data

MDPI

Data Descriptor

Panel Regression Modelling for COVID-19 Infections and Deaths in Tamil Nadu, India

Rajarathinam Arunachalam

Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli 627012, India; rajarathinam@msuniv.ac.in

Abstract: The impacts of the coronavirus disease 2019 (COVID-19) pandemic have been extremely severe, with both economic and health crises experienced worldwide. Based on the panel regression model, this study examined the trends and correlations in the number of COVID-19-related deaths and the number of COVID-19-infected cases in all 37 regions of the Tamil Nadu state in India, in August 2020. The fixed effects model had the greatest R^2 value of 78% and exhibited significant results. The slope coefficient was also highly significant, showing a considerable variation in the relationship between new COVID-19 cases and deaths. Additionally, for every unit increase in COVID-19-infected cases, the death rate increased by 0.02%.

Dataset: https://stopcorona.tn.gov.in/

Dataset License: National Data Sharing and Accessibility Policy (NDSAP)

Keywords: Hausman test; random effects model; Wald test; fixed effects model; least squares dummy variable

Citation: Arunachalam, R. Panel Regression Modelling for COVID-19 Infections and Deaths in Tamil Nadu, India. *Data* **2023**, *8*, 158. https://doi.org/10.3390/data8100158

Academic Editor: Henning Müller

Received: 17 August 2023
Revised: 19 October 2023
Accepted: 20 October 2023
Published: 23 October 2023

1. Introduction

1.1. Study Background

The coronavirus epidemic began in Wuhan City (China) on 31 December 2019, and became a pandemic. The incidence of novel COVID-19 infections dramatically increased due to the absence of antiviral medications and vaccines, resulting in enormous economic losses, panic, and many deaths.

Using different statistical models to analyse epidemic data has emerged as a critical study field for predicting the number of COVID-19 deaths and infected individuals.

Statistical models represent the numerical data relevant to specific samples or groups. To assess trends in the data shown, these models frequently take the form of line graphs and scatterplots. While statistical models may display data in various scenarios, those dealing with COVID-19 are particularly valued at present since they provide numerical information about this pandemic, such as the number of cases and deaths brought on by COVID-19. These models have also proved very helpful in localizing cases to specific nations, regions, cities, and specific areas within cities, enabling the authorities in these locations to respond appropriately to the infection. Additionally, models have focused on various crucial traits among individuals who present with COVID-19, such as age, race, sex, and preexisting diseases. This enables researchers to determine which populations are most at risk of infection [1].

Artificial intelligence (AI) techniques built on machine learning (ML) and mathematical models have been utilized to evaluate the epidemic's progress throughout each country and identify any potential amplifying factors that might mitigate its effects [2].

1.2. Literature Review

To examine the relationship between dependent and independent variables and determine the current rate of spread of COVID-19, [3] sought to build on earlier research. This research statistically analysed the relationship of factors such as region, sex, birth year, infection date, and recovery or symptom relief date with the noted number of recovered and dead patients. The findings revealed that region, infection date, and sex were associated with the number of recovered and deceased patients, whereas birth year was only associated with the number of deceased patients. Furthermore, no deaths from COVID-19 were noted among recovered patients, whereas 11.3% of patients who died were confirmed to be COVID-19 positive after their deaths. In South Korea, the main factor associated with the number of infections was the number of patients infected by an unknown source, representing more than 33% of the total number of infected patients.

The association between the overall number of COVID-19 infections and recovered people in various countries were studied and analysed by [4] using the chain-binomial variant of Bailey's model. They also noted that most studies have investigated COVID-19 cases with different regression and time series models commonly used to assess the trend or growth of any illness.

The relationship between the transmission of viral infections and human migration was investigated by [5]. They concluded that the intensity of pedestrian traffic in the research period impacted virus spread after 15–20 days on average.

A time series-based system to track epidemics is a system that [6] aimed to create. Utilizing univariate time series models, the author showed the evolution of the reported incidents in the first stage. Additionally, he combined the models to offer more precise and reliable findings and analysed statistical probability distributions to create hypothetical futures. The "time series susceptible-infected-recovered" (tsiR) model was developed and used in the last stage, and its epidemiological ratio (R_0) was calculated to determine when the epidemic ended. The time series models comprised traditional exponential smoothing, ARIMA techniques, feed-forward artificial neural networks (ANNs), and multivariate adaptive regression splines (MARS) from the ML toolbox. The primary mean and Granger–Newbold and Bates–Granger techniques were included in the combinations. To assess the spread and containment of the epidemic, the tsiR model, as well as the R_0 ratio, was applied. The recommended method was used to monitor the COVID-19 outbreak in Greece.

