University of Mila
UFR of Sciences

2024-2025

Bachelor's degree in Science, Technology, and Health, all specializations - Semester 2
Probability and Statistics

Regression - Least Squares Line

The previous chapter dealt with **univariate descriptive statistics,** that is, the description of a statistical series based on a single characteristic (size, for example). We now want to study, visualize, and measure (if applicable) the links between two variables: this is the purpose of **bivariate descriptive statistics.**

We consider a population on which we study two quantitative variables (or characteristics) *X* and Y. We will therefore study **statistical series with two variables ;** in other words a pair of variables *X,Y.* We want to know if the two variables are linked by a functional link of the type *Y fX* (that is to say that we can predict the values of *Y* from the values of X), or *X fY* (that is to say that we can predict the values of *X* from the values of Y).

Let us now clarify that the existence of such a link between the two variables *X* and *Y* does not mean necessarily a cause and effect link between them.

Fundamental example : $Y = aX + b$ (affine functional connection).

Graphical representation : scatter plot.

On a sample of *n* individuals taken from the population, we observe *n* pairs $(x_1,y_1), (x_2,y_2),\dots,(x_n,y_n)$ of values of *X* and Y.

These observations can be represented in the plane. To each pair (x_i,y_i), $i = 1,\dots,n$ we make a point *Mi* correspond . We obtain a point cloud.

The shape of the cloud obtained can indicate the type of possible dependence between *X* and Y. If the points are "rather" aligned, we can consider a relationship of type*Y=aX+b* (equation of a straight line). If the cloud "forms" a parabola, we can consider a relationship of type $Y = aX^2 + bX + c$.

We say that we are trying to fit a curve to the point cloud.

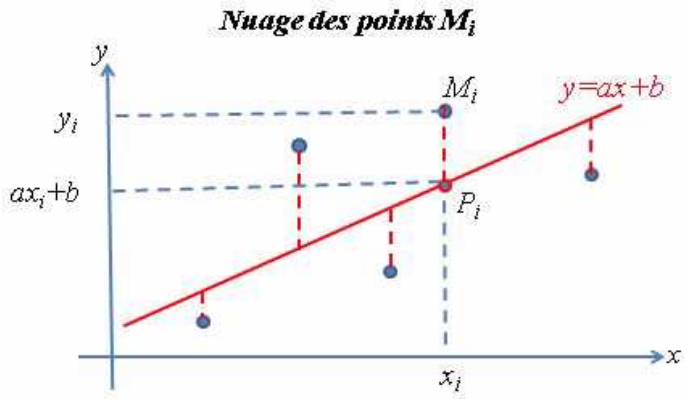
1. **Least squares (or regression)** line **from** *y* **to** *x* We are trying to fit a line with equation $y = ax + b$ to the cloud of points.

The adjustment criterion is the total distance between the points of the $M_i(x_i,y_i)$ cloud and the points $P_i(x_i, ax_i + b)$ corresponding to the adjustment line.

We are therefore looking for the pair $\left(\hat{a},\hat{b}\right)$ which minimizes $f(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2$.

It can be demonstrated (we will admit it here) that there exists a unique couple $\left(\hat{a},\hat{b}\right)$ making $f(a,b)$ minimum, and therefore only one straight line answering the problem.

It is the least squares line from *y* to *x ;* it is also called the regression line from *y* to x.



Equation of the least squares line from *y* **to** *x :*

$D_{y/x}$: $y = \hat{a}x + \hat{b}$, with $\hat{a} = \frac{cov(x,y)}{s_x^2}$ et $\hat{b} = \bar{y} - \hat{a}\bar{x}$.

Ratings:

Averages: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i.$ *Covariance:* $cov\,(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\bar{y}.$

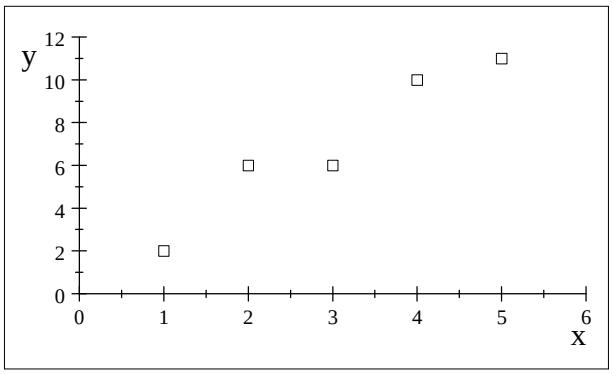
Variances: $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2, s_y^2 = \frac{1}{n}\sum_{i=1}^{n}y_i^2 - (\bar{y})^2.$

Example.

