Investment (Expenditure on) in Education and National Income (GDP) in India

Time Series Analysis for Cointegration and Causality

Venkatanarayana Motkuri

Research Cell on Education (RSEPPG)
Centre for Economic and Social Studies
(ICSSR, Ministry of Education, Government of India)
Hyderabad-16

November 2020

Research Cell on Education

Research Cell for Studies in Education Policy, Planning and Governance (RSEPPG) in Telangana State at the **Centre for Economic and Social Studies** (CESS) is set up in July 2020 with the support of **Telangana State Council for Higher Education** (TSCHE). The predominant objective of the ***Research Cell on Education*** (RSEPPG) is to comprehensively study the gamut of issues and challenges in the education system/sector at all levels in a holistic perspective while ***focusing on higher education*** in the state and in the country. Research studies are to be thematically organized around five dimensions or foundational pillars of education system: *Access*, *Quality*, *Equity*, *Affordability and Accountability*. The impact of *state policy*, *funding*, *regulatory framework*, *educational standards* and *governance* on achieving these five dimensions of education system would be the focus of research studies. The Research Cell while providing policy inputs and support derived from evidence-based policy-oriented research output, assists the *TSCHE* and thereby the *Telangana State Government* in their endeavour for educational development in the state.

Activities:

- To organize seminars, workshops, discussions and conferences on topical issues related to Education;
- To conduct research studies focusing on education policy, funding, and governance;
- To conduct assessment and evaluation studies on initiatives and programmes with respect to education development; and
- To bring out research reports, policy briefs, and working papers along with research publications

Contact Details:
Research Cell on Education
Centre for Economic and Social Studies
(ICSSR, Ministry of Education, Government of India)
N.O. Campus, Begumpet,
Hyderabad – 16
Phone: 040 2340 2789 / 23416780 / 23416610-13
post@cess.ac.in

https://cess.ac.in/

Citation:
Motkuri, V. (November, 2020). Investment (Expenditure on) in Education and National Income (GDP) in India: Time Series Analysis for Cointegration and Causality, *CESS-RSEPPG Background Paper Series (BPS) #6*, Research Cell on Education (RSEPPG), Centre for Economic and Social Studies, Hyderabad.

Investment (Expenditure on) in Education and National Income (GDP) in India: Time Series Analysis for Cointegration and Causality

Venkatanarayana Motkuri[≠]
Research Cell on Education (RSEPPG), CESS, Hyderabad

Abstract
This paper examine the long-rum equilibrium relationship between investment (expenditure) on education in India and its GDP using their time series for the period of last seven decades (between 1950-51 and 2019-20) and applying cointegration procedures of time series analysis. The exploratory data analysis and the objective tests of cointegration (Granger and Johansen procedure) indicate that these times series in integrated of order two are cointegrated. It indicates that the annual changes in these time series are cointegrated rather the actual amounts. Causality test has shown that while public expenditure on education has an impact on GDP, the latter affects the private expenditure on education.

\- - -

I Context

Education has a value for its own sake (intrinsic) as well as an instrumental role/value in economic growth and development. It is crucial in the perspectives of human capital and human development and human rights. While the growth theories and human capital literature have shown considerably significant contribution of education and training of human resources to national income, returns to education literature has shown both the social and private returns to education (Balug, 1972). Higher rates of economic growth and higher levels of income along with individuals' well-being (material as well as non-material) are predicated on the educational development and the resultant increase in labour productivity along with certain other externalities of it (easy diffusion and adaptation of technology and changing skills needs) (Nelson and Phelps, 1966; Barro and Sala-i-Martin, 1995; Krueger and Lindahl, 1999; Hanushek and Kimko, 2000). Thus, the human capital perspective and the evidence of returns to education indicate that the expenditure on education (private and public) is more than consumption in-itself as it was considered in the past, beyond that it turned out to be an investment good (Schultz, 1961; Nelson and Phelps, 1966; Jorgenson and Fraumeni, 1992).

On the other hand the conventional wisdom has revealed that the expenditure (investment) on education depends on the level and growth of income. Higher levels of incomes and higher rate of growth in such incomes would facilitate better and more investment in education, sparing the children to study (in the private domain) and spending on their education (in both the public and private domains). The conventional wisdom considered that economic growth is a pre-condition for educational development and expenditure on education is consumption good. But they evolved over a period as a contributing factor in economic growth and as an

---

[≠] **Venkatanarayana Motkuri,** Ph.D., (venkatanarayan@gmail.com) is an Associate Professor and Coordinator of Research Cell on Education (RSEPPG), Centre for Economic and Social Studies (CESS), Hyderabad. This Research Study at RSEPPG of CESS is conducted with financial support of the ***Telangana State Council for Higher Education*** (TSCHE).

investment good. The expenditure on education as an investment gained economic importance, in the human capital perspective, with the Solow's growth model of 1950s and its growth accounting procedure followed by the growth puzzle studies of next two decades (1960s and 1970s) and the new growth theories of 1980s, all of which are complemented by the studies, since 1980s, on returns to education (Solow, 1956; Schultz, 1961; Becker, 1964; Dennis, 1972; Romer, 1988). Human capital was factored in the 'Residual Growth Hypothesis' that emerged with the Solow's growth accounting model. In other words, the residual of growth after accounting for two conventional factors of production (capital and labour), was explained by the factor of human capital (Dennis, 1972). Although human capital conceptually in general consists of education along with health and social capital, education is largely represents the human capital in growth accounting exercises. As mentioned above, there exists reasonably abundant literature of research studies based on cross-sectional and/or panel analysis evaluating relationship between educational development and economic growth across countries.

Therefore, the symbiotic relationship between the two appeared to be that it is not of one-way linear causality. There is a two-way or cyclic causality (each is a cause as well as an effect, of the other) between investment in education and income. A cyclic causality can be a sequential or simultaneous one. A sequential two-way or cyclic causality with amplification (or de-amplification) effects can lead to spiralling-up (down) causality, one factor causes the other one to move up (down) cyclically. Various methods and procedures for cross-sectional as well as time series along with that of panel-data analysis have been developed in understanding such relationships.

In this backdrop, the present paper examines the long-rum equilibrium relationship between investment (expenditure) on education in India and its GDP using their time series for the period of last seven decades (between 1950-51 and 2019-20) and applying cointegration procedures of time series analysis. There exist a reasonably abundant literature of research studies based on cross-sectional and/or panel analysis evaluating relationship between educational development and economic growth across countries and provinces within countries. Most of these studies have focussed on the outcome variable of education i.e. levels of schooling. Very few studies are based on the time series analysis evaluating the long-rum equilibrium relationship between investment (expenditure) on education and GDP in a specific country context or multiples. Importance of the present study can be seen in this context.

Data and Methods

For the cointegration analysis of investment in education and GDP, the ***public (budgeted)*** and ***private expenditure on education*** along with ***GDP*** estimates of Government of India are used. The public expenditure on education that is compiled by Ministry of Education (MoE), Government of India in its ***Analysis of Expenditure on Education*** is used. For the private expenditure on education it is Private Final Consumption Expenditure (PFCE) on education as estimated by the National Accounts Statistics (NAS), Government of India. These three time series in ***constant (2011-12) prices*** for the period of last seven decades (between 1950-51 and 2019-20) is used for the analysis. The analysis of time series checking the stationarity and cointegration is based on using the statistical computing environment that is **R**: *A language* *and environment for statistical computing* that is developed by ***R Foundation for Statistical Computing*** (R Core Team, 2020).

