Journal of Statistical Theory and Applications (2024) 23:44–66
https://doi.org/10.1007/s44199-024-00072-0

RESEARCH ARTICLE

Understanding of Causes of Spurious Associations: Problems and Prospects

Ghulam Ghouse[1] · **Atiq Ur Rehman**[2] · **Muhammad Ishaq Bhatti**[3]

Received: 30 March 2023 / Accepted: 24 January 2024 / Published online: 4 March 2024

Abstract

This paper contributes to the social science literature when analyzing survey or time series data social scientists use spurious regression without due consideration of its assumptions and the data structure. This results in misinterpretation and misleading conclusions about the population. The paper reviews basic statistical and econometrics literature which led to the development of modern time series analysis in the presence of spurious regression. It concludes that the term 'Spurious' was well known before the Granger and Yule's work in time series context rather than cross-sectional data. The same reasons can produce spurious regression today and surely the solution doesn't exist in the cointegration analysis. Social scientists and applied econometrician investigators need more serious thinking and care to avoid spurious regression, if it is necessary even if data is stationary or cross-sectional. In this study, we extended the Ghouse experiment which is based on simulated data by employing real-world data to assess the effectiveness of the newly proposed Ghouse Equation in comparison to conventional approaches. The findings demonstrate that the Ghouse Equation produces the lowest probability of spurious regression as compared to its counterparts. Moreover, in forecasting performance, Ghouse Equation outperformed its counterparts. These results highlight the Ghouse Equation as a valuable and better tool for econometric analysis for nonstationary time series.

Keywords Spurious Regression · Spurious Correlation · Stationarity · Unit root testing · Cointegration · ARDL · Ghouse Equation

Extended author information available on the last page of the article

JEL Classification C18 · C22 · C50

1 Introduction

Spurious regression is a very famous phenomenon in Econometrics [13]. The struggle to avoid the existence of spurious regression led to the development of modern time series analysis. The core objective of unit root and cointegration procedures, which are now the starting point of time series analysis, is to differentiate between genuine and spurious regression [3]. However, despite this level of the popularity of the term, the concept is quite misunderstood. The popular understanding of the term today is quite different from the understanding of the term used by its inventors and early users. The objective of this paper is to discuss the popular understanding of this term and its meaning. The implications of the poor understanding of the term are also discussed.

Spurious regression is one of the most popular concepts in econometrics and at the same time, it is the most misunderstood concept, even by the topmost professionals. Clive Granger is the person who wrote a highly cited paper explaining the spurious regression in time series coauthored with Paul Newbold and he won the Nobel Prize in Economic Sciences in 2003 for introducing a method to avoid spurious regression. While defining the spurious correlation, Granger writes:

> *"A spurious regression occurs when a pair of independent series but with strong temporal properties, are found apparently to be related according to standard inference in an OLS regression. [Granger Hyung and Geon, (2001)]".*

The word '*temporal dependence'* used by Granger et al. [6] indicates that the term describes a time series phenomenon that occurs when a series has temporal dependence. Therefore, the concept is necessarily linked with the time series. Not only Granger, but many other top economists have also used the term spurious regression and/or spurious correlation only in a time series context. Ventosa-Santaularia (2009) wrote an article titled 'Spurious Regression' in which he wrote:

> *"The spurious regression phenomenon in Least Squares occurs for a wide range of Data Generating Processes, such as drift less unit roots, unit roots with drift, long memory, trend and broken-trend stationarity".*

This means that, though not explicitly stated, the authors assume spurious regression to be a time series phenomenon because all reasons mentioned for spurious regression are in the time series context. The legendary econometrician David F. Hendry wrote a paper entitled 'Econometrics-Alchemy or Science', where he gives two examples of spurious regression, and both examples come from time series data [8]. Peter Phillips defines spurious regression more realistically. Philips (1998) writes:

"In a prototypical spurious regression, the fitted coefficients are statistically significant when there is no "true relationship" between the dependent variable and the regressors".

There is no mention of the time series context in Philips's definition of spurious regression. However, all of the work on spurious regression by Peter Philips revolves around time series spurious regression. He is the person who for the first time analytically explained the spurious regression phenomenon in non-stationary time series.

Given these circumstances, it is really hard to imagine the spurious regression in a context not involving time series. All standard econometrics textbooks discuss the term in a time series context and more particularly in the context of non-stationary time series.

2 Meaning of Spurious Regression

The terms *spurious regression and/or spurious correlation* have roughly the same history as the term regression itself. The correlation and regression analysis were invented by Sir Francis Galton in around 1888 and were popularized by Karl Pearson and George Undy Yule. Pearson wrote a paper in 1897 with the following title, '*Mathematical Contributions to the Theory of Evolution: On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs'* [10]. This title indicates a number of important things about the term spurious correlation: (a) the terms spurious regression was known as early as 1897, that is, less than 10 years after the invention of correlation analysis (ii) there were more than one types of spurious correlation known to the scientists of that time, therefore, the author used the phrase 'On a Farm of Spurious Regression' (iii) the cause of spurious correlation mentioned in the title of paper has nothing to do with the time series properties/context. Instead of this, the '*use of indices'* is considered as the reason for spurious regression which is indicative of cross-sectional context. Pearson in the same paper, defines the spurious correlation as follows:

"As a matter of fact, since the coefficients of variation for femur, tibia, and humerus are approximately equal, there would be, as we shall see later, a correlation of about 0.4 to 0.5 between these indices had the bones been sorted absolutely at random. I term this a spurious organic correlation, or simply a spurious correlation [10]*".*

This paragraph indicates that the variables where spurious regression was seen by Pearson were *random*; having no relationship between them, but the calculated correlation coefficient was seen between 0.4–0.5, an exaggerated number indicating high correlation. This is how the term spurious correlation was understood by Karl Pearson, one of the founding fathers of econometrics. Brown et al. [2] wrote a paper with the title *'A study of Index Correlation'* in which they note:

"But we know that the mixing of heterogeneous records having entirely different mean values leads to the production of correlations which are "spurious" and do not measure any real association between the variable".

This excerpt again clarifies that (i) the term *spurious correlation* was well known to the experts in 1914 (ii) the term was used for unrealistic outcomes of correlation coefficient in cross-sectional data (iii) the reason for spurious correlation as they understand was not any kind of temporal dependence but the mixing of heterogeneous records.

