

# ELEMENTARY STATISTICS COURSE

---

Alfredo Sánchez Alberca ([asalber@ceu.es](mailto:asalber@ceu.es))

Feb 2016

Department of Applied Math and Statistics  
CEU San Pablo



**CEU**  
*Universidad  
San Pablo*

# LICENSE TERMS

This work is licensed under an Attribution-NonCommercial-ShareAlike 4.0 International Creative Commons License. <http://creativecommons.org/licenses/by-nc-sa/4.0/>

You are free to:

- Share – copy and redistribute the material in any medium or format
- Adapt – remix, transform, and build upon the material

Under the following terms:



**Attribution.** You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.



**NonCommercial.** You may not use the material for commercial purposes.



**ShareAlike.** If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

## 1. Regresión y Correlación

# REGRESIÓN Y CORRELACIÓN

---

1. Regresión y Correlación
  - 1.1 Joint frequency distribution
  - 1.2 Covariance
  - 1.3 Regression
  - 1.4 Regression line

In the last chapter we saw how to describe the distribution of a single variable in a sample. However, in most cases, studies require to describe several variables that are often related. For instance, a nutritional study should consider all the variables that could be related to the weight, as height, age, gender, smoking, diet, physic exercise, etc.

To understand a phenomenon that involve several variables is not enough to study every variable by its own. We have to study all the variables together to describe how they interact and the type of relation among them.

Usually in a *dependency study* there is a **dependent variable**  $Y$  that it is supposed to be influenced by a set of variables  $X_1, \dots, X_n$  known as **independent variables**. The simpler case is a *simple dependency study* when there is only one independent variable, that is the case covered in this chapter.

## JOINT FREQUENCIES

To study the relation between two variables  $X$  and  $Y$ , we have to study the joint distribution of the **two-dimensional variable**  $(X, Y)$ , whose values are pairs  $(x_i, y_j)$  where the first element is a value of  $X$  and the second a value of  $Y$ .

### Definition (Joint sample frequencies)

Given a sample of  $n$  values and a two-dimensional variable  $(X, Y)$ , for every value of the variable  $(x_i, y_j)$  is defined

- **Absolute frequency**  $n_{ij}$ : Is the number of times that the pair  $(x_i, y_j)$  appears in the sample.
- **Relative frequency**  $f_{ij}$ : Is the proportion of times that the pair  $(x_i, y_j)$  appears in the sample.

$$f_{ij} = \frac{n_{ij}}{n}$$

*Watch out! For two-dimensional variables it make no sense cumulative frequencies.*

## JOINT FREQUENCY DISTRIBUTION

The values of the two-dimensional variable with their frequencies is known as **joint frequency distribution**, and is represented in a joint frequency table.

$X \backslash Y$	$y_1$	$\cdots$	$y_j$	$\cdots$	$y_q$
$x_1$	$n_{11}$	$\cdots$	$n_{1j}$	$\cdots$	$n_{1q}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_i$	$n_{i1}$	$\cdots$	$n_{ij}$	$\cdots$	$n_{iq}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_p$	$n_{p1}$	$\cdots$	$n_{pj}$	$\cdots$	$n_{pq}$



## JOINT FREQUENCY DISTRIBUTION

The height (in cm) and weight (in kg) of a sample of 30 students is:

(179,85), (173,65), (181,71), (170,65), (158,51), (174,66),  
(172,62), (166,60), (194,90), (185,75), (162,55), (187,78),  
(198,109), (177,61), (178,70), (165,58), (154,50), (183,93),  
(166,51), (171,65), (175,70), (182,60), (167,59), (169,62),  
(172,70), (186,71), (172,54), (176,68), (168,67), (187,80).