The pretesting exercise for time series analysis is to check the stationarity of the time series. Based on exploratory data analysis through spread plots of time series and their autocorrelation function (ACF) one could understand that the pattern of the time series whether it is non-stationary and reflecting stationarity. Further it can be checked with objective testing procedure. The present exercise has followed both the procedures exploratory analysis observing the pattern through graphic presentation as well as the testing objectively for the stationnarity as well for the cointegration (see Appendix I and II for details).

II Results and Observations

Stationarity is the pre-condition for modelling the time series, a non-stationary time series is not suitable for predictions. As very often we find most of the time series are non-stationary, it needs to be transformed into a stationary series to make use of such series for any further analysis. Differencing is one procedure to transform a non-stationary series into a stationary one, following the Box-Jenkin method (Box and Jenkin, 1970).

In this regard, the results of stationarity tests reported here are that stationarity/unit root tests applied to first difference of log transformed three univariate time series: public and private expenditure on education and GDP. All the tests (of independence, unit root or trend stationarity) applied to said time series indicate that stationarity is achieved with the first difference of log transformed time series (Table 1). It means that all the three time series is integrated of order one: I(1) processes.

For any statistical test, significance (at <=5% confidence level) of a test-statistics means to reject the null hypothesis and accept the alternative hypothesis. It means that when a test-statics confirm us to reject the null hypothesis at <=5% confidence level, there is 5% or less chance that we reject a true null hypothesis. Herein what is that we have framed in our null hypothesis is a matter of concern. In respect of stationarity tests, while variants of DF/ADF test have null hypothesis that assumes time series is a non-stationary process because whether the autoregressive (AR) residuals are serially correlated (autocorrelated) or there is a unit root in the process. The alternative hypothesis for these tests is that time series is a stationary process. Here significance of a test-statistics (i.e. computed value is greater than reference statistics in the distribution) means to reject the null hypothesis of non-stationarity and accept the alternative hypothesis of stationarity. In this regard all the tests and their model variants (constant and/or trend) have confirmed the stationarity of the three times series process

In case of Ljung-Box and KPSS test, it is opposite. The Ljung-Box has a null hypothesis that AR residuals are independent (not serially (or auto) correlated) and the alternative hypothesis is *they are not independent*. For KPSS test, the null hypothesis is that a time series is stationary process and the alternative hypothesis is that it has a unit root in the process (i.e. non-stationary). In case of Ljung-Box test as it can be observed that p-values of three time series are slightly higher than 5% indicating that not to reject null hypothesis (of independence) and not to accept the alternate hypothesis (of no independence). Similarly is the case of KPSS test-statistics significance is not allowing to reject null hypothesis.

Table 1: Results of Stationarity Tests: Independence, Unit Root and Level and Trend Stationary – Result for First Differenced Series

Sno	Test Variants	Model	GDP		PFCE-Edn		PE-Edn	
			t	p	t	p	t	p
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	Ljung-Box	None	13.002	0.072	20.254	0.005	13.922	0.052
2	ADF	None	-1.5437	0.100	-2.2585	0.010	-1.1871	0.050
		C	-5.4556	0.001	-4.1537	0.001	-3.9467	0.001
		C&T	-7.2981	0.001	-5.5171	0.001	-3.9120	0.001
3	PP	None	-11.000	0.020	-11.200	0.019	-19.100	0.010
		C	-71.900	0.010	-46.948	0.010	-60.500	0.010
		C&T	-72.544	0.010	-46.948	0.010	-68.417	0.010
4	ADF-GLS	C&T	-5.3283	0.001	-3.0564	0.001	-4.5947	0.001
		C	-2.9665	0.001	-3.0204	0.001	-31331	0.001
5	KPSS	Level	1.1112	0.010	0.0983	0.100	0.9156	0.010
		Trend	0.0747	0.010	0.0976	0.010	0.0836	0.010

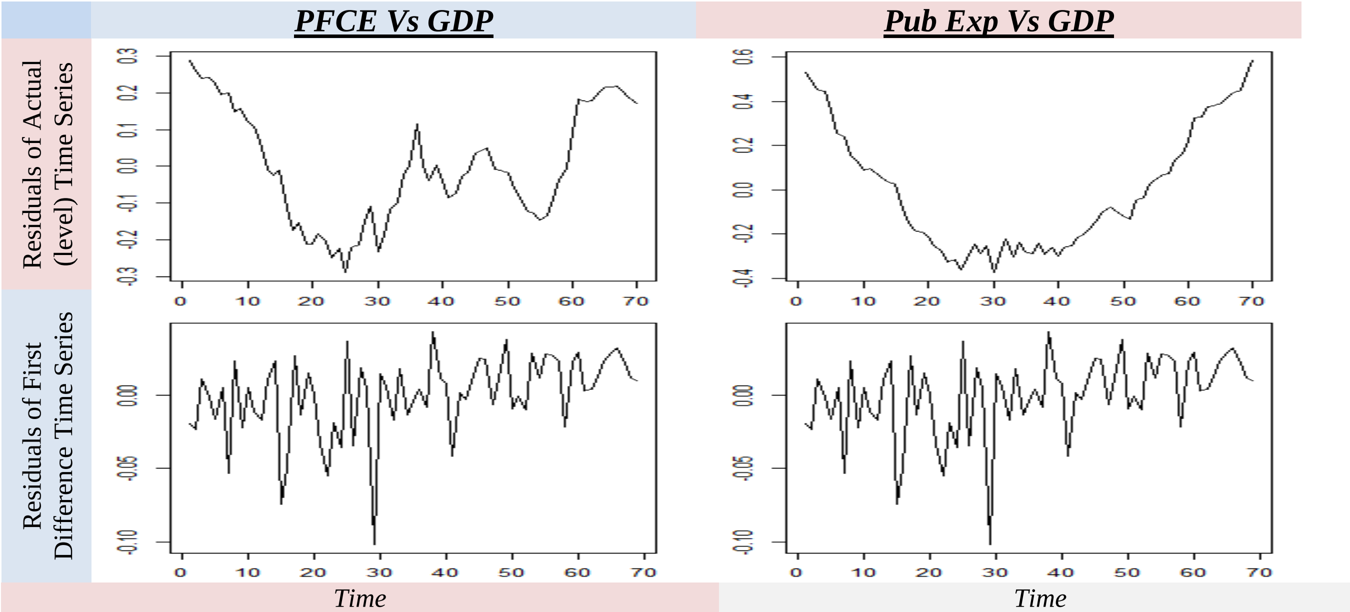
***Notes**: 1. ADF – Augmented Dickey-Fuller; KPSS – Kwiatkowski–Phillips–Schmidt–Shin test; PP – Philip-Perron test; 2. GDP – Gross Domestic Product of India; PFCE-Edn – Private Final Consumption Expenditure on Education; PE-Edn – Public (Budgeted) Expenditure on Education; 3. C – Constant; T – Trends; C&T – Constant and Trends; t – computed test statistics; p – probability or significance level (p-value).*

***Source**: Author's estimate.*

To reiterate once again, all of the objective tests confirm the stationarity of the series with first difference in all the three time series: private and public expenditure on education and GDP. These tests in fact indicate that these three series individually are integrated of order one: I(1). It means that they are all likely to be cointegrated. The visual graphic presentation (in Appendix I), however, indicated second difference of all these three time series are more concrete in confirming stationarity in their process. In fact it is reflected in testing cointergration where I(1) series have failed to confirm their cointegration but the I(2) series have confirmed.