Similarly, one can find that many kinds of spurious correlations were known to experts in the first two decades of the twentieth century. These kinds of spurious correlations include the correlation due to the use of indices [10], spurious correlation due to variations in the magnitude of the population [15], spurious correlation due to the mixing of heterogeneous records [2], etc. The most important reason, however, was the missing third variable. Yule, in his 1926 paper 'Why do we sometimes get non-sense regression in time series….', explains the occurrence of spurious regression as:

"I cannot regard time per se as a causal factor; and the words only suggest that there is some third quantity varying with the time to which the changes in both the observed".

3 Granger and Newbold Experiment and Spurious Regression

Granger and Newbold [5] performed a simulated study in which they generated two independent random walk time series $x_t = x_{t-1} + \varepsilon_t$ and $y_t = y_{t-1} + v_t$. The two series are non-stationary in the sense that their second moment doesn't remain constant over time. The correlation of error terms of both series is zero so both series are independent of each other. The two variables don't have any common missing factor to which the movement of the two series can be attributed. Now the regression of the type $y_t = \alpha + \beta x_t + \epsilon_t$ should give an insignificant regression coefficient, but the Monte Carlo experiment of Granger and Newbold yielded a very high probability of getting a significant coefficient. The probability of getting this spurious significance did not reduce with the increase in sample size. Therefore, Granger and Newbold concluded that spurious regression occurs due to non-stationarity. This explanation was taken by the profession and now spurious regression has become a synonym of time series spurious regression.

Three factors are important to be considered regarding the study of Granger and Newbold. First, the above cited literature indicates that the spurious correlation in the cross-sectional data was quite well known to the practitioners in 1910's and 1920 and the Granger-Newbold experiment is not capable of explaining this cross-sectional spurious correlation. Second, even for the time series spurious correlation, the existing understanding was that it was due to missing variable(s). The Granger Newbold experiment shows that spurious correlation can occur due to non-stationarity, it does not deny that missing a third variable can also generate spurious correlation.

Third, the experiment does not prove that stationary series cannot produce spurious correlation.

However, the profession adapted three misunderstandings from the paper which were actually not implied by their experiment. Now, in most modern econometric textbooks, you would find the discussion on spurious correlation/spurious regression only in the time series context, and that too, in combination with the non-stationarity.

4 Spurious Correlation in Time Series Phenomenon

A Spurious correlation is often observed in time series data. This section attempts to answer a major question, 'Is spurious correlation needed in time series phenomenon?'. To answer this question, we begin by considering the following small data set (a sample of 16 observations) in Table 1 is given to justify the question raised:

This is a mixture of two different data sets. The first eight observations contain two columns of independent random numbers from Gaussian distribution with a mean 3 and variance of 1. The last eight observations contain similar columns of Gaussian random numbers with a mean 8 and variance of 1. The X and Y are independent of each other. There is absolutely no temporal property in the two variables and the data does not have any characteristic of typical time series data. The data is plotted in the following graph:

The graph shows that there are two clusters with each cluster showing no or very weak correlation between X and Y. But if we take the two clusters together, it will show a strong correlation. The correlation coefficient calculated for this data set was 88% which is no doubt very high, even though the two columns are independent. This shows that spurious correlation can occur even when there is no time series structure in the data.

Table 1 Data Form Gaussian Distribution with Different Mean and Same Variance

Mean = 3 and variance = 1			Mean = 8 and variance = 1		
S.N	X	Y	S.N	X	Y
1	2.35	1.43	9	6.11	8.81
2	2.80	2.95	10	8.20	8.21
3	3.36	1.25	11	7.14	8.32
4	4.60	3.09	12	7.73	8.75
5	2.80	2.91	13	8.34	9.19
6	3.81	4.34	14	7.60	8.76
7	2.60	3.14	15	8.62	7.94
8	0.74	4.59	16	9.70	9.79

5 Sources of Spurious Correlation?

There are numerous sources of spurious correlation, some of which are mentioned by Aldrich. Few common sources of spurious correlation are described as under.

5.1 Spurious Correlation due to Mixing Non-homogenous Groups

This form of spurious correlation was known to experts as early as 1914, as reported by Brown et al. [2] in their paper:

> *"Mixing of heterogeneous records having entirely different mean values leads to the production of correlations which are "spurious" and do not measure any real association between the variable".*

The example cited in the previous section is an example of the correlation that occurs due to mixing two non-homogenous groups. The first eight observations belong to one group whereas the last eight observations belong to the second group.

This kind of situation can frequently occur in the analysis of real data. Consider the data on the heights of individuals including male and female. It is well known that female usually has shorter height and less weight, and that there exists medium level correlation between height and weight of females. The male individuals, on the other hand, have relatively more height and more weight. The data on heights and weights of randomly selected individuals from the two genders will form two separate clusters similar to the clusters shown in Fig. 1 and when taken together, the data shall show a very strong correlation between heights and weights.

'Shoe size and intelligence' is an example often quoted in the literature related to the teaching of correlation coefficient (e.g. [4]. Consider a researcher going to a high school and taking a random sample of the students present in the school so that every student of the school is having an equal chance of being selected in

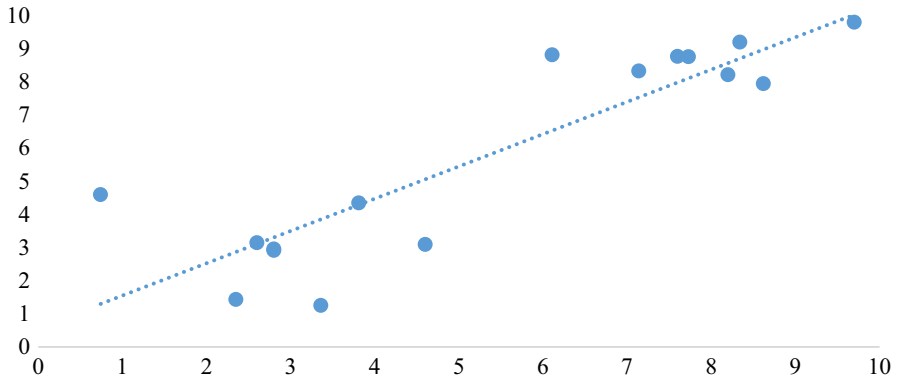

Fig. 1 Scatter plot of X and Y

the sample. The researcher is taking readings on the ability to solve mathematical problems and the shoe size of students selected in the sample. The sample selection is unbiased, and taking a large sample may reduce the sampling error. The researcher finds that there is a very high correlation between the size of the shoe and the ability to solve mathematics. Would this be sufficient to argue that the admission policy of the school should be based on the measurement of shoe size?

