Journal of Statistical and Econometric Methods, Vol 13, No 1, 2024, 41-56
ISSN: 2241-0384 (print), 2241-0376 (online)
https://doi.org/10.47260/jsem/1312
Scientific Press International Limited

Cointegration Modeling for COVID-19 Infected Cases and Deaths in the US

Rajarathinam Arunachalam[1]

Abstract

This paper aims to study the short-run and long-run cointegration relationships between the total population, the cumulative number of new COVID-19-infected cases, and the cumulative number of deaths due to COVID-19 in different states in the US. The short-run relationship is assessed using the ARDL model, and the long-run relationship is assessed using the ARDL bounds test. To assess the consistency of the model parameters, the cumulative sum of recursive residuals test and the cumulative sum of recursive residuals squares tests are used.

JEL classification numbers: E18, HO, I1, J64, J88.
Keywords: Autoregressive distributed lag model, Error correction model, Unit root tests, Residual diagnostics, Bounds cointegration test, Stability tests.

[1] Department of Statistics, Manonmaniam Sundaranar University.

Article Info: Received: March 15, 2024. Revised: April 16, 2024.
Published online: April 22, 2024.

1. Introduction

1.1 Background of the study

In the United States, the worldwide pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in 103,436,829 confirmed cases with 1,177,223 all-time deaths, the most of any country, and the 20th-highest per capita worldwide. The COVID-19 pandemic ranks first on the list of disasters in the United States by death toll; it was the third-leading cause of death in the U.S. in 2020, behind heart disease and cancer. From 2019 to 2020, U.S. life expectancy dropped by three years for Hispanic and Latino Americans, 2.9 years for African Americans, and 1.2 years for white Americans. These effects persisted as U.S. deaths due to COVID-19 in 2021 exceeded those in 2020, and life expectancy continued to fall from 2020 to 2021.

On December 31, 2019, China discovered a cluster of pneumonia cases in Wuhan. The first American case was reported on January 20, and President Donald Trump declared the U.S. outbreak a public health emergency on January 31. Restrictions were placed on flights arriving from China. Still, the initial U.S. response to the pandemic was otherwise slow in preparing the healthcare system, stopping other travel, and Testing.

The first known American deaths occurred in February. On March 6, 2020, Trump allocated $8.3 billion to fight the outbreak and declared a national emergency on March 13. The government also purchased sizeable medical equipment, invoking the Defense Production Act 1950 to assist. By mid-April, disaster declarations were made by all states and territories as they all had increasing cases. A second wave of infections began in June, following relaxed restrictions in several states, leading to daily cases surpassing 60,000. By mid-October, a third surge of cases started; there were over 200,000 new daily cases in December 2020 and January 2021.

1.2 Objectives of the present study

The main objectives of the present study are to investigate the short-run and long-run cointegration relations between the total population, the cumulative number of new COVID-19 infected cases, and the cumulative number of deaths due to COVID-19 to estimate the long-run equilibrium relationship between these using an autoregressive distributed lag model (ARDL) and bounds cointegration tests, and to study the stability of the model parameters.

1.3 ARDL model

A $ARDL(p,q)$ model has p lags of the dependent variable and q lags of the independent variable:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \alpha_0 x_t + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_q x_{t-q} + \mu_t \quad (1)$$

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=1}^{q} \alpha_i x_{t-i} + \mu_t \quad (2)$$

where μ_t is a random "disturbance" term. These $\beta_1, \beta_2, \beta_3, ..., \beta_p$ are called long-run dynamics and $\alpha_1, \alpha_2, \alpha_3, ..., \alpha_q$ are short-run coefficients.

The model is *"autoregressive"* in the sense that y_t it is "explained (in part) by lagged values of itself. It also has a *"distributed lag"* component in the form of successive lags of the "x" explanatory variable. Sometimes, the current value of x_t itself is excluded from the distributed lag part of the model's structure (Soharwardi, Khan, and Mushtaq,2018).

2. Materials and Methods

2.1 Materials

The cumulative number of COVID-19 infections and deaths as of October 25, 2023, starting on February 15, 2020, were collected from the website. https://www.worldometers.info/coronavirus/country/us/. Several econometric statistical methodologies have been employed to achieve the stipulated objectives of the present study. Here, the cumulative number of new COVID-19 cases and each state's total population are considered the independent variables, and the cumulative total number of deaths due to COVID-19 infections is considered the dependent variable. EViews Ver. 11 software was used to estimate the model parameters and error diagnostics and to study the stability of the estimated model.

2.2 Methods

To apply the ARDL model, the study variables should fulfil certain stationarity conditions. That is, the variables should be purely I(0), purely I(1) or I(0)/I(1), Alimi (2014). Three different tests, Dickey and Fuller (1979), Phillips and Perron (1988), and Kwiatkowski et al. (1992), were used to test this. The Akaike information criterion (AIC) was used to select the optimal lag. The Jarque-Bera test (Jarque and Bera 1980) is used to test the normality of the residual. The Ljung-Box test (Ljung and Box 1979) and the Breusch-Godfrey test (Breusch 1978; Godfrey 1978) were used for autocorrelation and serial correlation testing. To test for heteroscedasticity, the Breusch-Pagan-Godfrey heteroscedasticity test (Godfrey 1978; Breusch and Pagan 1979) was used. Model stability was studied based on the cumulative sum of recursive residuals (CUSUM) and cumulative sum of recursive residuals squares

(CUSUMSQ) tests (Brown et al. 1975). Finally, to test the cointegration (long-run relationship), the bounds test (Pesaran et al. 2001) was employed.
Details of these methods have been omitted in this paper and are available extensively in the literature.

