Heliyon 6 (2020) e03747

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Investigating the cases of novel coronavirus disease (COVID-19) in China using dynamic statistical techniques

Samuel Asumadu Sarkodie *, Phebe Asantewaa Owusu

Nord University Business School, Norway

ARTICLE INFO

Keywords:
COVID-19
Novel coronavirus disease
Cases of novel coronavirus
Modelling COVID-19
China
Econometrics
Economics
Environmental economics
Environmental science
Health economics
Public health

ABSTRACT

The initial investigation by local hospital attributed the outbreak of the novel coronavirus disease (COVID-19) to pneumonia with unknown cause that appeared like the 2003 severe acute respiratory syndrome (SARS). The World Health Organization declared COVID-19 as public health emergency after it spread outside China to several countries. Thus, an assessment of the novel coronavirus disease (COVID-19) with novel estimation approaches is essential to the global debate. This study is the first to develop both time series and panel data models to construct conceptual tools that examine the nexus between death from COVID-19 and confirmed cases. We collected daily data on four health indicators namely deaths, confirmed cases, suspected cases, and recovered cases across 31 Provinces/States in China. Due to the complexities of the COVID-19, we investigated the unobserved factors including environmental exposures accounting for the spread of the disease through human-to-human transmission. We used estimation methods capable of controlling for cross-sectional dependence, endogeneity, and unobserved heterogeneity. We predicted the impulse-response between confirmed cases of COVID-19 and COVID-19-attributable deaths. Our study revealed that the effect of confirmed cases on the novel coronavirus attributable deaths is heterogeneous across Provinces/States in China. We found a linear relationship between COVID-19 attributable deaths and confirmed cases whereas a nonlinear relationship was confirmed for the nexus between recovery cases and confirmed cases. The empirical evidence revealed that an increase in confirmed cases by 1% increases coronavirus attributable deaths by ~0.10%–~1.71% (95% CI). Our empirical results confirmed the presence of unobserved heterogeneity and common factors that facilitates the novel coronavirus attributable deaths caused by increased levels of confirmed cases. Yet, the role of such a medium that facilitates the transmission of COVID-19 remains unclear. We highlight safety precaution and preventive measures to circumvent the human-to-human transmission.

1. Introduction

On 31 December 2020, the World Health Organization (WHO) received information on an outbreak with unknown aetiology detected in a seafood market located in the city of Wuhan, Hubei Province, China. The 2019 novel coronavirus was detected in 44 case-patients with pneumonia with unknown cause between 31 December 2019 to 3 January 2020 by the Chinese authorities [1]. On 11 February 2020, WHO named the novel coronavirus disease as COVID-19 and declared the infectious disease as a public health emergency, after spreading from China to other 24 countries [2]. As of 20 February 2020 (04:00 GMT), 76,498 cases had been reported globally including from China (75,245), "Diamond Princess" cruise ship and others (634), South Korea (104), Japan (94), Singapore (84), Hong Kong (67), Thailand (35), Taiwan (24), Malaysia (22), Germany (16), Vietnam (16), Australia (15), the US (15), France (12), Macau (10), United Arab Emirates (9), UK (9), Canada (8), Italy (3), Philippines (3), India (3), Iran (2), Russia (2), Spain (2), Nepal (1), Cambodia (1), Belgium (1), Finland (1), Sweden (1), Egypt (1), and Sri Lanka (1).

Following the emergence of COVID-19, several studies have examined the transmission dynamics of the infectious disease [3]. While clinical, epidemiological, laboratory, and radiological features of COVID-19 [4] have been reported, phenomenological models using statistical methods have been used to examine epidemiological data [5, 6]. The COVID-19 is reported to have spread through human-to-human transmission [3]. However, it might be possible that other unobserved environmental exposures may have facilitated the rate the disease spreads through human-to-human transmission. Earlier studies based on phenomenological models fail to capture unobserved factors and heterogeneity, which are useful in understanding cases with limited

* Corresponding author.
E-mail address: asumadusarkodiesamuel@yahoo.com (S.A. Sarkodie).

https://doi.org/10.1016/j.heliyon.2020.e03747
Received 5 March 2020; Received in revised form 8 March 2020; Accepted 1 April 2020