The fact is that if the sample is selected from a high school having classes from grades 0 to 10, this kind of observation is almost sure to occur. The pupils in higher classes have larger shoe sizes and have more mathematical skills compared to students in lower grades. Therefore, a high correlation is expected. However, if we take data from only one class, say grade III, we will not see such a high correlation.

This phenomenon is shown in the Fig. 2. The lower-most ellipse indicates the relationship between shoe size and mathematical skills for students of grade 1. The students are the youngest, having very small mathematical skills and very small shoe sizes. The ellipse does not have any slope, indicating an insignificant relationship between the two variables. The second ellipse corresponds to student of class II and it again shows almost zero correlation between variables on two axes. Similarly, for every class separately, there is no relationship between the two variables. But if we take all these ellipses together, we will get a very significant slope and consequently a very high correlation between two variables.

How this correlation has emerged? This can be explained in many ways: First, the apparent high correlation is due to the mixing of non-homogenous groups. Each class is a homogenous group, but class I and class IX put together make a non-homogenous group leading to a spurious correlation. Second, since the days of Yule, missing common cause is considered an important reason for spurious correlation.

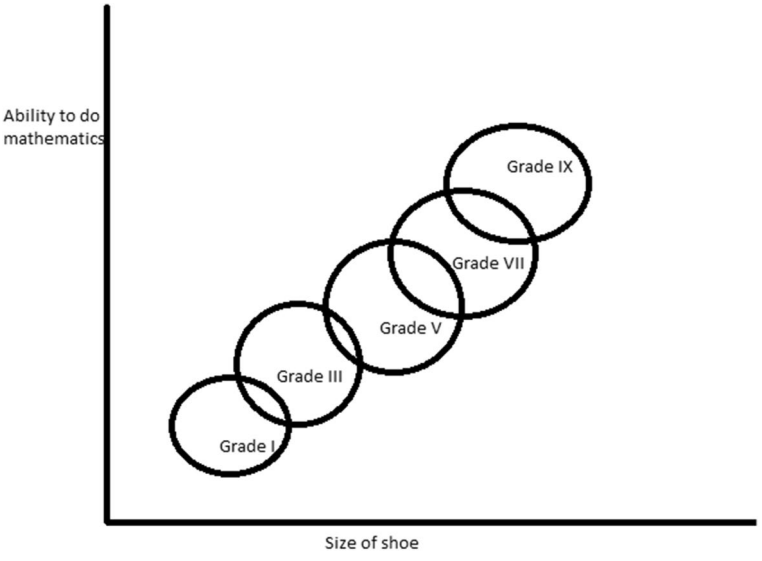

Fig. 2 Correlation between Size of Shoe and Ability to do Mathematics. Source: Goodwin and Leech [4]

The class/age of the students is a significant determinant of their mathematical skills, as well as of their shoe size. This common cause governs the correlation between two variables and the missing common cause has made the two variables have very high correlation. Third, in the regression context, the omitted variable bias can be held responsible for spurious regression. It is well known that if the true data generating process is $y_i = \alpha + \beta x_i + \gamma z_i + \epsilon$, but the researcher estimates the model $y_i = \alpha + \beta x_i + \epsilon$ in which the variable z is missing. In this case, it is very easy to show that the estimate $\widehat{\beta}$ would be biased. In our example, the data generating process for mathematical skills contains the variable age. Any regression missing this variable would be biased, and this bias can lead to spurious regression.

5.2 Missing Common Cause

In his highly cited paper, Yule [16] considers missing common cause as a reason for the spurious correlation. Therefore, Yule writes:

> *"I cannot regard time per se as a causal factor; and the words only suggest that there is some third quantity varying with the time to which the changes in both the observed variables are due."*

Yule presented this common cause diagnosis of the spurious regression problem for the time series data. Though there were many other explanations for spurious regression problems, the missing common cause was the popular explanation for non-sense correlation from the times of Yule [16] till 1974, when Granger and Newbold wrote their seminal paper explaining the spurious regression. Granger and Newbold generated two independent random walk series $x_t = x_{t-1} + \varepsilon_t and y_t = y_{t-1} + \epsilon_t$ where $\varepsilon_t and \epsilon_t were two$ IID series with no mutual correlation. In this case, the two series are independent and there is no mutual correlation. There is no third variable that can be attributed to as common cause. In such a scenario, the regression of the type $y_t = \alpha + \beta x_t + u_t$ should give insignificant results. But the Granger and Newbold's Monte Carlo experiment showed that the probability of getting a significant coefficient is very high and this probability increases with the increase in sample size. The series having this type of DGP are non-stationary, therefore Granger and Newbold concluded that the non-stationarity is a reason for spurious regression.

Granger and Newbold [5]'s findings are often misunderstood. There is no doubt that these findings imply that non-stationarity may lead to spurious regression, but these findings never imply that non-stationarity is the 'only' reason for spurious regression. In addition, neither do these findings deny missing common cause as a reason for spurious regression nor do they imply that spurious regression is only a time series phenomenon. Even today, the common cause can be shown as an extremely important reason for the spurious regression.

Consider the following scenario: a professor leaves home early in the morning, drops his son at school and then he/she comes to the university to deliver a lecture. The data on travel to school and travel to university, which is a time series data, will have quite a high correlation, but this is obviously due to a common cause, that is the timing which makes two variables have such a high correlation. Missing

common cause will make us believe that there is a high association between children proceeding to school and the arrival of a professor. But the professor has a son who also goes to university and the son whose father is not a professor also travels to university.

This kind of situation can occur frequently in economics. Consider two firms A and B, both having high dependence on the policy rate announced by the central bank, but having nothing in common with each other. A lower policy rate enhances the profits of the two firms, therefore, the profits of the two firms will appear to have quite a high correlation when in fact there is no direct link between the two firms.