3. Results and Discussion

In this section, the empirical findings, and their interpretations are discussed in sequence.

3.1 Unit root test

The results presented in Tables 1(a) and 1(b) are the ADF and PP test results. Test statistics values are significant at a 1% level of significance, and hence, the null-hypothesis of presence of unit roots are rejected and hence all the study variables under study are found to be stationary without differencing and are therefore they are stationary at level (I(0)).

Table 1(a): Unit root rest Augmented Dickey-Fuller test at level

Variables	Intercept	Intercept & Trend	None
Cases	-5.5470** (0.000)	-7.0688** (0.000)	-5.0597* (0.000)
Death	-3.7455** (0.000)	-3.4624** 0.0578	-4.2909* (0.000)
Population	-12.8019** (0.000)	-15.2894** (0.000)	-10.4484* (0.037)

** 1% level of significance;*5% level of significance ;
Figures in the () represent p -values.

Table 1(b): Unit root test Phillips-Perron test at level

Variables	Intercept	Intercept & Trend	None
Cases	-8.8556** (0.000)	-13.1261** (0.000)	-6.9213* (0.000)
Death	-7.9847** (0.000)	-8.2362** (0.000)	-4.5055* (0.000)
Population	-7.8198** (0.000)	-7.8878** (0.000)	-4.6068** (0.000)

** 1% level of significance;*5% level of significance ;
Figures in the () represent p -values.

3.2 Summary Statistics

The results presented in Table 2 reveals that all three study variables are not normally distributed since the Jarque-Bera (Jarque and Bera, 1980) statistics values are highly significant. The number of infected cases has a higher range value than the deaths due to infections. The skewness values fall between the acceptable range (− 2 and + 2). All the study variables are positively skewed. The pattern of COVID-19-infected cases is skew-symmetric since all the skewness values are positive. The skewness values are the same in the total population and the number of infected cases. Since all the kurtosis values are more significant than one, the distribution is too peaked and leptokurtic. The standard deviation value is higher in the total number of infected cases than in the total number of deaths.

Table 2: Summary Statistics

	POPULATION	CASES	DEATHS
Mean	7953673.	2562513.	27873.23
Median	5790585.	1819782.	19218.50
Maximum	39512223	12488495	105383.0
Minimum	1415872.	405653.0	2056.000
Std. Dev.	7647879.	2488549.	25122.23
Skewness	2.484933	2.365561	1.764854
Kurtosis	9.626196	8.682617	5.430023
Jarque-Bera	114.3434	91.12609	30.60642
Probability	0.000000	0.000000	0.000000
Sum	3.18E+08	1.03E+08	1114929.

Figure 1 depicts the geographical area of different states of the U.S. Figure 2 depicts the total state population size, the highest population are registered in California, Texas, Florida, New York, etc. Figures 3 and 4 represent the cumulative COVID-19 infected cases and deaths, respectively. In California, the highest number of 1,24,88,495 infected cases have been registered, which is the most populated state. In Texas, Florida, and New York, the total number of infected cases is 88,61,046, 78,25,982, and 72,16,960, respectively. In California, the very highest number of 1,05,383 deaths due to COVID-19 infections have been registered, followed by Texas (95,162), Florida (91,590), New York (78417), etc.