Using Bailey's model and secondary data, [7] calculated the removal rate, or the percentage of deceased individuals in the infected population. Additionally, regression analysis was performed to demonstrate the linear association between this indicator and the frequencies of all infections. Finally, they discussed the connection between the model and decision-making.

By carefully analysing the cases reported in the country up to 22 April 2020, [8] used exploratory data analysis to create a statistical model to help people understand COVID-19 in India. The study's findings illustrated the daily and weekly effects of COVID-19 in India and drew comparisons between that nation, its neighbours, and other badly afflicted nations.

The impact of travel history and interaction with travellers on the dissemination of COVID-19 in Nigeria was evaluated by [9] using the ordinary least squares (OLS) estimator. They created predictions by extracting data from the Nigeria Centre for Disease Control (NCDC) website from 31 March 2020 to 29 May 2020. The model evaluated the time before and after the Nigerian federal government imposed travel restrictions. Based on the diagnostic checks performed, the fitted model exhibited an excellent fit for the dataset with no validity violations. With travel history and contact with travellers observed to increase the likelihood of COVID-19 infection by 85 and 88%, respectively, the results demonstrated that the government made the right choice in enforcing travel restrictions. The authors concluded that the government must enforce this policy to contain the spread of COVID-19.

Using stochastic modelling, [10] forecasted the prevalence of COVID-19 trends in East African countries, focusing on Somalia, Sudan, Djibouti, and Ethiopia. The study's

findings indicated that, under the average rate scenario, the number of COVID-19-positive individuals in Ethiopia would increase, ranging between 5846 and 56,610 within four months after 30 June 2020.

An autoregressive distributed lag model and limited cointegration tests were used by [11] to evaluate the long-term equilibrium relationship between the cumulative number of new COVID-19 infections (X) and the cumulative number of deaths due to COVID-19 (Y). The stability of the calculated model was also assessed. The consistency of the model parameters was evaluated using the cumulative sum of the recursive residuals and squares tests.

The dynamic relationship between the number of cases and deaths was examined by [12] using the vector error correction model (VECM), the Johnsen–Fisher cointegration test, and the Granger causality test. From 1 April 2020 to 26 December 2020, data on daily new COVID-19 cases and COVID-19-related deaths in India, Ukraine, Canada, and the USA were obtained from the website. Summary figures showed that the United States had the most significant COVID-19 cases, followed by India, Canada, and Ukraine. The USA also had the highest number of COVID-19-related deaths, followed by India, Ukraine, and Canada. Canada led all other countries regarding the death rate, followed by the USA, Ukraine, and India. The results of the Johnsen–Fisher cointegration test indicated that there was only one cointegration equation. The Granger causality test and the VECM demonstrated short- and long-term causal correlations between COVID-19 infection and mortality. The rate of adjustment was 9.9%.

1.3. Objectives of the Present Study

This study aimed to determine the relationships and trends between the number of COVID-19 deaths (DEATH) and the number of new COVID-19 infections (NCASE) in all 37 regions of Tamil Nadu (India) using the number of daily COVID-19-infected cases and deaths in August 2020 based on the preceding discussion. A panel regression model was used, with DEATH as the dependent variable and NCASE as the independent variable.

1.4. Panel Data Model

These data include observations of events gathered over various time scales for the same group of people, entities, or units. Econometric panel data, in a nutshell, are multidimensional data collected over a certain period.

A simple regression model of panel data is defined as

$$Y_{it} = \alpha + \beta X_{it} + v_{it}$$

where $v_{it} = \gamma_i v_{i(t-1)} + \mu_{it}$ represents the predicted residuals obtained from panel regression analysis, Y represents the dependent variable, X denotes the explanatory or independent variable and indicates the intercept and slope, respectively, t represents the tth period, i represents the ith cross-sectional unit, and X is considered to be non-stochastic as well as an error term to follow the classical assumptions, i.e., $v_{it} \sim N(0, \sigma^2)$. In the present research paper, the number of cross-sections (districts) was 37 (i = 1, 2, 3, . . ., 37), and the number of time points was 1, 2, 3, . . ., 30.