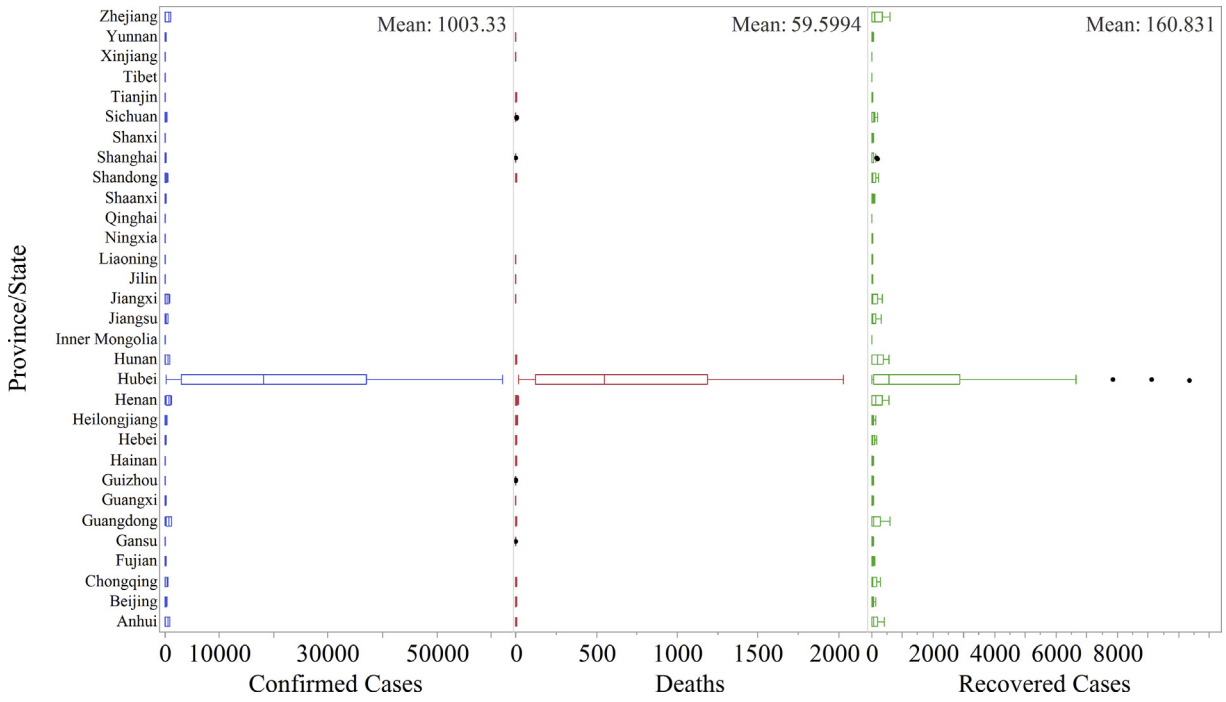

Figure 1. Descriptive statistics of COVID-19 across Provinces/States in China.

