible to obtain the curve coefficients that is better adapted to the noticed data, with any dispersion points shape.
- Anyway, the **curve coefficients calculation** is not enough for the statistician. It would be possible to have:
  - a **significative relation** (or "real") between the independent variables and the explanatory one, if the point dispersion around the curve is small;
  - a **not significative relation** (or random), when the dispersion points around the curve is almost equal to the one around the mean.
- The **hypothesis test** allows to verify in which of the two conditions we are working. It is possible thanks to the assumptions that have been made on the errors distribution.

The **hypothesis test** on the linear regression model can be:

- ⋆ Hypothesis test **on the coefficient linked to a single explanatory variable**. With this test it is possible to check the hypothesis that the coefficient of the variable taken into consideration can be considered statistically equal to a specific value. For this kind of hypothesis test, the error normality allows the use of the **Student's t** distribution.
- ⋆ Nullity hypothesis test of **more than one explanatory variable coefficient** (possibly on all of them). The error normality allows the use of the **F of Fisher-Snedecor** distribution.
- ⋆ Testing **on the assumptions of the gaussian linear model**. There are tests to check: normality of errors, the independence of errors, the constant variability of the errors (homoscedasticity).

- The aim of the **hypothesis test on a coefficient** is to check the **hyphotesis** that the "true" value of a parameter β_j is **equal to an hypothesized value** β_{H_0}, against the hypothesis that this **different** from β_{H_0}.
- Setting up $\beta_{H_0} = 0$, it is verified if the estimated effect on the sample for the analysed parameter is "real" and "concrete" for the whole population (I accept $\beta_{H_0} \neq 0$), or it is due to random fluctuations in the analysed sample (I accept $\beta_{H_0} = 0$). This is the **statistical significativity test** of a coefficient.
- If we want to formulate an hypothesis test in a formal way, the hypothesis (called **null hypothesis**) is

$$H_0 : \beta_j = \beta_{H_0}$$

against the **alternative hypothesis**

$$H_A : \beta_j \neq \beta_{H_0}$$

The statistical hypothesis testing about a parameter is based on the following statistical basis, that can be demonstrated:

- If H_0 is true, and σ_ε is known, then the **standard coefficient** $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ follows a standard normal distribution:

$$\frac{\hat{\beta}_j - \beta_{H_0}}{\sigma_\varepsilon \sqrt{g_j(X)}} \sim \; N(0,\, 1)$$

- As the **standard deviation** of the regression error **is not known** and has to be estimated from the data, the used index is the following one:

$$t = \frac{\hat{\beta}_j - \beta_{H_0}}{\hat{\sigma}_\varepsilon \sqrt{g_j(X)}}$$

- With H_0, the index t follows a **Student's t** distribution with $(n - (p+1))$ degrees of freedom, where $p+1$ is equal to the number of estimated parameters.

- The H_0 "null" hypothesis is accepted if $|t|$ is "small", and therefore, if the difference $|\hat{\beta}_j - \beta_{H_0}|$ is "small" compared to the standard error of $\hat{\beta}$.

- Affirming that a regression coefficients is different from the β_{H_0} hypothesize value means that $\hat{\beta}_j$ has an "high" distance from β_{H_0} compared to its standard error, measured by the element in the denominator $\hat{\sigma}_\varepsilon\sqrt{g_j(X)}$.

- ⋆ When significativity of the coefficient β_j is verified, it is check if the jth explanatory variable does not influence the dependent variable, and then if the β_j value can be null ($\beta_{H_0} = 0$).

- ⋆ The significativity of the coefficient β_j is confirmed if:

$$|t| = \frac{|\ \hat{\beta}_j\ |}{\hat{\sigma}_\varepsilon \sqrt{g_j(X)}} > t_{n-p-1,\,\frac{\alpha}{2}}$$

or rather if the absolute value of the t statistics is greater than the tabulated value of the t of Student's distribution with $n - p - 1$ degrees of freedom for the significativity level that has been required (α).

