Econometrics visualization

Lucas Chaveneau, Thibault Fuchez, Allan Guichard

2021-12-21

Contents

Chapter 1

Introduction

1.1 What is Econometrics?

«««< HEAD Econometrics is a branch of economics that aims to estimate and test economic models (simplified representation of reality). Thus, the econometer tries to identify the parameters of a model by means of statistical estimation, and thus tries to induce the characteristics of a general group (the population) from those of a particular group (the sample). ======= BLABLABLA

Econometrics is a branch of economics that aims to estimate and test economic models (simplified representation of reality). Thus, the econometrician tries to identify the parameters of a model by means of statistical estimation, and thus tries to induce the characteristics of a general group (the population) from those of a particular group (the sample). »»»> deedb0dc1cd0094bc1d3cb0eaf99e5e078b9f27a

Three essential words in the language of the econometer are: correlation, regression, and causality.

To better understand the relationships between different variables, visualization is a key tool to help interpret mathematical and statistical results.

1.2 What about econometric visualization?

We can start by saying what it is not: it is not strictly speaking data visualization. Data visualization "the graphical display of data" is often seen as something trivial, to be quickly scrolled through to show stakeholders simple but salient facts of a possibly very complex problem…

The analyst consciously chooses what to include in a visualization in order to identify intuitively relevant patterns and trends in the data in the most efficient way. The analyst then makes choices in this representation. Modern data sets tend to be very large (in terms of number of observations) and broad (in terms of number of variables). Therefore, it is hard to simply represent all the data. The difficulty is that everything must be made as simple as possible - but not simpler.

Visualization is used to explain, notably to teach, econometrics in order to facilitate the understanding of the fundamental concepts without necessarily using a mathematically complex corpus. Diagrammatic representations of models, methods and diagrams can facilitate the understanding and reading of results.

1.3 State of the art

Data visualization is an important part of any statistical analysis. Any good statistician will tell you that it is dangerous to undertake an econometric analysis without first examining the data. Scatterplots are used as a tool to diagnose the overall trend, the relationship between variables, the variation around the perceived trend, the exceptions whether they are outliers or distinct groups of observations, the discontinuities. The analysis of residuals largely uses the same tools.

A bibliographic search on the theme of visualization in econometrics does not yield much. Nevertheless, we will mention in this document some articles that deal with the subject in a direct or indirect way, and that have been an anchor point in our reflections.

Moreover, some packages related to econometric visualization have already been developed in R (or other languages). At the end of this document, we will briefly describe the state of the art of these packages.

1.4 Our project, an R package

Our project stands to propose a library developed under R which proposes different ways to visualize the essential elements of the econometer.

Our package will therefore focus on the representation of correlations. We will try to put forward a clear and synthetic representation of the covariance, a representation of the elements that allow to visualize the quality of a regression, as well as the part of each variable in the explanation of the variable to predict.

This document is intended to be a guide for the package. It proposes a definition of the econometric tools usefull to understand the package. It also proposes an overview of the different visualization methods that already exist.

Chapter 2

Covariance

2.1 Variance definition reminder

In the fields of statistics and probability theory, the variance can be define as a measure of a sample's values dispersion or a measure of a probability distribution. According to the König-Huygens theorem, the variance show the gap between the squares of the values of the variable and the square of the mean.

- Classical formula of variance:

$$\sigma_x^2 = \frac{1}{N}\sum_{x=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{x=1}^{N}x_i^2 - \bar{x}^2$$

- A new proposition for variance formula(Heffernan, 1988):

$$\sigma_x^2 = \frac{1}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j>i}^{N}(x_i - x_j)^2$$

This formulation (extended to covariance) will be discussed again in this study. Intuitively, we can see here the geometrical aspect of the variance of x perceived as the expression of a square.

- Variance versus Covariance:

Variance and covariance are mathematical terms frequently used in statistics and probability theory. Variance refers to the spread of a data set around its mean value, while a covariance refers to the measure of the directional relationship between two random variables.

2.2 Covariance standard definition

Covariance as an extension of variance notion. The covariance between two random variables is a number which provide us the joint difference to there respective means in quantity.

A covariance refers to the measure of how two random variables will change when they are compared to each other.

Intuitively, covariance is a measure of the simultaneous variation of two random variables. That is, the covariance becomes more positive for each pair of values that differ from their mean in the same direction, and more negative for each pair of values that differ from their mean in the opposite direction.

The sign of the covariance therefore shows the tendency in the linear relationship between the variables. The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables.

The covariance of two independent random variables is zero, although the converse is not always true.

When we have more than 2 variables, the concept naturally generalizes through the covariance matrix (or the variance-covariance matrix). Let be a set of p real random variables $X_1, X_2, ..., X_p$, then the covariance matrix is the square matrix of which the term line i and the term column j is the covariance of the variables X_i and X_j. This matrix quantifies the variation of each variables compared to each of the others.

2.3 Mathematical standard and alternatives covariance definition

2.3.1 Standard definition

- Covariance formula:

For two jointly distributed real-valued random variables X and Y with finite, the covariance is defined as the expected value (or mean) of the product of their deviations from their individual expected values.

$$Cov(X,Y) = \mathbb{E}((X - \mathbb{E}(X)) \times (Y - \mathbb{E}(Y)))$$

Where $\mathbb{E}(X)$ is the expected value of X, also known as the mean of X. The covariance is also sometimes denoted $\sigma_{X,Y}$, in analogy to variance. By using the linearity property of expectations, this can be simplified to the expected value of their product minus the product of their expected values:

$$\begin{aligned} Cov(X,Y) &= \mathbb{E}((X - \mathbb{E}(X)) \times (Y - \mathbb{E}(Y))) \\ &= \mathbb{E}(XY - X\mathbb{E}(Y) - \mathbb{E}(X)Y + \mathbb{E}(x)\mathbb{E}(Y)) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(x)\mathbb{E}(Y) + \mathbb{E}(x)\mathbb{E}(Y) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \end{aligned}$$

The empirical covariance of a sample is defined by:

$$Cov(X,\ Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

With, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_j$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_j$ an unbiased estimator of the population.

Covariance is defined by:

$$Cov(X,\ Y) = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

or equivalently:

$$Cov(X,\ Y) = \frac{n}{n-1}(\overline{xy} - \overline{x}\ \overline{y})$$

2.3.2 Alternative definition, a story of rectangles

- Formula from Heffernan definition of covariance:

$$Cov(X,Y) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^{n} \frac{1}{2}(x_i - x_j)(y_i - y_j)$$

Let be two random variables $(X,\ Y)$, and a sample of N pairs of independent observations.

$$(x_1,\ y_1), (x_2,\ y_2), \cdots, (x_N,\ y_N)$$

Let us consider two obervations in a chart k and l, and let us calculate the mathematical expectation of the area of the rectangle formed by both points.

$$\mathbb{E}(x_k - x_l)(y_k - y_l)$$

So

$$\mathbb{E}(x_k y_k - x_k y_l - x_l y_k + x_l y_l)$$
$$\mathbb{E}(x_k y_k) - \mathbb{E}(x_k)\mathbb{E}(y_l) - \mathbb{E}(x_l)\mathbb{E}(y_k) + \mathbb{E}(x_l y_l)$$
$$2\mathbb{E}(XY) - 2\mathbb{E}(X)\mathbb{E}(Y)$$

Hence, covariance equal to:

$$\mathbb{E}(x_k - x_l)(y_k - y_l) = 2\ \ Cov(X,Y)$$

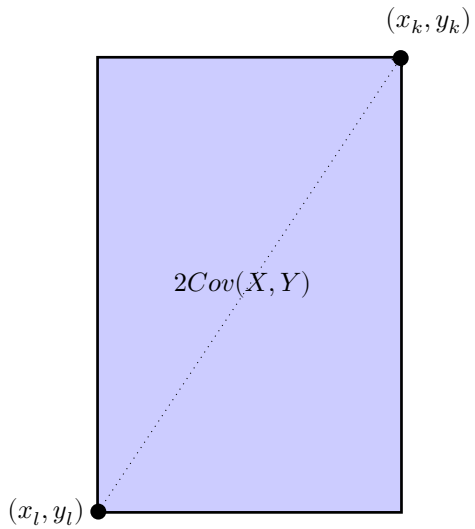


Figure 2.1: Positive Correlation

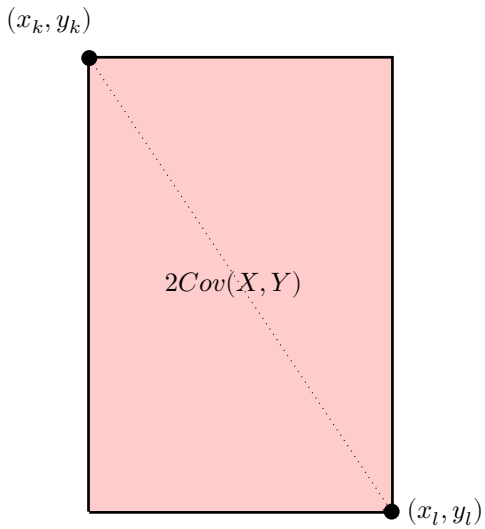


Figure 2.1: Negative Correlation

On a sample of n pairs of observations (x_i, y_i) with $i = 1, \ldots, n$ the empirical covariance can also

be computed as the average of the $n(n-1)$ areas of the (x_i, y_i) and (x_j, y_j) vertex rectangles for $i = 1, \dots, n-1$ and $j = 1, \dots, n$
Note that the n rectangles (degenerate, of zero area) of vertex (x_i, y_i) and (x_j, y_j) with $i = j$ should not be taken into account in the calculation.

$$Cov(X,\ Y) = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (x_i - x_j)(y_i - y_j)$$

An equivalent definition is to consider half (due to symmetry) of the average of the $n(n-1)$ areas of the rectangles of vertices (x_i, y_i) and (x_j, y_j) for $i = 1, \dots, n$ and $j = 1, \dots, n$ (evidence in annexes).

$$Cov(X,\ Y) = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)(y_i - y_j)$$

Chapter 3

Correlation

3.1 Correlation overview

3.1.1 From covariance to correlation

The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

- Both covariance and correlation measure the relationship and the dependency between two variables.
- Correlation values are standardized while covariance values are not.
- Covariance indicates the direction of the linear relationship between variables as well as the correlation but the latter measure also the strengh of the relationship.

The normalized form of the covariance matrix is the correlation matrix.

3.1.2 Correlation definition

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data.

In the broadest sense correlation is any statistical association, though it actually refers to the degree to which a pair of variables are linearly related.

There are several correlation coefficients, often denoted ρ or r, measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may be present even when one variable is a nonlinear function of the other). Other correlation coefficients – such as Spearman's rank correlation – have been developed to be more robust than Pearson's, that is, more sensitive to nonlinear relationships.

3.1.3 Galton, a pioneer in the history of correlation

> "I can only say that there is a vast field of topics that fall under the laws of correlation, which lies quite open to the research of any competent person who cares to investigate it." (Galton, 1890)

Galton's 1888 paper (Galton, 1888), presented to the Royal Society in London, defines correlation as follows:

> "Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction.... It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes... If they were in no respect due to common causes, the co-relation would be nil."