Detailed discussions of panel data modelling can be found in [13–17].

Panel data provide "more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency" because they combine time series of cross-sectional observations [14].

2. Materials and Methods

2.1. Materials

The COVID-19 infection and death dataset for August 2020, which included data on all 37 regions of Tamil Nadu, India, was gathered from the Tamil Nadu government's official website. The current study's research objectives were examined using various econometric methodologies linked to panel data regression modelling. The techniques section discusses

several panel data regression modelling strategies. Model and parameter estimates were performed using EViews Ver. 11.

Models based on panel data provide descriptions of individual behaviours across time and individuals. Pooled models (OLS regression) or constant coefficient models (CCMs), RE (random effects), and FE (fixed effects) models are the three different types of models.

2.1.1. Unit Root Tests

Lagrange multiplier (LM) stationarity [18] or the [19] test may be used to check for unit roots inside panel data. The alternative hypothesis is that the panels are stationary, whereas the null hypothesis is that they have unit roots. Based on these findings, one could accept the alternative hypothesis and reject the null hypothesis if the *p*-value is <0.05.

2.1.2. OLS Regression (Pooled Model) or CCM

Cross-sectional analysis often makes the following assumptions about the pooled model with constant coefficients:

$$Y_{it} = \alpha + \beta X_{it} + v_{it}$$

where i = 1, 2, 3, . . ., 37, and t = 1, 2, 3, . . ., 31; here, i represents the ith cross-sectional unit, t represents the tth period, and X indicates a non-stochastic error term that follows the classical assumptions, i.e.,

$$v_{it} \sim N(0, \sigma^2)$$

2.1.3. Individual-Specific Effects Model

We assumed that, for the people who were assessed, α_i exhibits unobserved heterogeneity. The fundamental question is that whether there is a relationship between the individual-specific effects and the regressor. An FE model is used if they are linked. An RE model is used if they are not correlated.

2.1.4. FE Least Squares Dummy Variable (LSDV) Model [17]

The phrase "fixed effects" is applied since every entity's intercept does not fluctuate with time; it is, therefore, time-invariant, although the intercept might change among districts.

$$y_{it} = \alpha + x_{it}\beta + \gamma_{it}$$

After estimating, the individual-specific result is obtained as

$$\hat{\alpha_i} = \overline{y}_i - \overline{x}_i\hat{\beta}$$

In other words, individual-specific impacts are the residual variance in the dependent variable that the regressor cannot account for. The fixed effects intercept might differ among the districts when utilizing the dummy variable approach.

2.1.5. RE Model

It is assumed that the regressor is not affected by the individual-specific effects α_i, which are included as α_i in the error term. The composite error term and slope parameters are the same for each person, i.e.,

$$y_{it} = x_{it}\beta + (\alpha_i + v_{it})$$

Here $\text{var}(\varepsilon_{it}) = {\sigma_\alpha}^2 + {\sigma_v}^2$ and $\text{cov}(\varepsilon_{it}, \varepsilon_{is}) = {\sigma_\alpha}^2$, so $\rho_\varepsilon = cor(\varepsilon_{it}, \varepsilon_{is}) = \frac{{\sigma_\alpha}^2}{{\sigma_\alpha}^2 + {\sigma_\varepsilon}^2}$.

Rho indicates the error's interclass correlation or the percentage of its variation accounted for by person-specific effects. If the individual effects exceed the idiosyncratic mistake, it becomes closer to 1.

2.1.6. Hausman Test

The RE model is favoured, and the null hypothesis of the given test and that of the FE model are selected, with the latter being the alternative hypothesis. The null hypothesis is the one that assumes there is no correlation between the regressor and (α_i), and the Hausman test [20] checks for their relationship. If the Hausman test suggests using the RE estimator, it should be used because it is more effective. Only the time-varying regressors could be used to calculate the Hausman test statistic.

$$H = (\hat{\beta}_{RE} - \hat{\beta}_{FE})^{|}(V(\hat{\beta}_{RE}) - V(\hat{\beta}_{FE})(\hat{\beta}_{RE} - \hat{\beta}_{FE})$$

2.1.7. Wald Test

The Wald test [21] determines which model variables significantly contribute to the observed impact. The test, also known as the Wald chi-squared test, can be utilized to examine whether explanatory variables within a model are important, namely, whether they add to the model's explanatory power. Variables with no explanatory power can be removed from the model without having any significant influence. One parameter that equals some value is the test's null hypothesis.

