Munich Personal RePEc Archive

ARDL model as a remedy for spurious regression: problems, performance and prospectus

Ghouse, Ghulam and Khan, Saud Ahmed and Rehman, Atiq Ur

Pakistan Institute of Development Economics

10 January 2018

Online at https://mpra.ub.uni-muenchen.de/83973/
MPRA Paper No. 83973, posted 19 Jan 2018 02:37 UTC

ARDL model as a remedy for spurious regression: problems, performance and prospectus

(1) Ghulam Ghouse
Ghouserazaa786@gmail.com
PhD scholar (Department of Econometrics and Statistics)
Pakistan Institute of Development Economics, Islamabad, Pakistan.

(2) Saud Ahmed Khan
saudak2k3@yahoo.com
Assistant Professor (Department of Econometrics and Statistics)
Pakistan Institute of Development Economics, Islamabad, Pakistan.

(3) Atiq Ur Rehman
atiq@pide.org.pk
Assistant Professor (Department of Econometrics and Statistics)
Pakistan Institute of Development Economics, Islamabad, Pakistan.

Abstract

Spurious regression have performed a vital role in the construction of contemporary time series econometrics and have developed many tools employed in applied macroeconomics. The conventional Econometrics has limitations in the treatment of spurious regression in non-stationary time series. While reviewing a well-established study of Granger and Newbold (1974) we realized that the experiments constituted in this paper lacked Lag Dynamics thus leading to spurious regression. As a result of this paper, in conventional Econometrics, the Unit root and Cointegration analysis have become the only ways to circumvent the spurious regression. These procedures are also equally capricious because of some specification decisions like, choice of the deterministic part, structural breaks, autoregressive lag length choice and innovation process distribution. This study explores an alternative treatment for spurious regression. We concluded that it is the missing variable (lag values) that are the major cause of spurious regression therefore an alternative way to look at the problem of spurious regression takes us back to the missing variable which further leads to ARDL Model. The study mainly focus on Monte Carlo simulations. The results are

providing justification, that ARDL model can be used as an alternative tool to avoid the spurious regression problem.

Keywords: *Spurious regression, Stationarity, unit root, cointegration and ARDL.*

1. Introduction

The most important feature that led to development of new time series econometrics was spurious regression. Spurious regression is a phenomena known to econometricians since the times of Yule (1926). Spurious regression was attributed to missing variable until Granger and Newbold (1974) who showed that spurious regression could be found with nonstationary time series even with no missing variable. Nelson and Plosser (1982) argued that most of the time series are better characterized as nonstationary. Spurious regression have performed a vital role in the construction of contemporary time series econometrics and have developed many tools employed in applied macroeconomics. However, the widespread literature considers the non-stationarity as the only reason for spurious regression. To evade the problem of spurious regression caused by the non-stationarity, researchers frequently employed unit root and co-integration testing.

Supposing that the spurious regression occurs due to non-stationarity and unit root and cointegration testing are used as the remedy, even then it is very hard to find reliable inference. There is no test of unit root with good size and power in small sample. The unit root and cointegration procedures involves many prior specification decisions e.g. lag length, trend and structural stability etc. If we do a data based decision making, it will involve a large battery of tests. Each test is having specific statistical error (type I, II error). The cumulative probability of error in all tests leave the results of unit root test unreliable. Because, of these reasons, the literature is still underdevelopment after four decades without of reaching any conclusion.

It is a common fallacy that the unit root only cause of spurious regression. Nonetheless, the missing relevant variable is a major cause of spurious regression. Even it can be shown that the spurious regression in Granger and Newbold (1974) experiment was also due to missing variable see, (section, 5.1).

So, an alternative way to look at the problem of spurious regression takes us back to missing variable which further leads as to ARDL. Suppose, we have two independent autoregressive nonstationary series

$$Y_t = \rho Y_{t-1} + \varepsilon_{yt} \quad \text{.......... (1)} \qquad \rho = 1$$

$$X_t = \rho X_{t-1} + \varepsilon_{xt} \quad \text{.......... (2)} \qquad \rho = 1$$

Where $X_t$and Y_t both are expressed by their own lag values. There is no third variable involved in the construction of both variables. Granger and Newbold (1974) shown that the spurious regression by estimating of regression of the type

$$Y_t = a + \beta_1 X_t + \varepsilon_{yt} \quad \text{.......... (3)}$$

But we know that true data generating process (DGP) of Y and X contain lag of values, including the lag of Y and X we get

$$Y_t = a + \beta_1 X_t + \beta_1 X_{t-1} + \beta_3 Y_{t-1} + \varepsilon_{yt} \quad \text{.......... (4)}$$

Which is an ARDL model. It is observed in our study (section 4) that this kind of model significantly reduce the probability of spurious regression in case of nonstationary series. This indicates that spurious regression occur due to missing variable and can be avoided by including missing lag see, (section, 5).

The objective of this study is to explore an alternative solution that is expected to perform for nonstationary series. This study will investigates that, is it possible to use ARDL model to evade the spurious regression bypassing the very complicated and ambiguous unit root testing, cointegration analysis and other treatments. We will generate the autoregressive (nonstationary, stationary and negative moving average) series and investigate that how the probability of spurious increase dramatically in nonstationary case by ignoring the lag dynamics through Monte Carlo simulations.

2. Literature review

An immense amount of studies are available on spurious regression topic in time series econometric literature. In this section we briefly discuss the proposed theoretical and empirical methods for the treatment of spurious regression in literature. The literature review is arranged as follows

2.1 Spurious Regression in Classical Econometrics

There is long historical debate on nonsense correlation (spurious regression) issue in econometrics literature, at least seeing back to the well-known study of Yule (1926). In his study, he presented the presence of a strong correlation of 0.95 between mortality rate and proportion of marriages of the Church of England to all marriages during 1866 to 1911. Yule (1926) thought that the spurious regression is a consequence of relevant missing variables.

Simon (1954) also supported the idea that the missing variable is a source of spurious correlation. Simon described that if we are uncertain that the perceived correlation is spurious, we have to introduce extra variable which could be observed the genuine correlation.

2.1.1 Granger and Newbold's Experiment

Granger and Newbold (1974) showed that if the series are nonstationary then the results would be significant. In their experiment they generated independent autoregressive series like, X_t and Y_t. Where $X_t$and Y_t both are expressed by their own lag values.

$$Y_t = Y_{t-1} + \varepsilon_{yt} \quad \text{……….. (5)}$$

$$X_t = X_{t-1} + \varepsilon_{xt} \quad \text{……….. (6)}$$

There is no third variable involved in the construction of both variables. They regressed $X_t$on Y_t and $Y_t$on X_t.

$$Y_t = a + \beta_1 X_t + \varepsilon_{yt} \quad \text{……….. (7)}$$

$$X_t = a + \beta_1 Y_t + \varepsilon_{xt} \quad \text{……….. (8)}$$

They came up with spurious results. This alternative explanation of spurious regression become more popular in literature and other explanations went to the darkness.