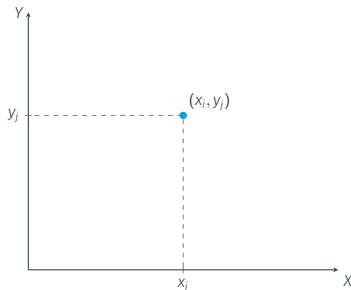
The joint frequency table is

X/Y	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100)	[100, 110)
(150, 160]	2	0	0	0	0	0
(160, 170]	4	4	0	0	0	0
(170, 180]	1	6	3	1	0	0
(180, 190]	0	1	4	1	1	0
(190, 200]	0	0	0	0	1	1

# SCATTER PLOT

The joint frequency distribution can be represented graphically with a **scatter plot**, where data is displayed as a collections of points on a  $XY$  coordinate system.

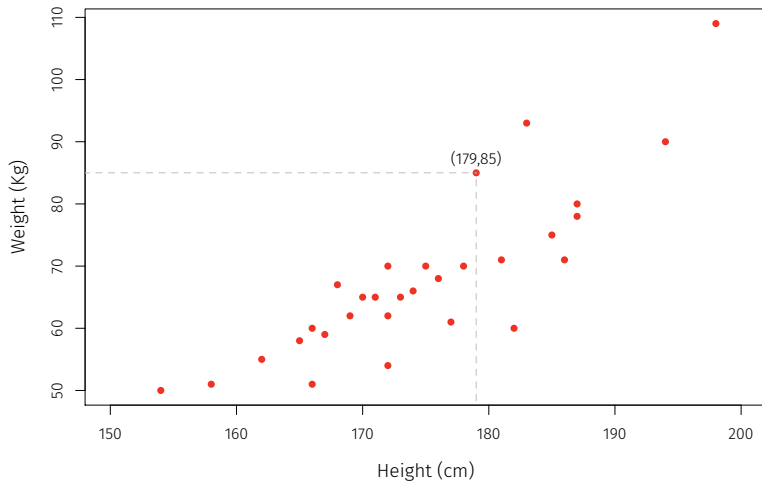
Usually the independent variable is represented in the  $X$  axis and the dependent variable in the  $Y$  axis. For every data pair  $(x_i, y_j)$  in the sample a dot is drawn on the plane with those coordinates.



The result is a set of points that usually is known as a *point cloud*.

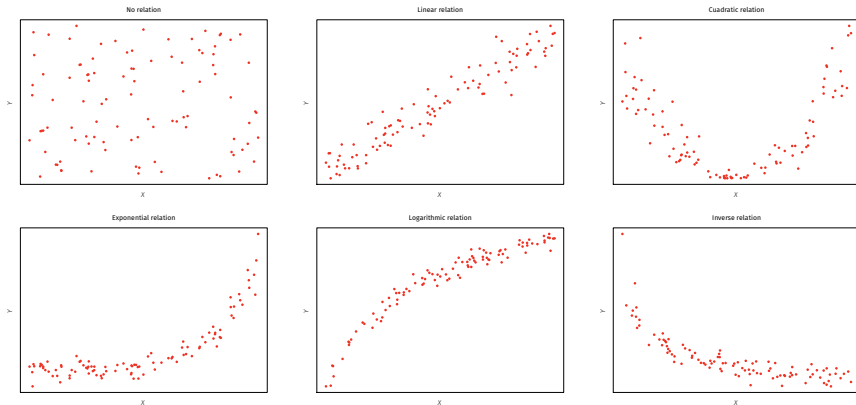
# SCATTER PLOT

Height and weight scatter plot



# SCATTER PLOT INTERPRETATION





The shape of the point cloud in a scatter plot gives information about the type of relation between the variables.



## MARGINAL FREQUENCY DISTRIBUTIONS

The frequency distributions of each variable of the two-dimensional variable are known as **marginal frequency distributions**.

We can get the marginal frequency distributions from the joint frequency table summing frequencies by rows and columns.