3. Results and Discussion

3.1. Unit Root Tests

It is crucial for time series data studies that the research variables remain stationary, which indicates that the variable data's variances and means are the same. "Levin–Lin–Chu unit root tests" were performed to determine if the research variables—NCASE and DEATH—were stationary. Table 1 presents the findings.

Table 1. Unit root test outcomes for variables DEATH and NCASE.

Variables	NCASE	DEATH
Method	Levin, Lin, and Chu t	
Statistic	−8.6252	−8.6611
Prob **	0.0000	0.0000

** Probabilities were calculated assuming asymptotic normality.

The NCASE and DEATH variables are shown to be stationary in Table 1 because the method used was highly significant ($p < 0.0000$). As a result, the analysis's variables were stationary.

3.2. Summary Statistics

The number of COVID-19-infected cases reported in the various regions of Tamil Nadu in August 2020 is shown in Figure 1. The most significant numbers of COVID-19 infections were noted in Chennai (35,491), followed by Coimbatore (11,504), Thiruvallur (11,334), Chengalpattu (10,517), and Tirunelveli (8393). The smallest numbers of new COVID-19 infections were reported in Krishnagiri (917), Dharmapuri (802), and Nilgiris (502). Overall, in August 2020, 181,817 COVID-19 infections were reported in Tamil Nadu.

The above Figure 2 shows the maximum number of fatalities due to COVID-19 in Chennai (663), followed by Coimbatore (250), Thiruvallur (138), Chengalpattu (156), Tirunelveli (138), and Kanyakumari (135). Nine deaths were registered in Dharmapuri and Nilgiris, the lowest number among the districts. In August 2020, 3387 deaths were reported due to COVID-19 in Tamil Nadu, for a monthly death rate of 0.02%.

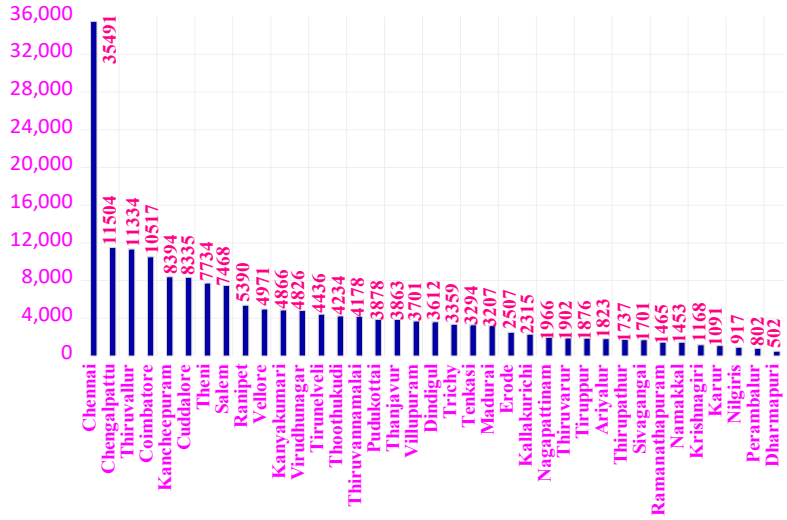

Figure 1. Total number of new COVID-19-infected cases in August 2020.

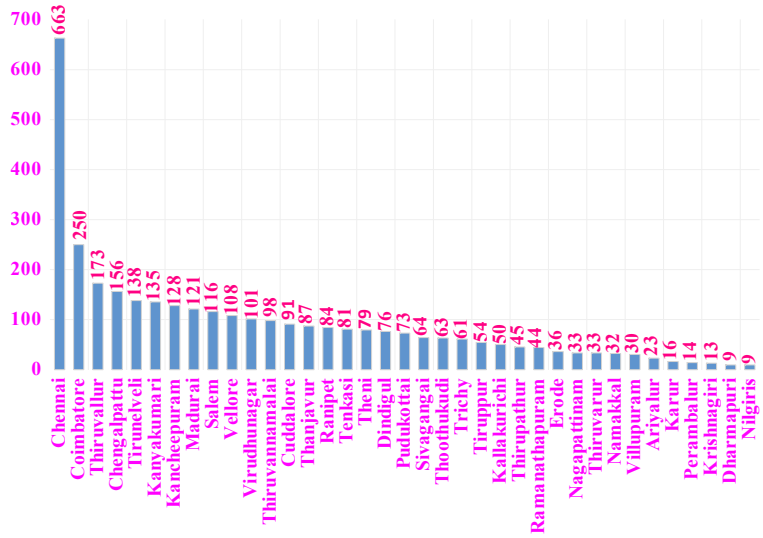

Figure 2. Total number of COVID-19-related deaths in August 2020.