Galton's definition reveals the properties of the correlation coefficient. It is a measure of the strength of a linear relationship; the closer it is to 1, the more two variables can be predicted from each other by a linear equation. It is a measure of direction: a positive correlation indicates that X, Y increase together; a negative correlation indicates that one decreases as the other increases. Note that Galton does not claim that co-relation implies cause and effect (it would be absurd to assume that the size of one organ determines the size of another). Galton hypothesized that the correlation indicated the presence of "common causes" for the observed relationship between the variables (the size of each organ respectively).

More technically Galton continues his presentation as follows:

> "Let y = the deviation of the subject [in units of the probably error, Q], whichever of the two variables may be taken in that capacity; and let x1, x2, x3, &c., be the corresponding deviations of the relative, and let the mean of these be X. Then we find: (1) that y = rX for all values of y; (2) that r is the same, whichever of the two variables is taken for the subject; (3) that r is always less that 1; (4) that r measures the closeness of co-relation."

Galton particularly liked the correlation coefficient because it could be used to predict deviations Y from X or X from Y. Thus, from the beginning, the correlation coefficient was closely related to the regression line. Originally, r meant the regression slope, but there was a problem in that the regression line of the slope was partly a function of the units of measurement chosen. Galton perceived the correlation coefficient as a unitless regression slope and appropriated the label r.

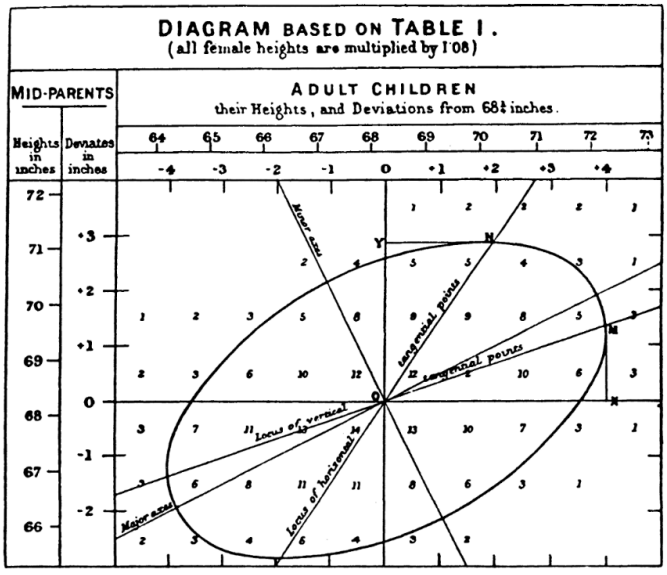


Figure 3.1: The first Bivariate Scatterplot

«««< HEAD ### Thirteen ways to see correlation (Rodgers and co, 1988)

> In 1885, Sir Francis Galton first defined the term "regression" and completed the theory of bivariate correlation. A decade later, Karl Pearson developed the index that we still

use to measure correlation, Pearson's r . Our article is written in recognition of the 100th anniversary of Galton's first discussion of regression and correlation.

According Joseph Lee Rodgers and co article, Several ways to interpret the correlation:

- As standardized covariance
$$r = \frac{Cov(x,y)}{\sigma_x^2 \sigma_y^2}$$
- As standardized slope of regression line
$$r = b_{Y.X}\left(\frac{\sigma_x^2}{\sigma_y^2}\right) = b_{X.Y}\left(\frac{\sigma_y^2}{\sigma_x^2}\right)$$
- As geometric mean of the two regression slopes
$$r = \pm\sqrt{b_{Y.X} \times b_{X.Y}}$$
- As the square root of the ratio of two variances
$$r = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)}{\sum(Y_i - \bar{Y}_i)}} = \sqrt{\frac{SS_{reg}}{SS_{tot}}}$$
- And so many other…

3.1.4 Pearson correlation coefficient

- Some notations that may be useful:

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

$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

- Pearson correlation coefficient:
$$\rho = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

The Pearson correlation coefficient is a bounded index (i.e., $-1 \leq \rho \leq 1$) that provides a unitless measure for the strength and direction of the association between two variables.

3.2 Alternatives correlation measurement

3.2.1 Spearman's rank correlation coefficient

Measures the association based on the ranks of the variables.

$$\hat{\theta} = \frac{\sum_{i=1}^{n}(R_i - \bar{R}(S_i - \bar{S}))}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(S_i - \bar{S})^2}}$$

Where R_i and S_i are the rank of the x_i and y_i values, respectively.

Note that this is just the estimated Pearson's correlation coeffcient, but the values of the variables have been replaced by their respective ranks.

The Spearman correlation is the non-parametric equivalent of the Pearson correlation. It measures the relationship between two variables. If the variables are ordinal, discrete or do not follow a normal distribution, the Spearman correlation is used. This correlation does not use the values of the data but their rank. In fact, nothing changes, everything is the same as calculating a Pearson correlation but on transformed variables. The interest of establishing a correlation on the ranks of the variables is to detect if there is a monotonic relationship, which may not be linear.

3.2.2 Partial Correlation

The partial correlation coefficient, noted here $r_{AB.C}$, allows us to know the value of the correlation between two variables A and B, if the variable C had remained constant for the series of observations considered.

Put another way, the partial correlation coefficient $r_{AB.C}$ is the total correlation coefficient between variables A and B when we have removed their best linear explanation in terms of C. It is given by the formula :

$$r_{AB.C} = \frac{r_{AB} - r_{AC} \cdot r_{BC}}{\sqrt{1 - r_{AC}^2} \cdot \sqrt{1 - r_{BC}^2}}$$

Let's go a little further in understanding this coefficient:

The OLS estimator of β is written

$$\hat{\beta} = \frac{Cov(y, x_1)}{\mathbb{V}ar(x_1)}$$

The estimator β' is written

$$\begin{aligned}
\hat{\beta}' &= \frac{Cov(y, x_1)\mathbb{V}(x_2) - Cov(y, x_2)Cov(x_1, x_2)}{\mathbb{V}(x_1)\mathbb{V}(x_2) - Cov(x_1, x_2)^2} \\
&= \hat{\beta}' = \frac{\rho_{y1}\sigma_y\sigma_1\sigma_2^2 - \rho_{y2}\sigma_y\sigma_2\rho_{12}\sigma_1\sigma_2}{\sigma_1^2\sigma_2^2 - \rho_{12}^2\sigma_1^2\sigma_2^2} \\
&= \hat{\beta}' = \frac{\rho_{y1} - \rho_{y2}\rho_{12}}{1 - \rho_{12}^2} \quad \frac{\sigma_y}{\sigma_1}
\end{aligned}$$

After some transformation we have:

$$\hat{\beta}' = \underbrace{\underbrace{\frac{\rho_{y1} - \rho_{y2}\rho_{12}}{\sqrt{1 - \rho_{12}^2}\sqrt{1 - \rho_{y2}^2}}}_{\text{Partial correlation}}}_{\rho_{y1|2}} \quad \frac{\sigma_y\sqrt{1 - \rho_{y2}^2}}{\sigma_1\sqrt{1 - \rho_{12}^2}}$$

«««< HEAD In order to understand this expression, consider the following two regressions:

$$x_1 = \kappa + \tau x_2 + \varepsilon_{1|2}$$

$$y = \delta + \gamma x_2 + \varepsilon_{y|2}$$

We have:

$$\begin{aligned}
Cov(e_{1|2}, e_{y|2}) &= Cov(x_1 - \hat{\kappa} - \hat{\tau}x_2,\ y - \hat{\delta} - \hat{\gamma}x_2) \\
&= Cov(x_1, y) - \hat{\gamma}Cov(x_1, x_2) - \hat{\tau}Cov(x_1, y) + \hat{\gamma}\hat{\tau}\mathbb{V}ar(x_2) \\
\mathbb{V}ar(e_{y|2}) &= \mathbb{V}ar(y - \hat{\delta} - \hat{\gamma}x_2) \\
&= \mathbb{V}ar(y) - 2\hat{\gamma}Cov(x_1, y) + \hat{\gamma}^2\mathbb{V}ar(x_2) \\
\mathbb{V}ar(e_{1|2}) &= \mathbb{V}ar(x_1 - \hat{\kappa} - \hat{\tau}x_2) \\
&= \mathbb{V}ar(x_1) - 2\hat{\tau}Cov(x_1, x_2) + \hat{\tau}^2\mathbb{V}ar(x_2)
\end{aligned}$$

The linear correlation coefficient between $e_{1|2}$ and $e_{y|2}$ corresponds to the correlation between y and x_1 after taking into account the linear influence of x_2 on these two respective variables:

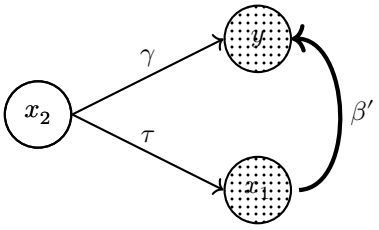
$$\rho_{yx_1|x_2} = \frac{Cov(e_{1|2}, e_{y|2})}{\sqrt{\mathbb{V}ar(e_{y|2})\mathbb{V}ar(e_{1|2})}}$$

After simplification we obtain the expression of the partial correlation:

$$\rho_{y1|2} = \frac{\rho_{y1} - \rho_{y2}\rho_{12}}{\sqrt{(1 - \rho_{y2}^2)(1 - \rho_{12}^2)}}$$

And so, the estimator $\hat{\beta}'$ can thus be written as that of a simple linear regression where the variables are the residuals of prior regressions of y respectively x_1 on x_2.

$$\hat{\beta}' = \rho_{y1|2}\frac{\sigma_y\sqrt{1 - \rho_{y2}^2}}{\sigma_1\sqrt{1 - \rho_{12}^2}}$$



«««< HEAD

When the number of variable is quite high, computing the partial correlation coefficient can be quite laborious. It is advised to use some regression methods when there are more than 3 variables. The alternative is to compute regression's residuals of both chosen variables on the other variables. This approach leads to the same results. Let's remind that partial correlation measuring the link between the residual information of X and Y which is not already explained by the other variables. The j-order partial correlation amounts to calculating the correlation between the regression's residuals.

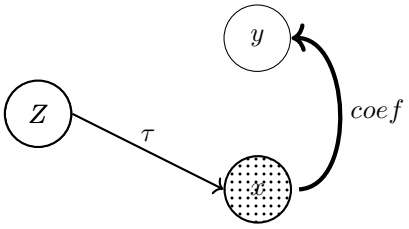
$$\rho_{x_1 y.x_2,\ldots,x_j} = \rho_{e_{x_1} e_y}$$

3.2.3 Semi partial Correlation

Unlike to partial correlation, semi-partial correlation is asymetrical. it get closer from multiple regression. We try to quantify, for one variable, its additional ability to explain.