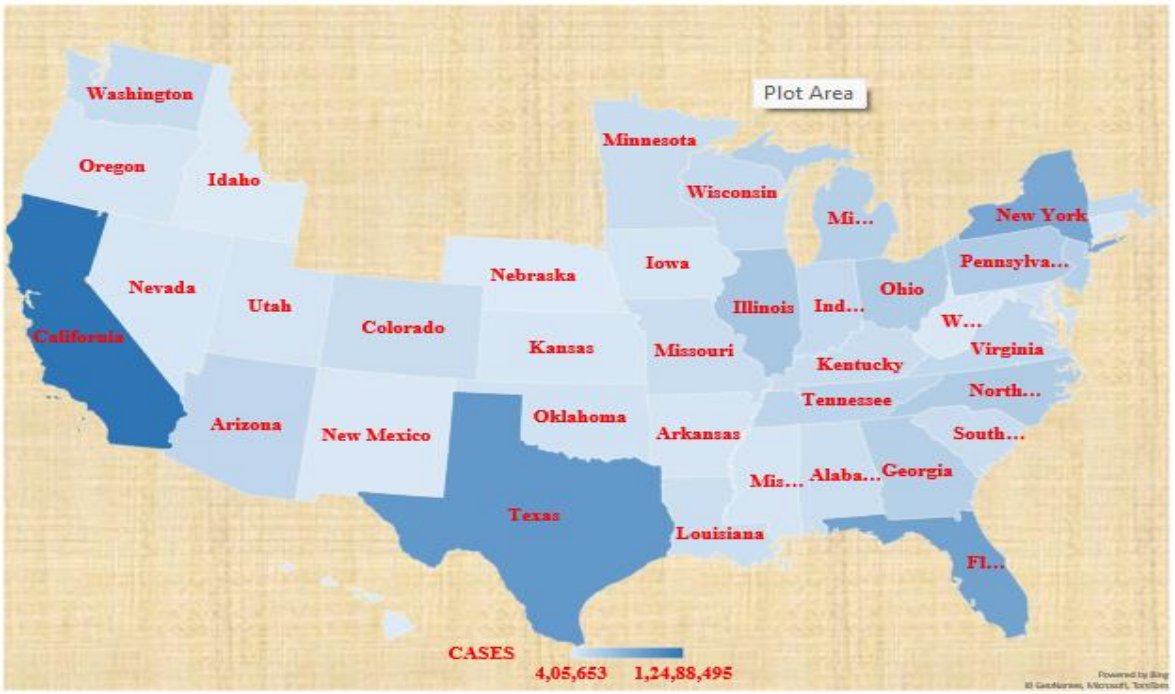

Figure 1: Details different states of the US considered for the study

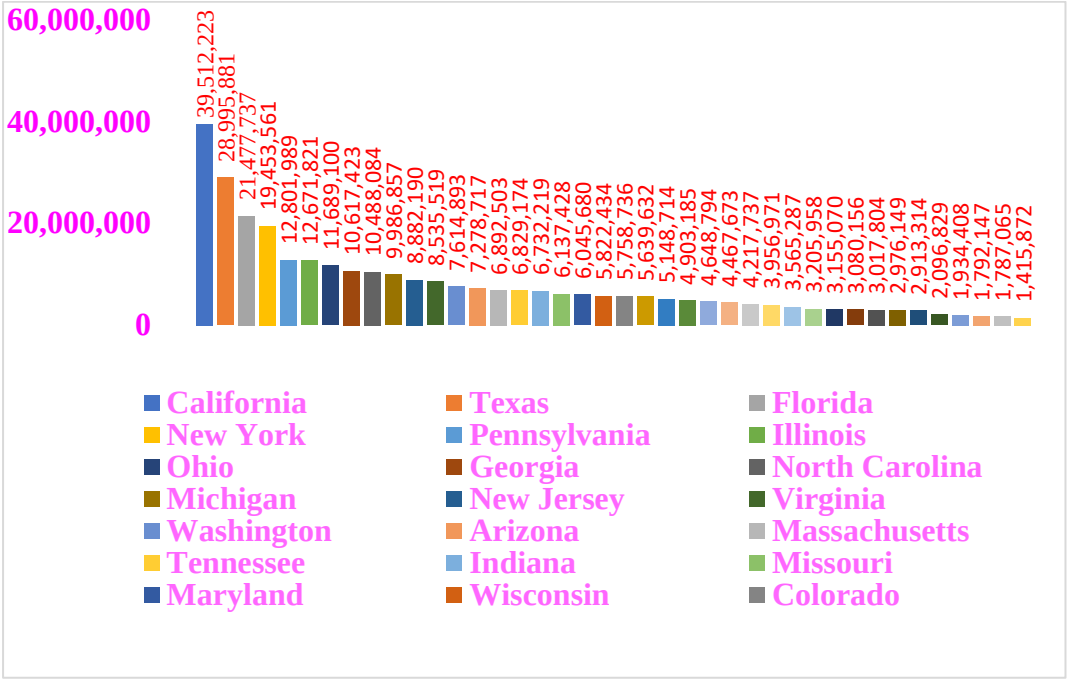

Figure 2: State-wise total population

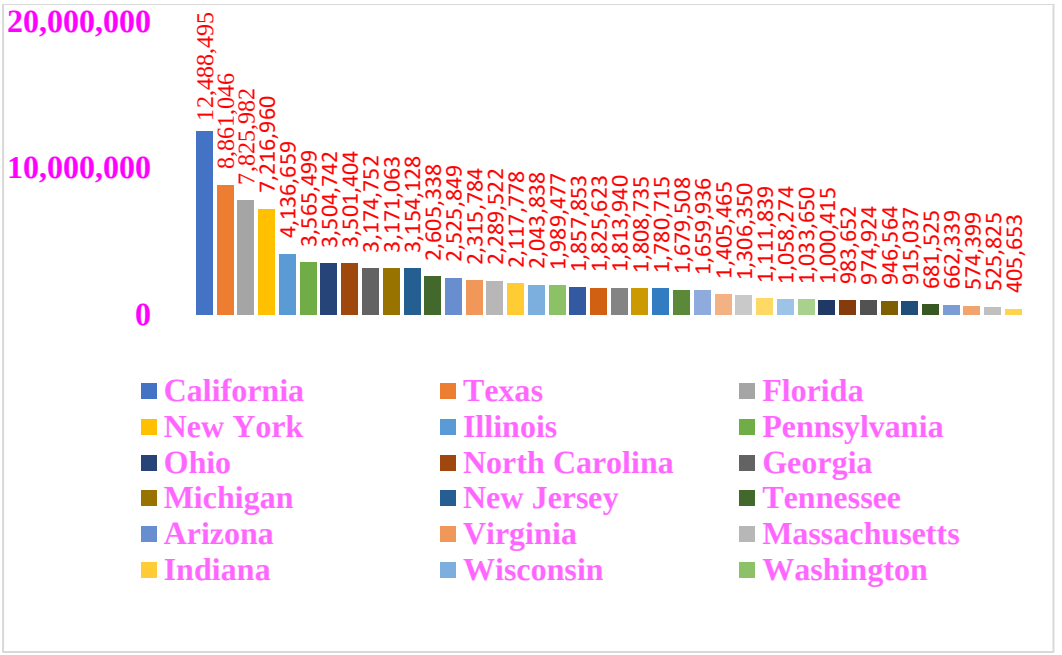

Figure 3: State-wise number of COVID-19-infected cases

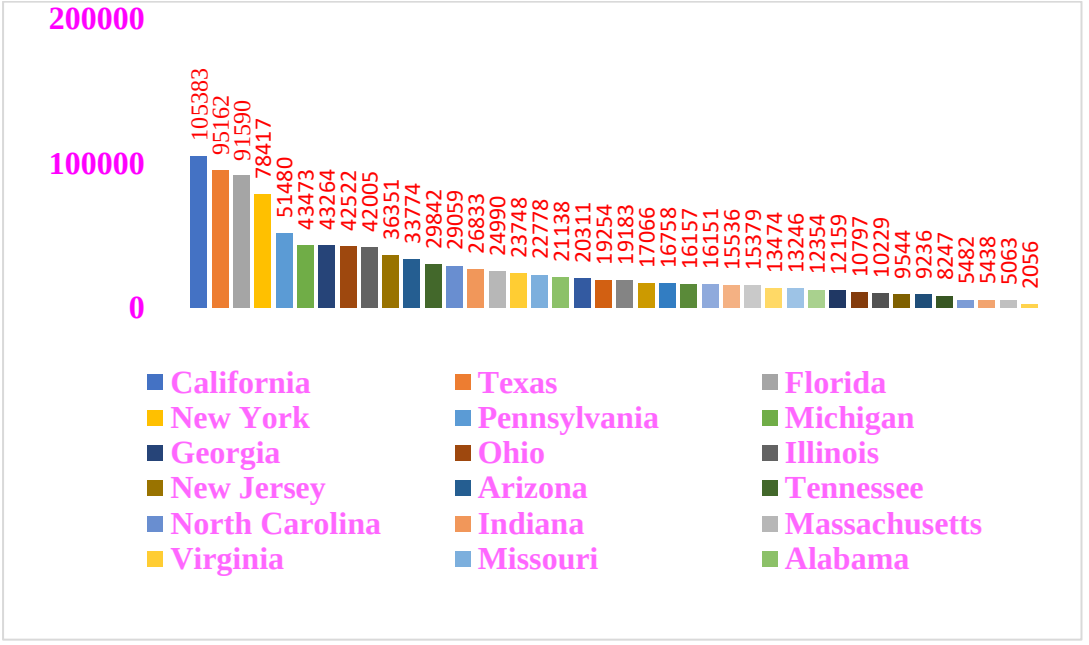

Figure 4: State-wise number of deaths due to COVID-19 infections

3.3 Model selection

To choose the optimal lag values, p and q, the Akaike information criterion (A.I.C.) was calculated for the different values of p and q. The lower the A.I.C. values are, the better the lag values for p and q. Figure 5 illustrates that the A.I.C. value is meager for the lags p=1 and q=0. Accordingly, the ARDL(1,4,3) model is the best among the 20 models investigated with different lag values.