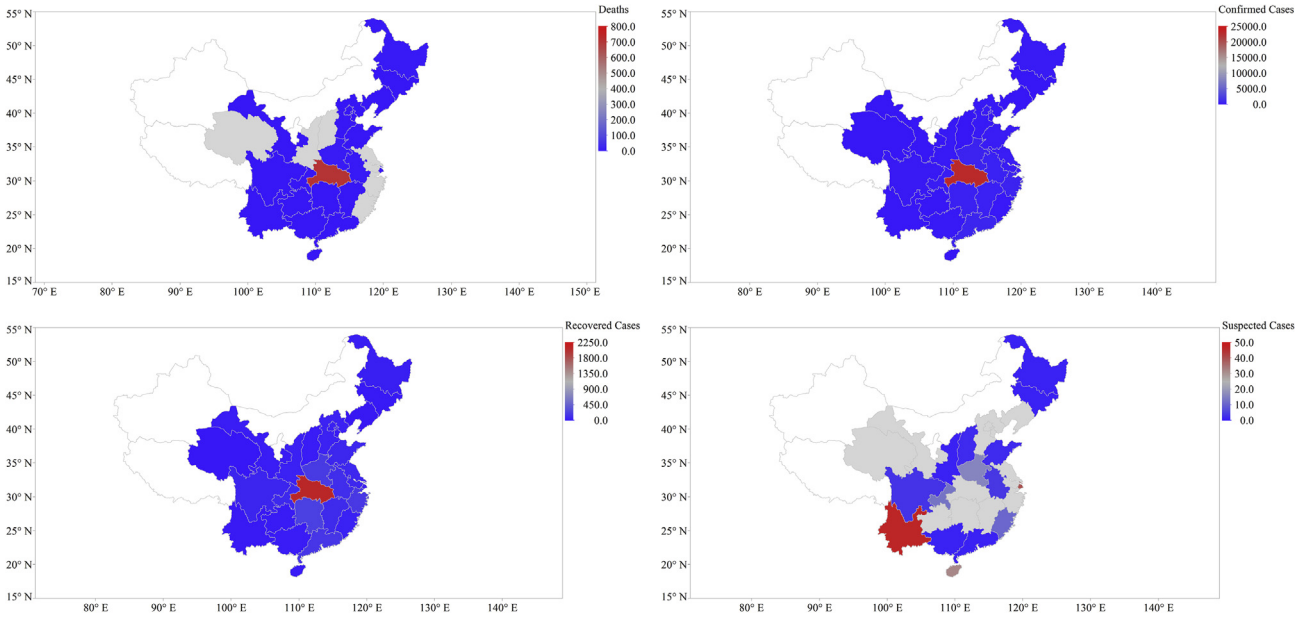

Figure 2. Provinces/States distribution of COVID-19 across China (a) deaths (b) Confirmed cases (c) Recovery cases (d) Suspected cases.

epidemiological data. The complexities of the unobserved factors accounting for COVID-19 underpin this study. Using publicly available data for 31 Provinces/States across China, this study is the first to develop both time series and panel data models to examine the nexus between the novel coronavirus attributable deaths and confirmed cases of COVID-19. We use novel estimation methods capable of accounting for Provinces/States-specific fixed-effects and unobserved heterogeneity of the human-to-human transmission.

2. Materials & method

2.1. Data description

Data were collated on 20 February 2020 from the Center for Systems Science and Engineering at John Hopkins University[1]. The data spans

[1] https://systems.jhu.edu/research/public-health/ncov/.