As the linear combination of two or more time series integrated of same order are considered to be cointegrated, if their residuals are found to be stationary. In this regard, the exploratory analysis modelled the time series with bivariate OLS regression equation that regressing ***investment in education against GDP*** that generated residuals in the process. The visual signal-spread (graph) of residuals have not indicated stationarity (see first panel of Figure 1), so that their cointegration is not confirmed. Stationarity of model residuals is a pre-condition for the cointegration of time series in the model. The objective tests (ADF) applied to the residuals generated also has not confirmed stationarity (Table 2).

Figure 1: Signal-Spread of Residuals



Note: **PFCE** *and* **Pub Exp** *are respectively private and public investment in education;*

Source: *Author's Estimate,*

But when the first differenced series is modelled, regressing the first differenced time series of ***investment in education against that of GDP***, the distribution of residuals of first differenced time series have indicated the stationarity. Both the visual signal-spread (see second panel of Figure I) and objective test (ADF) have confirmed their stationarity hence they are cointegrated (Table 2). In fact it is indicating that rather than the level (actual value) of the time series (GDP and Education Expenditure), the time series with annual change (first difference) in the selected time series (public and private expenditure on education and GDP) have certain long-run relationship. Their second differences are more stationary and cointegrated.

Table 2: Testing Bivariate Residuals for Cointegration: ADF Test

Sno	Series Tested	Residuals of GDP and Private Investment		Residuals of GDP and Public Investment	
		t	p	t	p
1	*2*	*3*	*4*	*5*	*6*
1	Residuals of Actual Time Series (Level Data)	-1.7261	0.055	-0.4053	0.100
2	Residuals of First Differenced Time Series	-5.3571	0.001	-5.6734	0.001

Note: *t is t-test statistic; p represent significance level*

Source: *Author's Estimate.*

The Jahansen test based on Vector Autoregressive (VAR) model which is an extension of vector-valued AR (p), has a null hypothesis that there is no cointegration among the time series and the alternative is there is such cointegration. Unlike Engle-Granger procedure, an advantage with Johansen method is that fixing a prior the dependent variable is not necessary. Significance (i.e. p-value <5%) of test statistics is anyway to reject the null hypothesis and accept the alternative, it means that there exists cointegration. Results of both the variants of Johansen method test (Trace and Max. Eigen Value tests) in fact suggest to reject null hypothesis of r=0 but not to reject the null hypothesis of r<=1 especially in case of cointegration of public investment in education and GDP. It means there is at least one cointegrating vector of the time series even in the latter case.

Table 3: Testing Cointegration: Johansen Test

Sno	Test Variant		GDP and Private Investment		GDP and Public Investment	
			T	P	t	p
1	*2*		*3*	*4*	*5*	*6*
1	Johansen Trace Test	r=0	27.23	0.010	33.50	0.010
		r<=1	8.07	0.100	16.68	0.010
2	Johansen Maximum Eigen Value Test	r=0	19.66	0.050	16.82	0.050
		r<=1	8.07	0.100	16.62	0.010

***Note**: T- Test Statistics; p – significance level*

***Source**: Author's Estimate.*

Along with these standard cointegration tests that procedures (developed by Engle-Grenger and that by Johansen), the Autoregressive Distributed-Lag (ARDL) model based Bound Test for cointegration developed by Pesaran and others (2001) is considered to be an advanced test and has certain advantages. Unlike previous method, the Bound Test for cointegration does not require pre-testing of time series level variables for their stationarity and order of integration (Pesaran *et al.*, 2001). It also confirms the same, cointegration of time series of interest here.

In all, with the first difference of all three time series are integrated of order one and are stationary. By the Granger procedure they are cointegrated (for the series integrated of order one: I(1)). It indicates that the annual change in these time series are cointegrated rather the actual values/amounts (level). Johansen method also confirms it for the time series integrated of order one only. These results are, in fact, indicating the long-run equilibrium relationship between the change in educational investment and national income (GDP).

Causality: Granger's 'No Causality' (Null) Hypothesis Test

Cointegration testing (both the Engel-Granger and Johansen procedures) has established that there is a long-run equilibrium relationship between public and private investment in education to that of country GDP. Given such confirmation it is also important to understand the direction of causality. None of the cointegration tests establishes the direction of causality. In this regard, Granger Causality test is performed for three time series and the resulted are presented in the following Table-4.

Table-4: Granger Causality Test Results and Decision

Sno	Causality	F	p	Decision
1	PFCE does not cause GDP	0.7582	0.522	Do not reject
2	PEE does not cause GDP	4.6479	0.005	Reject
3	GDP does not cause PFCE	3.3940	0.023	Reject
4	GDP does not cause PEE	2.3041	0.859	Do not reject
5	PFCE does not cause PEE	1.3689	0.261	Do not reject
6	PEE does not cause PFCE	1.1259	0.346	Do not reject

***Note: 1**. PFCE – Private Final Consumption Expenditure on Education; PEE – Public Expenditure on Education; GDP – Gross Domestic Product; **2**. Both the direct Granger Causality test and the VAR based test for the same is performed and both have shown same results.*

***Source**: Author's estimation*

Granger causality test is performed for six combinations of three times series. It indicates us that country GDP has an impact on private investment (expenditure) in education whereas public expenditure in education has an impact on GDP (Table-4). However, there is no direct causality in either direction found between private and public investment. The direction of causality observed provides a tip/clue for the *path analysis* wherein while the *public expenditure on education influences the country's GDP which in turn influences the private expenditure on education* in the country.

GDP and Investment in Education: Long-Run Equilibrium/Relationship and Short-run Dynamics

For understanding impact or effect of one time series on the other (cointegrated ones), although the ordinary least square (OLS) linear models may fit better if is a stationery time series, it is not, however, considered as a better method as many a times non-stationarity is the usual feature of time series and hence such model results must be spurious ones (Granger, 1969; Granger and Newbold, 1978; Engle and Granger, 1987). Further, it is also possible to fit OLS by transforming the level data with methods like differencing or de-trending but such filtering causes loss of information (Jenkin and Box). Again, OLS by differencing presents only short-run relationship of time series and misses out the long-run equilibrium relationship. However, based on the least square estimation (LSE) methodology and OLS model, a two-step error correction model (ECM) procedure as suggested by Engle-Granger is found to be useful method. Representation theorem of Engle-Granger facilitates estimating cointegration relationship using the ECM.

Further, a Vector Auto Regressive (VAR) model procedure requires that all the time series included in the model must be stationery and of the same order of integration. There is also the Vector Error Correction Model (VECM) which is an extension of VAR to cointegrated non-stationery time series while incorporating the error correction feature of ECM. Similarly, based on LSE methodology and based on OLS system, the Auto-Regressive Distributed-Lag (ARDL) model is not only useful for non-stationery times series and that of mixed orders of integration but also different lag lengths of the dependent and independent variables are accommodated in this model. Through a simple linear transformation of an ARDL can also derive the Error Correction Model (ECM) which integrates the short-run dynamics with long-run equilibrium (Shreshta and Bhatta, 2018).