3.4 The ARDL(1,4,3) model

The ARDL(1,4,3) model is employed to study the short-run relationship between the cumulative number of COVID-19-infected cases, the total population as the independent variables, and the cumulative deaths due to COVID-19 as the dependent variable. The statistical findings are reported in Table 3. The results reveal that the overall goodness of fit of the model, as shown by the coefficient of determination, R^2 = 96 %, is extremely high and highly significant, implying that almost 96% of the variation in the dependent variable is explained by the model and the rest is explained by the error term. The value of the D-W statistic is nearly equal to two, confirming no spurious results.

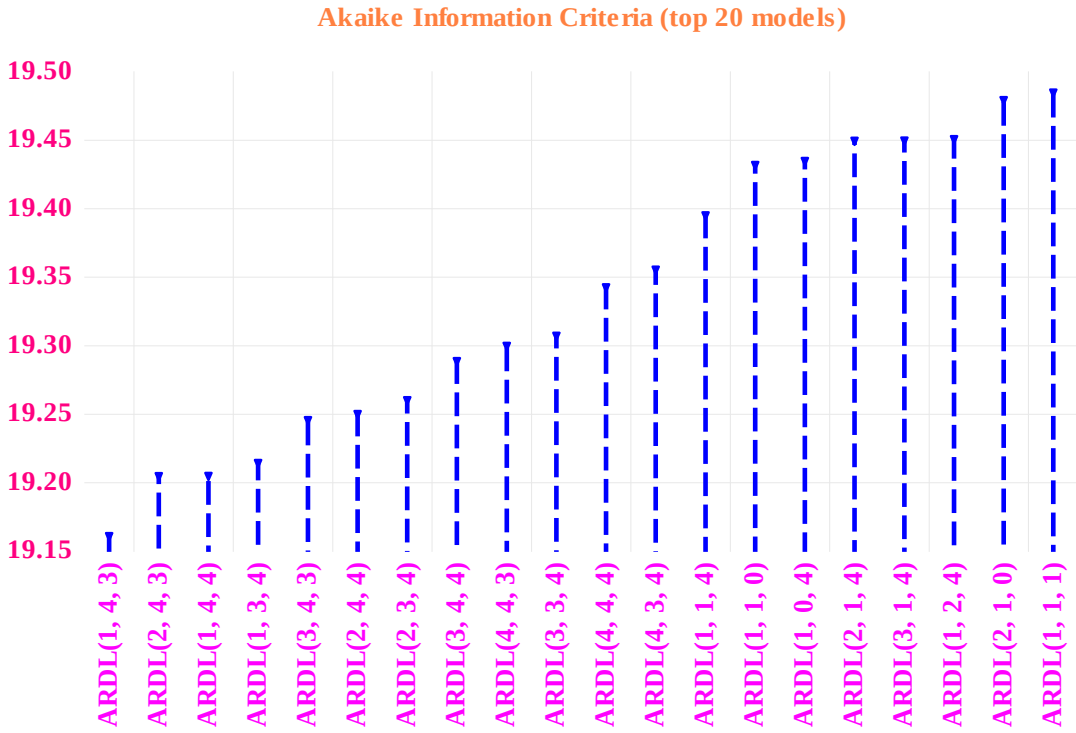

Figure 5: Selection of the appropriate model based on the A.I.C.