- As the distributive family $\hat{\beta}_j$ is the Student's t, some confidance intervals for the parameter β_j can be easily obtained if it is based on the distribution characteristics themselves.

- The confidence interval with interval $1 - \alpha$ for the parameter β_j is expressed by:

$$\hat{\beta}_j - t_{n-p-1;\,\frac{\alpha}{2}} \cdot \widehat{if(\hat{\beta}_j)} \leqslant \beta_j \leqslant \hat{\beta}_j + t_{n-p-1;\,\frac{\alpha}{2}} \cdot \widehat{if(\hat{\beta}_j)}$$

where $\widehat{se(\hat{\beta}_j)} = \hat{\sigma}_\varepsilon \sqrt{g_j(X)}$

- Another possible hypothesis test on the gaussian linear model is the **test on the simultaneous nullity of more coefficients** linked to specific variables.
- This statement can be expressed, without loosing the generality, like $H_0 : \beta_{p-J} = \cdots = \beta_p = 0$, where J is the number of null parameters. The alternative hypothesis will state that at least a β_j $(p - J \leqslant j \leqslant p)$ is different from.
- The **uses** of this kind of hypothesis test can be:
  - to verify if an alternative regression model, with a smaller number of explanatory variables, has the same predictive ability than the most complex model, in favor of the interpretative parsimony;
  - to verify if **the model that has been realised explains the phenomenon better than the sample mean** of the dependent variable, or rather to verify if the developed model is better than regression model where all the coefficients linked to the explanatory variables are null.

- S_0 is the **residuals deviance** of regression of the estimated model with the hypothesize **null coefficients** of the J chosen variabiles.

- S_1 is the residuals deviance of regression of **complete estimated model**, then, with the hypothesis H_0 that all the hypothesized coefficients are null.

- Then, with the H_0 null hypothesis, the ratio $(S_0 - S_1)/\sigma_\varepsilon^2$ follows a χ^2_J distribution.

- It is possible to demostrate that, with the H_0 hypothesis, the ratio

$$F = \frac{S_0 - S_1}{J} / \frac{S_1}{n-p-1} \sim F_{J,(n-p-1)}$$

- The nullity hypothesis of the J coefficients is “translated” in statistical words with the hypothesis that **the ratio F is “small”**.

- This ratio, with H_0, is distributed like a F of Fisher-Snedecor with parameters $J, (n-p-1)$. The null hypothesis will be refused if the F ratio exceedes the critical value, or rather if:

$$F > F_{J,\, n-p-1;\, 1-\alpha}$$

- ⋆ If the hypothesis test regards the coefficient significativity linked to **all** the explanatory variables that gave been used, the J parameter has to be replaced with p.

- ⋆ In conclusion, if the hypothesis test regards a **single parameter**, it is demonstrated that the F test is equal to the **t test**.

- ⋆ The **F test** is used for **automatic selection method for the explanatory variables** (**stepwise** regression algorithm). These algorithms add and remove variables comparing alternative models (with diffent variables included) according to the F test values.

⋆ The last step, fundamental, of the regression analysis is **testing the assumption of the gaussian linear model**. This test is done studying the residuals with graphic tools and eventually with statistical tests.

⋆ The **residuals analysis** allows to verify the main assumptions for the applicability of the regression techniques:

- **gaussianity** of the residuals;
- **independence** of the residuals;
- **homoscedasticity** (uniform variance) of the residuals.

⋆ Beyond these assumpotions, the residuals analysis allows to evaluate, within some limits, if the chosen model is adequate and if other explanatory variables are needed.

