Quantitative Tools - Level 2 - Fall 2017
Louis de Charsonville

Lecture 2

Describing bivariate data (1/2)
OLS, correlation

Introduction Up to this point, the lectures focus on analysis of a single variable. It could be a quantitative variable like the income in a given firm or even in a gang dealing crack, or a qualitative variable like the colour of candies in a standard MMs Bag. But what most of statisticians are interesting in, is the analysis of the link between multiple variables. How is GDP of Uruguay related to the consumption in that country, how sales of a razor are related to expenditures in broadcast ads of that razor product, how grades in Statistics are linked to the hours spent studying (very deeply actually...) ? Assessing the link between two variables in this lecture's goal.
More precisely, we will do two things :

- establish a link between two given variables
- measuring the correlation between these two variations.

1 Linear adjustments

1.1 Brief history

Sir Francis Galton (1822-1911) is an English statistician and the father of concepts of correlation and regression towards the mean. Galton related the heights of sons with the heights of fathers and found a linear relation[1] in the following form :

$$\text{height}_{sons} = a * \text{height}_{fathers} + b + \epsilon_i$$

Indeed, the height of sons is no fully determined by the height of the father but is somehow determined by the height of the father. The a being a constant term and ϵ is an unobservable disturbance term accounting for noise or errors.

1.2 Method of Ordinary Least Squares (OLS)

Preliminaries The problem is the following, we have n individuals[2], each individual i having a set of two properties (x_i, y_i). We call X and Y, respectively, the variable that takes the values $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$. And we want to find the best linear relation between the variables X and Y. Or in other words, we want to explain Y with X. We call X, an **exogenous variable** (which explains) and Y an **endogenous variable** (which is explained). And we wish to investigate the relation between :

- An *outcome* or *exogenous* variable Y
- One explanatory variable X

Formally, That is, we want to find a a, and a b so that, for each i :

$$y_i = a * x_i + b + \epsilon_i \quad \text{with } \epsilon_i \text{ the smallest possible}$$

This method is called the **Ordinary Least Square method** or in abbreviate OLS.

[1] In French textbooks a linear relation is noted $y = ax + b$ whereas in English ones it is the inverse $y = bx + a$

[2] in Francis Galton experience, for each son i, x_i would his father's height and y_i his own height

Graphical intuition Plotting the y_i, x_i on a scatter plot (see Figure below), we want to find the straight line that fit the most to the data. That is for which the distance of each point to the straight line is minimal.

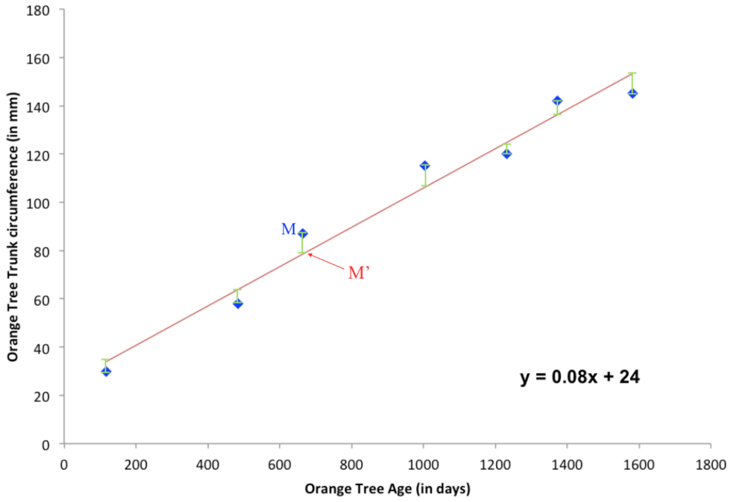

Figure 1: Orange Tree Circumferences according to its Age

This is equivalent to minimise the sum of the distance between each point of coordinate (x_i, y_i) and the straight line. That is on Figure 1, minimising the sum of the green segments.
The straight line has the functional form : $y = ax + b$.
So each point on the line has the coordinate $(ax_i + b, x_i)$.
The length of green segments are then, for a given point (x_i, y_i), equal to $|y_i - ax_i - b|$.
We want to choose a and b so that the sum of the squared length of green segments is minimal, that is :

$$\min_{a,b} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

One can find that the solution is :