3.3. Differences between Districts

ANOVA tests were performed separately for NCASE and DEATH to assess the differences across districts regarding the number of COVID-19-infected cases and COVID-19-related fatalities. The findings are shown in Tables 2 and 3.

Table 2. Analysis findings of the mean equality of COVID-19 infections.

Method	df	Value	Probability
ANOVA F test	(36, 1110)	364.6168	0.0000
Welch F test	(36, 389.769)	191.3449	0.0000
	Analysis of Variance		
Between	36	41,239,909	1,145,553.00
Within	1110	3,487,398	3141.80
Total	1146	44,727,306	39,029.06

Table 3. Test findings of mean equality of COVID-19-related deaths.

Method	df	Value	Probability
ANOVA F test	(36, 1110)	105.9176	0.0000
Welch F test	(36, 390.212)	44.55014	0.0000
Analysis of Variance			
Between	36	14,006.17	389.06
Within	1110	4077.290	3.67
Total	1146	18,083.46	15.78

The findings showed significant differences between the districts, as the ANOVA tests were highly significant ($p < 0.0000$) for the research variables. This indicated that there were considerable disparities in the number of infections reported in various areas, as well as the number of deaths.

3.4. A Model with Constant Coefficients or Pooled OLS Regression

In a panel least squares analysis, NCASE and DEATH were the dependent and independent variables. Table 4 displays the regression findings based on EViews, Version 11.

Table 4. Findings from a model with constant coefficients or pooled OLS regression.

Variable	Coefficient	Std Error	T Statistic	Prob.
C	0.2940	0.0830	3.5432	0.0004
NCASE	0.0168	0.00033	51.1907	0.0000
Durbin-Watson stat	1.4981	Prob. (F-Statistic)		0.0000
Hannan-Quinn criterion	4.4121	F-Statistic		2620.49
Schwarz criterion	4.4175	Log-likelihood		−2526.41
Akaike info criterion	4.4087	Sum squared resid.		5498.77
SD dependent var.	3.9724	SE of regression		2.19
Mean dependent var.	2.9529	Adjusted R-squared (%)		0.70
Root MSE	2.1895	R-squared (%)		0.70

According to the findings, the slopes and intercept were highly significant, and the model F-statistic was also quite substantial, with an extraordinarily high R^2 of 70%. This demonstrated a direct correlation between an increase in the number of COVID-19 cases and a variation in COVID-19-related deaths. Additionally, as previously mentioned, the DEATH rate increased by 0.02 percent for every unit increase in NCASE.

The main issue with this model is that it did not differentiate between the various districts or inform us whether the overall COVID-19 mortality response to the explanatory variable over time was consistent across all districts. As a result, there is a good chance that the error term and the model's regressor could be associated. If this is the case, the calculated coefficients in the abovementioned model could be biased and inconsistent.

3.5. FE LSDV Model

The dummy variable approach was applied to create this FE model. The model is expressed as

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_{3i} D_{3i} + \alpha_4 D_{4i} + \ldots + \alpha_{37} D_{37i} + \beta_2 X_{it} + \nu_{it}$$

where $D_{2i} = 1$ if the observation was from the Chengalpattu region and 0 otherwise, $D_{3i} = 1$ if it was from Chennai and 0 otherwise, $D_{4i} = 1$ if it was from Coimbatore and 0 otherwise, and so on. In this case, the baseline or reference category was the district of Ariyalur. As a result, the intercept shows the intercept value for the Ariyalur region. In contrast, the other coefficients of α show how much the intercept values for the different regions deviate from

the Ariyalur district's intercept value. Therefore, α_2 indicates how much the intercept's 2nd district value, Chengalpattu, differs from α_1. The sum $(\alpha_2 + \alpha_2)$ provides the intercept's actual value for Chengalpattu. Similar calculations may be performed for the intercept values of the remaining districts.