For more accuracy, we take off the third variable information from one of both variables. Thanks to it, We are seeking to quantify the link between y and the residuals parts of x in relation to the third variable.

$$\rho_{xy.z_1,\dots,z_j} = \rho_{e_x e_y}$$



«««< HEAD

$$r_{y(x.z)} = \frac{r_{yx} - r_{yz}r_{xz}}{\sqrt{1 - r_{xz}^2}}$$

It is obvious that if X and Z are indepedent so $r_{y(x.z)} = r_{yx}$. Unlike, If X and Z are perfectly correlated, $r_{y(x.z)}$ is undefined, which means nothing remains in the X-residuals to explain Z.

3.2.4 Transitivity correlation

Let ρ_{xy} be the correlation between the variables X and Y. Let ρ_{xz} and ρ_{yz} be the correlations of variables X and Y with respect to a third variable Z.

Given ρ_{xz} and ρ_{yz}, can we deduce the possible values for ρ_{xy}?

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}}$$

$$\begin{aligned}\rho_{XY} &= \left(\rho_{XY|Z} - \frac{-\rho_{XZ}\rho_{YZ}}{\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}}\right)\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2} \\ &= \rho_{XY|Z}\sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2} + \rho_{XZ}\rho_{YZ}\end{aligned}$$

ρ_{xy} is in the range $\rho_{XZ}\rho_{YZ} \pm \sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2}$

$$\begin{aligned}\rho_{XZ}\rho_{YZ} - \sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2} > 0 \\ \rho_{XZ}\rho_{YZ} + \sqrt{1 - \rho_{XZ}^2}\sqrt{1 - \rho_{YZ}^2} > 0.\end{aligned}$$

$$\begin{aligned}\rho_{XZ}^2\rho_{YZ}^2 &> (1 - \rho_{XZ}^2)(1 - \rho_{YZ}^2) \\ &= 1 - \rho_{XZ}^2 - \rho_{YZ}^2 + \rho_{XZ}^2\rho_{YZ}^2.\end{aligned}$$

Si $\rho_{xy} > 0$ alors on a

$$\begin{aligned}\rho_{XZ}^2\rho_{YZ}^2 &> (1-\rho_{XZ}^2)\,(1-\rho_{YZ}^2)\\ &= 1-\rho_{XZ}^2-\rho_{YZ}^2+\rho_{XZ}^2\rho_{YZ}^2.\end{aligned}$$

$$\text{sign}(\rho_{XY}) = \begin{cases}\text{sign}(\rho_{XZ})\text{sign}(\rho_{YZ}) & \text{if } \rho_{XZ}^2+\rho_{YZ}^2 > 1\\ \text{not known} & \text{if } \rho_{XZ}^2+\rho_{YZ}^2 \leq 1.\end{cases}$$

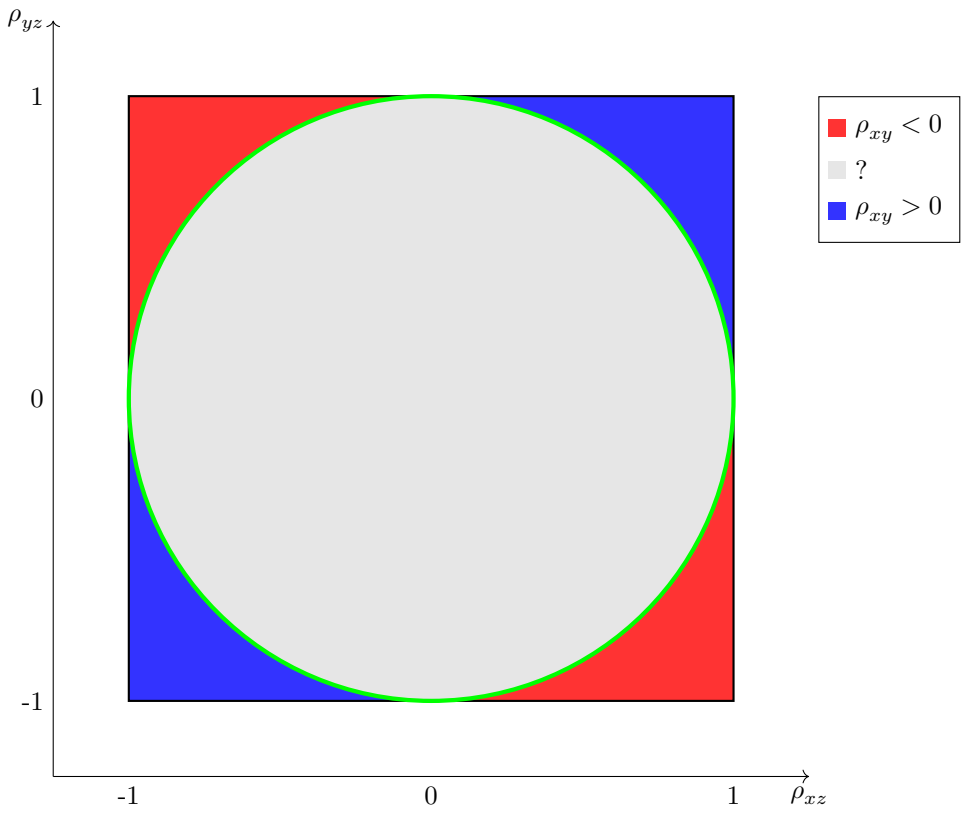


Figure 3.2: Value of $\rho_{x,y}$ correlated with a third variable

3.3 Distance correlation

The so-called association measures are an active and recent field of research that renews the well established and old field of correlation. The *energy* package developed under R as well as Gabor's article (**?**) are good references.

Let (x_i, y_i), $i = 1, 2, \ldots, N$ be a sample of pairs of observations of the variables X and Y.

We compute successively

$$dx_{ij} = \|x_i - xj\|$$

$$dy_{ij} = \|y_i - yj\|$$

$$\overline{\overline{dx_{ij}}} = dx_{ij} - \overline{dx_{i.}} - \overline{dx_{.j}} + \overline{dx_{..}}$$
$$\overline{\overline{dy_{ij}}} = dy_{ij} - \overline{dy_{i.}} - \overline{dy_{.j}} + \overline{dy_{..}}$$

with

$\overline{dx_{i.}} = \frac{1}{N} \sum_{j=1}^{N} dx_{ij}$

and

$\overline{dx_{..}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} dx_{ij}$

$$dCov(X,Y) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \overline{\overline{dx_{ij}}}\ \overline{\overline{dy_{ij}}}$$

$$dVar(X) = dCov^2(X,X) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \overline{\overline{dx_{ij}}}^{\,2}$$

$$d\rho = \frac{dCov(X,Y)}{d\sigma_X\ d\sigma_Y}$$

3.3.1 Other Correlations

Here we introduced some correlation type (Makowski and co, 2020):

- Kendall's rank correlation: In the normal case, the Kendall correlation is preferred to the Spearman correlation because of a smaller gross error sensitivity (GES) and a smaller asymptotic variance (AV), making it more robust and more efficient. However, the interpretation of Kendall's tau is less direct compared to that of the Spearman's rho, in the sense that it quantifies the difference between the % of concordant and discordant pairs among all possible pairwise events. Confidence Intervals (CI) for Kendall's correlations are computed using the Fieller et al. (1957) correction (see Bishara & Hittner, 2017).
- Hoeffding's D: The Hoeffding's D statistic is a non-parametric rank based measure of association that detects more general departures from independence (Hoeffding 1948), including non-linear associations. Hoeffding's D varies between -0.5 and 1 (if there are no tied ranks, otherwise it can have lower values), with larger values indicating a stronger relationship between the variables.
- Gamma correlation: The Goodman-Kruskal gamma statistic is similar to Kendall's Tau coefficient. It is relatively robust to outliers and deals well with data that have many ties.
- Gaussian rank correlation: The Gaussian rank correlation estimator is a simple and wellperforming alternative for robust rank correlations (Boudt et al., 2012). It is based on the Gaussian quantiles of the ranks.
- Winsorized correlation: Correlation of variables that have been Winsorized, i.e., transformed by limiting extreme values to reduce the effect of possibly spurious outliers.
- Biweight midcorrelation: A measure of similarity that is median-based, instead of the traditional mean-based, thus being less sensitive to outliers. It can be used as a robust alternative to other similarity metrics, such as Pearson correlation (Langfelder & Horvath, 2012).

- Percentage bend correlation: Introduced by Wilcox (1994), it is based on a downweight of a specified percentage of marginal observations deviating from the median (by default, 20 percent).
- Shepherd's Pi correlation: Equivalent to a Spearman's rank correlation after outliers removal (by means of bootstrapped Mahalanobis distance).
- Point-Biserial and biserial correlation: Correlation coefficient used when one variable is continuous and the other is dichotomous (binary). Point-Biserial is equivalent to a Pearson's correlation, while Biserial should be used when the binary variable is assumed to have an underlying continuity. For example, anxiety level can be measured on a continuous scale, but can be classified dichotomously as high/low.
- Tetrachoric correlation: Special case of the polychoric correlation applicable when both observed variables are dichotomous.

Chapter 4

Visualization of covariance

4.1 State of the art: different attempts to represent the covariance

4.1.1 Venn diagramm

A Venn diagram is a widely used diagram style that shows the logical relation between sets, popularized by John Venn in the 1880s. The diagrams are used to teach elementary set theory, and to illustrate simple set relationships in probability, logic, statistics, linguistics and computer science. A Venn diagram uses simple closed curves drawn on a plane to represent sets. Very often, these curves are circles or ellipses.

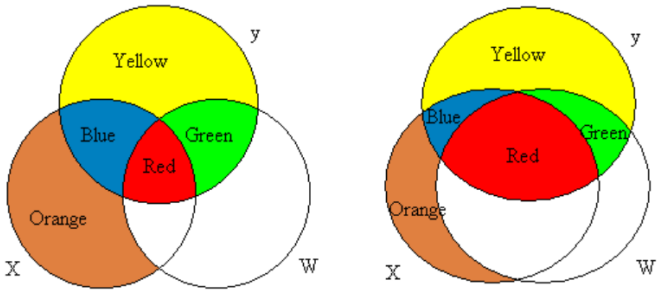


Figure 4.1: Ballentine Venn diagrams displaying modest and considerable collinearity

4.1.2 Visualizing distributions of covariance matrices

Covariance matrices and their corresponding distributions play an important role in statistics. To understand the properties of distributions, we often rely on visualization methods. (Tokudaa and co, 2011)

Visualizing a distribution in a high-dimensional space is a challenge, with the additional difficulty that covariance matrices must be positive semi-definite, a restriction that forces the joint distribution of the covariances into an oddly-shaped subregion of the space.