Having said, an attempt is made at the movement to estimate the relationship between the time series variables, that is GDP and Expenditure on Education (public and private separately). For this, we have used a basic and simple version of ARDL model with error correction (ECM) representation. Model estimates are derived as presented in the Table-5 below. Causality and dependent variables are decided based on Granger Causality test as presented above. The ARDL equation used to estimate the relationship is as follows.

$$\Delta \ln Y_t = B_0 + \{(B_1 * \Delta \ln Y_{t-1}) + (B_2 * \Delta \ln X_t) + (B_3 * \Delta X_{t-1})\} + \{(A_1 * \ln Y_{t-1}) + (A_2 * \ln X_t) + (A_3 * \ln X_{t-1})\} - \{(C_1 * EC_t)\} + v_t$$

Δ is ***delta*** indicating change or first difference; B_0 is intercept or constant; B_1 to B_3 are coefficient of short-run dynamics; A_1 to A_3 are coefficients of long-term dynamics; C_1 is coefficient of ECT indicating the speed of adjustment when short-run deviation from the long-term equilibrium condition (speed at which the dependent variable returns to its long-run equilibrium); t and t-1 no lag and lag terms of dependent and independent variables. Expected sign of the C_1, the coefficient of the ECT is negative.

Table-5: Results of a ARDL Estimate

Model 1: Dependent Variable is ΔlnGDP ***Model 2: Dependent Variable is ΔlnPFCE***

Sno	***Model-1: GDP on PFCE***			***Model-2: PFCE on GDP***		
	Variables	Estimate(b)	p	Variables	Estimates(b)	p
1	*2*	*3*	*4*	*5*	*6*	*7*
1	Intercept	7.07	0.00	Intercept	-6.64	0.00
2	$\Delta lnGDP_{t-1}$	0.00	0.16	$\Delta lnPFCEE_{t-1}$	0.00	0.62
3	$\Delta lnPEE_t$	0.00	0.33	$\Delta lnGDP_t$	0.00	0.61
4	$\Delta lnPEE_{t-1}$	0.00	0.72	$\Delta lnGDP_{t-1}$	0.00	0.01
5	$lnGDP_{t-1}$	-1.00	0.00	$lnPFCEE_{t-1}$	-1.00	0.00
6	$lnPEE_t$	0.68	0.00	$lnGDP_t$	1.18	0.00
7	$lnPEE_{t-1}$	0.00	0.10	$lnGDP_{t-1}$	0.00	0.10
8	ECT_t	1.00	0.00	ECT_t	1.00	0.00
	R^2: 0.99 (p < 0.05)			R^2: 0.99 (p<0.05)		

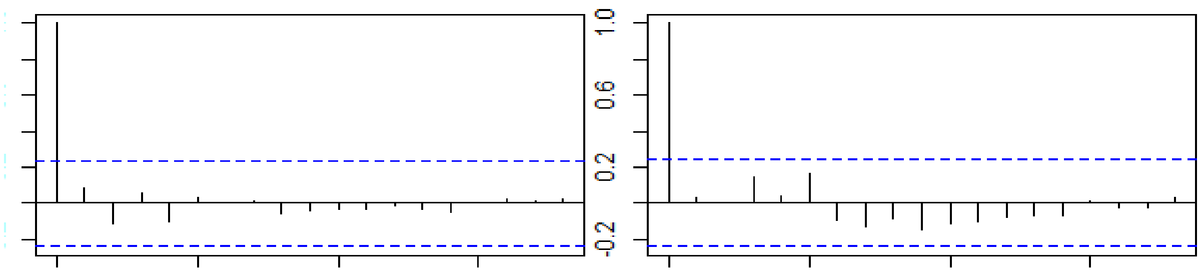
Notes: *Model 1 regressed GDP on Public Expenditure and Model regressed Private Expenditure on GSDP as indicated by Granger Causality; Significance of *** at less than 1% and * at less than 5%; 2. d- delta indicating change or first difference; ln – logarithmic transformed; GDP – Gross Domestic Product; PEE – Public Expenditure on Education; PFCE – Private Final Consumption Expenditure on Education; ECT – Error Correction Term.*

Source*: Author's Estimate.*

ARDL estimates indicate that the short-run dynamics between GDP and public expenditure on education appeared to be very negligible and insignificant whereas long-run relationship /equilibrium condition between them persisted and it is significant (Table-5). In respect of private expenditure on education and GDP as well, the short-run dynamics are little negligible whereas long-run equilibrium condition is significant (Table-5). Further, the speed of adjustment to long-run equilibrium from any short-run deviation appears to be very quick.

Figure-2: ACF for Autocorrelation of Residuals: Model 1 and 2

a) ACF for Residuals of Model 1 ***b) ACF for Residuals of Model 2***



Note*: ACF - Auto Correlation Function.*

Source*: Estimated from the above model.*

Diagnostics for model-residual autocorrelation is clear in indicating absence of such autocorrelation. Figure-2 exhibits the ACF of model-residuals for both the models. It indicates presence of such autocorrelation in the residuals is insignificant.

On the whole, it can be said that cointegration exercise implies a relationship and causality test has shown a direction, the estimates of relationship between the time series variables confirms such relationship with short-run and long-run dynamics.