The findings shown in Table 5 demonstrate that the FE model is highly significant, with an impressive R^2 of 78%. The slope coefficient for COVID-19 infections is also highly significant, indicating that COVID-19 infections displayed considerable fluctuations in the link to COVID-19-related deaths. Several negative dummy variable coefficients were discovered, but none were significant. The dummy variables for Nagapattinam, Karur, Kanyakumari, Erode, Dharmapuri, Cuddalore, Coimbatore, Chennai, Chengalpattu, Ramanathapuram, Salem, Sivaganga, Tenkasi, Thanjavur, Theni, Thiruvannamalai, Thiruvarur, Tirunelveli, Tiruppur, Trichy, Vellore, Virudhunagar, and Villupuram were highly significant, indicating that it is possible that these district changes were heterogeneous and that the results from the combined regression model may not be helpful. Moreover, the slope coefficient values in Table 5 are also different, which raises additional questions about the outcomes in Table 4. Furthermore, there is no autocorrelation in the FE model if the value of Durbin–Watson d is closer to 2. Therefore, the FE model is superior to the pooled regression paradigm.

Table 5. Regression model FE or LSDV results.

Coefficient	Estimated Coefficient	Std Error	t Statistic	Prob
C(1)	0.3972	0.2568	1.5468	0.1222
C(2)	0.0051	0.0010	5.1039	0.0000
C(3)	2.7289	0.5042	5.4122	0.0000
C(4)	15.1093	1.1413	13.2383	0.0000
C(5)	5.9247	0.4869	12.1681	0.0000
C(6)	1.1572	0.4543	2.5472	0.0110
C(7)	−0.1901	0.4232	−0.4491	0.6534
C(8)	1.4559	0.4177	3.4858	0.0005
C(9)	0.3487	0.4169	0.8365	0.4030
C(10)	0.8321	0.4170	1.9954	0.0462
C(11)	2.3410	0.4551	5.1438	0.0000
C(12)	3.1514	0.4223	7.4619	0.0000
C(13)	−0.0618	0.4203	−0.1471	0.8831
C(14)	−0.1714	0.4200	−0.4081	0.6833
C(15)	2.9747	0.4170	7.1332	0.0000
C(16)	0.3416	0.4176	0.8180	0.4135
C(17)	0.3943	0.4189	0.9412	0.3468
C(18)	−0.2588	0.4212	−0.6147	0.5389
C(19)	−0.0785	0.4216	−0.1861	0.8524
C(20)	1.3151	0.4183	3.1446	0.0017
C(21)	0.7794	0.4189	1.8607	0.0631
C(22)	1.4194	0.4254	3.3365	0.0009
C(23)	2.1073	0.4439	4.7475	0.0000
C(24)	1.3855	0.4182	3.3130	0.0010
C(25)	1.6699	0.4171	4.0034	0.0001
C(26)	1.7692	0.4183	4.2295	0.0000
C(27)	0.8697	0.4469	1.9460	0.0519
C(28)	0.7666	0.4181	1.8336	0.0670
C(29)	3.3055	0.5011	6.5960	0.0000
C(30)	2.0718	0.4193	4.9412	0.0000
C(31)	0.3522	0.4177	0.8431	0.3993
C(32)	0.9335	0.4195	2.2253	0.0263
C(33)	3.3194	0.4203	7.8978	0.0000
C(34)	1.0339	0.4178	2.4749	0.0135
C(35)	1.0140	0.4172	2.4303	0.0152

Table 5. *Cont.*

Coefficient	Estimated Coefficient	Std Error	t Statistic	Prob
C(36)	2.2630	0.4229	5.3513	0.0000
C(37)	2.0612	0.4221	4.8831	0.0000
Durbin-Watson stat	1.9148	Prob(F-statistic)		0.0000
Hannan-Quinn criterion	4.2091	F-statistic		109.10
Schwarz criterion	4.3104	Log-likelihood		−2341.69
Akaike info criterion	4.1477	Sum squared resid		3984.57
SD dependent var	3.9724	SE of regression		1.89
Mean dependent var	2.9529	Adjusted R-squared (%)		0.77
Root MSE	1.8638	R-squared (%)		0.78

3.6. Wald Test

We used the Wald test to examine whether the pooled OLS or FE model was more appropriate. The null hypothesis in this situation is that the OLS regression model is suitable (all dummy variables are equivalent to 0), and the alternative hypothesis is that the FE model is suitable (all dummy variables are not equivalent to 0). Thus, this test was performed, and the findings are expressed in Table 6.