Table 3: Characteristics of estimated ADRL(1,4,3) model

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
DEATHS(-1)	-0.2860	0.1466	-1.951	0.062
CASES	0.0160	0.0027	5.846	0.000
CASES(-1)	0.0098	0.0030	3.299	0.003
CASES(-2)	0.0019	0.0025	0.753	0.459
CASES(-3)	0.0077	0.0029	3.113	0.005
CASES(-4)	0.0046	0.0011	4.004	0.000
POPULATION	-0.0029	0.0016	-1.877	0.072
POPULATION(-1)	0.0018	0.0017	1.017	0.319
POPULATION(-2)	-0.0006	0.0017	-0.339	0.738
POPULATION(-3)	-0.0061	0.0016	-3.888	0.000
C	-1665.75	1256.09	-1.326	0.197
R-squared %	0.96	Mean dependent var		20677.14
Adjusted R-squared %	0.94	S.D. dependent var		12648.74
S.E. of regression	3097.52	Akaike info criterion		19.16
Sum squared resid	2.40E+08	Schwarz criterion		19.64
Log-likelihood	-333.90	Hannan-Quinn criteria.		19.33
F-statistic	55.86	Durbin-Watson stat		1.96
Prob(F-statistic)	0.0000			

* p-values and any subsequent tests do not account for model selection

3.5 Test for normality of the residuals

Figure 6 illustrates that the errors are normally distributed, as the Jarque-Bera test statistic's value is non-significant at a 5% significance level.

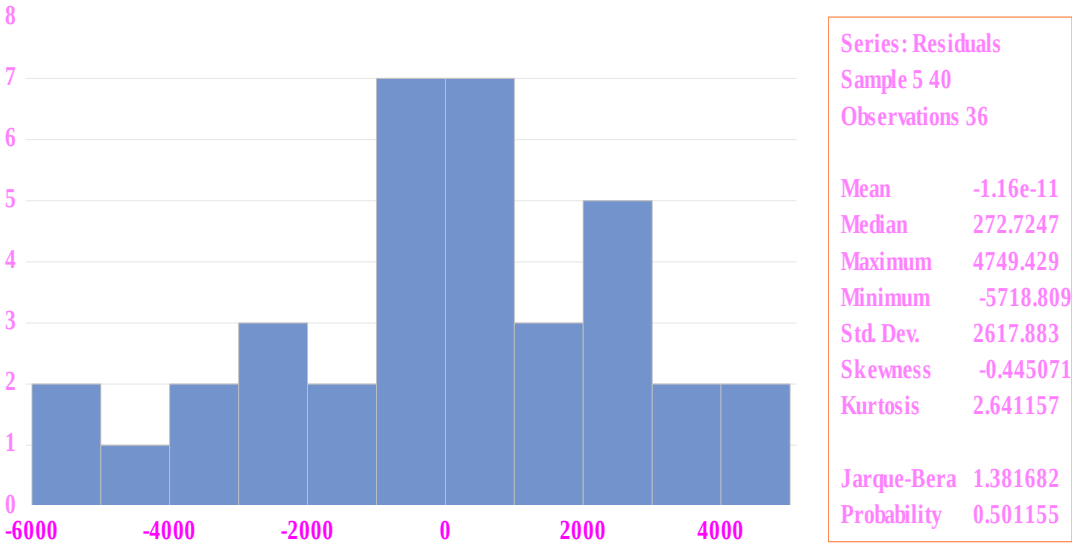

Figure 6: Test for normality of residuals

To ensure the consistency of the ARDL(1,4,3) model, the following residual diagnostic tests are carried out.

3.6 Ljung-Box test for autocorrelation

The results of the Ljung-Box test (Ljung and Box, 1979) indicate that the p-values of the Q statistics are non-significant at a 5% significance level and strongly suggest the absence of autocorrelation in the model error. If there is an autocorrelation of residuals, estimated parameters will not be consistent due to the lagged dependent variable appearing as an exogenous variable in the model.

Table 4: Characteristics of autocorrelation of residuals in different lags

Autocorrelation	Partial Correlation	Lag	A.C.	P.A.C.	Q-Stat	Prob*
. |** |	. |** |	1	0.266	0.266	2.7739	0.096
.*| . |	.*| . |	2	-0.106	-0.190	3.2244	0.199
. | . |	. |*. |	3	0.052	0.152	3.3351	0.343
. |** |	. |*. |	4	0.244	0.184	5.8875	0.208
. |*. |	. | . |	5	0.093	-0.017	6.2655	0.281
.*| . |	.*| . |	6	-0.101	-0.075	6.7304	0.346
**| . |	**| . |	7	-0.247	-0.249	9.6081	0.212
.*| . |	.*| . |	8	-0.136	-0.090	10.505	0.231
. | . |	. | . |	9	0.059	0.069	10.682	0.298
.*| . |	.*| . |	10	-0.126	-0.155	11.520	0.318

*Probabilities may not be valid for this equation specification

3.7 Breusch-Godfrey serial correlation LM test

Usually, when an analysis involves time series data, the possibility of autocorrelation is high. Therefore, it is necessary to test the residuals for autocorrelation using the Breusch-Godfrey (Breusch,1978; Godfrey, 1978) L.M. test. The results presented in Table 5 reveals that the F-statistic value is non-significant at a 5% significance level; hence, the null hypothesis of no serial correlation is accepted, and therefore there is no serial correlation.