2.1.2 Aftermath of Granger and Newbold's Experiment

2.1.2.1 Why is spurious regression a problem?

To find the relationship between the economic variables is the core objective of economic studies. The spurious regression offers deceptive statistical evidence of strong relationship even though the variables are independent. Hendry (1980) demonstrated a spurious correlation between cumulative rainfall and price level in UK. He inspected that all these time series were stationary at difference except unemployment rate. Plosser and Schwert (1978) claimed that, the regression without taking difference of nonstationary series most probably come up with invalid or nonsense results. The reasoning behind this claim is that if we run regression without taking difference of difference stationary series, the estimator properties and the distribution of test statistics are no more reliable. Phillips (1986) examined the asymptotic properties of spurious least square regression model and endorsed Granger and Newbold (1974) simulation results that the misspecification of level of series is the key element of spurious correlation.

2.1.2.2 Example of spurious regression in classical literature

Mostly, the nominal economic variables are correlated, even there is no relationship between them, and the mutual presence of price level in data series develops correlation between them. It was also shown that many time series are nonstationary that's why the probability of spurious regression is very high. We are presenting here some examples of spurious regression form time series econometrics literature.

Chaouachi (2013) inspected that Dar et al. (2012) in their study provided spurious strong positive relationship among usage of nass chewing, hookah smoking and many other habits with oesophageal squamous cell carcinoma (ESCC) risk. Dar et al. (2012) conducted a case control study in valley of Kashmir, India. They considered 702 historical cases of oesophageal squamous cell carcinoma (ESCC) and 1663 hospital based controls, exclusively matched to the cases for sex, age and residence district from Sep, 2008 to Jan, 2012. They used monthly data from Sep, 2008 to Jan, 2012. They concluded that nass chewing and hookah smoking are strongly positively associated with (ESCC) risk, which is based on severe misinterpretation. According to Chaouachi (2013) all the relevant studies showed that there is feeble or insignificant association among nass chewing, hookah smoking with (ESCC) risk. Chaouachi (2013) stated that Dar et al. (2012) came up with spurious results because they did not incorporate the very significant element which is filtering factor of water.

Roger and Jupp (2006) described an example of spurious positive relationship between human baby's birth and stork nesting in the sequence of spring, because these two variables are correlated to a third variable. According to the Roger and Jupp (2006) the sequence of Dutch statistics is showing a positive relationship between stork nesting in the sequence of spring and human baby's birth at that time, it is due to that the both variables are associated to the state of weather. It means that both variables are independent, but they have relation with the state of weather. This shows that both variables are spuriously correlated because of third missing variable. According to the Hofer et al. (2004) this spurious correlation is due to lack of statistical information.

2.1.2.3 Nelson and Plosser experiment and implications

Nelson and Plosser (1982) examined that most of the macroeconomics series of U.S.A economy are having unit root. Their study is generally acknowledged as a significant contribution with consequences for the theory and policy. They employed Dickey Fuller test for unit root detection on fourteen historical macroeconomics series for U.S.A economy, including GNP, wage, employment, prices, stock prices and interest rate and they found that twelve out of fourteen series were having unit root. In fact Nelson and Plosser (1982) study is a noteworthy contribution in time series econometric literature which enhanced the interest of researchers in unit root tests. That's why it has fashioned the development in the unit root theory.

2.1.2.4 Development in cointegration tests

Engle and Granger (1987) introduced the co-integration technique as a solution of spurious regression due to non-stationary time series. According to Granger the non-stationary time series are cointegrated, if their linear combination is a stationary process. Now the problem is that how to estimate the long run equilibrium relationship parameters for this Engle and Granger presented an Error Correction Mechanism. The residuals of equilibrium regression can be used for error correction model. The first drawback of EG (Engle and Granger) cointegration test is that it only deals with one cointegrated vector. Second, it depends upon two step estimator, first step is to produce series of residuals and second, to check the stationarity of residuals series. Third, the major limitation is the distributions of the estimators are non-standard. Phillips and Ouliaris (1990) proposed residual based tests under the null hypothesis of no cointegration in time series. In which the asymptotic distributions of residual based tests depend upon number of variables and deterministic trend terms. Engle and Yoo (1991) proposed three step procedure to evade the limitations of EG model, which is an extension of EG model. Engle and Yoo (EY) procedure confirms that the distributions of the estimators yield the normal distribution. It is also only useful for one cointegrated vector.

When we have more than one variable then there is the possibility of more than one cointegrated vector. EG and EY cointegration do not provide any solution in this situation. So, to overcome this problem Johansen and Juselius (1992) introduced the multivariate cointegration test. The Johansen and Juselius (JJ) test allows to find out more than one cointegrated vectors so, it is generally more applicable than EG and EY cointegration tests. We knew that EG and EY single equation procedures ignore short run dynamics, when the relationships are estimated. But, the JJ procedure also considers the short run dynamics. Pesaran et al. (1996) and Pesaran (1997) proposed a single equation ARDL (autoregressive distributed lag) approach for cointegration as an alternative of EG and EY. The first advantage is the ARDL cointegration approach provides explicit tests for the presence of a single cointegrating vector, instead of assuming uniqueness. Pesaran and Shin (1995) revealed that asymptotically valid inference on short run and long run parameters could be made by employing ordinary least square estimations of ARDL model. So, the ARDL model order is properly augmented to grant for contemporary correlation among the stochastic elements of the data generating processes involved in estimation.

2.1.2.5 Problems in cointegration analysis

The cointegration testing is involves many specification decisions which cut the reliability of results. The existing cointegration testing procedures do not provide any reasonable criteria regarding these specification decisions: choice of the deterministic part; the structural breaks; autoregressive lag length choice and innovation process distribution. For further detail see, (section, 2.3.2).

2.2 Conceptual Flaws in Understating of Spurious Regression

It is a common misconception that the spurious regression only prevails due to unit root. Nevertheless, the missing relevant variable is a major cause of spurious regression. Yule (1926) first time anticipated that the nonsense correlations could prevail due to missing variable.

Simon (1954) argued that the missing variable is a cause of spurious correlation. Simon has described this problem in following tactic that if we are uncertain that the observed correlation is spurious, we should introduce another (extra) variable which may observed the true correlation. Frey (2002) argued that the spurious regression could be probably due to missing variable.