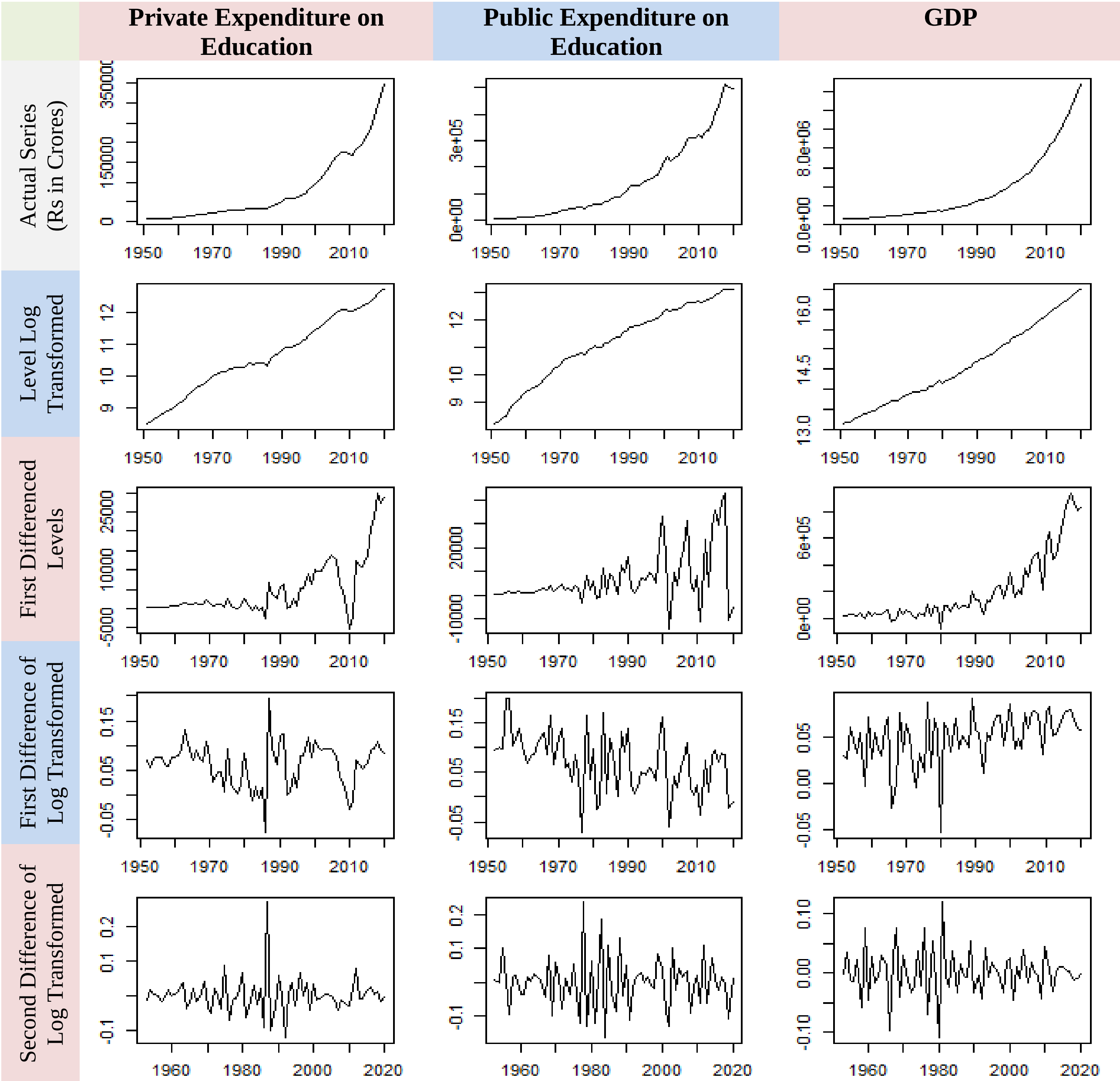
III Concluding Remarks

The long-rum equilibrium relationship between investment (expenditure) on education in India and its GDP analysed applying cointegration procedures for the analysis of time series for the period of last seven decades (between 1950-51 and 2019-20). The exploratory data analysis and the objective tests of cointegration (Granger and Johansen procedure) indicate that these times series individually integrated of same order are cointegrated. The Granger procedure confirms the same for the time series integrated of order two. It indicates that the annual changes in these time series are cointegrated rather than the actual level time series. Johansen method confirms it for the series integrated of order one only. Granger causality has shown that public expenditure in education has an impact on GDP while GDP has an impact on private investment (expenditure) in education (Table-4). But there is no direct causality in either direction found between private and public investment.

* * *

Appendix I

Figure 1A: Signals of Stationarity for Selected Time Series

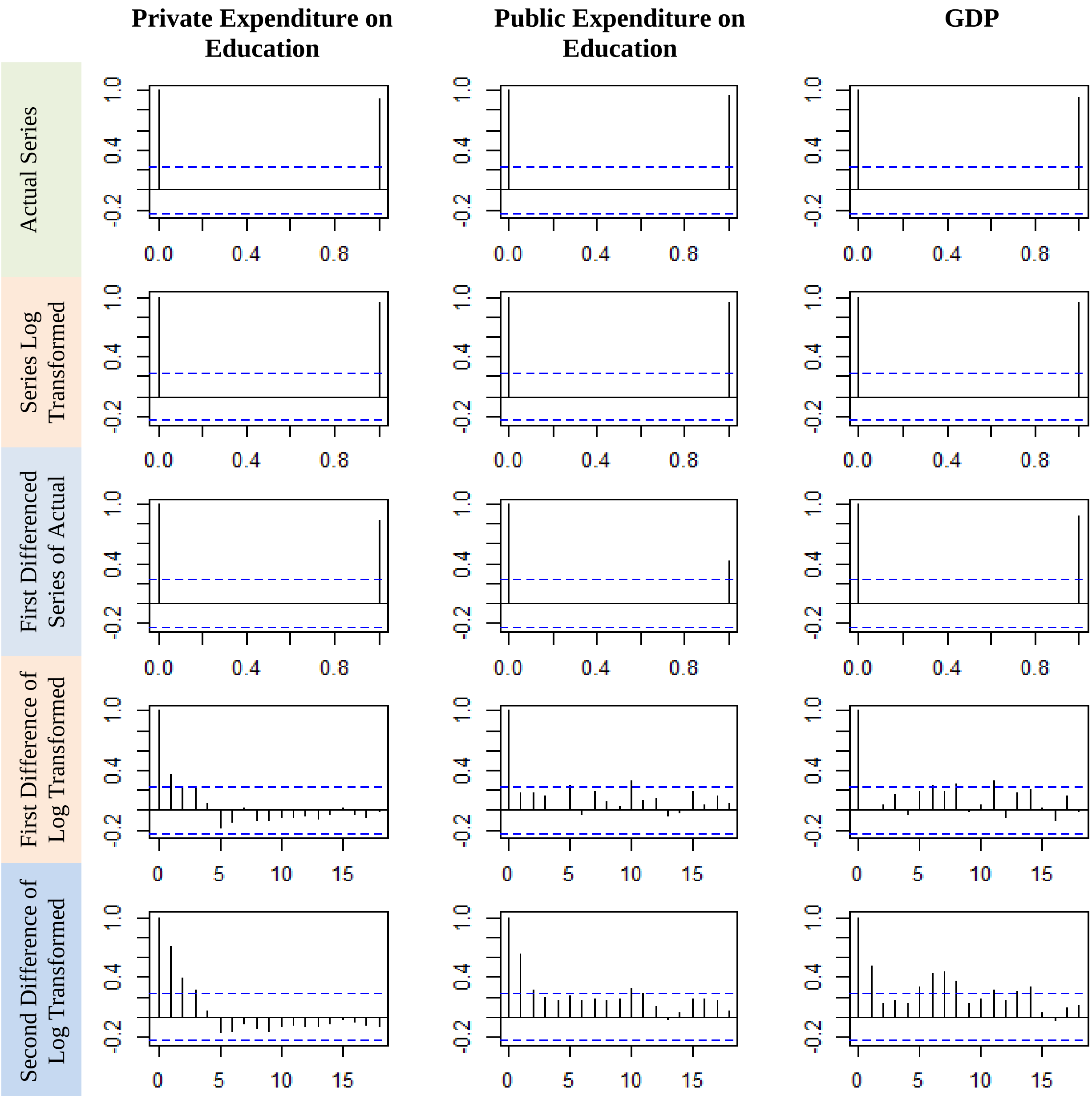


Notes:

Source: *Author's Computation based on R-Package.*

The ***first panel*** consisting spread of level (actual values) data of three time series, shows an exponential pattern of linear trend in all the three. It is in fact exhibiting the feature of these time series, a random walk with deterministic trend. The ***second panel*** consisting of log-transformed time series also exhibit a linear trend which is in fact another feature of time series, a random walk with drift. The first differenced series (of actual values without a log transformation) in the ***third panel*** exhibited again a random walk linear trend (exponential pattern). The first difference of log transformed time series in the ***fourth panel*** indicate a random walk but a kind of linear trend can be observed in case of two time series (Public Expenditure on Education and GDP). Finally, the spread of second difference of log transformed time series in the ***fifth panel*** which appears to be a random walk with white noise, that exhibits stationarity in all the three time series considered for analysis.

Figure 2A: Correlogram of Autocorrelation Function (ACF) for Stationarity Signals in Selected Time Series



Notes:

Source: *Author's.*

The autocorrelation function (ACF) exhibits, whether a time series has a trend, whether it is a white noise process etc. A time series with a trend always have a highly positive autocorrelation coefficient at lower lags and the value of such positive coefficients decreases slowly with higher order of lags and end up close to zero for the last possible lag. For a ***white noise*** time series process, the autocorrelation coefficients (positive or negative) are close to zero and more than 95% of spikes in autocorrelation function (ACF) represented by autocorrelation coefficients must be within the range of $\pm\, 2\,/\sqrt{T}$ where T is length of time series. Since our time series length is 70 (years), $\pm$0.24 is the range of the ACF for the time series that is being analysed.