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

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

$$\text{with} \begin{cases} \bar{x} \text{ the arithmetical mean of the } (x_i)_{i=1}^{n} \\ \bar{y} \text{ the arithmetical mean of the } (y_i)_{i=1}^{n} \end{cases}$$

Using the definition of variance and covariance :

$$var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

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

We have :

$$a = \frac{cov(x, y)}{var(x)}$$

Example 1.1. The Orange Tree We use the data of the orange tree presented in Figure 1, and derive a and b from their formula

Orange Tree age (in days) (x_i)	Orange Tree Trunk circumference (in mm) (y_i)	$x_i - \bar{x}$	$y_i - \bar{y}$
118	30	-804.14	-69.57
484	58	-438.14	-41.57
664	87	-258.14	-12.57
1004	115	81.86	15.43
1231	120	308.86	20.43
1372	142	449.86	42.43
1582	145	659.86	45.43
Mean	Mean	**a**	**b**
922.14	99.57	**0.08**	**24**

Table 1: The Lovely Orange Tree

2 Correlation

In the previous section, we show how to establish a linear relationship between two variables : one explained and one explaining. In this section, we are now seeking to establish how strong this link is. Indeed, when a linear relationship exists between two variables, there are said to be **correlated**, we want to measure how well there are correlated. For this, we compute a statistics named R^2

2.1 Sensitive approach and interpretation of results

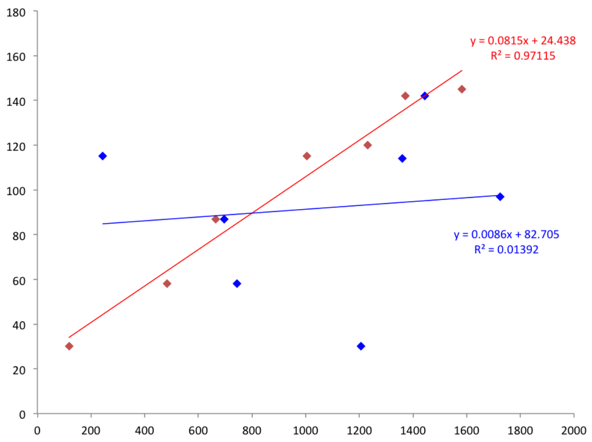

Figure 2: Correlation vs No-Correlation

[H] The Figure 2 shows two sets of data : the red one is the one of the orange tree, a linear relation can be seen by the data. On the contrary, no particular pattern arises from the blue set of data, the variable seems to be uncorrelated.

Sometimes, no linear relationship arises from the data because the relation between the data is not linear : it can be quadratic ($y = ax^2 + b$), like in Figure 3.

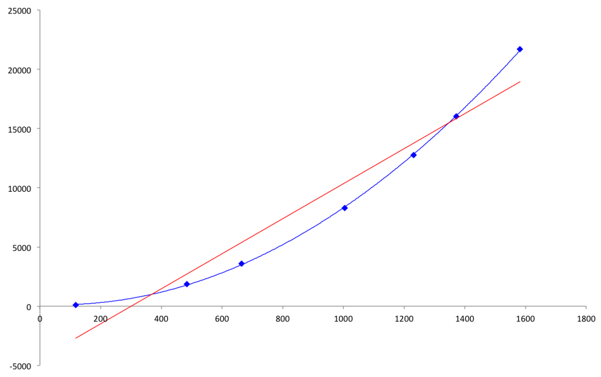

Figure 3: Quadratic correlation

Intuitively, we feel that the relationship between the two variables is great when the distance between the points and a straight line is little. Fortunately, we know how to calculate the distance between the straight line and the data points.

2.2 Total, explained and residual variance

Indeed, a measure of the distance between the straight line and the data points is what we minimise in the first section :

$$\sum_{i=1}^{n}(y_i - a * x_i - b)^2$$

We are going to use it in order to construct an estimator of "how good a correlation between the variable Y and X is".
Let's step back and try to sum up what we are willing to do. We have a variable Y that takes different values $(y_1, y_2, ..., y_n$; in other words, the variable Y varies, and we want to assess what part of this "variability" is explained by the "variability" of an other variable X. The variability of a variable is summed up in the *variance* of Y.
Then, what we are actually doing is computing **the variance of the variable Y explained by the variance X.**

Formally, let's denote $\hat{y}_i = a * x_i + b$.
We then have :

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Total Variance}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Explained Variance}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Residual Variance}}$$

Thus :

- **Total variance** is the variance of the endogenous variable
- **Explained variance** is the variance of the endogenous variable explained by the exogenous variable. Graphically, it is the variance of the projected y_i points on the straight line (the M' in Figure 1).
- **Residual variance** is the variance of the endogenous variable not explained by the exogenous variable. Graphically, it is also the distance between the data-points and the straight line.