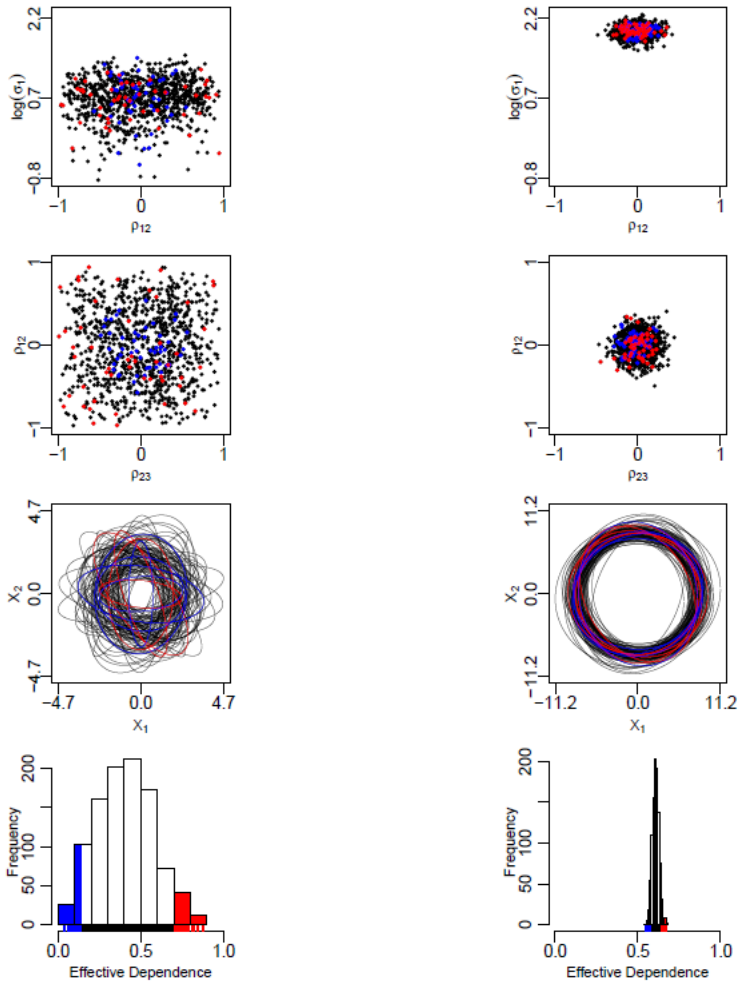


Figure 4.2: Visualize distributions of covariance matrices

4.1.3 A geometrical interpretation of an alternative formula for the sample covariance

Kevin Hayes (Hayes, 2011) proposes a new geometric and visual interpretation of covariance, based on the application of Hefferman's variance formula. He extends this formula to the covariance of a sample to extract his results.

- Formula from Heffernan definition of covariance:

$$cov(X,Y) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j>i}^{n} \frac{1}{2}(x_i - x_j)(y_i - y_j)$$

Geometrically, $\frac{1}{2}(x_i - x_j)(y_i - y_j)$ is ± 1 times the area right-triangle formed with the difference vector $(x_i—x_j, y_j)$ as its hypotenuse, where negatively sloped difference tors incur a (—1) sign and positively sloped difference vectors take a (+1) sign. (Hayes, 2011)

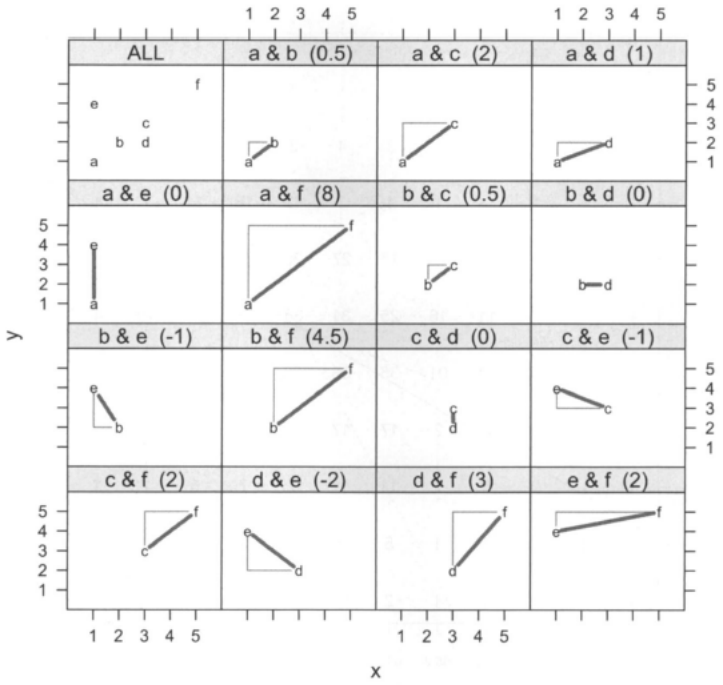


Figure 4.3: Some examples of covariance as area of rectangles

4.1.4 Covariance as signed area of rectangles

This article (Chudzicki, 2014) was written following a very interesting conversation on the **stats.statckexchange** site. The initial topic of this conversation was: "How to explain covariance to someone who only understands the notion of mean?".

Instructions for use:

- Draw all possible such rectangles. Color them transparently, making the positive rectangles red (say) and the negative rectangles "anti-red" (blue).

- The covariance is the net amount of red in the plot (treating blue as negative values).

Let's deduce some properties of covariance. Understanding of these properties will be accessible to anyone who has actually drawn a few of the rectangles.

- Bilinearity:

Because the amount of red depends on the size of the plot, covariance is directly proportional to the scale on the x-axis and to the scale on the y-axis.

- Correlation:

Covariance increases as the points approximate an upward sloping line and decreases as the points approximate a downward sloping line. This is because in the former case most of the rectangles are positive and in the latter case, most are negative.

- Relationship to linear associations:

Because non-linear associations can create mixtures of positive and negative rectangles, they lead to unpredictable (and not very useful) covariances. Linear associations can be fully interpreted by means of the preceding two characterizations.

- Sensitivity to outliers:

A geometric outlier (one point standing away from the mass) will create many large rectangles in association with all the other points. It alone can create a net positive or negative amount of red in the overall picture.

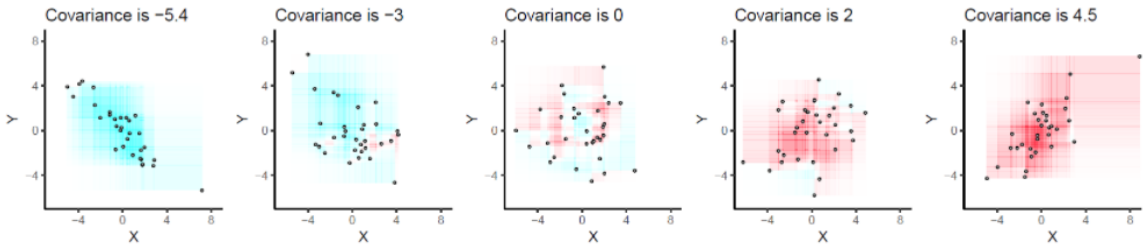
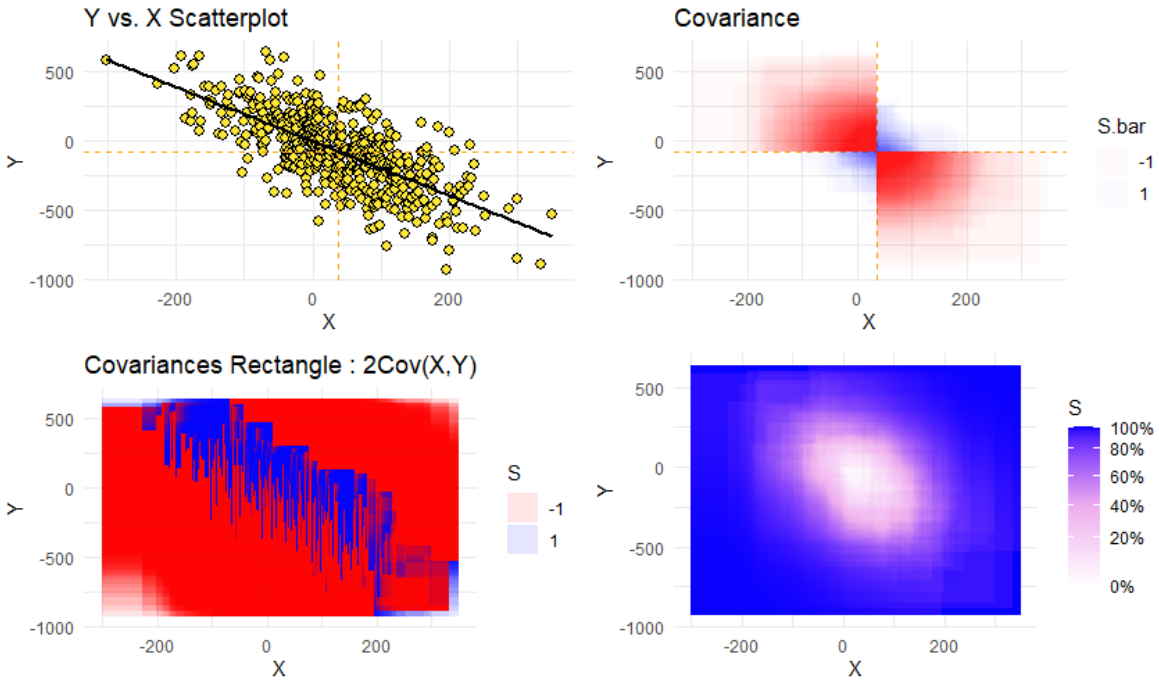


Figure 4.4: Covariance rectangle

4.2 Our current project: the package Plotnetrec

The Plonetrec package and its associated graphs is an alternative representation of a scatterplot. The aim is to detect certain particularities in a dataset.

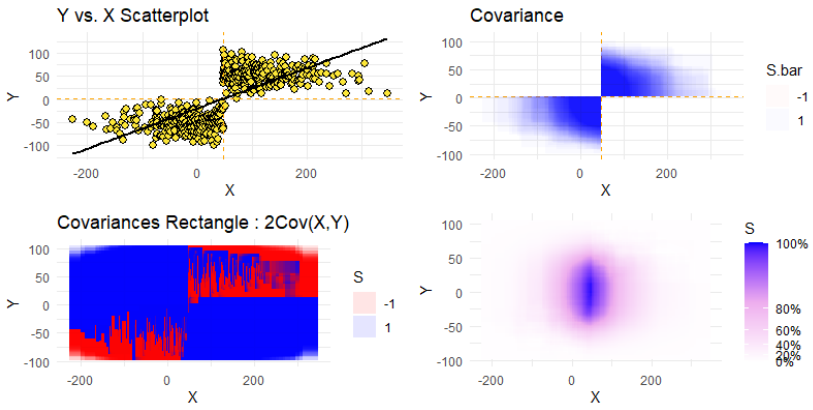


We locate in a standard case, whitout issues in he dataset.

- The graph at the top right relates all the observations to their respective averages and we draw all corresponding rectangles . If the slope of the diagonal is negative, the rectangle is red and conversely, the rectangle is blue.
- The bottom-left graphic relates all observations in pairs to draw all rectangle. In this figure, their is superposition of the rectangles. The sum of all rectangles areas gives two times the Covariance.
- The bottom-righ graphic differs from the bottom-left by the coloring. We apply a transparency effect according to the net amount of the two colors. A superposition of red and blue will cancel each other and it results a lack of color. The more a color is represented, the more blue is intense.