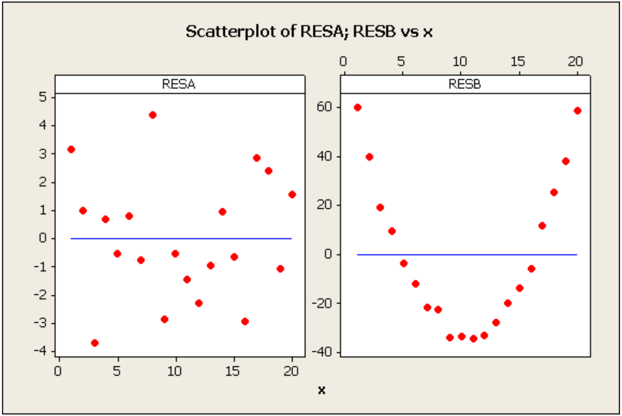
- As we have already shown, the **estimated residuals** $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}}$ are an estimation of the gaussian random errors (ε_i), and are defined like:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - f(\underline{\boldsymbol{x_i}}; \underline{\hat{\boldsymbol{\beta}}})$$

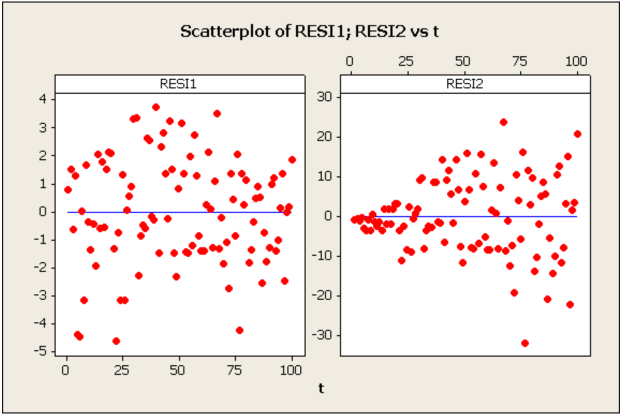
- For the residuals that come from **the least squares estimation**, it is always valid:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$$

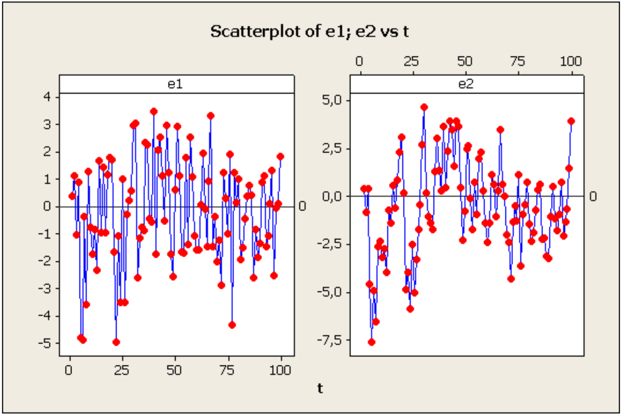
- Beyond the residuals that have just been mentioned, there are different versions of them (*deleted* residuals, Pearson residuals, *studentized* residuals, Cook residuals , etc. . .), with different finalities that will be partially explained later.
- In general, in order to evaluate if the applicability hypothesis of the model are adequate, the **instruments** are mainly **graphics**. Here there are some (not exhaustive) examples.



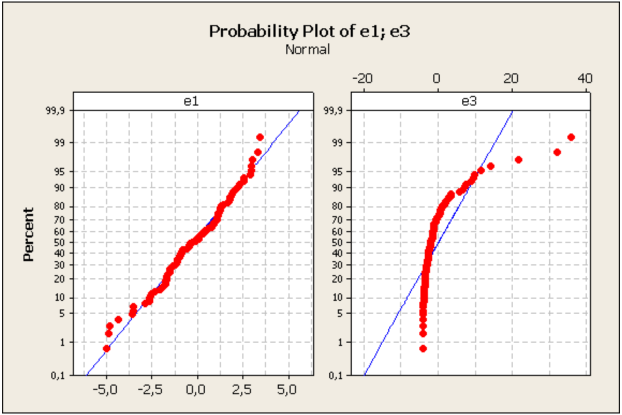
The graphics represents two case of linear regression residuals with an explanatory variable. The panel on the left shows an ideal behavior: the residuals are placed in a random way around the null value. The other plot shows a parabolic trend: a quadratic element for the x has to be added.