We consider the following double statistical series:

x_i	2	3	5	1	4
y_i	6	6	11	2	10

The corresponding scatter plot is shown in the graph below.



Point cloud

The regression line from *y* to *x* has the equation: $y = \hat{a}x + \hat{b}$, with $\hat{a} = \dfrac{cov(x,y)}{s_x^2}$ et $\hat{b} = \bar{y} - \hat{a}\bar{x}$.

x_i	y_i	$x_i y_i$	x_i^2
2	6	12	4
3	6	18	9
5	11	55	25
1	2	2	1
4	10	40	16
Total	**15**	**35**	**127**	**55**

We have $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \dfrac{1}{5} \times 15 = 3$, $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \dfrac{1}{5} \times 35 = 7$,

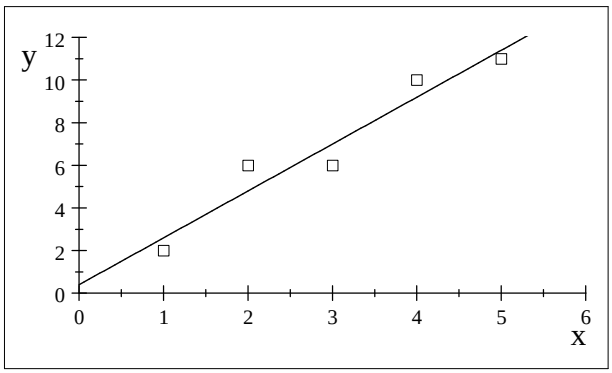
$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \frac{1}{5} \times 127 - 3 \times 7 = 4,4\,,$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{1}{5} \times 55 - (3)^2 = 2.$$

We deduce $\hat{a} = \dfrac{cov(x,y)}{s_x^2} = \dfrac{4,4}{2} = 2,2$

and $\hat{b} = \bar{y} - \hat{a}\bar{x} = 7 - 2,2 \times 3 = 0,4.$

The regression line from *y* to *x* therefore has the equation: $y = 2,2x + 0,4.$



Scatter plot and regression line from *y* to *x*

2. **Least squares line from *x* to** y.

We follow a procedure similar to that which gave the least squares line of *y* at *x :*

$$D_{y/x} : \; y = \hat{a}x + \hat{b}, \text{ with } \hat{a} = \frac{cov(x,y)}{s_x^2} \text{ and } \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

We are trying to fit a straight line Dx/y with equation *xayb* to the point cloud.
We obtain the least squares line from *x* to *y :*

$$D_{x/y} : \; x = \widehat{a'}y + \widehat{b'} \text{ ,with } \widehat{a'} = \frac{cov(x,y)}{s_y^2} \text{ and } \widehat{b'} = \bar{x} - \widehat{a'}\bar{y}.$$

Note. These equations can also be written:

$$D_{y/x} : \; y - \bar{y} = \hat{a}(x - \bar{x})$$

$$D_{x/y} : \; x - \bar{x} = \widehat{a'}(y - \bar{y})$$

The lines Dy/x and Dx/y therefore intersect at the point *Gx,* y.

Example.

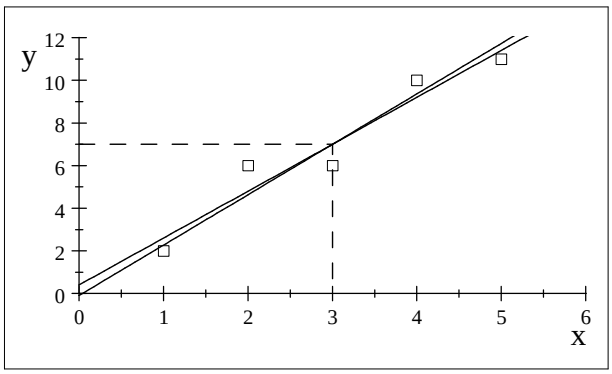
Let's take the previous example again. We still have $\bar{x} = 3, \bar{y} = 7, cov(x,y) = 4,4, s_x^2 = 2$ and $\hat{a} = 2,2.$

