Econometrics vizualisation

Lucas Chaveneau, Thibault Fuchez, Allan Guichard

2021-11-04

Contents

Chapter 1

Introduction

Nous avons comme projet de proposer une librairie développé sous R qui propose différents moyens de visualiser les éléments essentiels de l'économètre.

L'économétrie est une branche de la science économique qui a pour but de d'estimer et de tester les modèles économiques (représentation simplifiée de la réalité). Ainsi l'économètre essaie d'identifier les paramètres d'un modèle à l'aide d'estimation statistique, il cherche donc à induire les caractéristiques d'un groupe général (la population) à partir de celles d'un groupe particulier (l'échantillon).

Trois mots essentiels dans le langage de l'économètre sont: la corrélation, la regression, et la causalité.

Afin de mieux comprendre les liens entres différentes variables, la visualisation est un outil clé qui aide à interpréter les résultats mathématiques et statistiques.

Notre package s'intéressera donc à la représentation des corrélations. Nous essaierons de mettre en avant une représentation claire et synthétique de la covariance, ainsi qu'une représentation des éléments qui permettent de visualiser la qualité d'une regression ainsi que la part de chaque variable dans l'explication de la variable à prédire.

Ce document a pour vocation d'accompagner le package. Il propose une définition des outils économétrique nécessaires à la compréhension du package. Il propose également un état des lieux des différentes techniques de visualisation qui existent déjà.

Chapter 2

La covariance

2.1 Rappel de la définition de la variance :

En statistique et en théorie des probabilités, la variance est une mesure de la dispersion des valeurs d'un échantillon ou d'une distribution de probabilité. Elle exprime la moyenne des carrés des écarts à la moyenne, aussi égale à la différence entre la moyenne des carrés des valeurs de la variable et le carré de la moyenne, selon le théorème de König-Huygens.

- 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):

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

- Variance vs. 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 Définition littéraire usuelle de la covariance :

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

La covariance est une extension de la notion de variance. La covariance entre deux variables aléatoires est un nombre permettant de quantifier leurs écarts conjoints par rapport à leurs espérances respectives.

Intuitivement, la covariance est une mesure de la variation simultanée de deux variables aléatoires. C’est-à-dire que la covariance devient plus positive pour chaque couple de valeurs qui diffèrent de leur moyenne dans le même sens, et plus négative pour chaque couple de valeurs qui diffèrent de leur moyenne dans le sens opposé.

La covariance de deux variables aléatoires indépendantes est nulle, bien que la réciproque ne soit pas toujours vraie.

Ce concept se généralise naturellement à plusieurs variables (vecteur aléatoire) par la matrice de covariance (ou matrice de variance-covariance) qui, pour un ensemble de p variables aléatoires réelles X_1, etc.,Xp est la matrice carrée dont l’élément de la ligne i et de la colonne j est la covariance des variables X_i et X_j. Cette matrice permet de quantifier la variation de chaque variable par rapport à chacune des autres.

“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”.

2.3 Définitions mathématique usuelle et alternatives

- covariance formula :

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

- formula from heffernan definition of covariance :

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

- lister autres ...

Chapter 3

La corrélation

3.1 de la covariance à la corrélation

La covariance de deux variables aléatoires indépendantes est nulle, bien que la réciproque ne soit pas toujours vraie.

La forme normalisée de la matrice de covariance est la matrice de corrélation.

"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.
- Covariance indicates the direction of the linear relationship between variables.
- Correlation measures both the strength and direction of the linear relationship between two variables.
- Correlation values are standardized.
- Covariance values are not standardized.

3.2 la corrélation :

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.

3.3 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.

3.4 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.

3.5 Corrélation partielle

Le coefficient de corrélation partielle, noté ici $r_{AB.C}$, permet de connaître la valeur de la corrélation entre deux variables A et B, si la variable C était demeurée constante pour la série d'observations considérées.