The plot on the left panel shows residuals that have a constant variability compared to the appearance order. The residuals of the right panel show increasing variability compared to the appearance order.



The plot of the left panel shows not correlated residuals. The plot of the right panel shows residuals with autocorrelation compared to the appearance order.



The plot of the left panel shows residuals that follow a normal distribution.

The residuals in the right panel cannot be considered normally distributed.

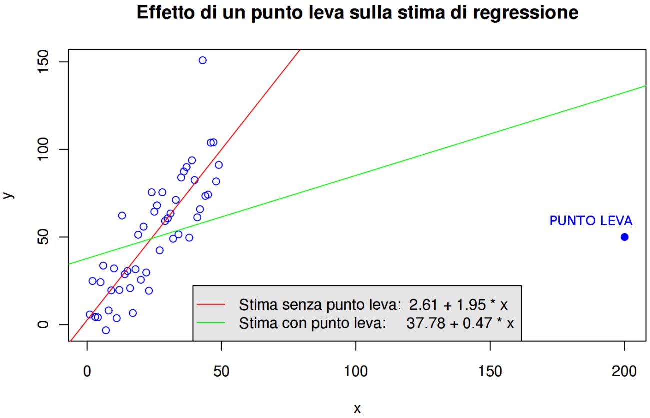
- ⋆ The **residuals plot towards the independent variables** (or prediction value) does not have to show an identifiable functional trend .
  - If the residuals are vaguely placed around the zero, it means that the side of the variability that has not been catched by the linear regression is not identifiable if we specify a different functional relation between the y and the explanatory variable.
  - If the residuals show a noticeable functional relation, it is possible that the linear model does not adequately catch the functional relation among the data.
- ⋆ The residuals must have constant variability (**homoscedasticity**). Or rather, the residual scatterplot compared to the values of the independent variables (or the prediction values) does not have to show an evident increase or decrease in the variability. If the homoscedasticity does not verify it can be because of:
  - the missing inclusion of an important explanatory variable;
  - the wrong specification of the functional relation among the data.

- ⋆ The residuals have to be **independent** among them, or rather they do not have to be influenced by the previous values. The presence of autocorrelation in the residuals means that there is dependence among the residuals. The **autocorrelation** of the residuals can be due to:
  - a wrong specification of the functional link among the variables;
  - not inclusion of a explanatory variable linked to the time in the model.

- ⋆ The residuals have to be distributed according to a **normal** law, with parameters $N(0, \sigma_\varepsilon)$. If this assumption is verified, then the hypothesis tests on the significativity of the model parameters are verified too.

- The residuals normality test can be done with **normal probability plot** and with **adaptation test** (e.g. **Anderson-Darling**).
- Of the residuals are not normally distributed, it could be appropriate to:
  - specify the model, changing the functional shape and/or the explanatory variables;
  - use more complex regression models (ex. GLM).

- ⋆ Other diagnostic instruments to verify the model are the "leverage points" or "influent points".
- ⋆ Please observe the following plot.



- In the previous plot 50 points have been drawn. One of them (highlighted with the “full” point) is an “outlier”, in the sense that it is really far from the others.

- The two regression lines have been estimated excluding (the red line) and including (the green line) the “outlier” point.

- The outlier point clearly “pushes down” the estimated regression line, modifying the parameters estimations.

- When a point (usually outlier) influences on the parameters estimation, then it is called **leverage point**.

- The leverage points come from the fact that the least squares techniques look for the curve that minimizes the **square** difference between the observed values and the estimated values of the curve itself.

- Not all of the outlier point are leverage points too.

- In order to identify the influent points, beyond the plots, two kinds of indicators can be used:
  - the **leverage** values;
  - the **Cook's distances**.

- The formula for both the indexes are quite complex.