To illustrate the importance of the common cause, I have performed a Monte Carlo experiment whose design is as follows:

Consider three series generated as follows.

$$x_t = 0.5x_{t-1} + \varepsilon_t \tag{1}$$

$$where x_0 = 0 \; and \; \varepsilon_t \sim IIDN(0, 1) \tag{2}$$

$$y_t = 0.5x_{t-1} + v_t \tag{3}$$

$$where y_0 = 0 \; and \; v_t = 0.2v_{t-1} + e_t \; and \; e_t \sim IIDN(0, 1)$$

$$z_t = 0.5x_{t-1} + u_t \tag{4}$$

$$where z_0 = 0 \; and \; u_t = 0.2u_{t-1} + \gamma_t \; and \; \gamma_t \sim IIDN(0, 1)$$

$$t = 1, 2, \ldots T$$

The series y_t *and* z_t have a common cause which is x_t, and there is no direct relationship between y_t *and* z_t. There is no non-stationarity anywhere in the data generating process. Given this data generating process with a length of time series T = 40, the regression of y_t *on* z_t has a 75% probability of being significant. Changing the magnitude of coefficients and length of time series changes the probability of getting a significant regression coefficient but it remains more than 60% in most of the scenarios. The series y_t *and* z_t can be seen as profits of firm's A and B which have mutually independent governance and business but the business depends on the central bank's policy rate. This common determinant of profit will make the two series appear to have a very high correlation.

5.3 Outliers

Outliers are the data points that seem quite different from the rest of the data set. Outliers can make a data set to show a high correlation when in fact there is a no correlation and conversely, outliers can make a data set to show that there is a weak correlation when in fact there is a high correlation. A popular example of such

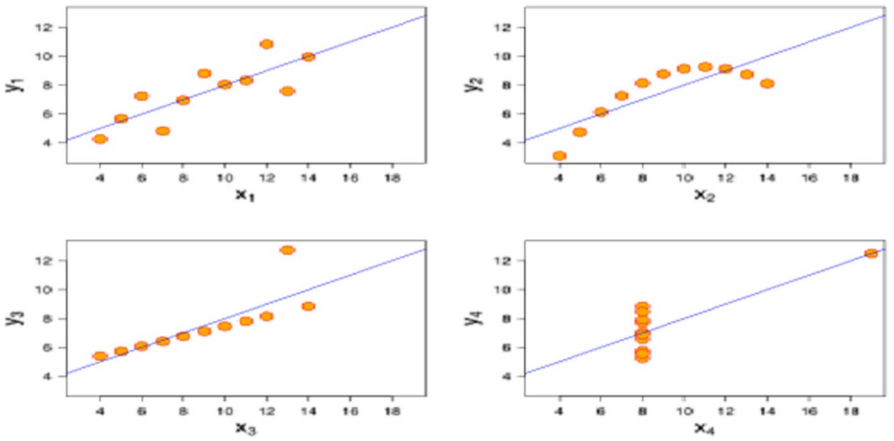

Fig. 3 Correlation in Presence of Outliers

correlation created by outliers is present in Anscombe data [1]. The Anscombe data consist of four data sets having same correlation coefficients and regression lines, but only one data set qualifies as a realistic description of data.

Consider the bottom right panel of Fig. 3 of the Anscombe data set. All data points form a vertical line showing that changes in Y are not attributable to changes in X. There is one outlier away from the rest of the data which makes the regression line appear to have a significant positive correlation. Obviously, this is not a genuine correlation and can be termed a spurious correlation because it exaggerates the strength of the true relation between X and Y. On the other hand, the data in the bottom left panel consists of a number of points that can be fitted to a single straight line so that the correlation should be 100%, but one outlier makes the correlation lose and the computed correlation coefficient is 82%. This phenomenon is also explained by [7].

5.4 Spurious Correlation due to the Use of Indices

This kind of spurious correlation was noted first by Pearson [10] and was explained by [15]. Probably the first time the term spurious correlation was used for this type of correlation. Karl Pearson explains this kind of correlation as follows:

If $u = f_1(x, y), and v = f_2(z, y)$ be two functions of three variables x, y, z, and these variables be selected at random so that there exists no correlation between $x, y, y, zorz, x$, there will still be found to exist a correlation between u *and* v. Thus real danger arise when a statistical biologist attributes the correlation between two functions like u *and* v to the organic relationship.

This kind of situation can occur very frequently in economics, the GDP per Capita and FDI per Capita, divided by the same denominator which is population, can easily produce such a correlation. Tax to GDP ratio and health expenditure to GDP, having the same denominator GDP can also be assumed to have such a relationship. Even with the cross-sectional data, one can find examples where two different

variables are divided by a third variable. The assets per employee and profits per employee of a firm can result in a situation similar to one observed by Pearson [10].

The reason why indices/ratios lead to such spurious correlation is explained by Yule [15]. He proves mathematically that in assuming the situation described by Pearson [10], the expectation of correlation between variables *u and v* is non-zero even when variables *x*, *y*, *andz* are independent of each other.

5.5 Ecological Correlation

Let $G_i i = 1, 2 \ldots N$ be a group of observations representing observation on two variables *xandy* and each group contains *n* observation. These observations can be represented as $x_{ij} and y_{ij}$ where *i* represents the number of groups and *j* represents the order of particular observations in his group. Let $\bar{x}_i = \frac{1}{n}\sum_j x_{ij}$ and $\bar{y}_i = \frac{1}{n}\sum_j y_{ij}$. The correlation between $\bar{x}_i$ and $\bar{y}_i$ is termed as ecological correlation. One can have a more general form of ecology where a number of observations in each group is different from one another. The ecological correlation would be the correlation between the group-average of variables *xandy*. It has been observed in a large number of studies that ecological correlation often exaggerates the correlation between *xandy*. Consider the example of the relationship between shoe size and mathematical skills.

Figure 4 represents the data on variable shoe size *xand* ability to do mathematics *y* discussed in Sect. 3 above. The left panel is a reproduction of Fig. 1 with the addition of group mean to the data for each class. The right panel plots only the class averages without going to the individuals. As discussed in section three, given any one class, there is no relationship between shoe size and mathematical ability and the result is an ellipse without any slope. However, all these ellipses taken together seem to have a strong relation. But when we plot the averages, they seem fitted to a straight line indicating near perfect correlation between *xandy*. Therefore, this example indicates that the correlation between averages overestimates the correlation variables.