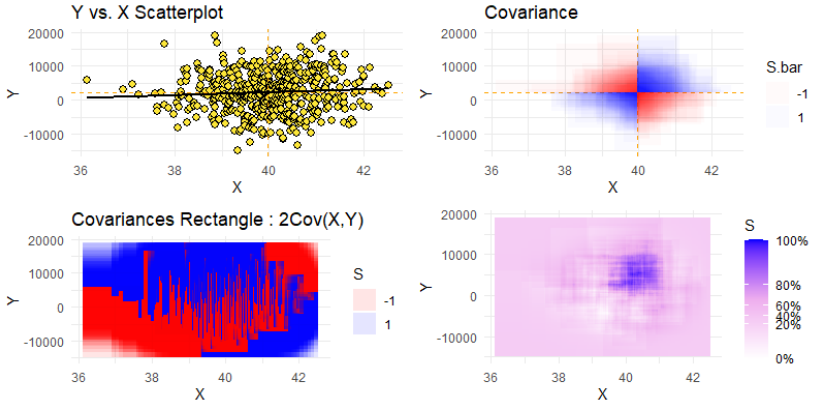
Let's compare with some problematic.

4.2.1 Heterogeneity



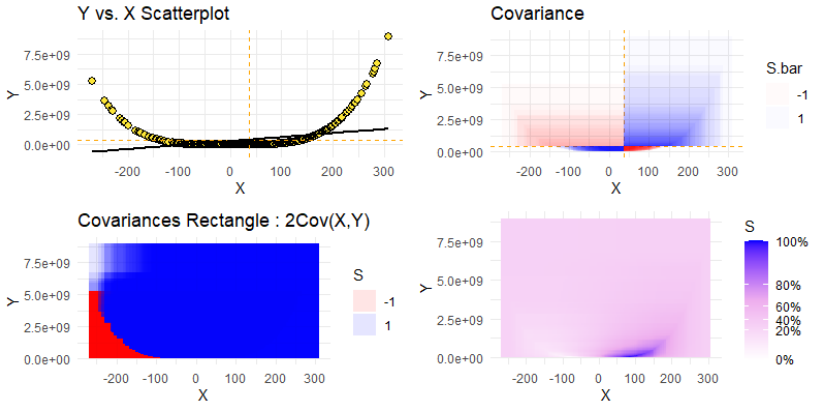
We clearly notice the heterogeneous character of Y. The Plotnetrec graphic seems to identify the split area of the groups.

4.2.2 Heteroskedasticity



The heteroskedasticity case has to be redesigned.

4.2.3 Non linear relationship



Non linear case seems very expressive in the panel of graphics beside.

Chapter 5

Linear regression and first reliability measure

5.1 Simple linear regression

5.1.1 Estimation by ordinary least squares

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where ϵ_i is the residual or deviation of y_i from the line $\beta_0 + \beta_1 x_i$.

- Ordinary least square

We need to find the values of β_0 and β_1 that minimizes the criterion:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Minimize this sum gives:

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$
$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Estimated simple linear regression model:

$$y_i = \hat{\beta_0} + \hat{\beta_1} x_i + e_i$$

From which we can calculate a few additionnal quantities:

- $\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i$; $\hat{y_i}$ is the **predicted value** (or predicted fit) of y for the i^{th} observation in the sample.
- $e_i = y_i - \hat{y_i}$; is the **observed error** (or residual) for the i^{th} observation in the sample.
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$; is the **sum of squared observed errors** for all observations in a sample of size n.

5.1.2 Measuring overall variation from the sample line

- $MSE = \frac{SSE}{n-p}$, where p is the number of parameters of the regression equation. $p = 2$ for regression with only one variable.
- $s = RMSE = \sqrt{(MSE)}$
- $SSTO = \sum_{i=1}^{n}(y_i - \bar{y_i})^2$
- Coefficient of determination:

$R^2 = \frac{SSTO-SSE}{SSTO} = \frac{SSR}{SSTO}$ is the proportion of variation in y that is explained by x.

In other words, the coefficient of determination is then the ratio of the variance explained by the SSE regression to the total SST variance.

The coefficient of determination is the square of the linear correlation coefficient R^2 between the predicted values $\hat{y}_i$ and the measurements y_i:

$$R^2 = corr(\hat{y}_i, y_i)$$

R^2 does not indicate whether:

- the independent variables are a cause of the changes in the dependent variable;
- omitted-variable bias exists;
- the correct regression was used;
- the most appropriate set of independent variables has been chosen;
- there is colinearity present in the data on the explanatory variables;
- the model might be improved by using transformed versions of the existing set of independent variables;
- there are enough data points to make a solid conclusion.

5.2 Multiple linear regression

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{1,1} + ... + \beta_p x_{1,p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{2,1} + ... + \beta_p x_{2,p} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n,1} + ... + \beta_p x_{n,p} + \varepsilon_n \end{cases}$$

We aim to determine the coefficients of this regression. We need to use matrix writting style in order to express the multiple linear regression.

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

With a stacked notation, it gives:

$$Y = \beta X + \varepsilon$$

And like simple linear regression, the ordinary least squares method search β's vector that minimize the criterion.

$$\min \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \min_{\hat{\beta}_0,..,\hat{\beta}_p} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_p x_{i,p})^2$$

And so the ordinary least squares estimators is determined by:

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

Chapter 6

Visualisation of regression and correlation beetween variable

6.1 State of the art

6.1.1 More on Venn Diagrams for Regression (figure 6.1)

Kennedy (Kennedy, 2002) extended the Venn diagram to the exposition of bias and variance in the context of the classical linear regression (CLR) model, written as $y = Xb + e$.

- Purple area: variation in y uniquely explained by variation in X
- A larger purple area means that more information is used in estimation, implying a smaller variance of the β_x estimate.
- The black area: the variation in y that cannot be explained by X

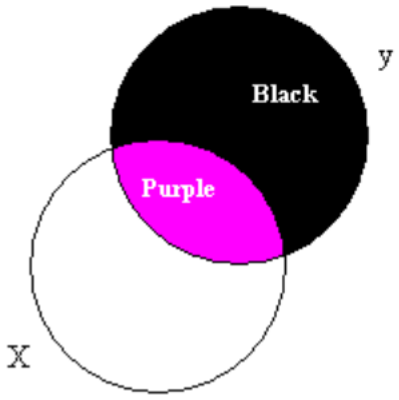


Figure 6.1: Venn Diagram in the context of classical linear regression

Another example of Venn Diagram: relating species richness to the structure of continuous landscapes (figure 6.2).

- Venn diagram representing the partition of the variance of the response variable Y (species richness) between two sets of explanatory variables, namely landscape data (a) and space (c).

- The variance jointly explained by landscape data and space is represented by (b) in the diagram. The rectangle represents the total variance in Y.

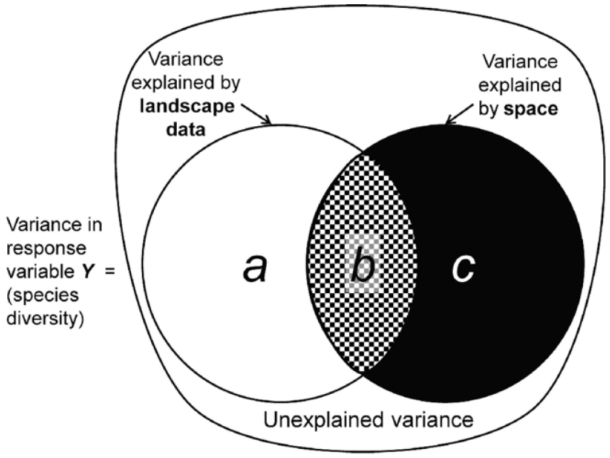


Figure 6.2: Venn Diagram relating species richness to the structure of continuous landscapes

6.1.2 A geometric approach to compare variables in a regression model

The article of Brings (Bring, 1996), proposes a geometric approache to compare variables in a regression model.

> "This article gives a brief introduction to the geometric approach in regression analysis, and then geometry is used to shed some light on the problem of comparing the"importance" of the independent variables in a multiple regression model. Even though no final answer of how to assess variable importance is given, it is still useful to illustrate the different measures geometrically to gain a better understanding of their properties".

The model vector, $\hat{y}$ is the perpendicular projection of y on x. Why it is the best estimate?

- The squared length of y, $||y||^2 = \sum y_i^2$ is the total sum of square, SS_{tot}
- The square length of $\hat{y}$, $||\hat{y}||^2 = \sum \hat{y}^2 = SS_{reg}$
- The square length of the vector $y - \hat{y}$, $||y - \hat{y}||^2 = \sum (y_i - \hat{y}_i)^2 = SS_{res}$

In other words, the observations vector is decomposed into a model vector and an error vector, and the shortest possible length of $y - \hat{y}$ is found by projecting y perpendicular in x.

- The shortest possible length of $y - \hat{y}$ is found by projecting y perpendicular on x.
- $R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{||\hat{y}||^2}{||y||^2}$ (the y vector is standardized to have lenght 1.)
-
$$r_{xy} = \begin{cases} ||\hat{y}|| & \text{if } \theta \leq 90° \quad \text{or } \theta \geq 270° \\ -||\hat{y}|| & \text{if } 90° \leq \theta \leq 270° \end{cases}$$

6.1.3 Two Additional Views of Linear Regression Coefficients

The author (Li, 1964) proposes an interesting interpretation of the slope in the least square method. The linear regression line of y on x, as determined by the method of least squares, passes through the central point with slope.

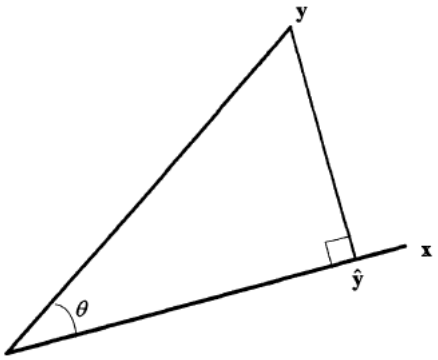


Figure 6.3: Geometrical representation of linear regression

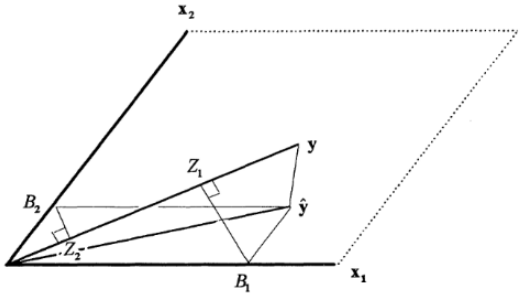


Figure 6.4: Geometrical representation of the product measure, $Z_i = B_i r_i$

A View of Linear Regression Coefficients as a weighted average slope

- Consider any point $P_i = (x_i, y_i)$ and $P_o = (\bar{x}, \bar{y})$, the slope of the line P_iP_o is: $b_i = \frac{y_i - \bar{y}}{x_i - \bar{x}}$.
- Let's attach a weight w_i to each b_i, if 2 points are very close, their slope is not very reliable (need little weight).
- The "distance" between two points $\rightarrow$ the distance projected on the x-axis. To avoid negative weight, we may then take the weight, $w_i = (x_i - \bar{x})^2$.
- Adopting this system of weighting, we see that b is the weighted mean of the b_i's.

$$b = \bar{b} = \frac{\sum b_i w_i}{\sum w_i} = \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{6.1}$$

> "This concept of weight for a slope is represented in the accompanying diagrams. (See Figure 1). The slope in the lefthand side diagram has a much larger weight than that in the righthand side for regression of y on x. If we were concerned with the regression of x on y, the reverse would be true. Note that the actual distance between the two points in the two diagrams is the same."