- ⋆ The leverage value for the ith ($i = 1, \ldots, n$) observation comes from the ith value of the diagonal of the (*hat matrix*): $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

- ⋆ The Cook's distance for the ith observation is proportional to

$$r_i = \sum_{k=1}^{n} (\hat{y}_k - \hat{y}_{k,(i)})^2$$

where $\hat{y}_k$ is the least squares estimation of the kth observation of the y. $\hat{y}_{k,(i)}$ represents the least squares estimation of the kth observation of the y obtained by removing the ith observation.

- ⋆ For both the indicators, a great value represents the potential possibility that the point taken into consideration is a leverage point. This is not sure for the leverage points.
- ⋆ In the case of the example, the leverage point obtains a *hat* (leverage) value equal to 0.7588, while all the other points obtain a value not greater than 0.04.
- ⋆ The Cook's distance value, for the leverage point is equal to 48.83, while all the other points obtain a value not greater than 0.15.
- ⋆ In order to define a threshold in which it is possible to establish if a point is or not a leverage point, with the Cook's distance it is possible to use the distribuition $F_{2,n-2}$ as reference point.
- ⋆ Alternatively, for the Cook's distance is usually considered "great" a value which is greater than 1.

- Towards the **prediction**, the developed and tested model can be used to predict **values of the dependent variable** (the y) for values of the **x** even though they have not been found in the experimental phase.

- The following statements are valid:
  - under the statistical point of view, every prediction or estimation of the y is valid only within the **experimental variation field** of the independent variable (the x);
  - in the linear regression, the regression model (or, to better say, the function) does not often correspond to the real mathematical relation that exist between the independent variables and the dependent variable.

- Most of the times, the chosen **model** represents **an approximation of the existing relation**. Furthermore, it is known that with a quite complex function, it is always possible to describe the phenomenon.

- **Extrapolating** the data out of the real observational field is an error of statistics technique. It can be accepted only in the specific context, if it is justifiable by a better knowledge of the phenomenon.

- ⋆ When a regression is estimated, given the experimental observations, the puntual estimations of the parameters represent the “best value”.
- ⋆ Anyway, as we saw before, the parameters estimations can change within a certain range, which usually depends on the size of the standard error of estimation (take a look to the slides about the confidence intervals of the parameters).
- ⋆ The estimation of the regression curve, can change within a certain (confidence) band.
- ⋆ When the α value (usually equal to 0.05) has been chosen, it is possible to estimate the confidence bands for the regression curve.
- ⋆ In those bands, we expected that the real regression curve “is present” with a confidence level equal to $(1 - \alpha)$.

- With the"synthetic" mathematical form (matrix), it is possible to demonstrate that the limits of the confidence intervals with level $(1-\alpha)$ for the regression curve, for a vector of explanatory (independent) variables values $\underline{x}_*$, are:

  LCL: $f(\underline{x}_*;\underline{\hat{\beta}}) - t_{n-p-1;\frac{\alpha}{2}} \cdot \sqrt{\underline{x}_*^T(X^TX)^{-1}\underline{x}_*} \cdot \hat{\sigma_\varepsilon}$

  UCL: $f(\underline{x}_*;\underline{\hat{\beta}}) + t_{n-p-1;\frac{\alpha}{2}} \cdot \sqrt{\underline{x}_*^T(X^TX)^{-1}\underline{x}_*} \cdot \hat{\sigma_\varepsilon}$

  where $\underline{\hat{\beta}}$ is the vector of the least squares parameters estimation and the operator ^{T} represents the matrix transposition.