The R^2 parameter A good estimator of the quality of the correlation, or the *goodness of fit* of our estimation, is then the explained variance normalised by the total variance, which we denote R^2 and is named the coefficient of determination. That is :

$$R^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$$
$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Remarks

- When the explained variance equals to the total variance : the variance of X explains perfectly the variance of Y and $R^2 = 1$
- In the contrary, the explained variance equals 0 and the residual variance equals the total variance, then $R^2 = 0$

$$\Rightarrow R^2 \text{ belongs to } [0,1] \text{ with } \begin{cases} 1 \text{ indicating perfect correlation} \\ 0 \text{ indicating no-correlation.} \end{cases}$$

2.2.1 Coefficient of correlation

In the previous sections, we've studied the strength of the correlation of the two variables. But the coefficient of determination, R^2 does not tell us whether the correlation is positive, as in Figure 4, or negative as in Figure 5.

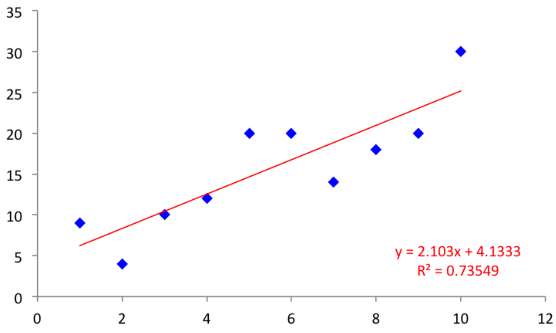

Figure 4: Positive correlation

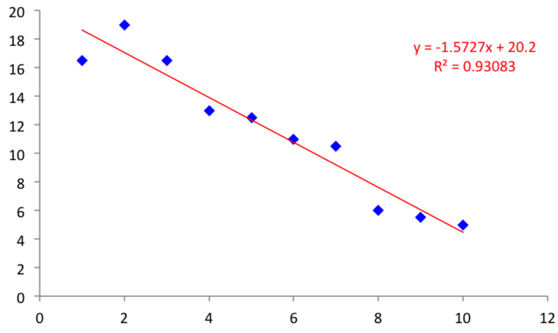

Figure 5: Negative correlation

Definition 2.1. Correlation coefficient
The *correlation coefficient*[3] is a measure of both the degree of correlation between two variables and the sign (positive or negative). It is usually denoted ρ.

Formally,

$$\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$
$$\rho^2 = R^2$$

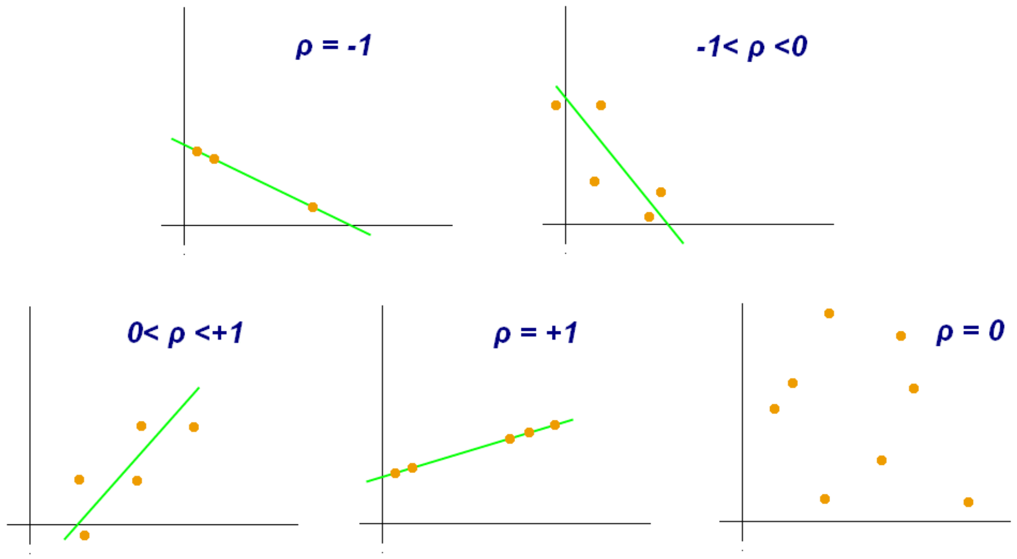

Figure 6: Scatterplots for different coefficient of correlation (source: Wikipedia)