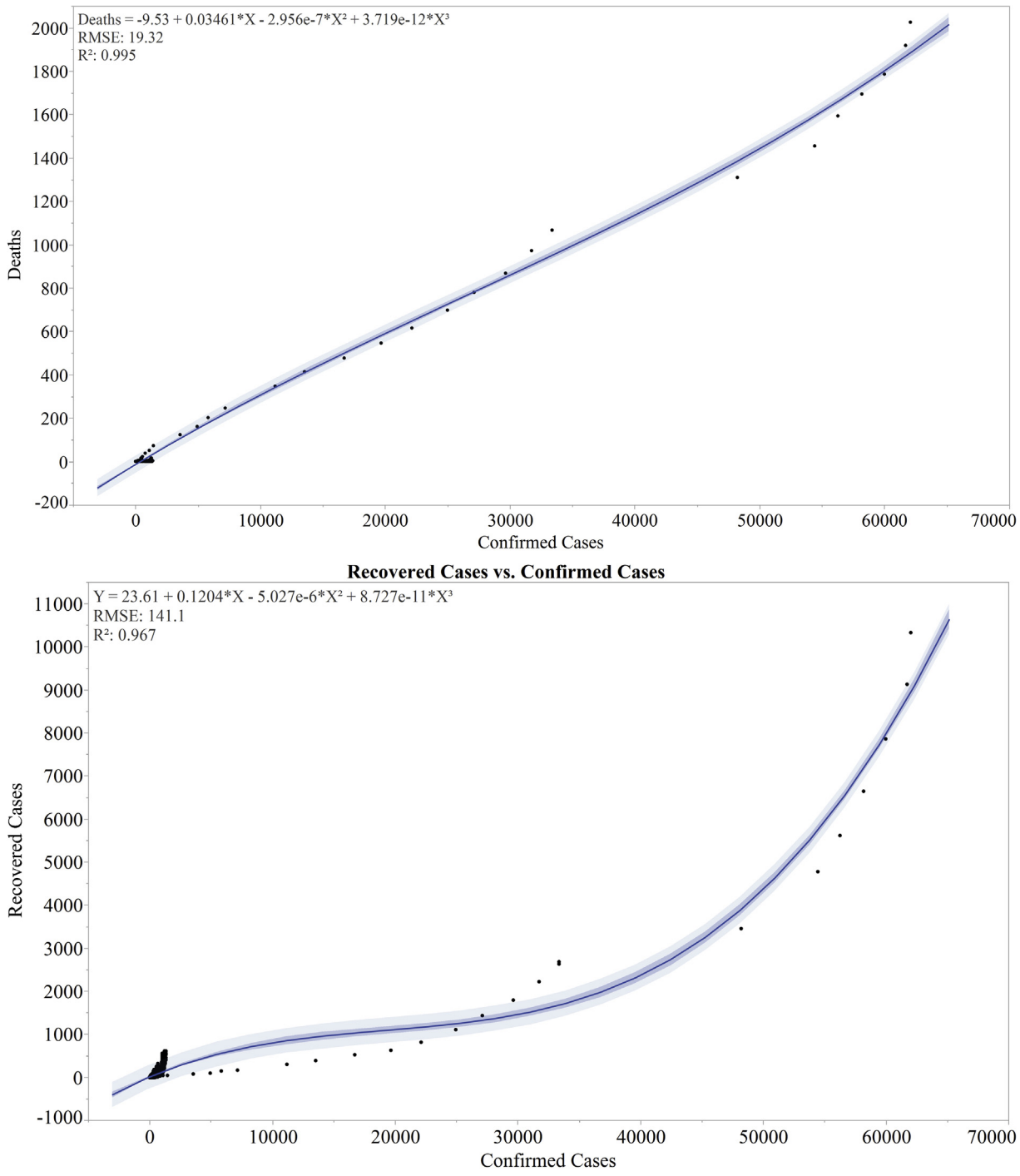

Figure 3. Relationship between (a) death and confirmed cases (b) recovery cases and confirmed cases.

from 21 January 2020 to 20 February 2020 and were preprocessed from wide to long, a replica of panel data and time series setting. The data consist of four health indicators such as deaths, confirmed cases, suspected cases, and recovered cases across 31 Provinces/States in China namely Anhui, Beijing, Chongqing, Fujian, Gansu, Guangdong, Guangxi, Guizhou, Hainan, Hebei, Heilongjiang, Henan, Hubei, Hunan, Inner Mongolia, Jiangsu, Jiangxi, Liaoning, Jilin, Ningxia, Qinghai, Shaanxi, Shandong, Shanxi, Shanghai, Tianjin, Tibet, Sichuan, Zhejiang, Yunnan and Xinjiang. Our intial observation of data available and presented in Figure 1 shows a widespread of case-patients in Hubei Province compared to other locations (Figure 2). This validates the exact location, the city of Wuhan, where the outbreak was first reported. We observe a daily average of about 1000 confirmed cases, 60 deaths and 161 recovered cases.

To use appropriate estimation methods, we examined the characteristics of the data series. We assessed whether the relationship between the novel coronavirus attributable deaths, recovery cases and confirmed cases of COVID-19 was linear or nonlinear. The plot presented in Figure 3

Table 1. Parameter estimation of the nexus between novel coronavirus attributable deaths and confirmed cases of COVID-19.