Appendix II

Testing Stationarity and Cointegration and Modelling Long-run Relationship of Time Series

I Testing Stationarity in the Time Series Process

Very often we find that time series have a non-stationary processes that exhibits ***trend***, ***seasonality***, ***cycles***, ***random walk*** or any *combination* of these. Further, a *linear trend* of the time series process can be a ***deterministic*** or ***stochastic*** one. Variability in statistical properties of the time series (its mean, variance and covariance) is main feature of the non-stationary process. Conventionally, to understand and identifying the underlying pattern it represents a time series process is decomposed (additive or multiplicative), depending on its features (trend: deterministic/stochastic, cycles/seasonality or their combination), into different components: trend-cycle, seasonality and residual/random components. Such decomposing is useful in removing deterministic trend and making it a stochastic process and also converting it into a stationary process.

$y_t = y_{t-1} + \varepsilon_t$ --- represents *pure random walk* time series process
$y_t = \propto + y_{t-1} + \varepsilon_t$ --- represents *random walk with drift*
$y_t = \propto + \beta_t + \varepsilon_t$ --- represents *linear trend* (deterministic)
$y_t = \propto + y_{t-t} + \beta_t + \varepsilon_t$ --- represents *random walk with drift and linear trend* (deterministic) time series process

Where α is an intercept (constant) represent drift in time series, y_{t-1} is the term representing lag of 'y_t', β_t represents time trend and ε_t represents the stochastic/random component (residual term). The time series (y_t) is regressed on its lag (y_{t-1}) when it exhibits the random walk process (AR - autoregressive model) and it is regressed on time trend (B_t) when it exhibits the time trend (ordinary regression) to get the random component (residual/error term) of the series. For most non-stationary time series, the random walk models are used. A typical random walk time series while consists of long-term trend (up/down), changes in any direction are sudden and unpredictable.

As it is the case in econometrics in general, the stochastic/random component (residuals or error term) of the time series model (equation) is critical for diagnostics of the times series to inform the pattern/behaviour or nature of the time series process. The best (model) fit condition is that the residual component as a whole should be independent of all the other components (or predicted values) of the model (equation) along with each error term is independent of the other (serially uncorrelated) and follow pattern of normal distribution. A ***pure random walk*** process consists of a non-systematic stochastic component (ε_t) with white noise. A ***white noise*** of a random walk time series process is a sequence of serially uncorrelated random (residual) observations in the time series. In other words an independent and identical distribution (i.i.d) of these time series observations represents the white noise.

While there are many forms of models representing different stochastic process of time series, broadly three types of basic linear models are very often used in modelling time series with its lag values (present value modelled/regressed on past value of time series): Autoregressive (AR), Integrated (I) and Moving Average (MA), and combinations of which are also used as ARMA and ARIMA. Further, the Vector Autoregressive (VAR) models are an extension to AR variants to deal with multivariate analysis of vector-valued time series. Further, there are certain non-linear time series models as well. The non-linear models that examine the changes in variance (in random component i.e. residuals/error) over time (i.e. heteroscedasticity): Autoregressive Conditional Heteroscedasticity (ARCH) and a Generalised ARCH (GARCH). There are different variants of GARCH. The non-linear models that accounts for an exogenous factor are Non-linear Autoregressive Exogenous (NARX) models which consisting of past of value (lab) of the same series along with past and current values of exogenous (driving force) series in the model (equation).

Stationarity of the time series process indicate that the statistical properties of the time series (its mean, variance and covariance) that do not change over time, they are time-invariant. In other words a change in time does not affect the shape of distribution of the series. Its mean and variance are constant and its covariance is independent of time. To make effective and more precise predictions of time series, stationarity of the process is an essential requirement of the time series analysis models. Especially in forecasting, requisite model can be applied for a stationery time series to predict more precisely the future values. A non-stationery series or one with unit root cannot be modelled and it does not make good predictions.

In an exploratory kind of analysis, qualitatively, the graphic signal spread of an actual time series, its transformative series (log, differenced and/or lag) and its prediction errors or residuals along with their correlograms of autocorrelation function (ACF) would present any indication of stationery and non-stationery process of a time series. Further to confirm stationarity or non-stationarity (or presence of unit root) of the time series process, quantitatively, there are different tests varying in their procedure of testing.

For testing stationarity, first of all, the time series needs to log-transformed that converts the exponential trend into a linear trend. Second, differencing and/or lag series of an order is required to convert a non-stationery series into a stationery one.

Test for Independence: Ljung-Box Test

The ***Ljung-Box test*** is a modified version of ***Box-Pierce test***, is to examine whether there is any significant evidence for non-zero correlations among prediction errors (residuals). In other words, whether the time series is an independently distributed (i.i.d) one having zero (there is no) serial or auto correlation. If there is any such correlation it must be due to randomness. The null hypothesis (H0) of the test is that series is independent and the alternative hypothesis (H1) is the series is not independent. Here the significant p-value is to reject the null hypothesis and accepting the alternative.

$$Q_m = n(n+2)\,.\sum_{j=1}^{m}\frac{r_j^2}{n-j}$$

Where the test statistics (of Ljung-Box) is function of accumulated autocorrelations (r) upto a lag (i.e. *m*) as specified, *n* is number of observations in a differenced time series. The test-statistics follows a chi-square distribution.

Unit Root Tests: Variants of Dickey-Fuller (ADF, PP and ADF-GLS)

Unit Root process in the theory probability, statistics and time series econometrics is a stochastic process with systematic pattern which is unpredictable. In the time series it is known as random walk with a drift.

Augmented Dickey-Fuller (ADF) test is to examine if the time series has a unit root. The *null hypothesis* (H0) of ADF test is that the time series has unit root in the process and the *alternative hypothesis* (H1) is there is not unit root. Usually the ADF test statistics is a negative number and if it is *more negative it is stronger* in rejecting the null hypothesis (i.e. presence of unit root at some level of confidence) and accommodating the alternative. When a computed p-value is lower than a significance level (5% or below) null hypothesis is to be rejected. The result is that there is ***no evidence for presence of unit root*** in the time series process. The ADF test is based on autoregressive (AR) model and t-test statistic for its null hypotheses. The original AR (1) model based on Dicky-Fuller (DF) test was later extended to AR(p) model test as an ADF. To determine the number of lags, different information criterion (AIC/BIC/HIC) can be used. The model equation for the test is as follows:

$$\Delta y_t = \propto + \beta t + \varphi . y_{t-1} + \delta . \Delta y_{t-1} + \cdots + \delta_{p-1} . \Delta y_{t-p+1} + \varepsilon_t$$

Where $\propto$ and $\boldsymbol{\beta . t}$ represent the constant and time trend. Model of the test vary with inclusion and exclusion of either or both these two parameters, in some models they may be restricted. It all depends on the patterns of time series, whether it exhibits a random walk with a drift (linear trend) and/or trend (deterministic trend or exponential pattern).

A modified version of ADF test is the *Phillips-Perron* (PP) *test* which takes into account necessary corrections for autocorrelation and heteroscedasticity in the errors (residuals) of predicted model; it is a non-parametric test. The null and alternative hypotheses of PP test and its interpretation are similar to that of ADF test.

In Augmented Dickey-Fuller Generalised Least Square (ADF-GLS) procedure test developed by Elliott, Rothernberg and Stock (ERS), a time series is transformed using generalised least square(GLS) regression for de-trending to prepare the time series and performing the unit root test (see Elliot *et al.*, 1996). In other words the GLS estimate of deterministic component is used to de-trend the time series and applied the ADF test for unit root. It is considered that ADF-GLS has the higher power and probability in rejecting the false null hypothesis (H0). The null hypothesis of the ADF-GLS is that y_t in the equation is a random walk, possibly with drift (linear trend). Alternative hypotheses are: y_t is stationary about a linear (time) trend; or it is stationary with nonzero mean but without any linear (time) trend.