Even it can be shown that the spurious regression in Granger and Newbold (1974) experiment was also due to missing variable. In their experiment they generated independent autoregressive series like, X_t and Y_t. Where $X_t$and Y_t both are expressed by their own lag values.

$$Y_t = Y_{t-1} + \varepsilon_{yt} \quad \ldots\ldots\ldots.. (9)$$

$$X_t = X_{t-1} + \varepsilon_{xt} \quad \ldots\ldots\ldots.. (10)$$

There is no third variable involved in the construction of both variables. They regressed $X_t$on Y_t or vice versa without involving their lag values in regression analysis.

$$Y_t = a + \beta_1 X_t + \varepsilon_{yt} \quad \ldots\ldots\ldots.. (11)$$

$$X_t = a + \beta_1 Y_t + \varepsilon_{xt} \quad \ldots\ldots\ldots.. (12)$$

They came up with spurious results due to missing variable because they did not include the lag values of variables as an independent variable. It is obvious that on determinant of Y_t that is $Y_{t-1}$is missing in equation (11) and similarly one determinant of X_t i.e. X_{t-1} is missing in equation (12). Taking these missing variables into account the equation shall become

$$Y_t = a + \beta_1 X_t + \beta_2 Y_{t-1} + \varepsilon_{yt} \quad \ldots\ldots\ldots.. (13)$$

Therefore, equation (13) shall not have spurious regression if our supposition if missing variable problem is true. It is shown in section (4) that it is actually true.

2.3 Problems in prevailing treatments

The most familiar procedures to evade the spurious regression are unit root and cointegrating testing. These methods are equally capricious because of some specification decisions like, choice of the deterministic part; the structural breaks; autoregressive lag length choice and innovation process distribution see, (section, 2.3.1.1). The cointegration analysis which is employed as a tool to avoid spurious regression, also experience with specification decisions problems see, (section, 2.3.2). It involves unit root testing which is also unreliable. The tests of unit root are so unreliable that is why it is very hard to conclude something reasonable see, (section 2.3.1).

2.3.1 Unit root testing

Numerous financial and economic series exhibit nonstationary or trending behavior like, Stock prices, exchange rate and Gross Domestic Product (GDP) and many others. It is unlikely to get accurate results from trendy series. The most common procedures to avoid the spurious regression are unit root and cointegrating testing. These procedures are equally unreliable due to specification decisions. The cointegration analysis which is used as a tool to avoid spurious regression, suffer numerous problems. It involves unit root testing and then testing for cointegration. The tests of unit root are so unreliable that is why it is very hard to conclude something reasonable. The US GNP is the series used by the large number of researchers as a guinea pig for the tests of unit root. However, nothing reasonable could be said about the unit root in series. Rehman and Zaman (2008) summarize findings of researchers in US GNP as follows.

"Trend Stationary: Perron (1989), Zivot and Andrews (1992), Diebold and Senhadji (1996), Papell and Prodan (2003),

Difference stationary: Nelson and Plosser (1982), Murray and Nelson (2002), Kilian and Ohanian (2002),

Don't know; Rudebusch (1993)".

2.3.1.1 Why unit root tests are so unreliable

So, the important task in econometrics is to determine the most suitable arrangement of trend in time series. There are two common procedures to eradicate the trend of data are regression with time trend and differencing. The unit root testing procedure offers an idea which procedure can be adopted to render the time series stationary. Besides, the precision and specification of unit root procedures are still a paradox, though, since mid-eighties the literature on unit root testing has been raised stormily.

Rehman and Zaman (2008) investigated that the two main causes for inadequate performance of unit root tests are observational equivalence and model misspecification. They mainly targeted four specification decisions: choice of the deterministic part; the structural breaks; autoregressive lag length choice and innovation process distribution, and examine their role in an inference from unit root tests. They explored that these specification decisions seriously impact the performance of unit root tests. Also investigated that the existing unit root tests do not provide any set criteria regarding these specification decisions, that is why they came up with unreliable results.

DeJong et al. (1992) found that Choi and Philips (1991) and Philips and Perron (1988) unit root procedures suffer from size distortion and low power issues in the presence of moving average (MA). While, Augmented Dicky Fuller (ADF) behaved well. Schwert (2002) Investigated that the Dicky Fuller (1979, 1981) is responsive to pure autoregressive process assumption means the data generating process of series is pure autoregressive (AR). When the moving average competent involves in fundamental process, then the Dicky Fuller reported distribution and test statistic distribution can be quite different. Many other unit root tests are being proposed, at some extent they all are facing similar problems.

2.3.2 Problems with Cointegration Testing

Like unit root tests the cointegration testing is also involves many specification decisions which cut the reliability of results. The existing cointegration testing procedures do not provide any reasonable criteria regarding these specification decisions, and that leads to their results are unreliable.

For example, Lag length specification is a significant practical question about the application of any econometric analysis. Like, in case of unit root test, if the lag length is too short then the serial correlation remains in errors and the results will be biased. If the lag length is too large this will reduce the power of the test. In the same way the cointegration tests are also very sensitive to lag length selection. Agunloye et al. (2014) explored that the Engle Granger (EG) cointegration test is extremely sensitive to lag length. Carrasco et al. (2009) examined that the lag length misspecification may significantly affect the cointegration results. In case of the under specification, it could undermine the cointegration results and in over specification, it may diminish the power of test. Similarly, trend specification is also a very significant issue in econometric literature.

Ahking (2002) explored that when the deterministic linear time trend included in Johansen's cointegration test it provides disproving results and after exclusion of deterministic linear time trend got robust results. He also suggested that great attention must be taken in trend specification in cointegration analysis. There are lot of studies are available in literature on this issue but most of them are with different results. Leybourne and Newbold (2003) used three cointegration test for independent integrated series and each series has a structural break. They found cointegration among them until structural break are not properly treated. Choi et al. (2004) examined that the economic models for cointegration are often provided erroneous results. The main reason is the errors are unit root non-stationary owing one of the variable has non-stationary measurement error. They stated that "If the money demand function is stable in the long-run, we have a cointegrating regression when money is measured with a stationary measurement error but have a spurious regression when money is measured with a nonstationary measurement error".

3. What is ARDL Model?

In ARDL model the dependent variable is expressed by the lag and current values of independent variable and its own lag value. Davidson et al. (1978) proposed ARDL methodology (DHSY hereafter) to model the UK consumption function. ARDL model normally starts from reasonably general and large dynamic model and progressively reducing its mass and altering variable by imposing linear and non-linear restrictions (Charemza and Deadman, 1997). Autoregressive distributed lag (ARDL) model is one of the most general dynamic unrestricted model in econometric literature. As we know ARDL methodology follows general to specific approach, that's why it could be possible to tackle many econometric problems like, misspecification and autocorrelation, and come up with a most appropriate interpretable model.

The ARDL (1, 1) is the simplest form of ARDL model. Consider an ARDL (1, 1) model

$$Y_t = a + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 Y_{t-1} + \varepsilon_{yt} \quad \text{……….. (14)}$$

Hendry and Richard (1983), Hendry, Pagan and Sargan (1984) and Charemza and Deadman (1997) argued that by imposing restrictions we can find out at least ten most appropriate economically interpretable models from ARDL (1, 1) model. We are giving hare some important cases of restriction

1. $\beta_2 = \beta_3 = 0$ Static regression,
2. $\beta_1 = \beta_2 = 0$ First order autoregressive process,
3. $\beta_3 = 1,\ \beta_1 = -\beta_2$ Equation in first difference,
4. $\beta_2 = 0$ Partial adjustment equation

As discussed, the spurious regression is may be a consequence of missing variable. ARDL is a general specification taking into account the lag structure. Therefore it could give better results.