Table 5: Characteristics of the Breusch-Godfrey serial correlation L.M. test of the residuals

F-statistic	0.623	Prob. F(2,32)	0.541
Obs*R-squared	1.379	Prob. Chi-Square(2)	0.501

3.8 Breusch-Pagan-Godfrey heteroscedasticity test

To ensure consistency, the study further employed the Breusch-Pagan-Godfrey heteroscedasticity test, and the test results are presented in Table 6. The test results reveals that the F-statistics value is significant at a 1 % level of significance, the null-hypothesis of Homoskedasticity is rejected. Hence, it shows that the error variance is not constant, which is the drawback of the quality of the fitted model.

Table 6: Characteristics of Breusch-Pagan-Godfrey heteroscedasticity test

F-statistic	20.284	Prob. F(3,45)	0.000
Obs*R-squared	28.169	Prob. Chi-Square(3)	0.000
Scaled explained SS	73.409	Prob. Chi-Square(3)	0.000

3.9 Fit of the model

The actual/fitted/residual plot of the ARDL(1,4,3) model depicted in Figure 6 showed that the model's fit is good enough to explain the cumulative total deaths.

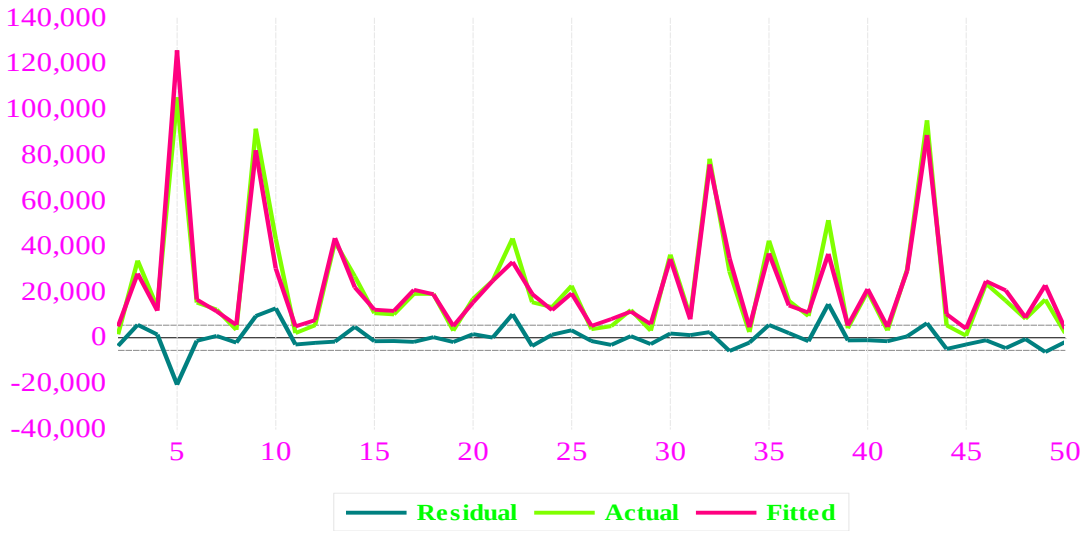

Figure 6: Model Fit

3.10 Model Stability

To check the robustness of our results, structural stability tests of the parameters of the long-run results are performed by the cumulative sum of recursive residuals (CUSUM) and cumulative sum of recursive residuals squares (CUSUMSQ) tests (Brown et al. 1975). This exact procedure has been utilized by Pesaran and Pesaran (1997) and Mohsen et al. (2002) to test the stability of long-run coefficients. A graphical representation of the CUSUM and CUSUMSQ statistics is shown in Figures 7 and 8, respectively. The plots of both the CUSUM and CUSUMSQ are within the boundaries (indicated by the dotted red lines) of the 5% significance level, and these statistics confirm the model's stability.