Variable	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[b]	Model 7[b]
$lnDeaths_{t-1}$	0.8487*** [0.0274]	0.8617*** [0.0381]	0.8617*** [0.0230]	0.8054*** [0.2906]	-0.3121*** [0.0703]	—	0.8080*** [0.0271]
lnConfirmedCases	0.1091*** [0.0273]	0.0961*** [0.0346]	0.0961*** [0.0209]	1.7075** [0.6739]	1.0252*** [0.3378]	0.9149*** [0.0384]	0.1329*** [0.0166]
constant	-0.4061*** [0.1260]	-0.3425** [0.1616]	-0.3425** [0.1054]	-6.1673 [4.8809]	—	-2.820*** [0.3843]	—
Prob > F	0.0000***	0.0000***	0.0000***	0.0113**	0.0000***	0.0000***	0.0000***
RMSE	—	—	—	0.1699	0.1600	0.0546	0.0877
R-squared	0.9877	0.9297	0.9865		0.6800	0.8091	0.9998
Obs	319	340	340	361	340	29	28
No of groups	21	21	21	21	21	—	—
F-test	0.0032***	—	0.0007***	—	—	—	—
MWALD	—	—	0.0000***	—	—	—	—
CD test	—	—	—	—	0.7075	—	—

Notes: Where [.] is the standard error; [a] denotes model estimation based on panel data setting; [b] represents modelling based on time series techniques; ***,** represent statistical significance at 1% and 5% level. $lnDeaths_{t-1}$ is the lagged dependent variable, RMSE is the Root Mean Square Error, R-squared explains the predictive power of the estimated model, Obs represents observations. MWALD is the modified wald statistic and CD test examines the independence of the residuals.

shows that the nexus between deaths and confirmed cases is perfectly linear, with a predictive power (R-squared) of almost 100% whereas the relationship between recovery cases and confirmed cases is nonlinear, with an R-squared of ~97%.

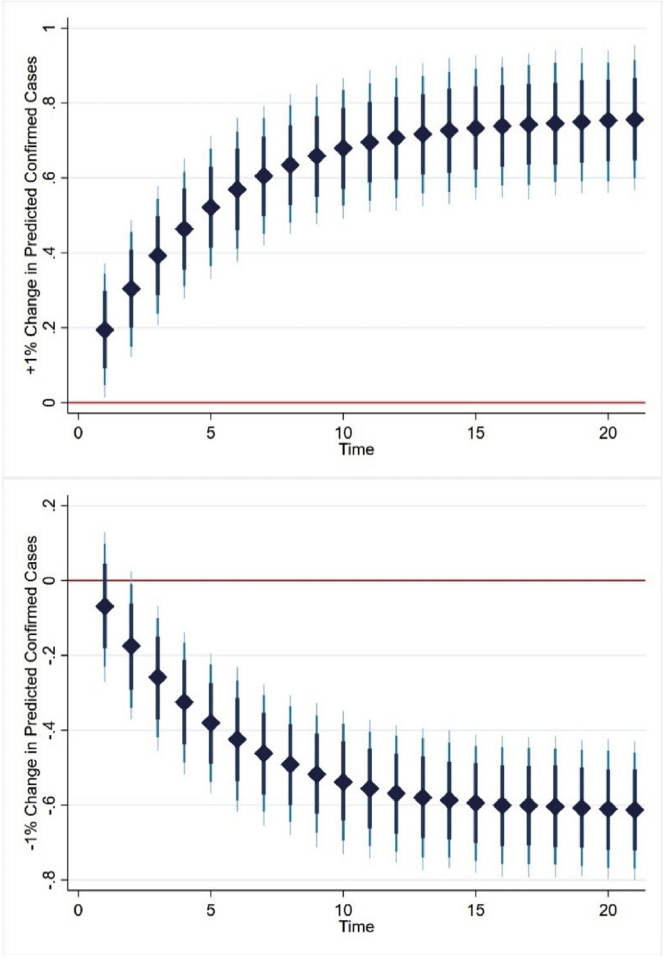

Figure 4. Impulse-Response of confirmed cases of COVID-19 attributable deaths. Note: The light blue spikes represent the 95% confidence interval.