We calculate $s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2 = \frac{1}{5} \times 297 - (7)^2 = 10,4,$ d'où $\widehat{a'} = \frac{cov(x,y)}{s^2} = \frac{4,4}{10,4} = \frac{1,1}{2,6}.$

The regression line from *x* to *y* therefore has the equation $x - \bar{x} = \widehat{a'}(y - \bar{y}),$ or $x - 3 = \frac{1,1}{2,6}(y - 7),$ that is, $y = 2,3637x - 0,0909.$

We also find an equation of the regression line from *y* to *x : yy ax* x, or $y - 7 = 2,2(x - 3),$ that is $y = 2,2x + 0,4.$

The lines Dy/x and Dx/y intersect at the point $G(\bar{x},\bar{y}) = G(3,7).$



Regression lines from *y* to *x* and from *x* to *y*

3. **Linear correlation coefficient between *x* and *y***

The linear correlation coefficient is given by: $r_{x,y} = \frac{cov(x,y)}{s_x s_y}.$

Quality of fit. .

We can show that $r_{x,y}^2 = 1 - \frac{1}{s_y^2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{a}x_i - \hat{b}\right)^2 \leq 1$ We deduce that $r_{x,y}^2 = 1$ if and only if $\sum_{i=1}^{n}\left(y_i - \hat{a}x_i - \hat{b}\right)^2 = 0,$ that is, $y_i - \hat{a}x_i - \hat{b} = 0,$ for all $i = 1, \ldots, n,$ that is $M_i(x_i, y_i) \in D_{y/x}$ So, $r_{x,y}^2 = 1$ if and only if the points *Mi* are aligned on Dy/x.

Generally speaking, the more $r_{x,y}^2$ is close to 1, the better the fit of the least squares line to the scatter plot. In practice, we say that there is a good linear correlation between *X* and *Y* if $\frac{\sqrt{3}}{2} \leq |r_{x,y}| \leq 1,$ c'est-à-dire si $r_{x,y}^2 \geq \frac{3}{4}.$

The sign of rx,y (same sign as that of a) indicates the direction of the bond (increasing if rx,y>0, decreasing if $r_{x,y} < 0$) between *X* and Y.

Meaning of

The question arises whether a large value of $r_{x,y}$ (in absolute value) or of $r_{x,y}^2$ proves that there is a strong correlation between the two characters *X* and *Y* (e.g. when the fit is good) or if it is due random sampling (e.g. when *n* is small). To obtain an answer, one can use statistical tests (issue not addressed here).

Decomposition formula.

The notion of a link between *X* and *Y* means that a variation in *X* leads to a variation in Y. The formula for decomposition makes it possible to specify the part of variation of *Y* explained by the variation of *X :*

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left(\widehat{y_i}-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2, \text{ with } \widehat{y_i}=\hat{a}x_i+\hat{b}.$$

The demonstration rests on THE do that the double product cancels out :

$\sum_{i=1}^{n}\left(\widehat{y_i}-\bar{y}\right)\left(y_i-\widehat{y_i}\right)=\hat{a}\sum_{i=1}^{n}(x_i-\bar{x})e_i=0$ with $e_i=y_i-\widehat{y_i}$ (observed error), and thanks to the equations defining *a* and b.

The **total sum of squares :** $\sum_{i=1}^{n}(y_i-\bar{y})^2$ measures the overall variation of the *yi* around their mean y.

The **sum of squares explained** by the variable *X :* $\sum_{i=1}^{n}\left(\widehat{y_i}-\bar{y}\right)^2=\hat{a}^2\sum_{i=1}^{n}(x_i-\bar{x})^2$ measures the variation of *Y* explained by the variable X. This term is only a function of the slope of the least significant line squares and X values.

The **residual sum of squares :** $\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2=\sum_{i=1}^{n}e_i^2$ measures the variation of *Y* not explained by the variable X.

Coefficient of determination.