$X \backslash Y$	$y_1$	$\cdots$	$y_j$	$\cdots$	$y_q$	$n_x$
$x_1$	$n_{11}$	$\cdots$	$n_{1j}$	$\cdots$	$n_{1q}$	$n_{x_1}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
$x_i$	$n_{i1}$		$n_{ij}$		$n_{iq}$	$n_{x_i}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
$x_p$	$n_{p1}$	$\cdots$	$n_{pj}$	$\cdots$	$n_{pq}$	$n_{x_p}$
$n_y$	$n_{y_1}$	$\cdots$	$n_{y_j}$	$\cdots$	$n_{y_q}$	$n$

## MARGINAL FREQUENCY DISTRIBUTIONS

The marginal frequency distributions for the previous sample of heights and weights are

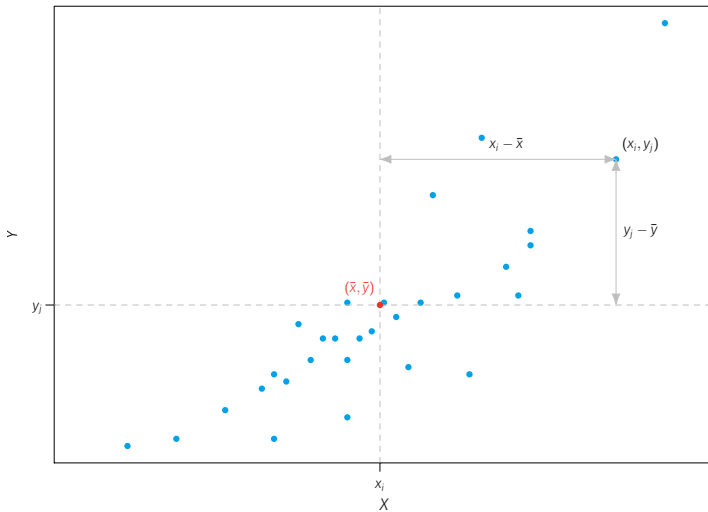
$X/Y$	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100)	[100, 110)	$n_x$
(150, 160]	2	0	0	0	0	0	2
(160, 170]	4	4	0	0	0	0	8
(170, 180]	1	6	3	1	0	0	11
(180, 190]	0	1	4	1	1	0	7
(190, 200]	0	0	0	0	1	1	2
$n_y$	7	11	7	2	2	1	30

and the corresponding statistics are

$$\begin{array}{lll}\bar{x} = 174.67 \text{ cm} & s_x^2 = 102.06 \text{ cm}^2 & s_x = 10.1 \text{ cm} \\ \bar{y} = 69.67 \text{ Kg} & s_y^2 = 164.42 \text{ Kg}^2 & s_y = 12.82 \text{ Kg}\end{array}$$

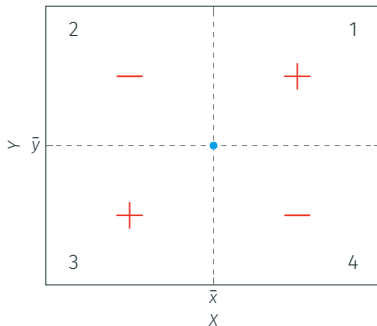
## DEVIATIONS FROM THE MEANS

To study the relation between two variables, we have to analyze the joint variation of them.



## SIGN OF DEVIATIONS FROM THE MEAN

Dividing the point cloud of the scatter plot in 4 quadrants centered in the mean point  $(\bar{x}, \bar{y})$ , the sign of deviations from the mean is:



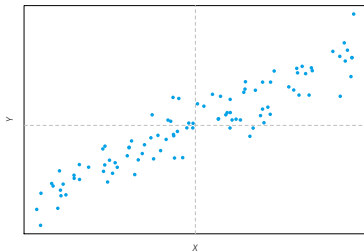
Quadrant	$(x_i - \bar{x})$	$(y_j - \bar{y})$	$(x_i - \bar{x})(y_j - \bar{y})$
1	+	+	+
2	-	+	-
3	-	-	+
4	+	-	-



## SIGN OF THE PRODUCT OF DEVIATIONS FROM THE MEAN

If there is an *increasing linear* relationship between the variables, most of the points will fall in quadrants 1 and 3, and the sum of the products of deviations from the mean will be positive.