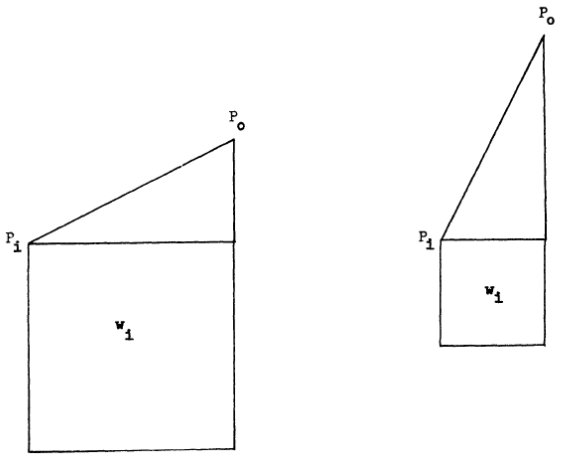


Figure 6.5: Representation of the weight equal to the square of the base, based on the slope from P_1 to P_0

6.2 Our current project, the Plotnetrec package:

6.2.1 OLS diagramatic representation

The principles of construction of the following diagram giving a diagrammatic representation of the regression and OLS are:

- We assume that the covariance (an average of all possible rectangles as seen previously among all pairs of points in a data sample) can be represented by a rectangle.
- This rectangle has by definition a known area given by the calculation of the empirical covariance $Cov(X, Y)$. But the covariance depends on three parameters since $Cov(X, Y) = \rho \sigma_x \sigma_y$. To represent the covariance and draw the corresponding rectangle, we have to make an additional assumption i.e. choose a normalization. We assume that one side of this rectangle is normalized by $sigma_x$. In this case, the other side necessarily has a length of $\rho \sigma_y$.

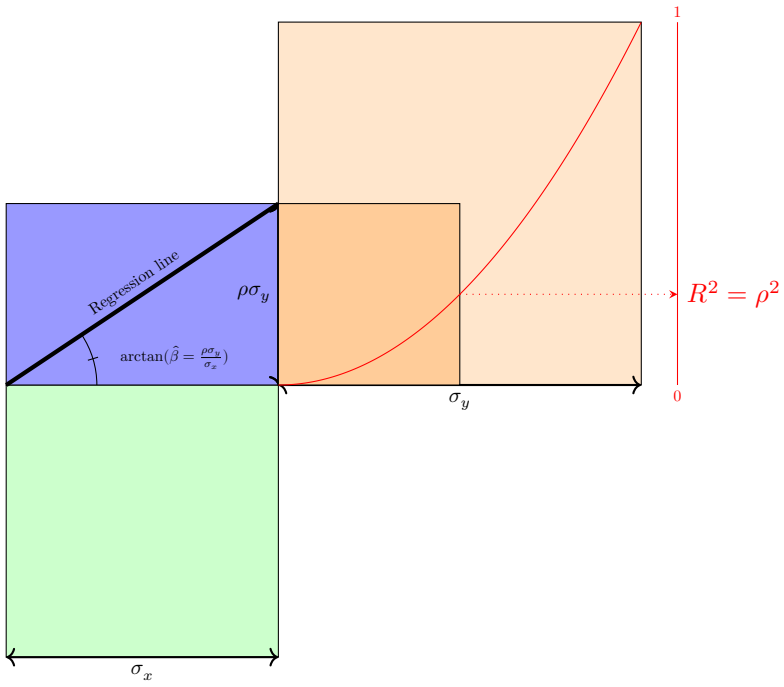
- Following the same approach, we now assume that the variances of X and Y can be represented by two squares whose sides are of course the standard deviations of X and Y respectively.

The diagram allows:

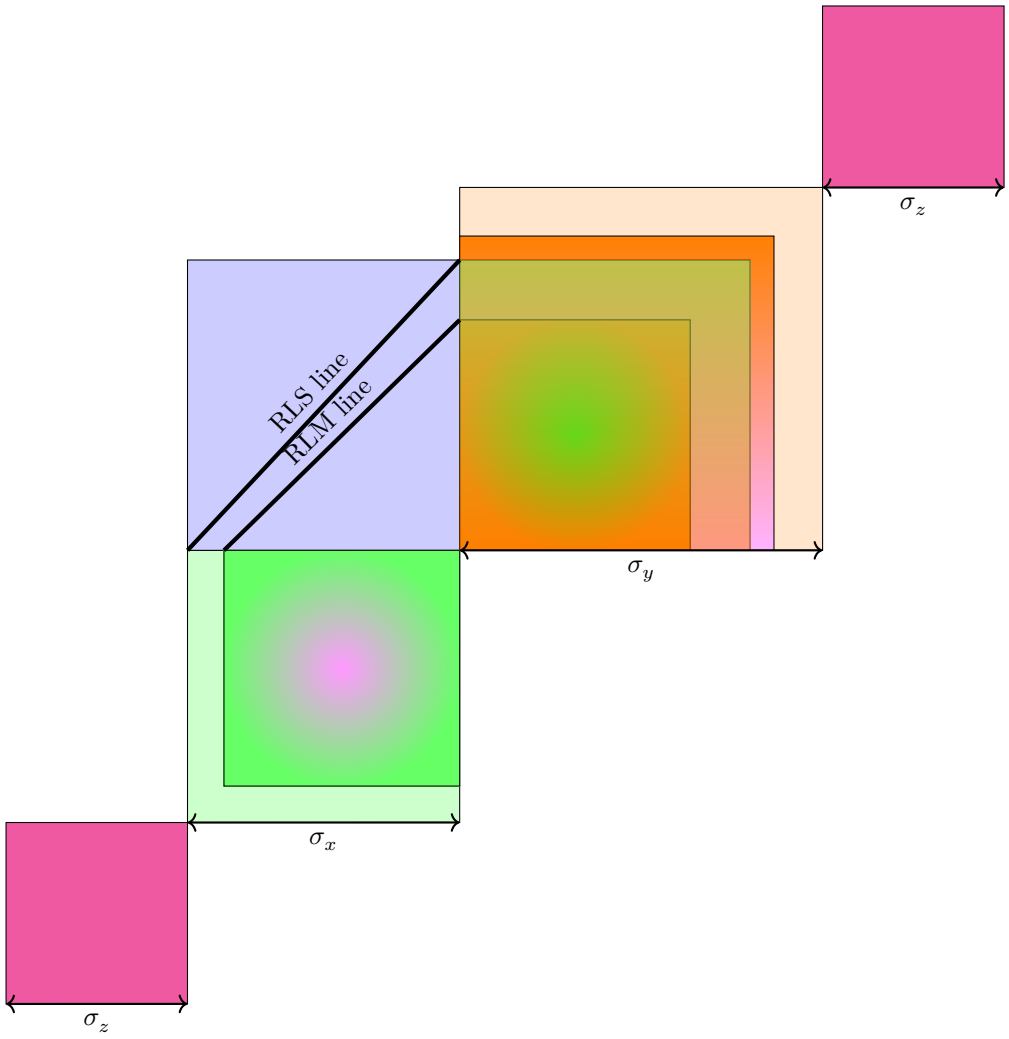
- To represent the OLS estimator by the slope of the diagonal of the rectangle representing the covariance. Geometrically, we obtain that if X increases by one standard deviation then Y increases by ρ standard deviation of the dependent variable Y.
- To represent the explained variance in the total variance.
- Correctly represent the coefficient of determination as the relative share of the explained variance in the total variance. With Venn diagrams sometimes used to represent the R^2, visual perception is based on the relative size of two surfaces and their intersection. The proposed diagram is not an ad hoc construction, it is based on a methodological and theoretical foundation allowing not only to represent the R^2 but also to read its value correctly.

Reminder:

$$\rho = \frac{COV(X,Y)}{\sigma_x \sigma_y}$$
$$\Leftrightarrow COV(X,Y) = \rho\sigma_y \times \sigma_x$$



6.2.1.1 Multiple linear regression diagram



6.2.2 Correlation representation

Keep going with alternative representations, we introduce visualization of the correlation coefficient.

First of all, we need some reminders:

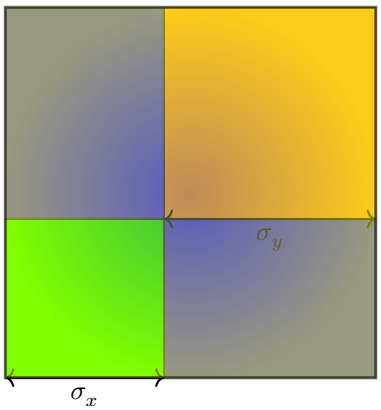
$$\rho = \frac{COV(X,Y)}{\sigma_x \sigma_y}$$

$$V(X+Y) = V(X) + V(Y) + 2COV(X,Y)$$

- Limits: $\rho = 1$

When $\rho = 1$ then $COV(X,Y) = \sigma_x \sigma_y$

By replacing, we found: $V(X+Y) = V(X) + V(Y) + 2\sigma_x \sigma_y$

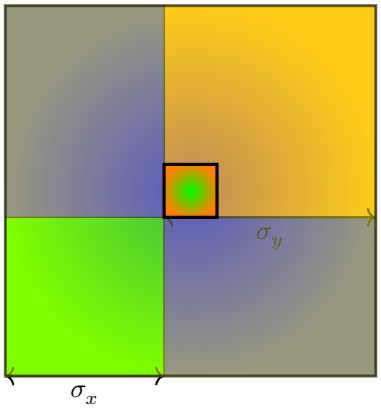


Here we decide to represent $V(X+Y)$ like a square in the middle. And due to this last equation, we represent $V(X+Y)$ as an area which equals to the sum of 4 rectangles area: $V(X)+V(Y)+2\sigma_x\sigma_y$

- Limits : $\rho = -1$

When $\rho = -1$ then $COV(X,Y) = -\sigma_x\sigma_y$

By replacing, we found: $V(X+Y) = V(X)+V(Y)-2\sigma_x\sigma_y$

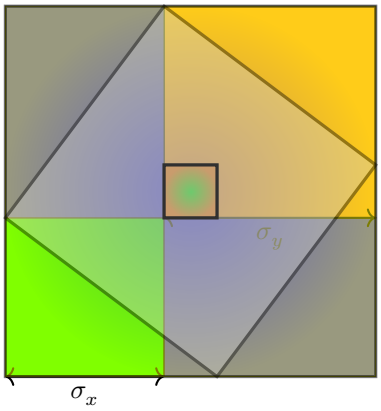


Here we decide to represent $V(X+Y)$ like a square in the middle. And due to this last equation, we represent $V(X+Y)$ as an area which equals to the sum of 4 rectangles area: $V(X)+V(Y)-2\sigma_x\sigma_y$

- Limits: $\rho = 0$

When $\rho = 0$ then $COV(X,Y) = 0$

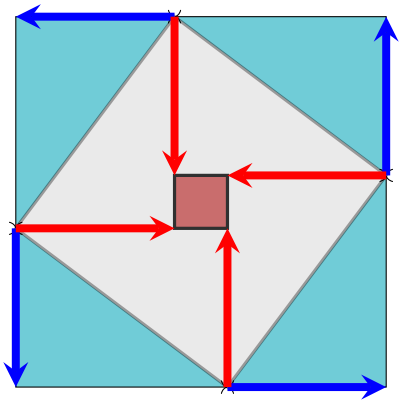
By replacing, we found: : $V(X+Y) = V(X)+V(Y)$



Here we decide to represent $V(X+Y)$ like a square in the middle. And due to this last equation, we represent $V(X+Y)$ as an area which equals to the sum of 2 rectangles area: $V(X)+V(Y)$

- Interest

Correlation matrix with distinct mathematical construction.



