Comparing correlation coefficient methods with R : Is there really a relationship between Statistics and Timothee ?

Thuy-My Ngo[*] Doriane Pfister[†] Léonie Lefebvre[‡]

May 23, 2021

Introduction

As an apogee of the introduction to statistics course and our final moments of being a first year undergraduate in International Relations, this project reflects all the knowledge acquired throughout the past few months attending this introductory class. Beginning with no history in programming nor experience in writing a scientific report on Rmarkdown, we were able to learn and develop step by step, our insight on the 'magic' of Data Analytics.

We are grateful to have had the chance to receive a rather large glance at basic R commands and to put into practice the notions seen during classe. Today, we are pleased to lead you through the outcome resulting from our small but still progressing abilities in statistics.

Description of the task

In question 1 of exercise 2, our topic is related to bias study, nonlinear relationship. We are asked to draw samples from a PDF for (X,Y), where (X, Y) have a nonlinear relationship, using the function gen_nonlinear. Using the parameters assigned to us as well as an angle parameter, we had to explain how the angle parameter visually affect the data by comparing two scatterplots. The first generated with angle = 0, and the second generated with angle = -0.45.

In question 2 of the same exercise, we study the Confidence Interval Coverage. We are given the task to study the coverage of different confidence intervals (CIs) for Px,y, with the following data settings: 1. the PDF of (X,Y) is a bivariate normal PDF 2. the PDF of (X,Y) is a bivariate normal PDF, but the observed sample contains outliers 3. the PDF of (X,Y) is a discrete PDF 4. X and Y have a nonlinear relationship

The types of confidence intervals used are the parametric bootstrap CI of level = 0.8 and the non-parametric bootstrap CI of level = 0.8.

[*] University of Geneva, ID 20-312-468

[†] University of Geneva, ID 20-325-122

[‡] University of Geneva, ID 20-302-063

Motivation

We decided to take part in this scientific report as we wanted to apply the numerous theoretical theories seen in the course provided by Profesor Victoria-Feser. We also saw this project as an opportunity to meet new people and experience the workgroup process, as we combine our knowledge and perspectives on the matter, while being able to discuss them within a group.

Analysis

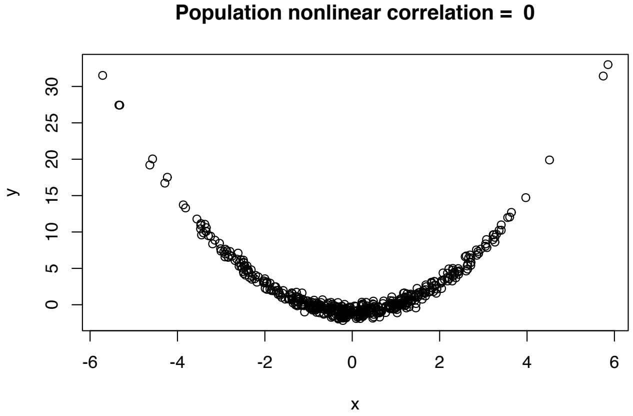

Figure 1: Hammock, this is the population correlation with angle = 0

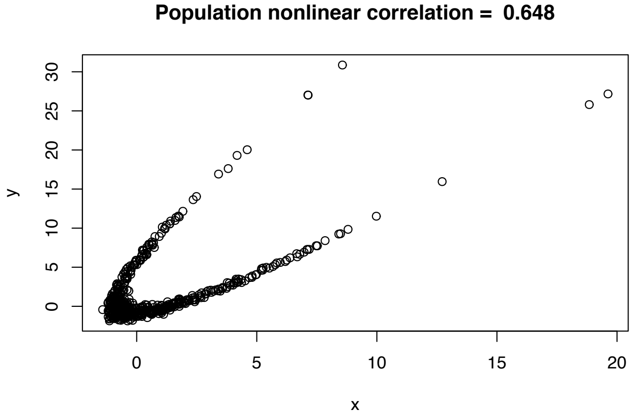

Figure 2: Comet, this is the population correlation with angle = -0.45

Here we see the two scatterplots of when the data is generated with angle = 0 (Hammock), and when it is generated with angle = -0.45 (Comet) from the samples from a PDF of (X, Y), where (X,Y) have a nonlinear relationship.