It is natural to measure the strength of the connection between variables *X* and *Y* using the coefficient of determination :

$$R^2=\frac{\sum_{i=1}^{n}\left(\widehat{y_i}-\bar{y}\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\text{sum of squares explained}}{\text{total sum of squares}}$$

We can verify that $R^2=r_{x,y}^2$. Which explains that $r_{x,y}$ measures the strength of the bond between *X* and Y.

Relative position of Dy/x **and** Dx/y.

If $r_{x,y}^2=0$, so $cov(x,y)=0$, and $\hat{a}=\widehat{a'}=0$. with, $D_{y/x}: y=\bar{y}$ et $D_{x/y}: x=\bar{x}$.

If $r_{x,y}^2\neq 0$, so $\hat{a}\neq 0$ et $\widehat{a'}\neq 0$. and we have $D_{y/x}: y-\bar{y}=\hat{a}(x-\bar{x})$ et $D_{x/y}: y-\bar{y}=\frac{1}{\widehat{a'}}(x-\bar{x})$.

On a : $\hat{a}\widehat{a'}=\frac{cov(x,y)}{s_x^2}\frac{cov(x,y)}{s_y^2}=\left(\frac{cov(x,y)}{s_x s_y}\right)^2=r_{x,y}^2$.

Si $r_{x,y}^2=1$, so $\hat{a}=\frac{1}{\widehat{a'}}$ and $D_{x/y}$ for equation $(y-\bar{y})=\frac{1}{\widehat{a'}}(x-\bar{x})=\hat{a}(x-\bar{x})$ $D_{y/x}$. with, $D_{x/y}=D_{y/x}$.

Si $0<r_{x,y}^2<1$, so $0<\hat{a}\widehat{a'}<1$ and two case possible :

-let $0<r_{x,y}<1$ and $\hat{a}>0, \widehat{a'}>0$ et $\hat{a}<\frac{1}{\widehat{a'}}$;

-let $-1<r_{x,y}<0$ and $\hat{a}<0, \widehat{a'}<0$ et $\hat{a}>\frac{1}{\widehat{a'}}$.

4. **Variable transformation**

When the linear correlation between *x* and *y* is poor, fitting a straight line with equation *y ax b* to the point cloud is not good. By observing the point cloud, we can then think of other types of relationship between *x* and *y ;* for example $y=\beta e^{\alpha x}, y=a\ln x+b$ By transformation of one of the variables *x* or y, or of the two variables, we can reduce ourselves to an affine relationship between the transformed variables and use the previous results.

On this subject, see the examples covered in class and exercises 2 to 4.

5. **Exercises**

Exercise 1.

In the following statistical series, *x* represents the number of days of sun exposure of a leaf and *y* the number of airy stomata per square millimeter:

x	2	4	8	10	24	40	52
y	6	11	15	20	39	62	85

1) Graphically represent the corresponding point cloud.
2) Determine an equation of the regression line from *y* to x.
3) Calculate the linear correlation coefficient between *x* and y. Comment on the result.
4) What number of stomata can be expected after 30 days of sun exposure? After 60 days?

Exercise 2. *(Based on the November 2007 exam)*

The table below gives an estimate of the amount of online purchases by French households:

Year	1998	1999	2000	2001	2002	2003	2004
Year Rank: x_i	0	1	2	3	4	5	6
Purchase amount in millions of euros: y_i	75	260	820	1650	2300	4000	5300

1) a) Specify the population, the variable(s) studied and the sample size.
b) Give an equation of the regression line from *y* to x.
c) Give the linear correlation coefficient between *x* and y. Interpret the result obtained.
d) What forecast of the amount of purchases can be made for the year 2005? Is it reliable?
2) We consider the new variable *zy* .
a) Determine an equation of the regression line from *z* to x, as well as the correlation coefficient linear between *x* and z. Interpret the result obtained.
b) Deduce an expression of *y* as a function of x, then a forecast of the amount of purchases for the year 2005.
3) From the data table, the Excel software proposes a polynomial adjustment by the equation $y = 130x^2 + 100x + 68$.
a) Is this the same adjustment as that obtained in 2)? Explain this situation.
b) Deduce from this adjustment a forecast of the amount of purchases for the year 2005.
4) The amount of online purchases in 2005 was 7,700 million euros. Which of the three adjustments Which of the preceding ones seems to you to be the most accurate? Justify your answer.