Let's observe what dynamically happens starting from the independant case ($\rho = 0$). If the correlation increases positively, the square increases its area and pivots toward the case where the correlation is perfectly positive. Conversely, if the correlation increases negatively, the area's square decreases and pivote towards the perfectly negative correlation place.

$$V(X+Y) = V(X) + V(Y) + 2COV(X,Y)$$

Chapter 7

Causality

In the basic model $y = \alpha + \beta x + \varepsilon$ the OLS estimator $\widehat{\beta}$ is unbiased if x is independent of ε. We cannot know a priori if this is the case and in any case not a posteriori at the end of the estimation since the residuals are by construction independent of X. We can easily show that there is no unique solution in $\widehat{\beta}$ of the maximum likelihood when we increase the model by a free parameter representing the correlation between ε and x. The prediction $\mathbb{E}(y\ /x) = \widehat{\beta}x$ is correct only if $\mathbb{E}(\varepsilon\ /x) = 0$. Otherwise $\mathbb{E}(\varepsilon\ /x) = 0$ there is an endogeneity and $\widehat{\beta}$ is not an exact measure of the marginal effect of an exogenous variation of x. The main situations of endogeneity encountered and studied are related to the problems:

- Simultaneous causality
- Omitted variables
- Measurement errors on the variables
- Selection bias
- Poorly specified functional form

7.1 The main cases

7.1.1 Simultaneous causality

Is it enough to simply show by a diagram that regressing y on x and then x on y does not give the same result? The answer is no, it is not enough!

7.1.2 Omitted variables

.... Repeat the previous results in a simplified way to illustrate the effect of omitting a variable and present the conditions under which the omission of a variable leads to a bias on the model parameters.

7.1.3 Measurement errors

Let's consider the following RLS model

$$y^\star = \alpha + \beta x^\star + \varepsilon$$

We do not observe $y^\star$ (but $y = y^\star + v_y$), nor $x^\star$ (but $x = x^\star + u_x$).

Consider the following assumptions:

$\mathbb{E}(u_x) = \mathbb{E}(v_y) = 0$

$\mathbb{E}(yv_y) = \mathbb{E}(yu_x) = \mathbb{E}(xu_y) = \mathbb{E}(xv_x) = 0$

$\mathbb{E}(\varepsilon u_y) = 0$

The measurement errors are of zero expectation, and independent of the variables and the error term. Under these assumptions, for two randomly chosen points k and l, the covariance between X and Y does not change.

$$\mathbb{E}[(x_k - x_l)(y_k - y_l)]$$

$$\mathbb{E}[(x_k^\star + u_k - (x_l^\star + u_l))(y_k^\star + v_k - (y_l^\star + v_l))]$$

$$\mathbb{E}[(x_k^\star + u_k)(y_k^\star + v_k)] - \mathbb{E}[x_k^\star + u_k]\mathbb{E}[y_l^\star + v_l] - \mathbb{E}[x_l^\star + u_l]\mathbb{E}[y_k^\star + v_k] + \mathbb{E}[(x_l^\star + u_l)(y_l^\star + v_l)]$$

$$2Cov(x^\star, y^\star)$$

Measurement errors are detrimental to accuracy but on average they do not affect the covariance. The only essential problem is therefore related to the correct estimation of the parameters, while the correlation between the dependent and independent variables does not change. On the other hand, one should not confuse the effect of x on y with the effect of the measurement error of x on y (which is zero).

7.2 Measurement error on the dependent variable

Let us consider a measurement error on the dependent variable y, i.e. let $\sigma_v^2 > 0$ and $\sigma_u^2 = 0$. The model becomes:

$$y = alpha + betax + v + \varepsilon$$

The measurement error on the dependent variable does not matter in the sense that it has no biasing effect on the estimated parameters of the model. In practice, it can be considered as contributing to the disturbance term of the model. It is obviously undesirable, as anything that increases the noise in the model will tend to make the regression estimates less accurate, but it has no impact in terms of bias in the estimates.

In this case, we do not make a diagram for a graphical proof because it is so obvious!

7.3 Measurement error on the independent variable

We consider the case where the measurement error is on the dependent variable x, i.e. Let $sigma_v^2 = 0$ and $sigma_u^2 > 0$. The model becomes:

$$y = alpha + betax - betau + \varepsilon$$

In the regression of y on x, the measurement error of x becomes part of the error in the regression equation related to the parameter β, thus creating an endogeneity bias. The OLS estimator $\widehat{\beta} = \frac{Cov(x,y)}{Var(x)}$ is biased towards 0. The unbiased estimator is $\beta = fracCov(x^\star, y)Var(x^\star)$.

Proof:

Since $Cov(x^\star, y) = Cov(x, y)$, we have:

$$\frac{\hat{\beta}}{\beta} = \frac{Var(x^\star)}{Var(x)} = \frac{\sigma_x^{\star 2}}{\sigma_u^2 + \sigma_x^{\star 2}} \implies plim\ \hat{\beta} = \lambda\beta$$

With $\lambda = \frac{\sigma_x^{\star 2}}{\sigma_u^2 + \sigma_x^{\star 2}}$. Since $0 < lambda < 1$ the coefficient $\hat{\beta}$ is biased towards 0. The bias depends on the level and the sign of β:

$$plim\ \hat{\beta} - \beta = \lambda\beta - \beta = -(1-\lambda)\beta = -\frac{\sigma_u^2}{\sigma_u^2 + \sigma_x^{\star 2}}\beta$$

7.3.1 Diagram of the model with measurement error on the independent variable

The following figure illustrates the problem in a very simple way.

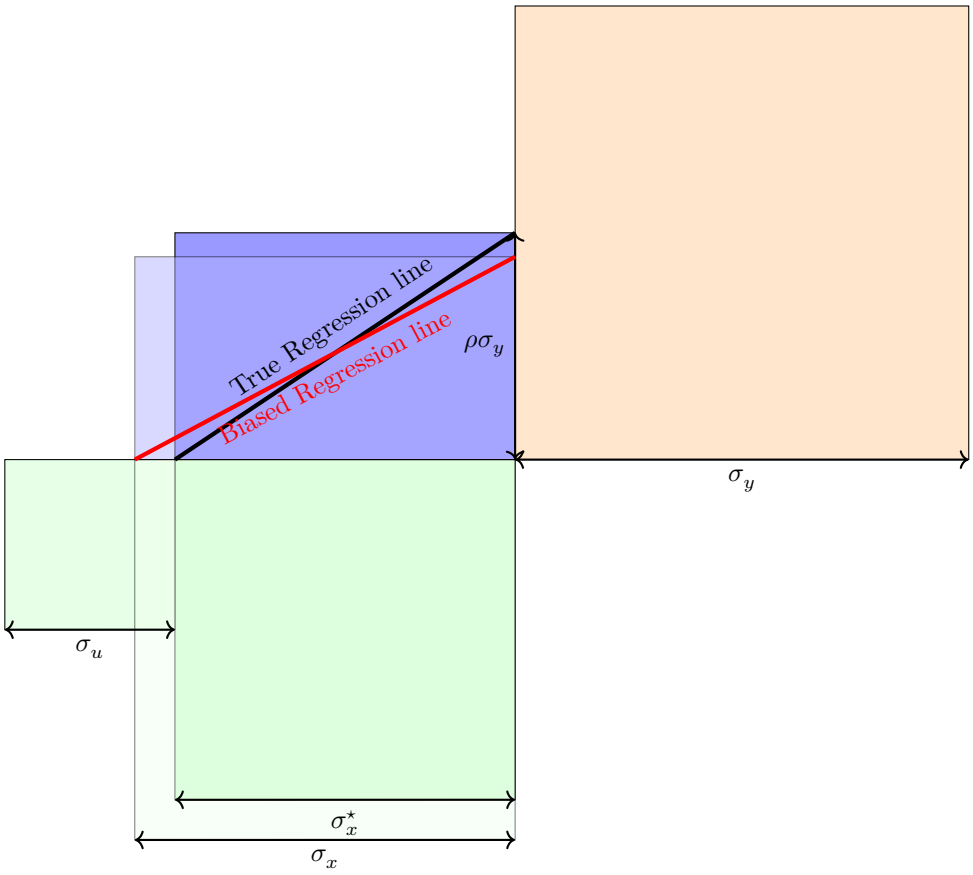


Figure 7.1: Regression of y on x with measurement error on X

7.4 Instrumental variables

$$Y = \alpha + \beta X + u$$
$$X = \theta + \delta Z + v$$
$$\beta_{IV} = \frac{Cov(Y, Z)}{Cov(X, Z)} = \frac{\rho_{zy}\sigma_y}{\rho_{zx}\sigma_x}$$

Chapter 8

State of the art of R packages in visualization econometrics (and data-viz more broadly)

- R energy

A compléter

- viscov

Our visualizations follow the principle of decomposing a covariance matrix into scale parameters and correlations, pulling out marginal summaries where possible and using two and three-dimensional plots to reveal multivariate structure. Visualizing a distribution of covariance matrices is a step beyond visualizing a single covariance matrix or a single multivariate dataset.

- Colourpicker

Colourpicker is a tool for Shiny framework and for selecting colours in plots. This tool supports various options, such as alpha opacity, custom colour palettes, and more. The most common uses of this tool include the utilisation of the colourInput() function to create a colour input in Shiny as well as the use of the plotHelper() function/RStudio Addin to select colours for a plot.

- Esquisse

The esquisse package allows a user to interactively explore data by visualising it with the ggplot2 package. It allows a user to draw bar graphs, curves, scatter plots, histograms, export the graphs, and retrieve the code generating the graph. With the help of esquisse, one can quickly visualise the data according to their type as well as export to PNG or PowerPoint, and retrieve the code to reproduce the chart.

- ggplot2

ggplot is a popular package that is based on the grammar of graphics. The idea behind this library is that one can build every graph from the same components, such as a dataset, a coordinate system, and more. The package provides graphics language for creating intuitive and intricate plots. It allows a user to create graphs that represent both univariate and multivariate numerical and categorical data.