Ecological correlation can also be seen very frequently in econometrics. In panel data, every cross-section can be regarded as a group and if the averages for

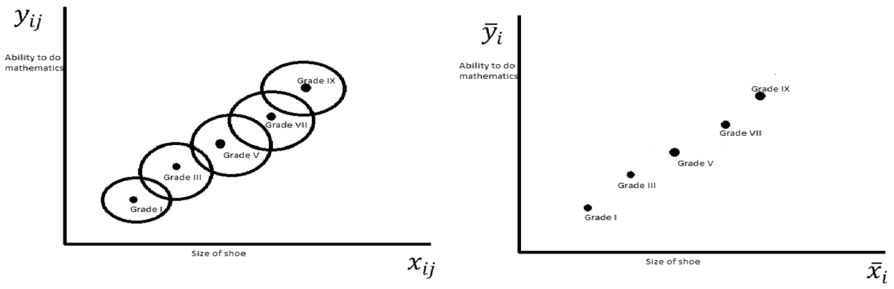

Fig. 4 Correlation Between Size of Shoe and Ability to Mathematics. Source: Goodwin and Leech [4]

Table 2 Data on GDP Growth and Inflation from 2010 to 2016

	Australia		Bahrain		Bulgaria		Canada	
Year	Growth	Inflation	Growth	Inflation	Growth	Inflation	Growth	Inflation
2010	2.05	2.85	4.33	1.96	1.32	2.44	3.08	1.78
2011	2.45	3.3	1.98	-0.4	1.91	4.22	3.14	2.91
2012	3.89	1.76	3.73	2.75	0.03	2.95	1.75	1.52
2013	2.64	2.45	5.42	3.31	0.86	0.89	2.48	0.94
2014	2.56	2.49	4.35	2.65	1.33	-1.4	2.86	1.91
2015	2.35	1.51	2.86	1.84	3.62	-0.1	1	1.13
2016	2.83	1.28	3.22	2.8	3.94	-0.8	1.41	1.43
Average	2.68	2.23	3.70	2.14	1.86	1.17	2.25	1.66

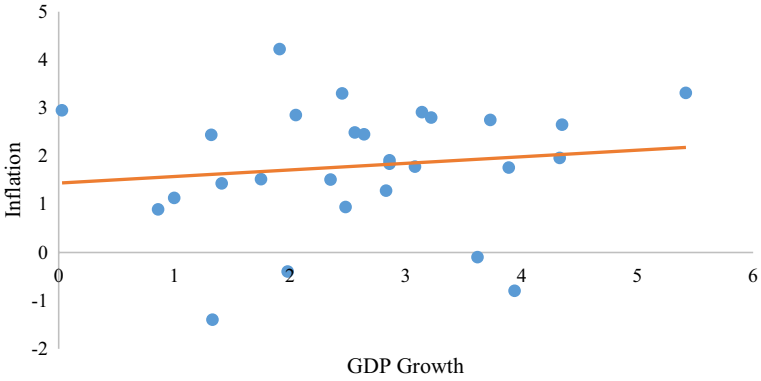

Fig. 5 Scatter Plot of GDP Growth and Inflation. Source: Author

the country are used, they will lead to ecological correlation. Consider the following data set in Table 2:

The data set is collected from WDI. It contains information on GDP growth and consumer price inflation for six countries including Australia, Bahrain, Bulgaria, and Canada. The last row of the Table gives the 7-year average of GDP growth and inflation for the sample countries. The data on the two variables for all samples is plotted in Fig. 5:

Figure 5 does not provide evidence of any significant correlation between the two variables and the calculated coefficient of correlation is just 12.7%. While Fig. 6 displays the 7 years averages for the sample countries.

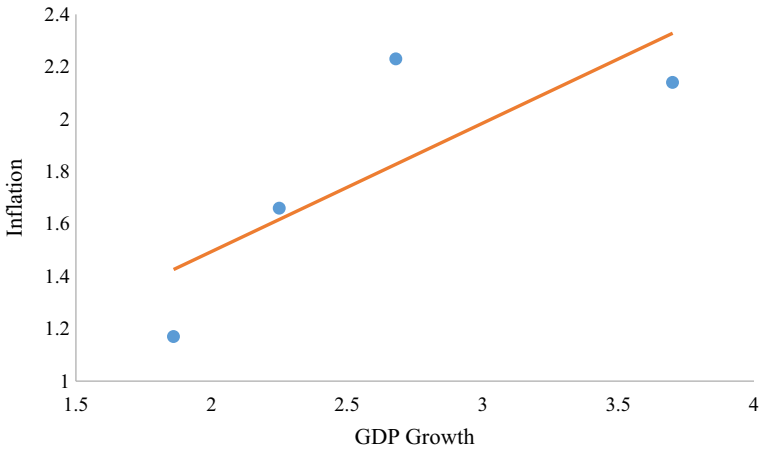

Fig. 6 Scatter Plot of Averages of GDP Growth and Inflation

Figure 6 shows that the averages are strongly correlated. The calculated correlation coefficient between the two variables is 79.5%, indicating a very strong association whereas actual correlation between the two variables is only 12.7%.

6 Spurious Regression with Stationary Variables

As stated earlier, Granger and Newbold [5] found that regression between two stationary series might be spurious. This research did not imply that non-stationary is the only cause of spurious regression, but most practitioners took this impression, and the word 'spurious regression'became synonymous with time series spurious regression. On the other hand, the fact is that stationary series may also produce spurious regression. This fact was later on recognized by Granger himself. Therefore, Granger et al. [6] writes:

> *"A spurious regression occurs when a pair of independent series, but with strong temporal properties, are found apparently to be related according to standard inference in an OLS regression. Although this is well known to occur with pairs of independent unit root processes, this paper finds evidence that similar results are found with positively autocorrelated autoregressive series or long moving averages. This occurs regardless of the sample size and for various distributions of the error terms".*

Granger et al. [6] performed a Monte Carlo experiment in which they generated independent stationary series with some positive autoregressive coefficient. They noted that given any pair of mutually independent stationary series with autoregression and with infinite time series length, the probability of spurious regression still

exists. Granger et al. [6] generated a two time series by following data generating process:

$$x_t = \theta_x x_{t-1} + \varepsilon_{xt} \tag{5}$$

$$y_t = \theta_y y_{t-1} + \varepsilon_{yt} \tag{6}$$

The two series would be stationary if $\theta_x < 1$ and $\theta_y < 1$. They took $\theta_x = \theta_y = 0.5$ and found that the probability of getting a significant coefficient in regression $y_t = \alpha + \beta x_t + \varepsilon_t$ is about 13% when the nominal size is 5%. They also note that increasing the time series length didn't decrease the probability of spurious regression. If the autoregressive coefficient is brought closer to 1, the probability of getting a significant coefficient also increases. With $\theta_x = \theta_y = 0.9$, the probability of getting a significant coefficient becomes more than 50%.