Dit autrement, le coefficient de corrélation partielle $r_{AB.C}$ est le coefficient de corrélation totale entre les variables A et B quand on leur a retiré leur meilleure explication linéaire en termes de C. Il est donné par la formule :

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

Chapter 4

Vizualization of covariance and correlation

4.1 Différentes tentatives pour représenter la covariance

4.1.1 A Geometrical Interpretation of an Alternative Formula for the Sample Covariance

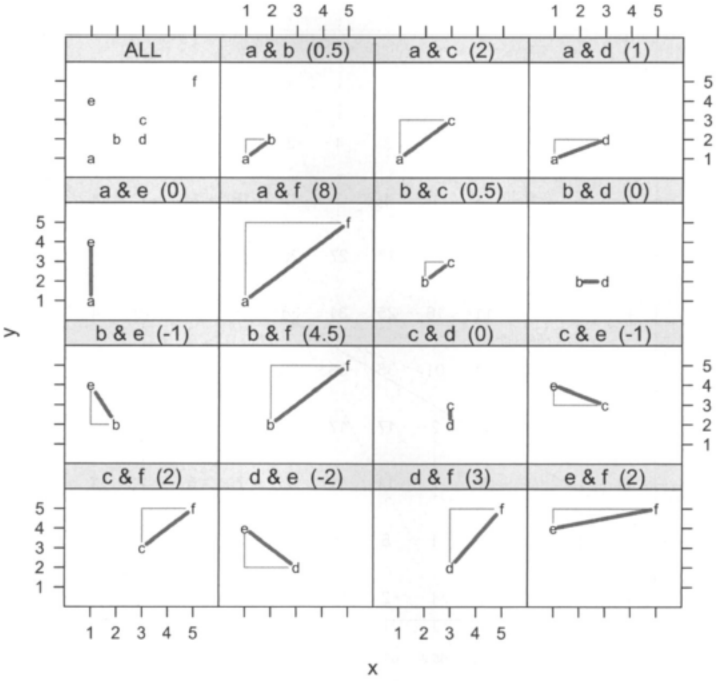
Kevin Hayes (Hayes, 2011) propose une nouvelle interprétation géométrique et visuelle de la covaraiance, à partir de l'application de la formule de la variance poposée par Hefferman. Il étend cette formule à la covariance d'un échantillon pour extraire ses résultats.

- 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)

A détailler et reformuler



4.1.2 Covariance as Signed Area of Rectangles :

Cet article (Chudzicki, 2014) a été écrit suite à une conversation trrès intéressante sur le site **stats.statckexchange** (https://stats.stackexchange.com/questions/18058/how-would-you-explain-covariance-to-someone-who-understands-only-the-mean.) Le sujet initial de cette conversation était : comment expliquer la covariance à quelqu'un qui ne comprends que la notion de moyenne?

mode d'emploi :

- 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.

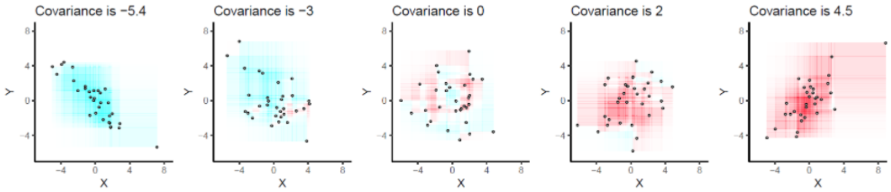
Covariance increases as the points approximate an upward sloping line and decreases as the points approximate a downward sloping line. This is because in the former case most of the rectangles are positive and in the latter case, most are negative.

- 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."



4.2 Notre projet : le package plotnetrec

Une part du package que nous en sommes en train de développer est de présenter différentes représentations de la covariance.

mettre schémas de représentation covariance issu de pltnetrec avec explications

Chapter 5

Diagramme de régression et représentation du coefficient de détermination

5.1 OLS deux variables

- quelques notations pouvant être utiles:

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

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

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

- simple linear regression model

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

where ϵ_i is te error 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 minimize 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

5.2 Measuring overall variation from the sample line

- $MSE = \frac{SSE}{n-p}$, where p is the number of parametrers 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.

c'est-à-dire que le coefficient de détermination est alors le rapport de la variance expliquée par la régression SSE sur la variance totale SST.

Le coefficient de détermination est le carré du coefficient de corrélation linéaire R entre les valeurs prédites $\hat{y_i}$ et les mesures 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 collinearity 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.

5.3 OLS 3 variables

5.4 OLS 3 variables et plus

Chapter 6

Reoprésentation de la régression linéaire

6.1 Etat des lieux de la visualisation de la régression

6.1.1 More on Venn Diagrams for Regression

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 .

6.2 Notre projet: le package plotnetrec

Chapter 7

Etat des lieux des package R en économétrie de la visualisation

- 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.

Bibliography

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

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

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

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