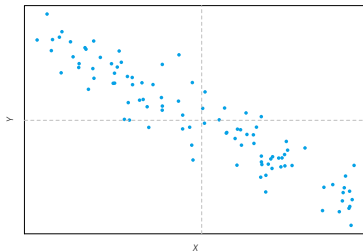
Increasing linear relation



$$\sum (x_i - \bar{x})(y_i - \bar{y}) = +$$

If there is an *decreasing linear* relationship between the variables, most of the points will fall in quadrants 2 and 4, and the sum of the products of deviations from the mean will be negative.

Decreasing linear relation



$$\sum (x_i - \bar{x})(y_i - \bar{y}) = -$$

Using the products of deviations from the mean we get the statistic

### Definition (Sample covariance)

The *sample covariance* of a two-dimensional variable  $(X, Y)$  is the average of the products of deviations from the respective means.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_j - \bar{y})n_{ij}}{n}$$

It can also be calculated using the formula

$$s_{xy} = \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y}.$$

The covariance measures the linear relation between two variables:

- If  $s_{xy} > 0$  there exists an increasing linear relation.
- If  $s_{xy} < 0$  there exists a decreasing linear relation.
- If  $s_{xy} = 0$  there is no linear relation.

## COVARIANCE CALCULATION

Using the joint frequency table of the sample of heights and weights

X/Y	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100)	[100, 110)	$n_x$
(150, 160]	2	0	0	0	0	0	2
(160, 170]	4	4	0	0	0	0	8
(170, 180]	1	6	3	1	0	0	11
(180, 190]	0	1	4	1	1	0	7
(190, 200]	0	0	0	0	1	1	2
$n_y$	7	11	7	2	2	1	30

$$\bar{x} = 174.67 \text{ cm} \quad \bar{y} = 69.67 \text{ Kg}$$

the covariance is

$$\begin{aligned} s_{xy} &= \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y} = \frac{155 \cdot 55 \cdot 2 + 165 \cdot 55 \cdot 4 + \dots + 195 \cdot 105 \cdot 1}{30} - 174.67 \cdot 69.67 \\ &= \frac{368200}{30} - 12169.26 = 104.07 \text{ cm} \cdot \text{Kg}, \end{aligned}$$

This means that there is a increasing linear relation between the weight and the height.

In most cases the goal of a dependency study is not only to detect a relation between two variables, but also to express that relation with a mathematical function,

$$y = f(x)$$

in order to predict the dependent variable for every value of the independent one.

The part of Statistics in charge of constructing such a function is **regression**, and the function is known as **regression function** or **regression model**.

## SIMPLE REGRESSION MODELS

Depending on the type of function there are a lot of types of regression models. The most common are shown in the table below.

Model	Equation
Linear	$y = a + bx$
Quadratic	$y = a + bx + cx^2$
Cubic	$y = a + bx + cx^2 + dx^3$
Potential	$y = a \cdot x^b$
Exponential	$y = a \cdot e^{bx}$
Logarithmic	$y = a + b \log x$
Inverse	$y = a + \frac{b}{x}$
Sigmoidal	$y = e^{a + \frac{b}{x}}$

The model choice depends on the shape of the points cloud in the scatterplot.

Once chosen the type of regression model, we have to determine which function of that family explains better the relation between the dependent and the independent variables, that is, the function that predict better the dependent variable.