4. The Methodology

This study mainly focuses on Monte Carlo Simulations. The data would be generated with pre decided specifications and the probability of spurious regression would be tested using classical methods and with ARDL model.

The Components of the methodology are as following

I. Data generating process (DGP) see (section 3.1)
II. Testing and Simulations see (section 4)

4.1 Data Generating Process (DGP)

Let's, we have a data generating process

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_{12} \\ \theta_{21} & \theta_2 \end{bmatrix} \begin{bmatrix} x_{t-i} \\ y_{t-i} \end{bmatrix} + \begin{bmatrix} \varepsilon_{xt} \\ \varepsilon_{yt} \end{bmatrix} \qquad \begin{bmatrix} \varepsilon_{xt} \\ \varepsilon_{yt} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

We can rewrite it as for simplification of notation

$$X_t = AX_{t-i} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \Sigma) \text{ .... (19)}$$

The data generating process equation (18) can generate data in quite large types of scenarios. Suppose, $\theta_{12} = \theta_{21} = 0$ and $\rho = 0$, the data generating process will generate two independent series and would be indication of spurious regression if the regression of x_t on y_t turns out to be significant. If A= 0, it indicates that there is no autocorrelation and cross autocorrelation in the series. If A is zero it means series would be IID (identically independently distributed). The value of degree of association depends upon only $\sum$.

5. Results and Inference

In this section we present inferences based on real and simulated data. The real data is based on Gross domestic product of thirty seven countries Albania, Antigua and Barbuda, Argentina, Austria, Bahamas, Bahrain, Barbados, Belgium, Botswana, Brazil, Brunei Darussalam, Cabo Verde, Canada, Comoros, Congo, Costa Rica, Denmark, Dominica, El Salvador, Fiji, Finland, France,Gabon,Gambia, Germany, Grenada, Guinea-Bissau,Guyana, Honduras, Hong Kong, Iraq Iceland, Ireland, Israel, Italy, Kiribati and Luxembourg from 1980 to 2014. We employed the ADF unit root test and come to know all the series are stationary at first difference. All the series are statistically independent of each other. We regress Antigua and Barbuda, Argentina, Austria, Bahamas, Bahrain, Barbados, Belgium, Botswana, Brazil, Brunei Darussalam, Cabo Verde, Canada, Comoros, Congo, Costa Rica, Denmark, Dominica, El Salvador, Fiji, Finland, France, Gabon, Gambia, Germany, Grenada, Guinea-Bissau, Guyana, Honduras, Hong Kong, Iraq Iceland, Ireland, Israel , Italy, Kiribati and Luxembourg on Albania and found that all regression come up with significant results. Even though all the series are independent of each other. As we can see in table 1 which is consists on linear regression results, all the GDP series are having statistically significant relations. Table consists on the coefficient values and the P values are in parenthesis. The P-values indicating that all the relation are highly significant even at 1% level of significance.

The table 3 shows the residual analysis of linear regression model. It shows that all the results of autocorrelation are significant at 1% level of significance. While the LM test for heteroskedasticity results are also significant, expect 15 cases. It means out of 36 regression only 15 regression residuals facing heteroskedasticity. Nonetheless, the table 4 is presenting the residual analysis of ARDL model. As we can see that the autocorrelation test are insignificant at 5% except Argentina and Brunei Darussalam, they are insignificant at 1%. The Hetroscedasticity test statistics are insignificant at 5% except Argentina, Canada but in case of Canada it is insignificant at 1%.

These results infer that ARDL model significantly reduced the probability of spurious regression from 100% to approximately 5%. It also rejects the common misconception about the spurious regression that it is only prevails due to unit root. Nevertheless, the missing relevant variable is a major cause of spurious regression. As we introduced the lag values the probability of spurious regression reduced significantly.

Table 1 Results after running Simple Linear Regression Model

Countries	ATG	ARG	AUT	BHS	BHR	BRB	BEL	BWA	BRA	BRN	CPV	CAN
Coeffi	173.456 [0.0000]	0.845535 [0.0000]	2.79251 [0.0000]	115.61 [0.0000]	55.0787 [0.0000]	1176.54 [0.0000]	2.44742 [0.0000]	6.94373 [0.0000]	0.374535 [0.0000]	89.9904 [0.0000]	3.37695 [0.0000]	0.355487 [0.0000]
Countries	**COM**	**COG**	**CRI**	**DNK**	**DMA**	**SLV**	**FJI**	**FIN**	**FRA**	**GAB**	**GMB**	**DEU**
Coeffi	9.22622 [0.0000]	0.558613 [0.0000]	0.245078 [0.0000]	0.299897 [0.0000]	682.991 [0.0000]	69.922 [0.0000]	158.877 [0.0000]	4.29561 [0.0000]	0.278377 [0.0000]	0.127738 [0.0000]	32.1748 [0.0000]	0.20898 [0.0000]
Countries	**GRD**	**GNB**	**GUY**	**HND**	**HKG**	**ISL**	**IRQ**	**IRL**	**ISR**	**ITA**	**KIR**	**LUX**
Coeffi	324.926 [0.0000]	1.98374 [0.0000]	2.03643 [0.0000]	3.9283 [0.0000]	0.2911 [0.0000]	0.374548 [0.0000]	0.0050 [0.0000]	2.78307 [0.0000]	0.630868 [0.0000]	0.319785 [0.0000]	5020.83 [0.0000]	13.7727 [0.0000]

Table 2 Results after employing ARDL model

Countries	ATG	ARG	AUT	BHS	BHR	BRB	BEL	BWA	BRA	BRN	CPV	CAN
Coeffi	2.2113 [0.1271]	2.3136 [0.0984]	1.9177 [0.1515]	3.3884 [0.024]*	2.3568 [0.0949]	0.97170 [0.4211]	1.5636 [0.2220]	1.2890 [0.2991]	2.6769 [0.0679]	2.9427 [0.0517]	2.5011 [0.0692]	1.7673 [0.1781]
Countries	**COM**	**COG**	**CRI**	**DNK**	**DMA**	**SLV**	**FJI**	**FIN**	**FRA**	**GAB**	**GMB**	**DEU**
Coeffi	2.5938 [0.0741]	1.0733 [0.3776]	2.4079 [0.0900]	0.55250 [0.6510]	1.3533 [0.2789]	2.5329 [0.0789]	3.9684 [0.018]*	2.7890 [0.0605]	2.4591 [0.0905]	0.75471 [0.4795]	1.7834 [0.1751]	1.2943 [0.2900]
Countries	**GRD**	**GNB**	**GUY**	**HND**	**HKG**	**ISL**	**IRQ**	**IRL**	**ISR**	**ITA**	**KIR**	**LUX**
Coeffi	2.5668 [0.0947]	1.8490 [0.1631]	2.7830 [0.0609]	2.2760 [0.1034]	1.4923 [0.2399]	2.1955 [0.1301]	2.1649 [0.1163]	2.8124 [0.0591]	2.2770 [0.1033]	0.19603 [0.8231]	2.5335 [0.0666]	3.0034 [0.0658]

The coefficient values are given in table 1 and 2. The P values are in square brackets. The table 2 consists on the F-stat coefficient value which is used to check the joint significance of independent variable and its lag values. Under null hypothesis H0: restrictions are valid. * shows the values which are significant at less than 5% level of significance.