[3] The extended name of this measure is "Pearson product-moment correlation coefficient"

2.3 Interpretations, spurious correlations

2.3.1 Interpretations

The calculus we made in the previous section are subject to a list of assumptions that when falses compromises the explicative power of the coefficients of the regression line and of the R^2.

Assumptions in Simple Linear Regression The most important assumptions are :

- We assume that the relation between the outcome variable and the explanatory variable is linear in parameters.
- $(x_i, y_i) : i = 1, ..., n$, where (x_i, y_i) are independent and identically distributed (i.i.d.). This means that the pair (x_i, y_i) is independent of (x_j, y_j) for $i \neq j$
- The errors are independent, identically distributed and have a zero conditional mean. This means that the errors ϵ are independent of X. This means that there is no third variable, not included in the model that could explained the variance of both the output and the explanatory powers (see below example of spurious correlations).
- We also assume that x is not constant, which means that we do not have $x_1 = x_2 = ... = x_n$

Common mistakes

- The size of the sample is important : there is always a perfect linear match when there are only two points of data. We want our estimators to be significant and we need large samples to do this.
- **Correlation is not causality** : the facts that two variables are correlated, does not mean that there is a causal link between two variables. For instance, sales of ice creams and number deaths by drowning are very correlated in seaside resorts. This is called a **spurious relationship**. Spurious relationship is the coined term for the phenomena when two variables are correlated due to an unseen factor or lurking variable rather than to direct causal connection.
- Correlation can also simply exists by chance... See the following graphs

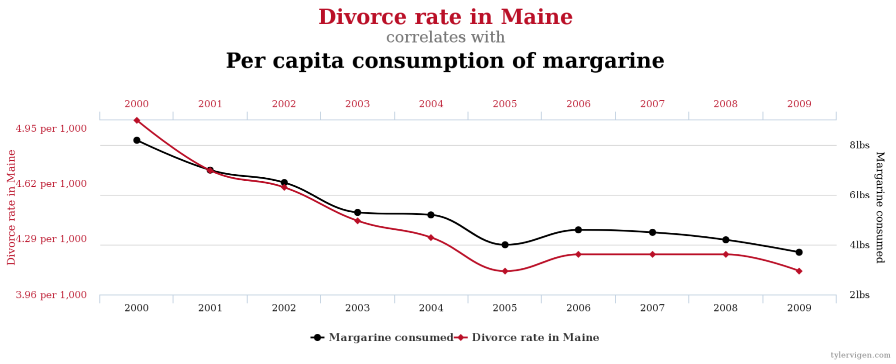

Figure 7: Divorce rate in Maine correlates with per capita consumption of margarine

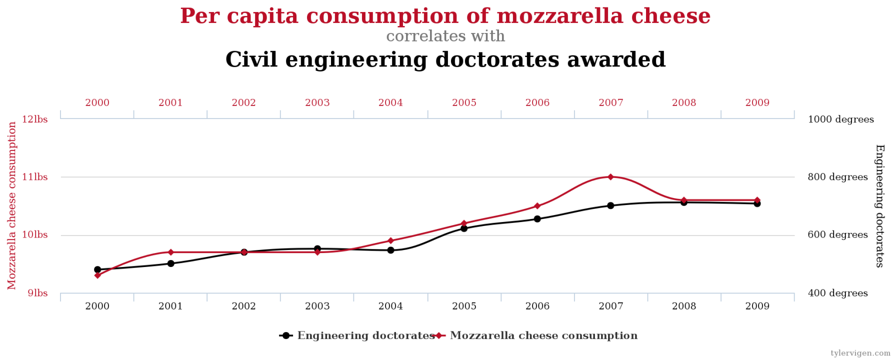

Figure 8: Per capita consumption of mozzarella cheese correlates with Civil engineering doctorates awarded

* * *