Exercise 3. *(Based on the May 2013 exam)*

The numerical results and the requested coefficients will be given with three decimal places.

The table below shows airline traffic, in millions of passengers, between Metropolitan France and foreign countries since 1980 (source INSEE).

Year	1980	1985	1990	1995	2000	2005	2008
Year Rank: x_i	0	5	10	15	20	25	28
Number of passengers (in millions): y_i	21,9	26,4	36,9	44,7	67,0	82,0	97,9

We are seeking to study the evolution of the number of passengers *between* metropolitan France and the countries foreigners according to the rank *x* of the year.

1) a) Graphically represent the statistical series *xi, yi.*
b) What adjustment curve(s) does this graphical representation suggest? Justify your answer.
2) A first adjustment.
a) Give an equation of the regression line from *y* to *x* (obtained by the method of least square).
b) Give the linear correlation coefficient between *x* and y. Interpret the result obtained.

3) A second adjustment. We consider the new variable *z lny.* a) Give an equation of the regression line from *z* to x, and the linear correlation coefficient between *x* and z. Interpret the result obtained.

b) Deduce a new expression for *y* as a function of x. c) Using this new model, determine an estimate of the number of passengers for the year 2011.

(d) Airlines expect the percentage increase between 2008 and 2011 to be 30%. Is this consistent with this second adjustment?

Exercise 4. *(Based on the November 2009 exam)*

The table below gives the experimental values of the volume *V* and the pressure *P* of a gas.

Volume (in cm^3) : v_i	620	890	1013	1186	1454	1944	2313	3179
Pressure (in Kg per cm^3) : p_i	6.7	4.3	3.48	2.644	1.997	1.35	1.1	0.7

According to Laplace's laws of thermodynamics for an ideal gas, we have the relation $PV^{\gamma} = C$ where and *These* are constants.

1) Specify the population, the variable(s) studied and the sample size.

2) Consider the variables $X = \ln V$ et $Y = \ln P$ Show that $Y = -\gamma X + \ln C$.

The table below gives the transformed experimental values:

$x_i = \ln v_i$	6,430	6.791	6.921	7.078	7,282	7,573	7,746	8,064
$y_i = \ln p_i$	1,902	1,459	1,253	0,956	0,693	0,336	0,095	–0,357

3) Give an equation of the regression line from *y* to x. Give the linear correlation coefficient between *x* and y. Interpret the result obtained.

4) Deduce, by justifying, the value of and of C, then an equation of *P* as a function of V.

5) Determine an estimate of the gas pressure for a volume of 2000 cm3 , then for 4000 cm3 . Are these two estimates reliable?

Exercise 5.

We select 12 people enrolled in a training course. Before the start of the training, these trainees take a test A marked from 0 to 20. At the end of the course, a test B identical to the first is also marked from 0 to 20. Considering the two variables *X* mark of A and *Y* mark of B, we obtained the following results:

intern	1	2	3	4	5	6	7	8	9	10	11	12
x_i	3	4	6	7	9	10	9	11	12	13	15	4
y_i	8	9	10	13	15	14	13	16	13	19	6	19

1) a) Represent these results by a scatter plot.

b) What fitting curve does this cloud suggest to you?

2) From the results obtained, we determined the regression line from *y* to x, thus the coefficient of linear correlation between *x* and y. We obtained the equation $y = 0,108\,x + 11,990$ and the coefficient $r = 0,101$.

Based on these results, explain why the fit is not good.

3) We decide to eliminate trainees 11 and 12, and therefore only take into account trainees 1 to 10.

a) Determine an equation of the regression line from *y* to *x* using the least squares method. b) Calculate the linear correlation coefficient between *x* and y. Interpret the result obtained.

Exercise 6.

We performed the affine adjustment of a point cloud. The equations obtained are as follows: - adjustment line from y to $x : D : y = x + 30$

adjustment line from x to $y : D' : x = 1/4y + 60$

1) Calculate the linear correlation coefficient.

2) Calculate the arithmetic means of x and y.

3) Calculate the covariance between *x* and *y* and the variance of x, knowing that the variance of *y* is equal to 40.