Table 3 Residual Analysis after simple linear regression Model

Countries	ATG	ARG	AUT	BHS	BHR	BRB	BEL	BWA	BRA	BRN	CPV	CAN
AR (1-2)	108.46 [0.0000]	37.166 [0.0000]	74.421 [0.0000]	50.957 [0.0000]	44.826 [0.0000]	28.088 [0.0000]	58.430 [0.0000]	47.607 [0.0000]	46.912 [0.0000]	70.425 [0.0000]	42.454 [0.0000]	93.299 [0.0000]
Hetro test	4.8584 [0.0144]	11.807 [0.0002]	3.0080 [0.0650]	3.4207 [0.0464]	10.664 [0.0003]	1.8093 [0.1818]	4.6607 [0.0176]	6.8516 [0.0037]	1.5076 [0.2383]	0.38325 [0.6850]	12.618 [0.0001]	0.42721 [0.6564]
Countries	**COM**	**COG**	**CRI**	**DNK**	**DMA**	**SLV**	**FJI**	**FIN**	**FRA**	**GAB**	**GMB**	**DEU**
AR (1-2)	23.463 [0.0000]	27.073 [0.0000]	48.132 [0.0000]	192.11 [0.0000]	49.202 [0.0000]	92.324 [0.0000]	70.445 [0.0000]	52.093 [0.0000]	179.65 [0.0000]	108.46 [0.0000]	37.166 [0.0000]	176.23 [0.0000]
Hetro test	3.9430 [0.0306]	0.47118 [0.6290]	12.139 [0.0001]	2.6543 [0.0874]	6.0150 [0.0065]	1.4328 [0.2550]	0.93896 [0.4026]	0.56898 [0.5723]	1.9298 [0.1634]	4.8584 [0.0144]	11.807 [0.0002]	3.0440 [0.0631]
Countries	**GRD**	**GNB**	**GUY**	**HND**	**HKG**	**ISL**	**IRQ**	**IRL**	**ISR**	**ITA**	**KIR**	**LUX**
AR (1-2)	51.204 [0.0000]	44.374 [0.0000]	60.715 [0.0000]	38.748 [0.0000]	43.418 [0.0000]	46.786 [0.0000]	17.337 [0.0000]	70.961 [0.0000]	56.707 [0.0000]	271.27 [0.0000]	36.165 [0.0000]	55.628 [0.0000]
Hetro test	0.84706 [0.4390]	1.8925 [0.1688]	1.7900 [0.1849]	9.6925 [0.0006]	6.4124 [0.0049]	0.32064 [0.7282]	8.8652 [0.0010]	0.0082 [0.9917]	9.2912 [0.0008]	4.0579 [0.0279]	0.54471 [0.5858]	4.9846 [0.0138]

Table 4 Residual Analysis after ARDL Model

Countries	ATG	ARG	AUT	BHS	BHR	BRB	BEL	BWA	BRA	BRN	CPV	CAN
AR (1-2)	3.2957 [0.0530]	4.8584 [0.0144]	2.8581 [0.0770]	1.8220 [0.1834]	1.9423 [0.1653]	1.5511 [0.2325]	4.0584 [0.1144]	2.4124 [0.1110]	2.5211 [0.1014]	3.5946 [0.0431]	2.0736 [0.1477]	0.91149 [0.4154]
Hetro test	0.32156 [0.9195]	173.456 (0.000)	1.7750 [0.1287]	1.7026 [0.1461]	1.4521 [0.2257]	1.0165 [0.4624]	173.456 (0.8006)	1.3759 [0.2572]	1.9732 [0.0911]	0.83462 [0.6020]	0.74033 [0.6804]	2.5478 [0.0341]
Countries	**COM**	**COG**	**CRI**	**DNK**	**DMA**	**SLV**	**FJI**	**FIN**	**FRA**	**GAB**	**GMB**	**DEU**
AR (1-2)	2.6788 [0.0891]	2.5176 [0.1017]	2.2340 [0.1289]	1.6776 [0.2080]	2.4530 [0.1073]	2.6688 [0.0898]	1.9338 [0.1665]	2.9251 [0.0730]	4.1468 [0.5284]	3.2957 [0.0530]	2.0250 [0.1539]	2.4418 [0.1067]
Hetro test	2.1383 [0.0684]	0.85294 [0.5871]	1.4666 [0.2201]	0.90939 [0.5422]	1.3667 [0.2612]	1.4615 [0.2221]	2.6555 [0.0285]	2.0190 [0.0841]	3.0658 [0.2147]	0.32156 [0.9195]	2.4831 [0.0380]	2.1177 [0.0869]
Countries	**GRD**	**GNB**	**GUY**	**HND**	**HKG**	**ISL**	**IRQ**	**IRL**	**ISR**	**ITA**	**KIR**	**LUX**
AR (1-2)	3.0651 [0.0638]	3.3840 [0.0507]	3.0038 [0.0670]	3.3708 [0.1499]	2.2546 [0.1267]	3.1490 [0.0596]	1.0909 [0.3520]	2.4140 [0.1109]	1.6316 [0.2166]	3.2739 [0.0539]	0.92427 [0.4117]	3.2180 [0.0564]
Hetro test	0.82311 [0.5628]	1.2434 [0.3214]	2.2719 [0.0691]	0.26387 [0.9486]	1.1274 [0.3885]	0.60599 [0.7231]	0.68400 [0.7276]	1.6800 [0.1520]	1.9253 [0.0990]	1.5242 [0.2112]	2.0197 [0.0848]	1.3186 [0.2857]

AR null hypothesis H0: There is autocorrelation. LM test for Hetroskedastic with null hypothesis H0: There is no hetroskedasticity

The reason behind the spurious regression is that when the potential variable is missing from the regression, then the irrelevant variable acts as a proxy of potential variable. It captures the effect of potential variables and then the results would be significant. If we start with ARDL model it will overtake the problem of missing variable. Even it can be shown that the results in Granger and Newbold (1974) experiments were significant only due to missing lag values. See, (section, 5.1).