2.2. Model estimation

We developed 7 models comprising of 5 panel data setting and 2 time series. The selection of estimation methods was based on real-time reporting of COVID-19 used as *a priori* expectation. By confirming a perfectly linear relationship between deaths and confirmed cases, our models were constructed on such tangent. Model 1 was developed using the fixed-effects linear model with first-order autoregressive [AR(1)] disturbances to accommodate for the unevenly spaced data across China, rendering the panel setting unbalanced. Model 2 was estimated based on a fixed-effects model with Driscoll-Kraay standard errors to account for possible heteroskedasticity, autocorrelation and cross-sectional dependence amid missing data and unbalanced panel setting [7]. Model 3 was estimated using a fixed-effects model with modified Wald (MWALD) statistic to examine heteroskedasticity in the residuals. Our model of interest with fixed-effects can be expressed as [8]:

$$lnDeaths_{i,t} = lnDeaths_{i,t-1} + \alpha + \beta^{*}lnConfirmedCases_{i,t} + v_i + \varepsilon_{i,t} \quad (1)$$

Where *ln* denotes logarithmic transformation to give the variable a constant variance, *Deaths* denotes the novel coronavirus attributable deaths, *ConfirmedCases* represents confirmed cases, α and β are the constant and coefficient to be estimated, v_i is the Provinces/States-specific fixed-effects and $\varepsilon_{i,t}$ is the independent and identically distributed error term across individual Provinces/States $i = 1, \ldots, N$ in time $t = 1, \ldots, T_i$.

Models 4 and 5 were estimated to account for heterogeneous slopes, after the parameters of Model 3 violated the normality assumption, hence, confirming the presence of heteroskedasticity. The common correlated effects mean group estimation can be specified as [9]:

$$lnDeaths_{i,t} = \beta_i^{*}lnConfirmedCases_{i,t} + u_{i,t} \quad (2)$$

Where $lnConfirmedCases_{i,t} = \alpha 2_i + \lambda_i^{*} f_t + \gamma_i^{*} g_t + \varepsilon_{i,t}$ and $u_{i,t} = \alpha 1_i + \lambda_i^{*} f_t + \varepsilon_{i,t}$. β_i denotes Provinces/States-specific slopes on confirmed cases and $u_{i,t}$ has unobservables and error term $\varepsilon_{i,t}$, $\alpha 1_i$ denotes the standard group fixed-effects that account for time-invariant heterogeneity across Provinces/States. f_t represents the unobserved common factor, λ_i, $\varepsilon_{i,t}$ and $\varepsilon_{i,t}$ are the white noise.

For brevity, the time series models follow a standard equation expressed as:

$$lnDeaths_t = \beta^{*}lnConfirmedCases_t + \varepsilon_t \quad (3)$$

The specification of Eqn. (3) follows the dynamic simulations of Autoregressive Distributed Lag model expounded in Ref. [11,12].

3. Results and discussion

The parameter estimation of the relationship between novel coronavirus attributable deaths and confirmed cases of COVID-19 is presented in Table 1. The estimated models are statistically significant at 5% level (95% CI) and a corresponding predictive power (R-squared) between 68%-100%. The modified wald statistic (MWALD) of Model 3 rejects the null hypothesis of homoskedasticity. Meaning that the effect of confirmed cases on the novel coronavirus attributable deaths is heterogeneous across Provinces/States in China. In both panel and time series models presented, the lagged-dependent variable (LDV) of coronavirus attributable deaths ($lnDeaths_{t-1}$) is positive and statistically significant at 1% level except Model 5 which shows a significant (99% CI) negative coefficient. LDV was introduced in the models to control for omitted variable bias and account for the inertia effects of the reported coronavirus attributable deaths. The positive coefficient of $lnDeaths_{t-1}$ in almost all the models reveals that the historical factors of coronavirus attributable deaths are persistent and likely to affect future reported deaths. On the contrary, when unobserved common factors affecting coronavirus attributable deaths are controlled in Model 5, the coefficient on LDV turns negative. Meaning that the inertia effect of historical deaths is curtailed, hence, reducing the impact of confirmed cases.

The coefficient on the estimated confirmed cases in Table 1 is positive and statistically significant (95% CI) in both estimated panel and time series models. The empirical evidence reveals that an increase in confirmed cases by 1% increases coronavirus attributable deaths by ~0.10%~1.71% (95% CI).