Test for Level and Trend Stationarity and Unit Root: KPSS

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test is to complement other unit root tests like ADF (Kwiatkowski *et al.*, 1992). As it is observed *absence of unit root* does not ensure that the time series is stationery. The non-stationary time series without a unit root in the series can be a trend-stationery series. The mean of a time series can be increasing or decreasing in both the processes of unit root and trend-stationery. However, the trend-stationary time series processes are mean-reverting in the presence of a shock (convergence of shock period mean with the trend overtime), whereas the unit root processes are not mean-reverting as they have permanent impact on the mean (no convergence). In other words, shocks have transitory effect in a stationery time series process with a deterministic trend where as they (shocks) have permanent effect in a time series process with stochastic trend (i.e. with unit root).

The KPSS test is to examine whether a time series is stationary around its mean or linear/deterministic trend or it is non-stationery due to the presence of unit root in the process. The linear regression equation (i.e. $\mathbf{y_t = z_t + B_t + e_t}$) for KPSS test has three components: deterministic trend (Bt), random walk (z_t which is again $z_t = z_{t-1} + u_t$) and stationery error (e_t). The time series is a sum of these three components (Kwiatkowski *et al.*, 1992). The one-sided Lagrange Multiplier (LM) statistics is the test statistics of the KPSS test. Null hypothesis is to be rejected when computed value of test statistics is greater than critical value at a given confidence level (5% or 1%).

The *null hypothesis* is that time series is stationery and the alternative is about presence of unit root (non-stationary) in the series. The KPSS is diametrically opposite in terms of hypothesis about unit root when compared to the other unit root tests which have a null hypothesis about it. A ***low p-value*** (below 5%) indicates that series is not trend-stationery (rejecting the null) and it has unit root whereas a ***high p-value*** (above 5%) indicates the series is stationery (cannot reject null) and it has ***no unit root*** in the process.

Therefore, unit root and stationarity tests in combination informs us about a time series whether it has unit root, whether it is stationary process, or whether the information is not sufficient to confirm the series is stationary and integrated.

II Testing Cointegration: Long-run Equilibrium/Relationship of Time Series

As mentioned above most often time series consists of trend (deterministic or stochastic). Any procedure evaluating or establishing relationships between such non-stationary time series having trend is considered to be not appropriate. Many a time in the past linear regression procedures were applied to time series as well while applying such procedure to de-trended time series but spuriousness of such regression procedures were observed as early as 1920s (see Yule, 1926). Later in 1970s Clive Granger and Paul Newbold have brought forth more emphatically the adverse implications of such approach and producing spurious correlations (see Granger and Newbold, 1974). Their work has shown that even the de-trended time series can continue to be a non-stationary one, hence the relationship established is spurious. After a decade Clive Granger with Robert Engle put forth a formal vector approach dealing with such problems and a procedure for it while coining the term cointegration for the same (see Engle and Granger, 1987).

The concept of cointegration in time series econometrics is to assess/examine long-rum equilibrium relationship between time series. In other words, if there is any correlation in the long run between two or more economic or other time series. Any such correlation indicates integration of two or more non-stationary time series. It means they do not deviate from the equilibrium in the long run. Testing for cointegration is to confirm presence of any such correlation between time series. Three most used tests in this regard are: Engle-Granger test, Johansen test and Phillips-Ouliaris test.

For cointegration, when of the time series individually are integrated of an order, *d*, and their linear combination of such time series is integrated of lower order (should be integrated of order less than *d*), then these time series are cointegrated. The time series is integrated of order one (I[1]) when a time series has a unit root (of AR). In other words when a first difference of a times series (ΔYt) is stationary it is integrated of order one or if the second difference of it is stationary it is integrated of order two (I[2]) and so on when it is differenced 'd' times to become stationary then the time series is integrated of 'd' (I[d]). Two time series are cointegrated in principle when these two times series (eg., Xt and Yt) share a common stochastic trend having same long-run behaviour and are integrated of same order 'd' (i.e. if both the Xt and Yt are I[1]), their difference is to be stationary and integrated of order zero (Zt = Yt – ∅Xt is to be I[0]). Relationship between two time series cointegrated must be justified theoretically or by common sense.

Engle-Granger Cointegration Test

Engle-Granger test is two-step method. First, using a static model regression equation it generates residuals. Second, the residual are tested for unit root. It is to examine whether the linear combination of two time series follows AR (1), i.e. autoregressive model of order one.

$$Y_i = \alpha + \beta . X_i + R_i$$

$$R_i = p . R_{i-1} + \varepsilon_i$$

The basic parameters to be estimated are: $\alpha, \beta \ and \ p$. Here if |p| < 1, the two time series (Y and X) are cointegrated.

Johansen Test

One of the limitations of Engle-Granger test is that it is applicable to two time series at a time. Johansen test overcomes such limitation and it can be applied several time series. Johansen test consists of two variants: one is maximum 'eigen value' test and the other is 'trace' test. Johansen tests are based on Vector Autoregressive (VAR) models which are multi-dimensional extension of AR model: VAR in general is similar to AR(p). In other words VAR is the starting point of Johansen methodology for cointegration analysis. It also allows us to perform Vector Error Correction Model (VECM).

Phillips-Ouliars Test

Phillips and Ouliars (1990) developed asymptotic theory for residual based tests for cointegration. As they observed most of the unit root tests are based on residuals of the time series as modelled. The residuals of cointegration model regression do not follow the distributions of usual unit root tests (say DF variants). Residuals of cointegration model regression have asymptotic distributions. While finding certain pitfalls in procedures designed to test a null of cointegration, owing to indiscriminate use, they suggested continuing use of such residual based unit root test. They provided a full set of critical values for such tests that allow demeaned and trended times series cointegration regression having variable upto five (see Phillips and Ouliars, 1990).

III Modelling Long-rum Equilibrium or Relationship of Time Series

The normal OLS is considered to produce spurious result for the non-stationery time series and only present short-run equilibrium for the stationery time series.

Error Correction Model (ECM) another method modified to fit for cointegrated non-stationery time series. Short-run dynamics of time series are influenced by divergence or deviations (error) from long-run equilibrium. Where it is important the speed at which any deviation in one time series returns to its long-run equilibrium following the changes in its cointegrated time series. Therefore, an ECM while capturing the both the short-run and long-run equilibrium relationships, it estimates the speed at which a time series after a deviation return to its equilibrium.

Engle-Granger (1987) procedure of ECM consist a ***two-step method,*** first step OLS estimation derives residuals and in the second step these residuals are modelled along with variables in OLS method.

The general OLS bi-variate model can be written as: $Y_t = a + bX_t + e$ ----- 1

Wherein $e = Y_t – a – bX_t$; and also for time series: $e = f(e_{t-1})$

In OLS systems it is assumed that e is random variable. Using the above OLS equation a change in time series while incorporating the residuals in the model can be constructed as:

$$\Delta y = a + b\Delta x - c(e_{-1}) + u \quad ---- 2$$

Here c is coefficient of the term and *e_{-1} is to represent residuals of equation 1* modelled in this construction for ECM, whereas the *u is the error term* for the equation 2. In this construction, if *c is equal to 0* and significant it in fact indicates cointegration between the variables modelled.

The equation 2 can be written as: $\Delta y = a + b\Delta x - c(Y – a - bX)_{-1} + u$ ---- 3

and further it can be written as: $\Delta y = a + b\Delta x + cY_{-1} + ca + cbX_{-1} + u$ ---- 4

In this construction an ECM represents both the short-run and long-run relationships (equilibrium) of the variables modelled. The equation 4 is single step ECM for the bi-variate model.