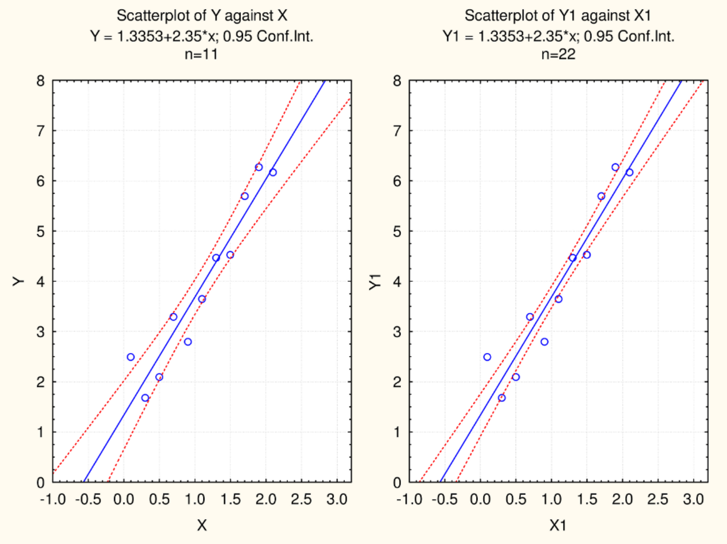
- The size of this interval is, then, equal to

  $2 \cdot t_{n-p-1;\frac{\alpha}{2}} \cdot \sqrt{\underline{x}_*^T(X^TX)^{-1}\underline{x}_*} \cdot \hat{\sigma_\varepsilon}$

- This size **increases** in a not linear way when the distance of the vector $\underline{x}_*$ from the vector of the mean values of the columns of the matrix X.

- The size, **decreases** in a not linear way when the numerosity of the sample, from which the estimations of $\underline{\beta}$ are obtained, increases.

- The following plot shows two regression lines with confidence band at 95%, with different sample numerosity. It has to be noticed how the bands “broaden” when they go away from the mean value of the independent variable, and how they “narrow” when the sample numerosity increases.


Scatterplot of Y against X
Y = 1.3353+2.35*x; 0.95 Conf.Int.
n=11
Y
X
0
1
2
3
4
5
6
7
8
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Scatterplot of Y1 against X1
Y1 = 1.3353+2.35*x; 0.95 Conf.Int.
n=22
Y1
X1

- The confidence bands can be interpreted like the bands we expected that the "real" regression line, with a certain confidence degree, stands.

- If we want to know (after having fixed the values of the independent variables) where it is expected that the potential future value of the y, we use the **prediction intervals (or bands)**.

- As for the confidence bands, in a "synthetic" (matrix) mathematical form, it is possible to demonstrate that the limits of the prediction interval with level $(1-\alpha)$ for the regression curve, for a vector of explanatory (independent) variables $\underline{x}_*$ values, are:

  LPL: $f(\underline{x}_*;\underline{\hat{\beta}}) - t_{n-p-1;\,\frac{\alpha}{2}} \cdot \left[\sqrt{\underline{x}_*^T (X^T X)^{-1} \underline{x}_* + 1}\right] \cdot \hat{\sigma_\varepsilon}$

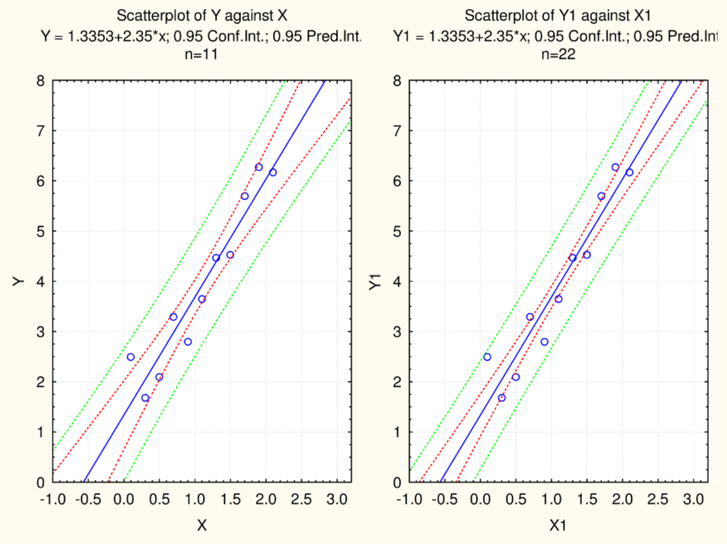
  UPL: $f(\underline{x}_*;\underline{\hat{\beta}}) + t_{n-p-1;\,\frac{\alpha}{2}} \cdot \left[\sqrt{\underline{x}_*^T (X^T X)^{-1} \underline{x}_* + 1}\right] \cdot \hat{\sigma_\varepsilon}$

  where $\underline{\hat{\beta}}$ is the vector of the least squares parameters estimation and the operator ^{T} represents the matrix transposition.