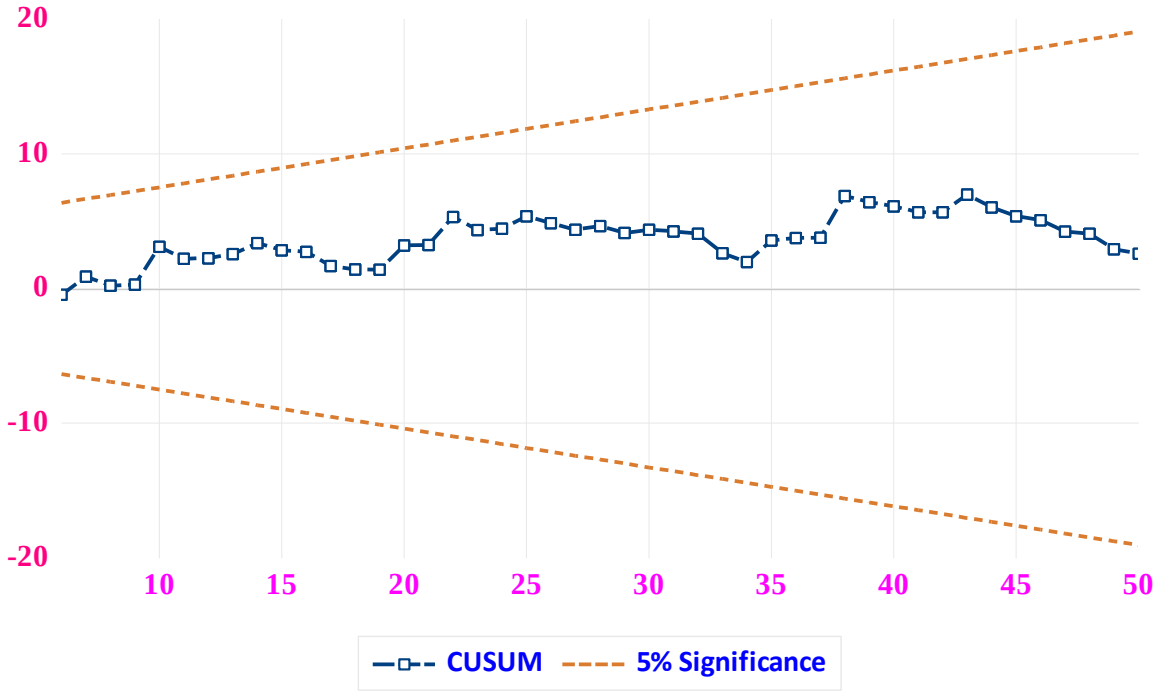

Figure 7: CUSUM stability test

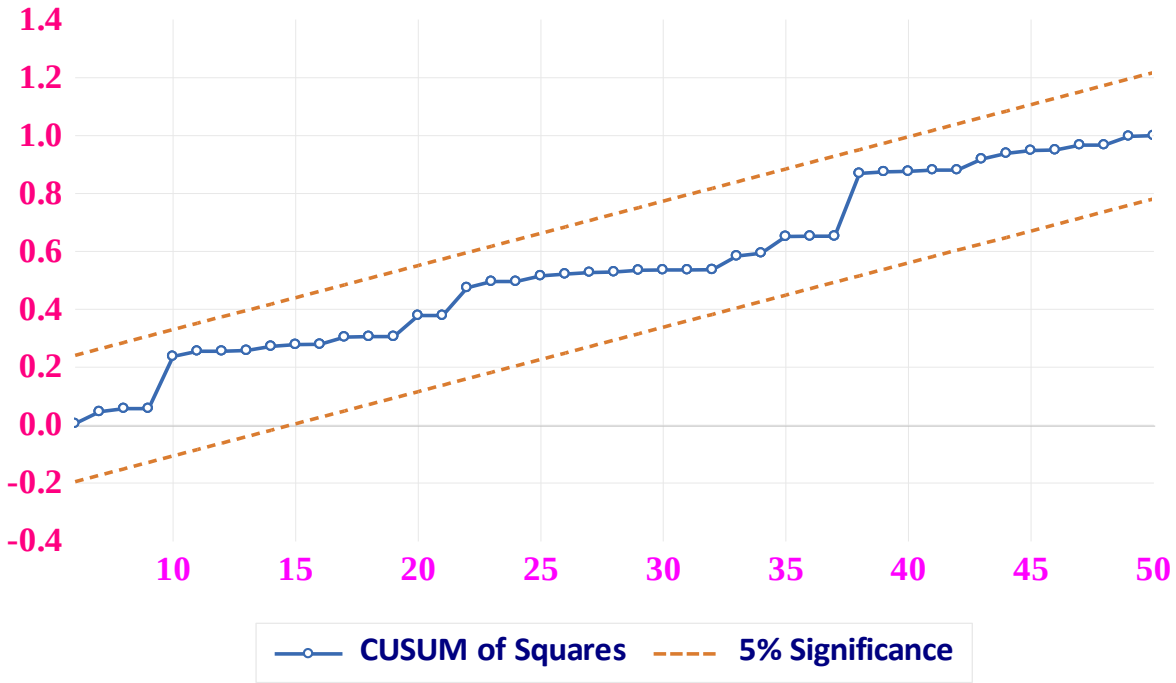

Figure 8: CUSUMSQ stability test

3.11 Bounds test for cointegration

The bounds test Pesaran et al. (2001) tests the cointegration (long-run relationship) between the study variables CASES and DEATHS and is presented in Table 7. The test results reveal that there exists a cointegration relationship between CASES and DEATHS, as the bounds test statistic is greater than the upper bound from I(1) (F-statistics = 20.87 > 3.35), and it is highly significant at a 1% level of significance. Hence, the null hypothesis of "No Levels Relationships" is rejected, which implies the possibility of estimating a log-run cointegration relationship between the study variables.

Table 7: Characteristics of the F-Bounds Test

Test Statistic	Value	Signif.	I(0)	I(1)
		Asymptotic: n=1000		
F-statistic	20.87	10%	2.63	3.35
k	2	5%	3.1	3.87
		2.5%	3.55	4.38
		1%	4.13	5
Actual Sample Size	36	**Finite Sample: n=40**		
		10%	2.835	3.585
		5%	3.435	4.26
		1%	4.77	5.855
		Finite Sample: n=35		
		10%	2.845	3.623
		5%	3.478	4.335
		1%	4.948	6.028

The conditional error correction regression model is presented in Table 8. All the estimated parameters are highly significant at a 1 % significance level except population, which is significant at a 5 % level. Here, the variable E.C.M. (-1) is called the error correction model, and its coefficient value should be negative and significant, which is one of the desirable qualities of the model. E.C.M. (-1) corresponds to the lagged error term equilibrium equation. The coefficient expresses the degree to which the variable DEATH will be recalled towards the long-term target. It is negative and significant at a 1 % level of significance, thus reflecting a relatively quick long-term target adjustment.