5.1 Simulation results with nonstationary series of integrated order 1

We have generated two independent autoregressive random nonstationary series of integrated order 1 by using our data generating process given above, imposing restrictions $\theta_{12} = \theta_{21} = 0$ and $\rho = 0$. Where $X_t$and Y_t both are expressed by their own lag values and the coefficients of lag values $\theta_1 = \theta_2 = 1$.

$$Y_t = \alpha_0 + \theta_1 Y_{t-1} + \varepsilon_{yt} \quad \text{……….. (15)}$$

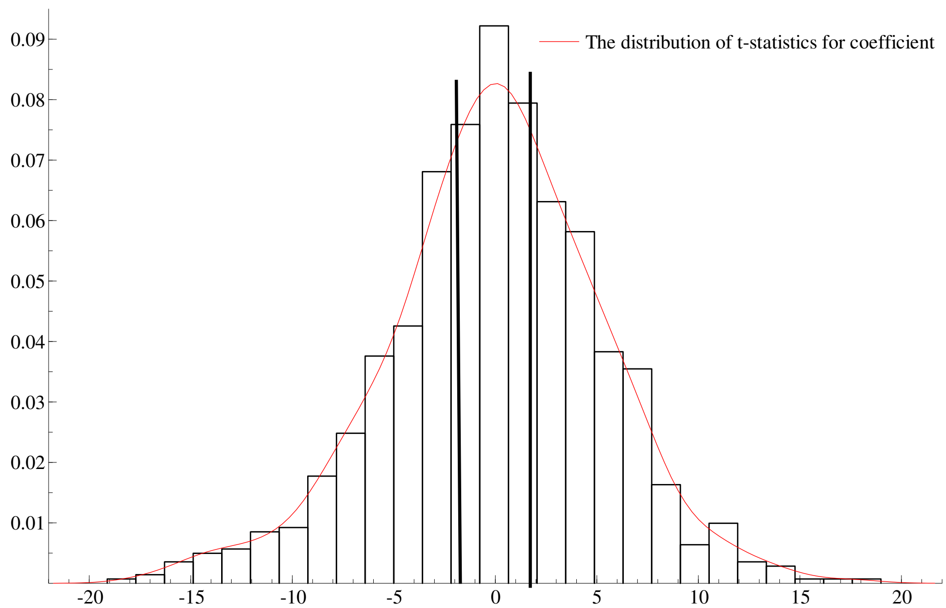
$$X_t = \alpha_1 + \theta_2 X_{t-1} + \varepsilon_{xt} \quad \text{……….. (16)}$$

We are using sample size of 50 observations. We regress X_t on Y_t by using simple linear regression model. The equation is following

$$Y_t = a + \beta_1 X_t + \varepsilon_{yt} \quad \text{……….. (17)}$$

Monte Carlo simulation are used for simulations of results. We simulated the t-stat value of X variable 1000 time and the results are explained through figure 1 given below. The vertical lines are indicating the asymptotic critical value at 5% nominal level of significance which is 1.96. It is noticeable that wider area of distribution lies in rejection region. The regression is estimated at 5% nominal level of significance but after 1000 time simulations of t-statistics for coefficient, we got the probability of spurious regression is increased from 5% to 67%. It means that we got 670 times significant results out 1000 instead of 50 times out of 1000.

Figure 1: The distribution of t-statistics for coefficient of X (t)

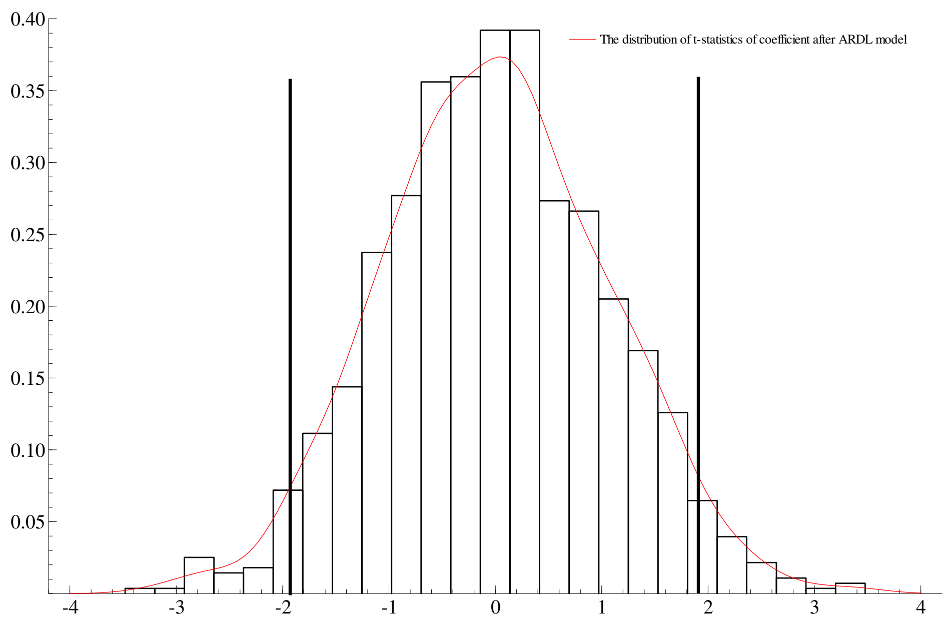


These spurious results are due to missing variable because we did not include the lag values of variables as an independent variable. Now, if we include the lag value as an independent variable then the model become ARDL (1, 1). We can see that the ARDL (1, 1) model reduce the probability of spurious regression and eliminate the chances of spurious regression. The equation is following

$$Y_t = a + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 Y_{t-1} + \varepsilon_{yt} \quad \ldots\ldots\ldots\ldots (18)$$

Figure 2 shows the distribution of t-statistics for coefficient X_t after ARDL (1, 1) model. The vertical lines are indicating the asymptotic critical value at 5% nominal level of significance which is 1.96. It is noticeable that smaller area of distribution lies in rejection region. The regression is estimated at 5% nominal level of significance, after 1000 time simulation of t-statistics for coefficient, the probability of spurious regression recorded to be approximately 5%. This directs that ARDL can be used as a treatment of spurious regression with nonstationary series. Same experiments were done in Granger and Newbold (1974) experiments and they did not consider the lag dynamic. That's why they got spurious results.

Figure 2: The distribution of t-statistics of coefficient of X(t) after ARDL model



5.2 Simulation results with nonstationary series of integrated order 2

We have generated two independent autoregressive random nonstationary series of integrated order 2 by using our data generating process given above, imposing restrictions $\theta_{12} = \theta_{21} = 0$ and $\rho = 0$. Where $X_t$and Y_t both are expressed by their own lag values and the coefficients of lag values $\theta_1 = \theta_2 = 1$. We are using sample size of 50 observations. We regress X_t on Y_t by using simple linear regression model. the equation is following