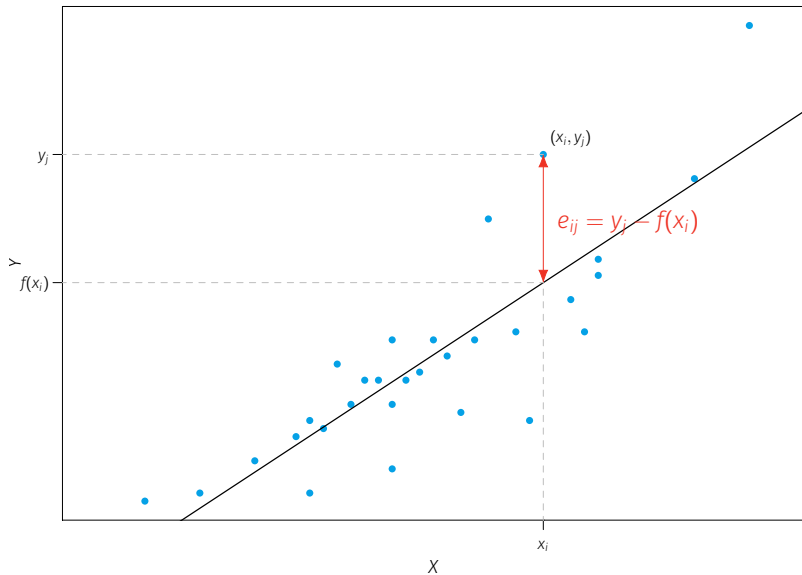
That function is the function that minimize the distances from the observed values for  $Y$  in the sample to the predicted values of the regression function. These distances are known as *residuals* or *predictive errors*.

### Definition (Residuals or predictive errors)

Given a regression model  $y = f(x)$  for a two-dimensional variable  $(X, Y)$ , the *residual* or *predictive error* for every pair  $(x_i, y_j)$  of the sample is the difference between the observed value of the dependent variable  $y_j$  and the predicted value of the regression function for  $x_i$ ,

$$e_{ij} = y_j - f(x_i).$$

## RESIDUALS OR PREDICTIVE ERRORS ON $y$



A way to get the regression function is the *least squares method*, that determine the function that minimize the squared residuals.

$$\sum e_{ij}^2.$$

For a linear model  $f(x) = a + bx$ , the sum depends on two parameters, the intercept  $a$ , and the slope  $b$  of the straight line,

$$\theta(a, b) = \sum e_{ij}^2 = \sum (y_j - f(x_i))^2 = \sum (y_j - a - bx_i)^2.$$

This reduces the problem to determine the values of  $a$  and  $b$  that minimize this sum.



## LEAST SQUARES FIT OF A LINEAR MODEL

To solve the minimization problem, we have to set to zero the partial derivatives with respect to  $a$  and  $b$ .

$$\frac{\partial \theta(a, b)}{\partial a} = \frac{\partial \sum (y_j - a - bx_i)^2}{\partial a} = 0$$
$$\frac{\partial \theta(a, b)}{\partial b} = \frac{\partial \sum (y_j - a - bx_i)^2}{\partial b} = 0$$

And solving the equation system, we get

$$a = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \quad b = \frac{s_{xy}}{s_x^2}$$

This values minimize the residuals on  $Y$  and give us the optimal linear model.

### Definition (Regression line)

Given a sample of a two-dimensional variable  $(X, Y)$ , the *regression line* of  $Y$  on  $X$  is

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

The regression line of  $Y$  on  $X$  is the straight line that minimize the predictive errors on  $Y$ , therefore is the linear regression model that gives better predictions of  $Y$ .

## REGRESSION LINE CALCULATION

Using the previous sample of heights (X) and weights (Y) with the following statistics

$$\begin{aligned}\bar{x} &= 174.67 \text{ cm} & s_x^2 &= 102.06 \text{ cm}^2 & s_x &= 10.1 \text{ cm} \\ \bar{y} &= 69.67 \text{ Kg} & s_y^2 &= 164.42 \text{ Kg}^2 & s_y &= 12.82 \text{ Kg} \\ & & s_{xy} &= 104.07 \text{ cm} \cdot \text{Kg}\end{aligned}$$

the regression line of weight on height is

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}) = 69.67 + \frac{104.07}{102.06}(x - 174.67) = 1.02x - 108.49.$$

And the regression line of height on weight is

$$x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y}) = 174.67 + \frac{104.07}{164.42}(y - 69.67) = 0.63y + 130.78.$$

*Observe that the regression lines are different!*

# REGRESSION LINES

Regression lines of weight and height