Hammock describes a positive relationship between X and Y as the graph shows a convex parabola (xˆ2). As we generate the data with an angle parameter of -0.45, we see that there is a rotation to the left. Thus, we deduce that the angle does visually affect the data, in that it imposes a rotation.

Boot Type	Sample Size	Normal PDF	With outliers	Discrete	Non Linear
Parametic	33	0.771	0.639	0.776	0.681
	110	**0.783**	0.257	0.776	0.630
Non-Parametric	33	0.758	0.758	0.753	0.725
	110	0.773	0.431	0.769	0.723

Furthermore, when the angle moves from $-\pi$ to $+\pi$, we observed a complete rotation of the parabola.

Results and discussion: description and interpretation of the results

The following discussion will be based on the results displayed in the Table. The use of colors to refer to the data is employed for the reader's convenience.

We observe a decrease in the interval confidence when the parametric boot is performed on a non-parametric functions like the discrete and nonlinear ones (blue). Furthermore, we expect the non-parametric bootstrap to perform better on a discrete PDF, as it is not a parametric function.

We assumed, that the bivariate normal PDF worked the best at approaching the confidence interval. We can see that 0.783 (bold, green) is the closest to the 0.8, particularly when operating with the parametric boot type.

We observe that the boot type parameter works the best when applied on a normal PDF (green data). As the boot type parameter is based on the mean and standard deviation, it is the most optimal approach to compute the normal PDF's confidence interval as the latter contains these two parameters. As for the non-parametric boot, it is the most optimal when applied on a no parameter function like the nonlinear function. It is displayed by the red data, which shows a greater confidence interval than the blue data, computed with a parametric boot.

Then with the outliers setting, we see that as the n is larger, the number of outliers increases (orange), thus making the result further than the confidence interval. The most surprising part of the table, is that the difference between the outliers setting when n = 33 and when n = 110, using the parametric boot type is quite large compared to when the data is performed with the non-parametric boot type.

Statistical methods used

The Pearson method is used to calculate a correlation coefficient that describes the strength relationship between two variables (Kenton (2021)), in our case X and Y, which ranges from 0 to 1. The closer the correlation coefficient is to 0, the more the variables are independent from each other. When equal to 1, the variables are equal to each other. The Pearson method can be used on empirical observations in daily life ("A Comparison of the Pearson and Spearman Correlation Methods" 2019). For instance, if we wanted to evaluate whether there is a positive or negative relationship between a student's age and their level of income when working at Starbucks.

The Spearman rank correlation method is the non-parametric version of the Pearson correlation coefficient method ("Spearman's Rank-Order Correlation" 2018). It is distinct in that the Spearman correlation coefficient is used to determine the strength of a monotonic relationship, which is a type of function that only increases or only decreases. In real life, we can imagine that the Spearman method would be used to see whether an Unige exam performance in Statistics is associated with the time students spent studying in the UniMail library.

Acquired skills during the term project

By accomplishing this report, we had an opportunity to broaden our understanding of the R language by computing various aspects of the topics seen in class. Despite a short learning period time, we are now able to compute various functions on R and interpret R codes. Furthermore, learning to code has opened new perspectives as well as revealed numerous ways of interpreting datasets. We will be able to apply these perspectives to diverse daily socio-political situations. For these reasons, the present report benefitted us greatly to acquire skills that will make our International Relations studies more valuable.

Conclusion

In conclusion, in the Nonlinear Relationship Study, we see that the angle parameter does affect the data by imposing a rotation, moving the parabola. As for the Confidence Interval Coverage Study, the Confidence Interval of non-parametric boot type and of parametric boot type, results in different outcomes depending on the setting and the sample size. Overall, most of the outcomes are close to the confidence interval with the highest value at 0.783 (bold, green), as one would expect.

Figure 3: Us three celebrating the finished report ("Giphy" 2015)

Use your mobile phone to scan the QR code below for a refreshing surprise!

Figure 4: Surprise!

Bibliography

"A Comparison of the Pearson and Spearman Correlation Methods." 2019. *Minitab Express*. https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression /supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/.

"Giphy." 2015. giphy.com. https://media.giphy.com/media/11sBLVxNs7v6WA/giphy.gif.

Kenton, Will. 2021. "What Is the Pearson Coefficient?" *Investopedia*. Investopedia. https://www.investopedia.com/terms/p/pearsoncoefficient.asp.

"Spearman's Rank-Order Correlation." 2018. Lund Research Ldt. https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide .php.