Granger et al. [6] make one conceptual correction that spurious regression is not bound to non-stationarity but still leaves behind one major misconception and one huge question mark. The misconception is to peg the spurious regression with time series context. As shown earlier, the term spurious regression was introduced much before the time of Granger and the early users of this term didn't have any time series context in mind. The question mark left behind is 'how to deal with spurious regression in stationary series?' The modern time series econometrics (that started after the seminal paper of Nelson and Plossor (1982) developed a remedy for spurious regression in the form of cointegration analysis. Cointegration analysis rests because spurious regression exists due to non-stationarity. But, as discussed by Granger et al. [6], if spurious regression can exist without non-stationarity, how to handle the problem? There is no satisfactory answer to the question. Rehman and Malik [12] show that spurious regression exists in stationary time series, regardless of sample size, specification of deterministic time trend, and distribution of error term.

7 Ghouse Experiment and Remedy for Spurious Regression

Ghouse et al. [3] investigated the existence of spurious regression in stationary and nonstationary variables, and found that the spurious regression can be found in stationery and unit root series. The unit root and cointegration are usually used to hold the problem of spurious regression due to the non-stationarity of series but due to the accumulation of type I and type II errors these tools lose their validity to some extent. While there is no valid tool to tackle the spurious regression in stationary series. According to Ghouse:

> *"The correct specification of a regression model is a crucial concern, and frequently, models encounter misspecification issues. This occurs when certain irrelevant variables gain significance by serving as proxies for the*

true variables, leading to a phenomenon known as spurious regression. To address this, missing variables can be substituted with lag values, acting as genuine proxies. The absence of these lag values is a primary cause of spurious regression in both stationary and nonstationary time series"

Ghouse et al. [3] proposed a methodology Ghouse Equation (GE) based on an autoregressive distributed lag mechanism. Suppose we have two variables x_t and y_t generated by following the Eqs. (5) and (6) with $\theta_x = \theta_y = 1$ for non-stationary series and with $\theta_x = \theta_y = 0.1 - 0.9$ for stationary series. Both variables are independent and the data generating process of both variables is autoregressive. Granger and Newbold [5] regressed nonstationary variables y_t on x_t and found spurious regression.

According to Ghouse et al. [3] the true determinant of y_t is the lag of y_t and by following the axiom of correct specification the y_t determinants of dependent must be in the equation then the equation will be:

$$y_t = \theta_y y_{t-1} + \theta_{1x} x_t + \theta_{2x} x_{t-1} + \varepsilon_{yt} \tag{7}$$

Equation (7) looks like the ARDL model and it significantly reduces the probability of spurious regression because it does not allow x_t to come up with significant results. This equation is equally applicable in the case of stationary series.

Ghouse et al. [3] introduced an alternative tool to deal with spurious regression with non-stationary time series but they used simulated data. This study used real data to test the validity of the Ghouse Equation (GE). We analyzed the size analysis forecast performance of GE and commonly used conventional cointegration procedures,Engle and Granger (EG), and Johansen and Juselius (JJ) cointegration tests. The forecast performance is tested based on real data. The real data consists of gross domestic product (GDP, at constant LCU) and household final consumption expenditures (HFC, at constant LCU) for the period of 1960 to 2019. The data is based on ten lower middle-income countries Pakistan, Bangladesh, India, SriLanka, Indonesia, Bolivia, Cameroon, Morocco, Nicaragua, and the Philippines.

7.1 Measuring the Probability of Spurious Regression among GE, GE, and JJ

The size analysis is performed to measure the probability of spurious regression. After running regression between independent series, if we got significant results, it counts as spurious regression. In this analysis, we used data given above of gross domestic product (GDP, at constant LCU) and Household Final consumption expenditures (HFC, at constant LCU) in all cases. However, we employed regression between cross-country series to ensure the independence of the series. Suppose the dependent variable is the GDP of Pakistan and the independent variable is HFC of other countries. It means, we used a statistically independent series because the HFC of one country has no relation with the GDP of any other country. After running regression, if we got significant results, it counts as spurious regression. The

Table 3 The Probability of Spurious Regression by using Engle and Granger (EG)Cointegration

Gross domestic product (at, constant LCU)										
Household final expenditure (at, constant LCU)		BGD	BOL	CMR	IDN	IND	LKA	MAR	NIC	PAK	PHL
	BGD	1	1	0	0	0	0	0	0	0	0
	BOL	1	1	0	0	1	0	0	0	0	0
	CMR	0	0	1	1	0	1	1	0	1	1
	IDN	0	0	0	1	0	0	0	0	1	0
	IND	1	1	0	0	1	0	0	0	0	0
	LKA	0	0	0	0	0	1	1	0	1	0
	MAR	0	0	0	1	0	0	1	0	1	0
	NIC	0	0	0	0	0	0	0	1	0	0
	PAK	0	0	1	1	0	1	1	0	1	0
	PHL	0	1	0	0	0	0	0	0	0	1
	Total	2	3	1	3	1	2	3	0	4	1

Table 3 shows the probability of spurious regression after employing the EG cointegration procedure. We used two Step procedure. In the first step, simple OLS regression is employed and generates a residual by using the parameters of OLS regression. In a second step we tested the stationary of the residual series. If series are stationary, it means series are cointegrated. 1's are showing series are cointegrated and 0's mean series are not cointegrated. Residual analysis is also employed for the validation of results

Table 4 The Probability of Spurious Regression by using Johansen and Juselius (JJ) Cointegration

Gross domestic product (at, constant LCU)										
Household final expenditure (at, constant LCU)		BGD	BOL	CMR	IDN	IND	LKA	MAR	NIC	PAK	PHL
	BGD	1	1	0	0	1	0	0	1	1	0
	BOL	1	1	0	0	1	0	0	0	0	0
	CMR	0	0	1	0	1	0	0	0	1	0
	IDN	0	0	0	1	1	0	1	0	1	0
	IND	1	1	1	1	1	0	0	0	1	1
	LKA	0	1	0	0	1	1	1	0	1	0
	MAR	0	0	1	0	1	0	1	0	1	1
	NIC	0	0	0	0	1	0	0	1	0	1
	PAK	1	0	0	0	1	0	1	0	1	0
	PHL	1	0	1	0	1	0	0	0	0	1
	Total	4	3	3	1	9	0	3	1	6	3