- The size of this interval, then, is equal to
  $2 \cdot t_{n-p-1;\,\frac{\alpha}{2}} \cdot \left[\sqrt{\underline{x}_*^T (X^T X)^{-1} \underline{x}_* + 1}\right] \cdot \hat{\sigma_\varepsilon}$

- It has to be noticed the strong "similarity" with the confidence intervals.

- As the quantity $\underline{x}_*^T(X^TX)^{-1}\underline{x}_*$ is often lesser than 1, the size of the prediction interval is quite always “overlooked” by the element “$+1$” within square brackets.
- The prediction bands, even though they have a “similar” behavior to those of the confidence bands, tend to be “wider” than the first ones, and to assume a less marked curvature than the one we highlighted before.
- The following plot shows the same regression lines of the previous graphic, with confidence band by 95% (in red) and prediction bands by 95% (in green).
- It has to be noticed how the prediction bands seems to be “parallel” to the regression line. Furthermore, the prediction bands are wider than the confidence ones.


Scatterplot of Y against X
Y = 1.3353+2.35*x; 0.95 Conf.Int.; 0.95 Pred.Int.
n=11
Y
X
0
1
2
3
4
5
6
7
8
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Scatterplot of Y1 against X1
Y1 = 1.3353+2.35*x; 0.95 Conf.Int.; 0.95 Pred.Int
n=22
Y1
X1

- We expect that the confidence bands tend to “meet” as the sample numorisity increases. On the contrary, this cannot happen for the prediction bands.

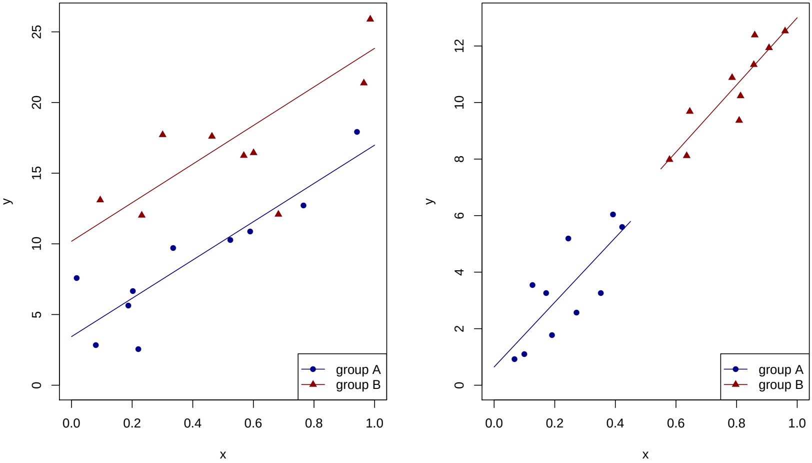
- This happens because the prediction bands express the variability of the single observations (y_i).

- If the specified regression model is correct, we will then expect that, about a fraction equal to α of the future observations “goes out” from the prediction bands.

- This cannot be said for the confidence bands.

- The analysis of the covariance, or **ANCOVA**, is a mix of the ANOVA and of the regression for quantitative variables.
- In the ANCOVA a phenomenon is explained:
  - by qualitative variables, the factors, as in the ANOVA;
  - by quantitative variables called **covariate**.
- In the easiest case, the ANCOVA is used when it is identified another qualitative variable linked to the dependent variable, beyond the qualitative dependent variable.
- If the link between the dependent variable and the covariate is strong, it is possible to extract the quota of variance due to the covariate from the MS_E.