Table 6. Findings of the Wald test.

Test Statistic	Value	df	Probability
F-statistic	12.05191	(35, 1110)	0.0000
Chi-square	421.8168	35	0.0000

The FE or LSDV regression model was more suitable than the panel pooled regression model as per the Wald test F-statistic, which was highly significant ($p < 0.0000$). Not every dummy variable had a value of zero.

3.7. RE Mode

Table 7 displays the test results for the RE model, which utilizes the number of COVID-19-related deaths as the dependent variable and NCASE as the independent variable. The RE model explains only 24 percent of the variance in DEATH compared to that in NCASE. The cross-sectional effects individually amount to 0.2 percent according to the rho value of 0.1839.

Table 7. Fitted RE model results.

Variable	Coefficient	Std. Error	t- Statistic	Prob.
C	0.9431	0.1883	5.0095	0.0000
NCASE	0.0127	0.0006	19.6753	0.0000
Effects Specification				
Cross-sectional random			0.8997	0.1839
Idiosyncratic random			1.8955	0.8161
Weighted Statistics				
Root MSE	1.9676	R-squared (%)		0.24
Mean dependent var	1.0450	Adjusted R-squared (%)		0.24
S.D. dependent var	2.2558	S.E. of regression		1.97
Sum squared resid	4440.53	F-statistic		358.63

Table 7. *Cont.*

Variable	Coefficient	Std. Error	t- Statistic	Prob.
Durbin-Waston stat	1.7848	Prob (F-statistic)		0.0000
Unweighted Statistics				
Sum squared resid	6248.85	Durbin-Watson stat.		1.2683
R-squared (%)	0.65	Mean dependent var		2.9529

3.8. Hausman Test

The RE model performed better. The FE and RE estimators were compared using the Hausman test to observe whether there was a significant variation. The statistic of the Hausman test was significant, and the null hypothesis was rejected, according to the findings shown in Table 8, demonstrating the suitability of the FE model. The Hausman test yielded an R^2 value of 80%, which was exceptionally high. This observation refuted the conclusion that the RE model was suitable. Additionally, the regressor variable's RE and FE coefficient values show high statistically significance in the final row of Table 8.

Table 8. Results of the Hausman test (test cross-sectional REs).

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-sectional random	91.94	1	0.0000
Cross-sectional random effects test comparisons:			
Variable	**Fixed**	**Random**	**Var (Diff.)**	**Prob.**
NCASE	0.005159	0.012679	0.000001	0.0000
Cross-sectional random effects test equation:			
Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**
C	2.1351	0.1704	12.5334	0.0000
NCASE	0.0052	0.0010	5.0832	0.0000
Effects Specification			
Durbin-Watson stat	1.9150	Prob (F-statistic)		0.0000
Hannan-Quinn criterion	4.2125	F-statistic		106.06
Schwarz criterion	4.3165	Log-likelihood		−2341.67
Akaike info criterion	4.1494	Sum squared resid		3984.46
S.D. dependent var	3.9724	S.E. of regression		1.90
Mean dependent var	2.9529	Adjusted R-squared (%)		0.77
Root MSE	1.8638	R-squared (%)		0.78

4. Conclusions

A pooled regression model was not appropriate for analysing trends and the link between new COVID-19 infections and COVID-19-related mortality. The ANOVA test results showed significant variation across districts. The most excellent numbers of new cases of COVID-19 were reported in Chennai (35,491), followed by Coimbatore (11,504), Thiruvallur (11,334), Chengalpattu (10,517), and Tirunelveli (8393). The lowest numbers of new cases of COVID-19 were reported in Krishnagiri (917), Dharmapuri (802), and Nilgiris (502). Overall, in August 2020, 181,817 COVID-19-infected patients were registered across Tamil Nadu. The most significant number of deaths due to COVID-19 occurred in Chennai (663), followed by Coimbatore (250), Thiruvallur (138), Chengalpattu (156), Tirunelveli (138), and Kanyakumari (135). Nine deaths were registered in Dharmapuri and Nilgiris, the lowest figure among the districts. In August 2020, 3387 deaths were reported due to COVID-19 in Tamil Nadu, India. The fixed effects model, which had the most incredible R^2 value of 78%, was significant. The slope coefficient was also highly significant, showing significant variation in the relationship between new COVID-19 cases and deaths due

to COVID-19. Additionally, for every unit increase in COVID-19 cases, the death rate increased by 0.02%.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available on Tamil Nadu's official website (https://stopcorona.tn.gov.in/) (accessed on 16 August 2023).