Table 8: Characteristics of Conditional Error Correction Regression

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1665.752	1256.090	-1.326	0.197
DEATHS(-1)*	-1.286	0.147	-8.771	0.000
CASES(-1)	0.040	0.007	5.545	0.0000
POPULATION(-1)	-0.008	0.002	-3.717	0.001
D(CASES)	0.016	0.003	5.846	0.000
D(CASES(-1))	-0.014	0.004	-3.411	0.002
D(CASES(-2))	-0.012	0.003	-4.044	0.000
D(CASES(-3))	-0.005	0.001	-4.004	0.000
D(POPULATION)	-0.003	0.002	-1.877	0.002
D(POPULATION(-1))	0.007	0.002	3.274	0.003
D(POPULATION(-2))	0.006	0.002	3.888	0.000

* p-value incompatible with t-bounds distribution

The results presented in Table 9 are the estimates of the long-run variables, and the Error Correction (EC) equation is given at the end of the table.

Table 9: Characteristics of Levels Equation

Variable	Coefficient	Std. Error	t-Statistic	Prob.
CASES	0.031	0.005	5.720	0.000
POPULATION	-0.006	0.002	-3.605	0.001
C	-1295.275	972.433	-1.332	0.195

EC = DEATHS - (0.031*CASES -0.006*POPULATION - 1295.275)

The results in Table 10 show that the error correction model estimates the speed of adjustment to equilibrium in a cointegration relationship. Here, the error correction term derived from the Levels Equation earlier is included among the regressors and is denoted as CointEq. The coefficient associated with this regressor is typically the speed of adjustment to equilibrium in every period. Here, the Coefficient of CointEq is negative and highly significant, which is one of the desirable qualities of the model. Thus, all the variables under study are moving in opposite positive directions.

Table 10: Characteristics of ADRL ECM Regression model parameters

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(CASES)	0.016	0.002980	8.065	0.000
D(CASES(-1))	-0.014	0.002560	-5.522	0.000
D(CASES(-2))	-0.012	0.002083	-5.893	0.000
D(CASES(-3))	-0.005	0.000913	-4.995	0.000
D(POPULATION)	-0.003	0.001343	-2.191	0.038
D(POPULATION(-1))	0.007	0.001200	5.561	0.000
D(POPULATION(-2))	0.006	0.001210	5.049	0.000
CointEq(-1)*	-1.286	0.132999	-9.670	0.000
R-squared %	0.89	Mean dependent var		-2121.139
Adjusted R-squared %	0.86	S.D. dependent var		7780.361
S.E. of regression	2926.882	Akaike info criterion		18.994
Sum squared resid	2.40E+08	Schwarz criterion		19.346
Log-likelihood	-333.899	Hannan-Quinn criteria.		19.117
Durbin-Watson stat	1.460			

* p-value incompatible with t-bounds distribution

4. Conclusion

The ARDL(1,4,3) model is found suitable for investigating the short-run relationships between the study variables. The model is highly significant, and the value of the coefficient of determination, $R^2 = 96\%$, implies that almost 96% of the variation in the dependent variable is explained by the model and that the error term explains the rest. The value of the D-W statistic is nearly equal to two, confirming no spurious results. The bounds test results reveal a long-run relationship between the study variables. The error correction term is negative and highly significant, reflecting a relatively quick adjustment to the long-term target.

References

[1] Alimi, R.S. (2014). ARDL Bounds Testing Approach to Cointegration: A Re-Examination of Augmented Fisher Hypothesis in an Open Economy, Asian Journal of Economic Modelling, 2, pp.103-114.

[2] Breusch, T.S. (1978). Testing for Autocorrelation in Dynamic Linear Models, Australian Economic Papers, 17, pp.334-355.

[3] Breusch, T.S. and Pagan, A.R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation, Econometrica, 47, pp.1287-1294.

[4] Brown, R.L., Durbin, J. and Evans, J.M. (1975). Techniques for Testing the Constancy of Regression Relationships over Time, Journal of the Royal Statistical Society. Series B (Methodological), 37, pp.149-192.

[5] Dickey, D.A. and Fuller, W.A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root, Journal of the American Statistical Association, 74, pp.427-431.

[6] Godfrey, L.G. (1978). Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables, Econometrica, 46, pp.1293-1301.

[7] Jarque, C.M. and Bera, A.K. (1980). Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals, Economics Letters, 6, pp.255-259.

[8] Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shin, Y. (1992). Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root, Journal of Econometrics, 54, pp.159-178.

[9] Ljung, G.M. and Box, G.E.P. (1979). The Likelihood Function of Stationary Autoregressive-Moving Average Models, Biometrika, 66, pp.265-270.

[10] Mohsen, Bahmani-Oskooee and Ng, R.W. (2002). Long run demand for money in Hong Kong: An application of the ARDL model, International Journal of Business and Economics, 1(2), pp.147-155.

[11] Pesaran, M.H. and Pesaran, B. (1997). Working with Microfit 4.0: Interactive Econometric Analysis, Oxford University Press, Oxford, U.K.

[12] Pesaran, M.H., Shin, Y. and Smith, R.J. (2001). Bounds Testing Approaches to the Analysis of Level Relationships, Journal of Applied Economics, 16, pp.289-326.

[13] Phillips, P.C.B. and Perron, P. (1988). Testing for a Unit Root in Time Series Regression, Biometrika, 75, pp.335-346.

[14] Soharwardi, M.A., Khan, R.E.A. and Mushtaq, S. (2018). Long run and short run relationship between financial development and income inequality in Pakistan. Journal of ISOSS, 4(2), pp.105-112.