- ggvis

ggvis is a data visualisation package for R that allows to declaratively describe data graphics with a syntax similar in spirit to ggplot2. It allows creating rich interactive graphics locally in Rstudio or in the browser as well as leverage the infrastructure of the Shiny package to publish interactive graphics usable from any browser. The goal of ggvis is to make it easy to build interactive graphics for exploratory data analysis.

- ggforce

The ggforce is a package aimed at providing missing functionality to ggplot2 through the extension system introduced with ggplot2 v2.0.0. The goal of this package is to provide a repository of geoms, stats, among others. Using ggforce, one can enhance almost any ggplot by highlighting data groupings and focusing attention on interesting features of the plot.

- lattice

Lattice is a powerful high-level data visualisation system for R that is designed with an emphasis on multivariate data and allows to create multiple small plots easily. The lattice package attempts to improve on base R graphics by providing better defaults and the ability to display multivariate relationships easily. Particularly, the package supports the creation of trellis graphs that show a variable or the relationship between variables, conditioned on one or more other variables.

- Plotly

Plotly is an open-source R package for creating interactive web-based graphs via the open-source JavaScript graphing library plotly.js. The Plotly's R graphing library helps in creating interactive, publication-quality graphs including line plots, scatter plots, area charts, bar charts, error bars, etc. One can use Plotly for R to make, view and distribute charts and maps online as well as offline.

- patchwork

patchwork is a package that expands the API to allow for the arbitrarily complex composition of plots by providing mathematical operators for combining multiple plots. The goal of patchwork is to make it simple to incorporate separate ggplots into the same graphic.

- quantmod

quantmod is an R package that provides a framework for quantitative financial modelling and trading. It provides a rapid prototyping environment that makes modelling easier by removing the repetitive workflow issues surrounding data management and visualisation.

- RGL

The RGL package is used to produce interactive 3-D plots using OpenGL. The library contains high-level graphics commands modelled loosely after classic R graphics and working in three dimensions. It also includes a low-level structure inspired by the grid package. RGL provides medium to high-level functions for 3D interactive graphics, including functions modelled on base graphics as well as functions for constructing representations of geometric objects.

- Highcharter

is an R wrapper for Highcharts, an interactive visualization library in JavaScript. Like its predecessor, highcharter features a powerful API.

Highcharter makes dynamic charting easy. It uses a single function, hchart(), to draw plots for all kinds of R object classes, from data frame to dendrogram to phylo. It also gives R coders a handy way to access the other popular Highcharts plot types, Highstock (for financial charting) and Highmaps (for schematic maps in web-based projects).

- Leaflet

Like highcharter, Leaflet for R is another charting packaged based on a hugely-popular JavaScript library of the same name.

Leaflet offers a lightweight but powerful way to build interactive maps, which you've probably seen in action (in their JS form) on sites ranging from The New York Times and The Washington Post to GitHub and GIS specialists like Mapbox and CartoDB.

The R interface for Leaflet was developed using the htmlwidgets framework, which makes it easy to control and integrate Leaflet maps right in R Markdown documents (v2), RStudio, or Shiny apps.

- RcolorBrewer

RColorBrewer makes it easy to take advantage of one of R's great strengths: manipulating colors in plots, graphs, and maps.

The package is based on Cynthia Brewer's work on the use of color in cartography (check out Colorbrewer to learn more), and it lets you create nice-looking sequential, diverging, or qualitative color palettes. It also plays nicely with Plotly, as these examples by Plotly demonstrate.

- dygraphs

This package provides an R interface for dygraphs, a fast, flexible JavaScript charting library for exploring time-series data sets. What's powerful about dygraphs is that it's interactive right out of the box, with default mouse-over labels, zooming, and panning. It's got lots of nifty other interactivity features, like synchronization or the range selector shown above.

But dygraph's interactivity doesn't come at the expense of speed: it can handle huge datasets with millions of points without slowing its roll. And you can use RColorBrewer with dygraphs to choose a different color palette for your time series— check out this example to see how.

- sunburst

pas forcément utile pour notre sujet

- VennDiagram

Create a Venn diagram and save it into a file. The function venn.diagram() takes a list and creates a file containing a publication-quality Venn Diagram.

Chapter 9

annexes

9.1 Correction of Bessel's proof

$$
\begin{aligned}
(n-1)S_{xy} &= \sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \\
&= \sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y} \\
&= \sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i \\
(N-1)\mathbb{E}(S_{xy}) &= \mathbb{E}\left(\sum_{i=1}^{N} x_i y_i\right) - \frac{1}{N}\mathbb{E}\left(\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i\right) \\
&= N\mu_{xy} - \frac{1}{N}[N\mu_{xy} + N(N-1)\mu_x\mu_y] \\
&= (N-1)[\mu_{xy} - \mu_x\mu_y] \\
&= (N-1)Cov(x,\ y)
\end{aligned}
$$

9.2 Proof for $Cov(x,\ y) = \frac{1}{2N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\ (x_i - x_j)(y_i - y_j)$

- Proof 1

$$Cov(x,\ y) = \frac{1}{2N(N-1)}\sum_{i=1}^{N}\sum_{j=1}^{N}\ (x_i - x_j)(y_i - y_j) =$$

$$\frac{1}{2N(N-1)}\sum_{i=1}^{N} N\ x_i\ y_i - x_i\ N\ \overline{y} - y_i\ N\ \overline{x} + N\ \overline{xy}$$

$$\frac{2N^2}{2N(N-1)}(\overline{xy} - \overline{x}\ \overline{y}) =$$

$$\frac{N}{N-1}(\overline{xy} - \overline{x}\ \overline{y}))$$

- Proof 2

another way, not uninteresting, to show the equivalence between the two formulations of the covariance but starting from the usual form.

$$Cov(x,\ y) = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})$$

$$Cov(x,\ y) = \frac{1}{2(N-1)} \left\{ \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}) + \sum_{j=1}^{N} (x_j - \overline{x})(y_j - \overline{y}) \right\}$$

$$Cov(x,\ y) = \frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\{ (x_i - \overline{x})(y_i - \overline{y}) + (x_j - \overline{x})(y_j - \overline{y}) \right\}$$

$$Cov(x,\ y) = \frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\{ (x_i - x_j)(y_i - y_j) + x_i y_j + x_j y_i - x_i \overline{y} - x_j \overline{y} - y_i \overline{x} - y_j \overline{x} + 2\ \overline{x}\ \overline{y} \right\}$$

$$Cov(x,\ y) = \frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\{ (x_i - x_j)(y_i - y_j) \right\}$$

9.3 a little bit of calculation

9.3.0.1 Variance of the sum of two random variables

$$\begin{aligned} Var(X+Y) &= Cov(X+Y, X+Y) \\ &= E((X+Y)^2) - E(X+Y)E(X+Y) \end{aligned}$$

en développant,

$$\begin{aligned} &= E(X^2) - (E(X))^2 + E(Y^2) - (E(Y))^2 + 2(E(XY) - E(X)E(Y)) \\ &= Var(X) + Var(Y) + 2(E(XY)) - E(X)E(Y)) \\ &= Var(X) + Var(Y) + 2Cov(X,\ Y) \end{aligned}$$

If the variables $X,\ Y$ are independent $E(XY) = E(X)E(Y)$ then $Var(X+Y) = Var(X) + Var(Y)$

$$\begin{aligned} \Delta &= 4Cov(X,\ Y)^2 - 4Var(X)Var(Y) \\ &= 4[Cov(X,\ Y)^2 - Var(X)Var(Y)] \leq 0 \end{aligned}$$

So $ Cov(X,Y)^2 Var(X) Var(Y)$ (In probability theory, the Cauchy-Schwarz inequality allows the proof of the result because of the inequality $E(XY) \leq \sqrt{E(X^2)E(Y^2)}$) or $-(X,Y) _x _y _x _y $. Let us notice that if $\Delta = 0$ then we have the equality $ Cov(X,Y)^2=Var(X) Var(Y)$, i.e. if there exists λ such that $Var(\lambda X+Y) = 0\backslash$ But to conclude, if we have $lambdaX+Y$ equal with probability 1 to a constant, let's say c, it is indeed that $Y = c - lambdaX$ almost surely, in other words a perfect linear link between the two variables.

9.3.0.2 Definition of the linear correlation coefficient (Bravais-Pearson)

- case $Cov(X,\ Y) \geq 0$

We define $\rho = \frac{2Cov(X,\ Y)}{Var_{max} - Var_{min}}$

$Var_{max} = \sigma_X^2 + \sigma_Y^2 + 2\sigma_X\sigma_Y$

$Var_{min} = \sigma_X^2 + \sigma_Y^2$

$Var_{max} - Var_{minx} = 2\sigma_X\sigma_Y$

From which, simplifying

$\rho = \frac{Cov(X,\ Y)}{\sigma_X\sigma_Y}$

- case $Cov(X,\ Y) \leq 0$

We define $\rho = \frac{2Cov(X,\ Y)}{Var_{max} - Var_{min}}$

$Var_{max} = \sigma_X^2 + \sigma_Y^2$

$Var_{min} = \sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y$

$Var_{max} - Var_{minx} = 2\sigma_X\sigma_Y$

from which, simplifying

$\rho = \frac{Cov(X,\ Y)}{\sigma_X\sigma_Y}$

9.3.0.3 Covariance framing

$$-\sqrt{(Var(X)Var(Y)} \leq Cov(X,Y) \leq \sqrt{(Var(X)Var(Y)}$$

With in the particular case of standardized variables

$$\begin{aligned} Var(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}) &= Var(\frac{X}{\sigma_X}) + Var(\frac{Y}{\sigma_Y}) + 2Cov(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}) \\ &= 2(1+\rho) \\ Var(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}) &= Var(\frac{X}{\sigma_X}) + Var(\frac{Y}{\sigma_Y}) - 2Cov(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}) \\ &= 2(1-\rho) \end{aligned}$$

One finds, knowing that $2(1+\rho) \geq 0$ and $2(1-\rho) \geq 0$ that one has $-1 \leq \rho \leq 1$

Bibliography

Bring, J. (1996). *Visualizing Distributions of Covariance Matrices*.

Chudzicki, D. (2014). *Covariance As Signed Area Of Rectangles*.

Galton, F. (1888). *Co-relation and their measurements*.

Hayes, K. (2011). *Geometrical Interpretation of an Alternative Formula for the Sample Covariance*.

Heffernan, P. M. (1988). *New Measures of Spread and a Simpler Formula for the Normal Distribution*.

Kennedy, P. E. (2002). *More on Venn Diagrams for Regression*. journal of statistics Education Vol 10, Number 1, Simon Fraser University, Canada.

Li, C. C. (1964). *Two Additional Views of Linear Regression Coefficients*.

Makowski, D. and co (2020). Methods and algorithms for correlation analysis in r. *Journal of Open Source Software*, 5(51):2306.

Rodgers, J. L. and co (1988). *Thirteen Ways to Look at the Correlation Coefficient*.

Tokudaa, T. and co (2011). *Visualizing Distributions of Covariance Matrices*. University of Leuven, Belgium, Columbia University, New York, NY, USA.