Conflicts of Interest: The author declares no conflict of interest.

References

1. Voyages, A. The Importance of Statistical Modeling for the COVID-19 Pandemic. *Young Sci. J.* 2021.
2. Santosh, K.C. AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data. *J. Med. Syst.* **2020**, *44*, 93. [CrossRef] [PubMed]
3. Al-Rousan, N.; Al-Najjar, H. Data Analysis of Coronavirus COVID-19 Epidemic in South Korea Based on Recovered and Death Cases. *J. Med. Virol.* **2020**, *92*, 1603–1608. [CrossRef]
4. Gondauri, D.; Mikautadze, E.; Batiashvili, M. Research on COVID-19 Virus Spreading Statistics Based on the Examples of the Cases from Different Countries. *Electron. J. Gen. Med.* **2020**, *17*, em209. [CrossRef]
5. Gondauri, D.; Batiashvili, M. The Study of the Effects of Mobility Trends on the Statistical Models of the COVID-19 Virus Spreading. *Electron. J. Gen. Med.* **2020**, *17*, em243. [CrossRef]
6. Katris, C. A Time Series-Based Statistical Approach for Out Break Spread Forecasting: Application of COVID-19 in Greece. *Expert Syst. Appl.* **2021**, *166*, 114077. [CrossRef]
7. Kumar, A. Application of Mathematical Modeling in Public Health Decision Making about Control of COVID-19 Pandemic in India. *Epidemiol. Int.* **2020**, *5*, 23–26.
8. Mittal, S. International Institute of Information Technology-Banglore. An Exploratory Data Analysis of COVID-19 in India. *Int. J. Eng. Res. Technol.* **2020**, *9*, 580–584. [CrossRef]
9. Ogundokun, R.O.; Lukman, A.F.; Kibria, G.B.M.; Awotunde, J.B.; Aladeitan, B.B. Predictive Modelling of COVID-19 Confirmed Cases in Nigeria. *Infect. Dis. Model.* **2020**, *5*, 543–548. [CrossRef] [PubMed]
10. Takele, R. Stochastic Modelling for Predicting COVID-19 Prevalence in East African Countries. *Infect. Dis. Model.* **2020**, *5*, 598–607. [CrossRef] [PubMed]
11. Rajarathinam, A.; Tamilselvan, P. Autoregressive Distributed Lag Model of COVID-19 Cases and Deaths. *Appl. Math. Inf. Sci.* **2021**, *10*, 767–777.
12. Rajarathinam, A.; Tamilselvan, P. Vector Error Correction Modeling of COVID-19 Infected Cases and Deaths. *J. Stat. Appl. Probab.* **2021**, *11*, 205–214.
13. Hsiao, C. *Analysis of Panel Data*; Cambridge University Press: Cambridge, UK, 2003.
14. Baltagi, B.H. *Econometric Analysis of Panel Data*, 4th ed.; Standards Information Network: New York, NY, USA, 2012.
15. Allison, P.D. *Fixed Effects Regression Models*; SAGE Publications: Thousand Oaks, CA, USA, 2009.
16. Biorn, E. *Econometrics of Panel Data: Methods and Applications*; Oxford University Press: London, UK, 2016.
17. Gujarati, D.N.; Porter, D.C. *Basic Econometrics*, 6th ed.; McGraw-Hill Education: Singapore, 2017.
18. Hadri, K. Testing for Stationarity in Heterogeneous Panel Data. *Econ. J.* **2000**, *3*, 148–161. [CrossRef]
19. Levin, A.; Lin, C.-F.; James Chu, C.-S. Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties. *J. Econ.* **2002**, *108*, 1–24. [CrossRef]
20. Hausman, J.A. Specification Tests in Econometrics. *Econometrica* **1978**, *46*, 1251. [CrossRef]
21. Wald, A. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large. *Trans. Am. Math. Soc.* **1943**, *54*, 426–482. [CrossRef]