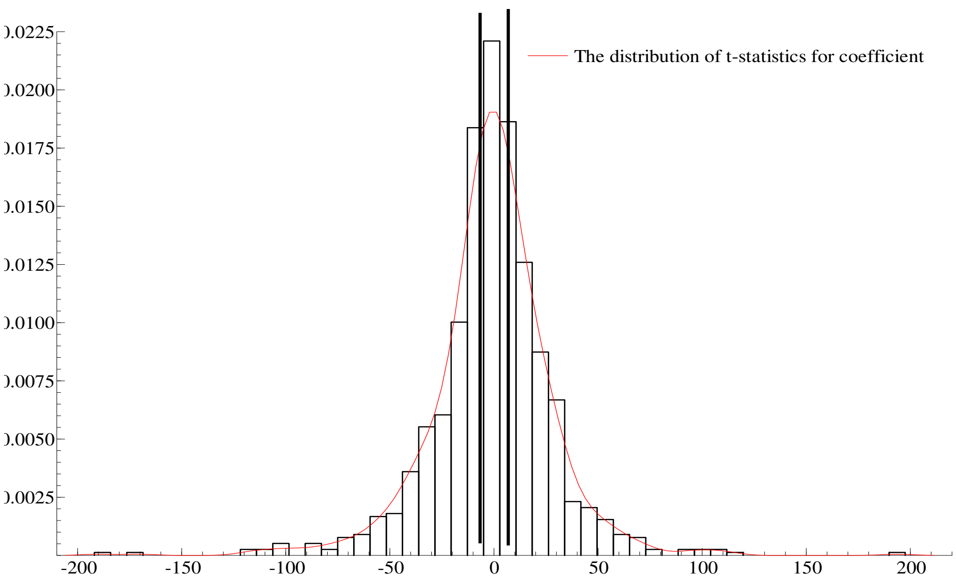
$$Y_t = Y_{t-1} + Y_{t-2} + \varepsilon_{yt} \quad \text{……….. (19)}$$

$$X_t = X_{t-1} + X_{t-2} + \varepsilon_{xt} \quad \text{……….. (20)}$$

There is no third variable involved in the construction of both variables. We regressed $X_t$on Y_t and $Y_t$on X_t without involving their lag values in regression analysis.

$$Y_t = a + \beta_1 X_t + \varepsilon_{yt} \quad \text{……….. (21)}$$

Figure 3: The distribution of t-statistics for coefficient of X (t)



The vertical lines are indicating the asymptotic critical value at 5% nominal level of significance which is 1.96. It is noticeable that wider area of distribution lies in rejection region. The regression is estimated at 5% nominal level of significance but after 1000 time simulation of t-statistics for coefficient, we got the probability of spurious regression is 92%. It means that the probability of spurious regression is increased 87%.

These spurious results are due to missing variable because we did not include the lag values of variables as an independent variable. Now at first, we include the one lag value of X and Y as an independent variables then the model become ARDL (1, 1). The equation is following

$$Y_t = a + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 Y_{t-1} + \varepsilon_{yt} \quad \ldots\ldots\ldots\ldots (22)$$

Figure 4: The distribution of t-statistics for coefficient of X (t) after ARDL (1, 1)

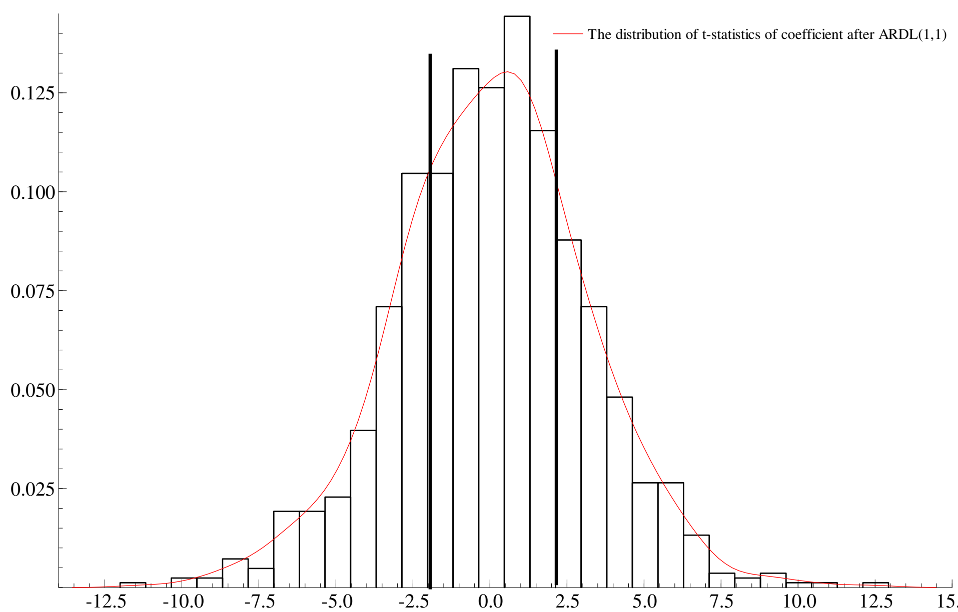
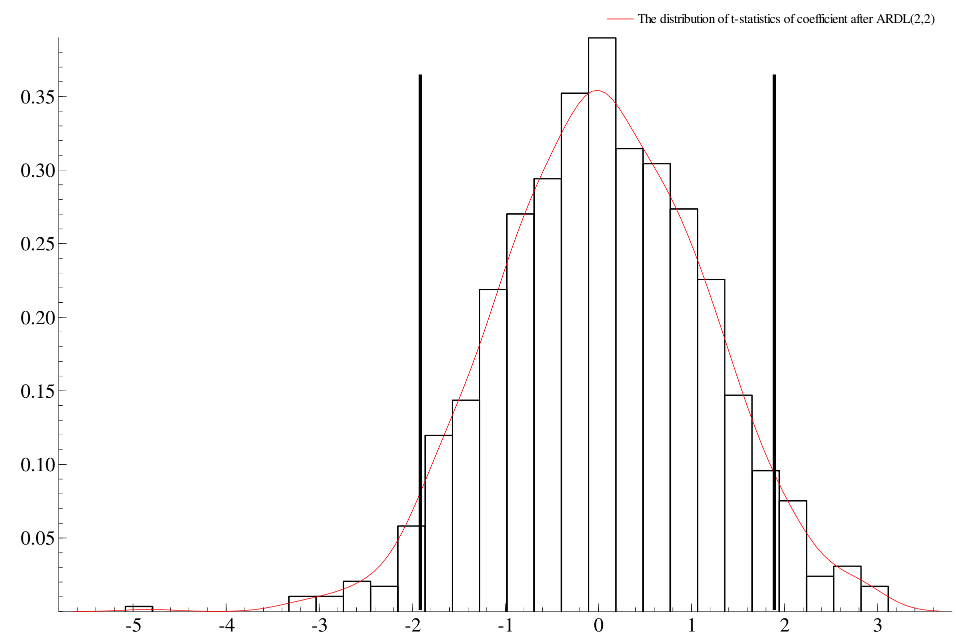


Figure 4 shows the distribution of t-statistics for coefficient of linear regression. The vertical lines are indicating the asymptotic critical value at 5% nominal level of significance which is 1.96. It is noticeable that wider area of distribution lies in rejection region. The regression is estimated at 5% nominal level of significance but after 1000 time simulation of t-statistics for coefficient, we got actual level of significance which is 50%. It means ARDL (1, 1) reduced the probability of spurious regression from 87% to 45%.