Table 4 shows the probability of spurious regression after employing the JJ cointegration procedure. We took three steps in this procedure. In the first step, VAR model is used. In a second step, we used the lag selection criteria for lag selection. In the third step, we employed the JJ procedure and made decision based on Unrestricted Cointegration Rank Test (Trace) statistics. 1's is showing series are cointegrated and 0's means not cointegrated. Residual analysis is also employed for validation of results

same procedure was done with a series of all countries to estimate the probability of spurious regression. The results are given in the following tables:

Table 3 shows the results of the Engle and Granger cointegration test. In this matrix "1" means the statistically independent variables are cointegrated and "0" means variables are not cointegrated. In this analysis, we are ignoring the diagonal 1's because these are the relationship between the same country series which means between dependent series. After employing EG cointegration test on independent series we got 20 significant relations out of 90 regressions. It means the probability of spurious regression after employing the EG procedure is 22.2%. It shows a 15.2% size distortion based on a 5% level of significance. It indicates that EG procedure is suffering from a size distortion problem.

Table 4 demonstrates the results of Johansen and Juselius's cointegration procedure. The "1" means the statistically independent variables are cointegrated and "0" means the variables are not cointegrated. In this analysis we are ignoring the diagonal 1's because these are the relationship between the same country series and, means between dependent series. After employing the JJ cointegration test on independent series we got 33 significant relations out of 90 regressions. It means the probability of spurious regression after employing the JJ procedure is 33.67%. It shows 28.67% size distortion based on a 5% level of significance. It indicates that the JJ procedure is suffering from a size distortion problem.

Table 5 represents the results of GE. The "1" means the statistically independent variables are cointegrated and "0" means the variables are not cointegrated. In this analysis we are ignoring the diagonal 1's because these are the relationship between the same country series and, means between dependent series. After employing GE on an independent series, we got 7 significant relations out of 90 regressions. It means the probability of spurious regression after employing GE is 7.78%. It shows a 2.78% size distortion based on a 5% level of significance which is negligible.

This analysis indicates that conventional cointegration procedures EG and JJ are suffering from a size distortion problem while GE tackles this problem by including lag values. It means the major cause of spurious regression is missing lag dynamics and by including the lag values, we can overcome the problem of spurious regression. It has theoretical justification, when we regress independent series, the independent variable starts working as a proxy of the relevant variable and captures the effect of the relevant variable which is why it becomes significant. But when we introduce the lag value of a dependent variable as an independent variable which is a potential determinant, it captures the effect and irrelevant variable becomes insignificant.

7.2 Forecasting Performance of GE, EG, and JJ

The Root Mean Square Error (RMSE) has been used to compare the forecast performance of GE, EG, and JJ. Figure 7 shows the RMSE statistics obtained after forecasting through these procedures:

Figure 7 shows that the forecast performance of GE is better as compared to conventional EG and JJ procedures. The RMSE statistics for GE in all cases remain

Table 5 The Probability of Spurious Regression by using Ghouse Equation (GE)

Gross domestic product (at, constant LCU)										
Household final expenditure (at, constant LCU)		BGD	BOL	CMR	IDN	IND	LKA	MAR	NIC	PAK	PHL
	BGD	1	0	0	0	0	0	0	1	1	0
	BOL	1	1	0	0	0	0	0	0	0	0
	CMR	0	0	1	0	0	0	0	0	1	0
	IDN	0	0	0	1	0	0	0	0	0	0
	IND	0	1	0	0	1	0	0	0	0	0
	LKA	0	0	0	0	0	1	0	0	0	0
	MAR	0	0	0	0	0	0	1	0	1	0
	NIC	0	0	0	0	0	0	0	1	0	1
	PAK	0	0	0	0	0	0	0	0	1	0
	PHL	0	0	0	0	0	0	0	0	0	1
	Total	1	1	0	0	0	0	0	1	3	1

Table 5 shows the probability of spurious regression after employing GE. We Used an F-stat to check the joint significance of lag and current values of the independent variable. All the decisions are taken based on F-stat. 1's is showing series are cointegrated and 0's mean not cointegrated. Residual analysis is also employed for the validation of results

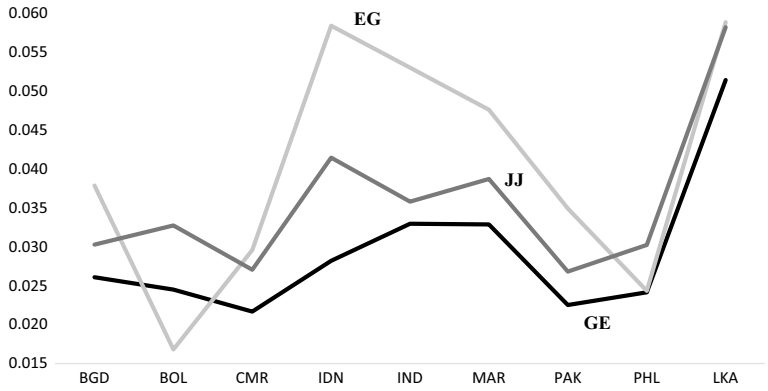

Fig. 7 The Root Mean Square Error (RMSE) after Forecasting

Table 6 The results of Root Mean Square Error (RMSE) after Forecasting

Countries	GE	EG	JJ
Bangladesh (BGD)	0.02607	0.03786	0.03027
Bolivia (BOL)	0.02450	0.01681	0.03273
Cameroon (CMR)	0.02167	0.02958	0.02704
Indonesia (IDN)	0.02818	0.05836	0.04142
India (IND)	0.03295	0.05298	0.03578
Morocco (MAR)	0.03286	0.04758	0.03869
Nicaragua (NIC)	0.04923	0.13033	0.17397
Pakistan (PAK)	0.02251	0.03491	0.02681
Philippines (PHL)	0.02415	0.02434	0.03023
Sri Lanka (LKA)	0.05140	0.05882	0.05817

smaller than the RMSE for EG and JJ procedures except only one case of Bolivia (BOL). Figure 7 also shows that the performance of the JJ cointegration procedure is better than the EG cointegration test. The RMSE statistics are given below in Table 6:

Table 6 illustrates the RMSE statistics of GE, EG, and JJ procedures. These RMSEs are calculated after forecasting. RMSE indicates the deviation of forecasted values from actual values. That is why the smaller value of RMSE shows less deviation from actual values and the higher value shows higher variation. The results in Table 6 indicate that the RMSE statistics of GE remain smaller in all cases apart from the Bolivia (BOL) case. In the case of EG, the second row has circle which indicates that in this particular case, EG performs well as compared to the GE model. In other cases, GE performs better as compared to the EG. In the case of JJ, rectangles indicte that EG performs better than the JJ procedure in only three cases. The overall condition is that GE performs well as compared to conventional cointegration procedures EG and JJ, and JJ cointegration procedure performs well

as compared to EG. Based on this analysis, we can express the performance of these procedures as:

$$\text{ForecastPerformance}$$

$$\text{GE} > JJ > EG$$

The results are theoretically admissible because GE has both contemporaneous and lag values of the independent variables. While JJ procedure contains only lag values of the independent variables and the EG procedure is based on static function. That is why, GE has more power to explain the relations instead of these conventional procedures.

The examination of spurious regression probability and forecast performance across the Ghouse Equation, Engle-Granger, and Johansen-Juselius methodologies reveals compelling insights. Notably, the Ghouse Equation, grounded in the ARDL mechanism, outperforms its counterparts in terms of size distortion, exhibiting the least vulnerability to generating misleading results. This underscores the Ghouse Equation's robustness and its capacity to provide more reliable estimates. Additionally, its superior forecasting ability further distinguishes it as a valuable tool in the realm of econometrics, offering enhanced precision and accuracy to researchers and analysts.

8 Conclusion: What All the Discussion Implies?

Huge science of unit root and cointegration analysis was developed during the last four decades that attempts to solve the problem of spurious regression in non-stationary. However, as revealed by the Nobel laureate Clive Granger, spurious regression can exist even if there is no non-stationarity. On the other hand, the literature days of econometrics reveal that the term was popular well before Granger and Yule, and those early users didn't use the term in time series context. They explored many reasons for spurious regression and all the reasons correspond to cross-sectional data. The same reasons can produce spurious regression today and surely the solution doesn't exist in the cointegration analysis. The Ghouse Equation is an alternative tool to reduce the probability of spurious regression in the case of spurious regression. The econometric investigation needs more serious thinking and the care for avoiding spurious regression is necessary even if data is stationary or cross-sectional.

Acknowledgements We Acknowledge the kind support and supervision of Prof. Dr. Ishaq Bhatti.

Author contributions AR Conceptualization and written original draft. IB supervised, edited, and reviewed the original draft. GG collected data and analyzed the data. All authors read and approved the final manuscript.

Funding There is no funding.

Data availability The data is freely available on the mentioned resources. Also, can be provided on request.

Declarations

Conflict of interest We all agree there is no competing interest.

References

1. Anscombe, F.J.: Graphs in statistical analysis. Am. Stat. **27**(1), 17–21 (1973)
2. Brown, J.W., Greenwood, M., Wood, F.: A study of index correlations. J. Roy. Stat. Soc. **77**(3), 317–346 (1914)
3. Ghouse, G., Khan, S.A., Rehman, A.U., Bhatti, M.I.: ARDL as an elixir approach to cure for spurious regression in nonstationary time series. Mathematics **9**(22), 2839 (2021)
4. Goodwin, L.D., Leech, N.L.: Understanding correlation: Factors that affect the size of r. J. Exp. Educ. **74**(3), 249–266 (2006)
5. Granger, C.W., Newbold, P.: Spurious regressions in econometrics. J. Econom. **2**(2), 111–120 (1974)
6. Granger, C.W., Hyung, N., Jeon, Y.: Spurious regressions with stationary series. Appl. Econ. **33**(7), 899–904 (2001)
7. Ghouse, G., Bhatti, M.I., Aslam, A., Ahmad, N.: Asymmetric spillover effects of Covid-19 on the performance of the Islamic finance industry: a wave analysis and forecasting. J. Econom. Asymmetries **27**, e00280 (2023)
8. Hendry, D.F.: Econometrics: alchemy or science? essays in econometric methodology. Oxford University Press (2000)
9. Nelson, C.R., Plosser, C.R.: Trends and random walks in macroeconmic time series: some evidence and implications. J. Monet. Econ. **10**(2), 139–162 (1982)
10. Pearson, K.: Mathematical contributions to the theory of evolution on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc. R. Soc. Lond. **60**(359–367), 489–498 (1897)
11. Phillips, P.C.: New tools for understanding spurious regressions. Econometrica **66**(7),1299–1325 (1998)
12. Rehman, A.U., Malik, M.I.: The modified R a robust measure of association for time series. Electron. J. Appl. Stat. Anal. **7**(1), 1–13 (2014)
13. Tu, Y., Wang, Y.: Spurious functional-coefficient regression models and robust inference with marginal integration. J. Econom. **229**(2), 396–421 (2022)
14. Ventosa-Santaulària, D.: Spurious regression. J. Probab. Stat. 1–27 (2009)
15. Yule, G.U.: On the interpretation of correlations between indices or ratios. J. Roy. Stat. Soc. **73**(6/7), 644–647 (1910)
16. Yule, G.U.: Why do we sometimes get nonsense-correlations between time-series? a study in sampling and the nature of time-series. J. Roy. Stat. Soc. **89**(1), 1–63 (1926)

Authors and Affiliations

Ghulam Ghouse[1] · Atiq Ur Rehman[2] · Muhammad Ishaq Bhatti[3]

✉ Ghulam Ghouse
ghulam.ghouse@econ.uol.edu.pk

Atiq Ur Rehman
ateeqmzd@gmail.com

Muhammad Ishaq Bhatti
ishaq.bhatti@ubd.edu.bn

[1] Department of Economics, University of Lahore, Lahore, Pakistan

[2] Kashmir Institute of Economics, University of Azad Jammu and Kashmir, Azad Jammu and Kashmir, Muzaffarabad, Pakistan

[3] School of Business and Economics, Universiti Brunei Darussalam, Bandar Seri Begawan, Brunei Darussalam