The following plot shows two possible settings that can happen after using ANCOVA: **to emphasize the differences among the groups** (on the left) or **to exclude possible differences among groups** (on the right).



Example.

We want to determine if two hypnotic medicine, A and B, have the same effectiveness in making the patients being hypnotized.

With this aim, we select 20 subjects:

- 10 subjects treated with the medicine A;
- 10 subjects treated with the medicine B.

The hypnothic suggestion is measured with a quantitative scale from 1 to 50 (the higher is the score, the higher is the suggestion).

The effectiveness of hypnotic suggestion is subjective because it depends on the susceptibility of the patient. Before the administration of the medicine, the subjects have to do a susceptibility test. The higher is the score of the test, the higher is the susceptibility of the hypnosis.

If we analyse the data with an ANOVA whose model is

$$x_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

we will obtain the following results:

	df	SS	MS	F	$P(F)$
$Medicine$	1	6.05	6.05	0.1644	0.69
$residuals$	18	662.50	36.81		
$Total$	19	668.55			

From the results of the ANOVA, there is no difference between the patients that have been treated with different medicines.

In this model the susceptibility of the hypnotic suggestion, different in every subject, was not taken into consideration.

The analysis of the covariance is able to use the link between the hypnotic suggestion of the medicine and the susceptibility of the patient.

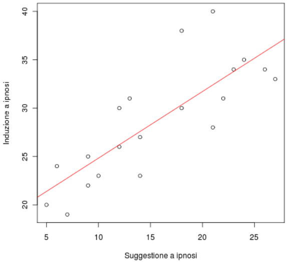
The mathematical model is the following one:

$$x_{ij} = \mu + \alpha_j + \beta \cdot c_{ij} + \varepsilon_{ij}$$

where c_{ij} is the value that has been measured by the covariate (the susceptibility to the hypnotic suggestion) for the ith subject of the jth level.

If we analyse the relation between the dependent variable and the covariate one, we will obtain the following formula:

$$y = 17.9 + 0.7 \cdot x$$



The model explains a part of the variance of the response variable (hypnosis induction). It is possible to identify, in the error variance, the element due to the covariate (hypnosis suggestion).

- **ANOVA**

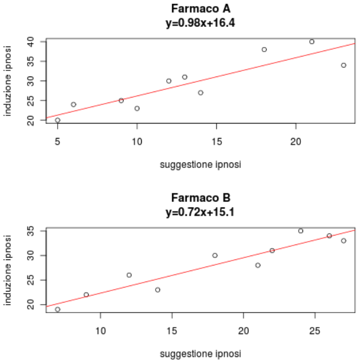
	df	SS	MS	F	$P(F)$
$Medicine$	1	6.05	6.05	0.1644	0.69
$residuals$	18	662.50	36.81		
$Total$	19	668.55			

- **ANCOVA**

	df	SS	MS	F	$P(F)$
$Medicine$	1	6.05	6.05	0.8403	0.37
$suggestion$	1	540.18	540.18	75.07	<0.0001
$residuals$	17	122.32	7.20		
$Total$	19	668.55			

- The analysis of covariance shows that the $SS_{suggestion}$ summed to the $SS_{residuals}$ is equal to $SS_{residuals}$ of the ANOVA.

For the classic ANCOVA, it is necessary to demonstrate that the slope of the covariate does not significantly change when the treatment levels change.
The following plot shows the regression lines for two kinds of medicine.



- We can use a t test to check the equality of the slopes of the model in order to test the homogeneity of the slopes.

- The null hypothesis, H_0, affirms that there is no difference between the coefficients of the various models. On the contrary, the alternative hypothesis, H_A, affirms that there is difference.

- The t value has to be calculated in order to verify or to refuse the null hypothesis:

$$t = \frac{\hat{\beta}_A - \hat{\beta}_B}{\sqrt{MSE_A + MSE_B}}$$

- The index t has to be compared to the cut-off value of the Student's t distribution: $t \sim t_{n_A+n_B-4}$.