In other words, a specification that allows a wide variety of dynamic patterns in the data and that captures the general dynamic relationship between variables (y and x) by including their lagged values in specification can be written as:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \alpha Y_{t-1} + u_t \qquad \text{---- 5}$$

When we are interested in long-run equilibrium relationships, the conditions under which the general dynamic equation (5) is consistence with the long-run equilibrium relationship requires to zero out factors that could cause divergence from equilibrium. Changes in x_t along with stochastic fluctuations (u_t) are key factors in this regard. Therefore, it is captured through the following specification:

$$\Delta y = \beta_0 + \beta_1 \Delta x + \alpha(X_{-1} - Y_{-1}) + u_t \qquad \text{---- 6}$$

Further, other than OLS, but following the least square equation (LSE) methodology, the Vector Auto Regressive (VAR) models are better and useful method for cointegrated time series. An VAR model developed by Sims (1980) is considered as useful method of understand dynamic relationship (both the short-run and long-run) in macroeconomic time series. VAR is basically relates the current value (observation) of a time series with its past value/observation (same variable) and past values/observations of other times series (Other variables in the model). It allows feedback or reverse causality of time series in the system. Made up of system of equations depending on number of endogenous time series variables and their lag length (p^{th} order), VAR is to represent relationships between the time series in the system. It requires that all the time series included in the model must be stationery and of the same order of integration.

Johansen (1988, 1991) procedure using a VAR representation of multi-variate time series model with error correction feature of ECM is considered as Vector Error Correction Model (VECM). In VECM it requires that all the time series included in the models must of the same order of integration.

$$\Delta y = \beta_0 + \beta_1 \Delta x + \alpha(X_{-1} - Y_{-1}) + u_t \qquad \text{---- 7}$$

The Auto Regressive Distributed Lag (ARDL) is proposed by Pesaran and Shin (1999) can be applied to time series integrated in different orders. It consists of two components: autoregressive (of dependent time series) and distributed lag (of explanatory time series). An ARDL accommodated different lag lengths of time series in the model. A bi-variate case of an unrestricted ARDL model with p (lags of dependent variable) and q (lags of independent variable[s]) order can be written as:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_s Y_{t-p} + \alpha_0 X_t + \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-q} + u_t \qquad \text{--- 8}$$

Further, for the multivariate (independent series), the ARDL (p, q) model of p and q order can be written as:

$$Y_t = \beta_0 + \sum_{s=1}^{\infty} \beta_1 Y_{t-s} + \sum_{i=0}^{k} \sum_{s=0}^{\infty} \alpha_{0i} X_{(t-q)i} + u_t \quad \text{--- 9}$$

Devised by Pesaran *et al* (2001), the Bounds Test for ARDL model indicates and estimate whether there is a long-term relationship between the variables.

The ARDL can have Error Correction component term (ECT) as well

$$Y_t = \beta_0 + \sum_{s=1}^{\infty} \beta_1 Y_{t-s} + \sum_{i=0}^{k} \sum_{s=0}^{\infty} \alpha_{0i} X_{(t-q)i} + c.ECT + u_t \quad \text{--- 10}$$

Error Correction terms is derived by running normal regression of the time series and the time series of their residual used as variable. Coefficient of the ECT presents the long-run adjustment of the model returning to its mean.

* * *

Reference

Barro, Robert J., and Xavier Sala-i-Martin (1995). ***Economic Growth***, MIT Press, Cambridge MA.

Box, George and Jenkins, Gwilym (1970). ***Time Series Analysis: Forecasting and Control***, Holden-Da, San Francisco.

Elliott, Graham; Thomas J. Rothenberg and James H. Stock (1996). Efficient Tests for an Autoregressive Unit Root, *Econometrica* (Econometric Society), Vol. 64(4), July: pp 813-836.

Engle, Robert F. and Clive W. J. Granger (1987). Co-integration and Error Correction: Representation, Estimation and Testing, *Econometrica*, Vol. 55(2): pp.251–276.

Granger, C. W.J. and P. Newbold (1974). Spurious Regressions in Econometrics, *Journal of Econometrics*, Vol. 2(2): pp.111–120.

Granger, C.W.J. (1969). Investigating Causal relations by Econometric Models and Cross-spectral methods, *Econometrica*, Vol. 37(3), pp. 424-438.

Ljung, G. M. and G. E. P. Box (1978). On a Measure of a Lack of Fit in Time Series Models, *Biometrika*, Vol. 65(2): pp. 297–303.

Hanushek, Eric, and Dennis Kimko (2000). Schooling, Labor Force Quality, and the Growth of Nations, *American Economic Review*, Vol. 90(5), December, pp. 1184-1208.

Johansen, Søren (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models, *Econometrica*, Vol. 59 (6): pp.1551–1580.

Johansen, Søren (1988). Statistical Analysis of Cointegration Vectors, *Journal of Economic Dynamics and Control*, Vol. 12 (2–3), pp. 231–254.

Jorgenson, Dale W. and Barbara M. Fraumeni (1992). Investment in Education and U.S. Economic Growth, *Scandinavian Journal of Economics*, Supp., 94, pp. 51-70.

Krueger, Alan B. and Mikael Lindahl (1999). *Education for Growth in Sweden and the World*, Working Paper No. 7190, National Bureau of Economic Research (NBER), Cambridge, MA.

Kwiatkowski, Denis; Peter C.B. Phillips; Peter Schmidt and Yongcheol Shin (1992) How sure are we that Economic Time Series have a Unit Root?, *Journal of Econometrics*, Vol. 54, pp.159-178. Accessed on 15/09/2020 at http://debis.deu.edu.tr/userweb//onder.hanedar/dosyalar/kpss.pdf.

Maddala, G.S. and I.M. Kim (2003). ***Unit Roots, Cointegration, and Structural Change***, Cambridge University Press, Cambridge.

Mincer, Jacob (1974). ***Schooling, experience, and earnings***, National Bureau of Economic Research (NBER), New York.

Nelson, Richard R. and Edmund Phelps (1966). "Investment in Humans, Technology Diffusion, and Economic Growth, *American Economic Review* (Papers and Proceedings), Vol. 56(2), May, pp. 69-75.

Pesaran, M. H.; Y. Shin and R. J. Smith (2001). Bounds Testing approaches to the Analysis of level Relationships, *Journal of Applied Econometrics*, Vol. 16(3), pp.289-326

Pesaran, M. Hasem and Yongcheol Shin (1999). *An Autoregressive Distributed Lag (ARDL) modelling approach to cointegration analysis*, in A. Holly Strom and P. Diamond (Eds.) ***Econometrics***

and Economic Theory in the 20th Century: The Ranger Frisch Centennial Symposium, Cambridge University Press, Cambridge.

Phillips, P. C. B. and S. Ouliaris (1990). Asymptotic Properties of Residual Based Tests for Cointegration, *Econometrica*, Vol. 58 (1): pp.165–193.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Shrestha, Min B. and Guna R. Bhatta (2018). Selecting appropriate methodological framework for Time Series data analysis, *The Journal of Finance and Data Science*, Vol. 4(2), June, pp. 71-89.

Sims, Christopher (1980). Macroeconomics and Reality, *Econometrica*, Vol. 48 (1): 1–48.

Welch, Finis (1970). Education in Production, *Journal of Political Economy*, Vol. 78(1), January-February, pp. 35-59.

Yule, 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*, Vol. 89 (1):pp. 11–63.