These spurious results are due to missing variable because we did not include the second lag values of variables as an independent variable. Now, we also include the second lag values of X and Y as an independent variable then the model become ARDL (2, 2). The equation is following

$$Y_t = a + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 Y_{t-1} + \beta_2 X_{t-2} + \beta_3 Y_{t-2} + \varepsilon_{yt} \quad \text{..........} \ (23)$$

Figure 5: The distribution of t-statistics for coefficient of X(t) after ARDL (2, 2)



It is noticeable that wider area of distribution lies in rejection region. The regression is estimated at 5% nominal level of significance but after 1000 time simulation of t-statistics for coefficient, we got the probability of spurious regression is 7%. This indicates a distortion of only 2%. It means that we got 70 times significant results out 1000 instead of 50 times out of 1000. This directs that ARDL can be used as a treatment of spurious regression in case of higher integrated order time series.

6. Conclusion

The Unit root and Cointegration analysis are the only ways to circumvent the spurious regression in case of nonstationarity in conventional Econometrics. Nevertheless, these procedures are equally unreliable because of some specification decisions like, autoregressive lag length choice, choice of the deterministic part, structural breaks and innovation process distribution. After having reviewed an excessive amount of available literature and inferences, we have been able to conclude that it is the missing variable (lag values) that are the major cause of spurious regression in all the cases therefore an alternative way to look at the problem of spurious regression takes us back to the missing variable which further leads to ARDL Model. The results are also providing justification that ARDL model can be used as a remedy of spurious regression.

Reference

Atiq-ur-Rehman, A. U. R., & Zaman, A. (2008). Model specification, observational equivalence and performance of unit root tests.

Ahking, F. W. (2002). Model mis-specification and Johansen's co-integration analysis: an application to the US money demand. *Journal of Macroeconomics, 24(1),* 51-66.

Agunloye, O. K., & Shangodoyin, D. K. (2014). Lag Length Specification in Engle-Granger Cointegration Test: A Modified Koyck Mean Lag Approach Based on Partial Correlation. *Statistics in Transition new series,* 15(4).

Charemza, W. W., & Deadman, D. F. (1997). New directions in econometric practice. *Books.*

Choi, C. Y., Hu, L., & Ogaki, M. (2004). A spurious regression approach to estimating structural parameters. *Ohio State University Department of Economics Working Paper,* (04-01).

Carrasco Gutierrez, C. E., Castro Souza, R., & Teixeira de Carvalho Guillén, O. (2009). Selection of optimal lag length in cointegrated VAR models with weak form of common cyclical features.

Chaouachi, K. (2013). False positive result in study on hookah smoking and cancer in Kashmir: measuring risk of poor hygiene is not the same as measuring risk of inhaling water filtered tobacco smoke all over the world. *British journal of cancer, 108(6),* 1389.

Davidson, J. E., Hendry, D. F., Srba, F., & Yeo, S. (1978). Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *The Economic Journal,* 661-692.

DeJong, D. N., Nankervis, J. C., Savin, N. E., & Whiteman, C. H. (1992). The power problems of unit root test in time series with autoregressive errors. *Journal of Econometrics, 53(1-3),* 323-343.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association, 74(366a),* 427-431.

Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica: *Journal of the Econometric Society*, 1057-1072.

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251-276.

Engle, R. and Yoo Sam (1991). Forecasting and Testing in Co-integrated Systems, In Engle and Granger (eds.), Long Run Economic Relationships. Readings in Cointegration, *Oxford University Press, New York*, 237-67.

Frey, B. S. (2002). Inspiring economics: Human motivation in political economy. *Edward Elgar Publishing.*

Granger, C. W., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of econometrics*, *2*(2), 111-120.

Hashimzade, N., & Thornton, M. A. (Eds.). (2013). Handbook of research methods and applications in empirical macroeconomics. *Edward Elgar Publishing*.

Hassler, U. (2003). Nonsense regressions due to neglected time-varying means. *Statistical Papers, 44(2),* 169-182.

Hendry, D. F. (1980). Econometrics-alchemy or science?. *Economica,* 387-406.

Hendry, D. F., & Richard, J. F. (1983). The econometric analysis of economic time series. *International Statistical Review/Revue Internationale de Statistique,* 111-148.

Hendry, D. F., Pagan, A. R., & Sargan, J. D. (1984). Dynamic specification. *Handbook of econometrics, 2,* 1023-1100.

Höfer, Thomas; Hildegard Przyrembel; Silvia Verleger (2004). New evidence for the Theory of the Stork. *Paediatric and Perinatal Epidemiology. 18 (1):* 18–22.

Juselius, K. (1992). Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for UK. *Journal of econometrics, 53(1-3),* 211-244.

Leybourne, S. J., & Newbold, P. (2003). Spurious rejections by cointegration tests induced by structural breaks. *Applied Economics, 35(9),* 1117-1121.

Nelson, C. R., & Plosser, C. R. (1982). Trends and random walks in macroeconmic time series: some evidence and implications. *Journal of monetary economics*, *10*(2), 139-162.

Plosser, C. I., & Schwert, G. W. (1978). Money, income, and sunspots: measuring economic relationships and the effects of differencing. *Journal of Monetary Economics, 4(4),* 637-660.

Perron, P. (1990). Testing for a unit root in a time series with a changing mean. *Journal of Business & Economic Statistics, 8(2),* 153-162.

Phillips, P. C. (1986). Understanding spurious regressions in econometrics. *Journal of econometrics, 33(3),* 311-340.

Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika,* 335-346.

Pesaran, M. H., Shin, Y., & Smith, R. J. (1996). Testing for the' Existence of a Long-run Relationship' (No. 9622). *Faculty of Economics, University of Cambridge.*

Pesaran, M. H. (1997). The role of economic theory in modelling the long run. *The Economic Journal, 107(440),* 178-191.

Pesaran, M. H., & Smith, R. (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of econometrics, 68(1*), 79-113.

Rehman, A. U., & Malik, M. I. (2014). The modified R a robust measure of association for time series. *Electronic Journal of Applied Statistical Analysis, 7(1),* 1-13.

Sapsford, Roger; Jupp, Victor, eds. (2006). Data Collection and Analysis. Sage. ISBN 0-7619-4362-5.

Simon, H. A. (1954). Spurious correlation: a causal interpretation. *Journal of the American statistical Association, 49(267),* 467-479.

Sun, Y. (2004). A convergent t-statistic in spurious regressions. *Econometric Theory, 20(05),* 943-962.

Su, J. J. (2008). A note on spurious regressions between stationary series. *Applied Economics Letters, 15(15),* 1225-1230.

Schwert, G. W. (2002). Tests for unit roots: A Monte Carlo investigation. *Journal of Business & Economic Statistics, 20(1),* 5-17.

Ventosa-Santaulària, D. (2009). Spurious regression. *Journal of Probability and Statistics*, 2009.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between Time-Series?--a study in sampling and the nature of time-series. *Journal of the royal statistical society*, *89*(1), 1-63.