Using the dynamic ARDL simulations estimation technique [11,12], we predicted the counterfactual change in COVID-19 attributable deaths in case of positive or negative shocks in confirmed cases. The plot presented in Figure 4 reveals that a positive shock (1%) in confirmed COVID-19-case-patients will increase attributable deaths from 0.2% to around 0.8% over the horizon. On the contrary, a 1% negative shock in confirmed cases of COVID-19 will decline death rates from 0.1% to 0.6%.

Several novel protocols for clinical and epidemiologic investigations have been outlined to ascertain the clinical features, the pattern of transmission, severity and risk factors of the novel coronavirus disease [10]. Our estimated results confirm the presence of unobserved heterogeneity and common factors that facilitates the novel coronavirus attributable deaths caused by increased levels of confirmed cases. However, the role of the unobserved heterogeneity and common factors that facilitate the transmission of COVID-19 remains unclear. This corroborates the findings of the Situation Report – 33 released by WHO. According to the report [10], the role of environmental risk factors in the COVID-19 transmission process is uncertain. However, confirms the human-to-human transmission through community spread, household, health facilities and environmental surfaces [3, 10]. In such a transmission process, our study reveals a perfectly linear relationship between confirmed cases and novel coronavirus attributable deaths, as such, safety precaution and preventive measures are required to circumvent human-to-human transmission.

4. Conclusions

Our study presented is based on phenomenological models but not a clinical procedure, hence, care should be taken in the interpretation of the outcome. We demonstrated that the effect of confirmed cases on COVID-19 attributable-deaths is perfectly linear whereas the impact of confirmed cases on recovery cases follows a nonlinear path. Our study suffers from the limitation of early case investigation and historical data, hence, our estimation results may change at the latter stage of the novel coronavirus disease (COVID-19). In view of this, we utilized a battery of estimation approach to increase the sensitivity and robustness of the models.

Declarations

Author contribution statement

S.A. Sarkodie: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

P.A. Owusu: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] WHO, Novel Coronavirus (2019-ncov) - Situation Report – 1, 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf?sfvrsn=20a99c10_4. (Accessed 20 February 2020).

[2] WHO, Novel Coronavirus (2019-ncov) - Situation Report – 22, 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=fb6d49b1_2. (Accessed 20 February 2020).

[3] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K.S. Leung, E.H. Lau, J.Y. Wong, Early transmission dynamics in wuhan, China, of novel coronavirus–infected pneumonia, N. Engl. J. Med. (2020).

[4] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, Clinical features of patients infected with 2019 novel coronavirus in wuhan, China, Lancet (2020).

[5] N.M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A.R. Akhmetzhanov, S.-m. Jung, B. Yuan, R. Kinoshita, H. Nishiura, Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data, J. Clin. Med. 9 (2020) 538.

[6] S. Zhao, S.S. Musa, Q. Lin, J. Ran, G. Yang, W. Wang, Y. Lou, L. Yang, D. Gao, D. He, Estimating the unreported number of novel coronavirus (2019-ncov) cases in China in the first half of january 2020: a data-driven modelling analysis of the early outbreak, J. Clin. Med. 9 (2020) 388.

[7] J.C. Driscoll, A.C. Kraay, Consistent covariance matrix estimation with spatially dependent panel data, Rev. Econ. Stat. 80 (1998) 549–560.

[8] B.H. Baltagi, P.X. Wu, Unequally spaced panel data regressions with ar (1) disturbances, Econom. Theor. 15 (1999) 814–823.

[9] M.H. Pesaran, Estimation and inference in large heterogeneous panels with a multifactor error structure, Econometrica 74 (2006) 967–1012.

[10] WHO, Coronavirus Disease 2019 (Covid-19) - Situation Report – 33, 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200222-sitrep-33-covid-19.pdf?sfvrsn=c9585c8f_2. (Accessed 24 February 2020).

[11] Jordan Soren, Andrew Q. Philips, Cointegration testing and dynamic simulations of autoregressive distributed lag models, Stata J. 18 (4) (2018) 902–923.

[12] Samuel Asumadu Sarkodie, et al., Environmental sustainability assessment using dynamic autoregressive-distributed lag simulations—nexus between greenhouse gas emissions, biomass energy, food and economic growth, Sci. Total Environ. 668 (2019) 318–332.