

利用面板数据分析检验新冠疫情数据的影响

王伟

东北师范大学, 长春 130024, 中国

摘要: 本研究以病例、检测、年龄、住院率和死亡率为输入, 考察了 G20 国家的 COVID-19 疫情传播函数之间的关系。研究还显示了各国根据建议的拥挤指数采取的 COVID-19 预防措施的程度。分析了 2020 年 3 月 12 日至 2020 年 5 月 29 日 G20 国家的数据, 并从 <https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.xlsx> 计算了描述性统计数据。面板数据分析用于研究基于所讨论变量的对一次性事件输出值的影响。在检验紧密度指数对死亡人数的影响时, 计算出的相关值为 0.7639。据观察, 硬度指数每变化一个单位, 产量就会增加 7.8017。与这些研究不同, 本研究考察了社会因素对病例数量的影响, 并使用 R Studio 应用了面板数据分析固定效应模型。同时, 还考察了各国采取的措施与病例数量/死亡率之间的关系。

关键词: COVID-19; 大流行; 数学模型; RStudio; 面板数据分析

Examining the effects of COVID-19 Data with Panel Data Analysis

Wei Wang

Northeast Normal University, Changchun 130024, China

Abstract: In this study, the relationship between the COVID-19 outbreak spreading function, where cases, tests, age, hospitalization rate and mortality were defined as inputs, was examined for G20 countries. It also shows the extent to which countries have taken precautions against COVID-19 with the recommended congestion index. The data of G20 countries between 12.03.2020 and 29.05.2020 were analyzed and descriptive statistics were calculated from <https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.xlsx>. Panel data analysis is used to investigate the effect on the output value based on the variables in question for an event occurring at once. When examining the effect of the tightness index on the number of deaths, the correlation value was calculated as 0.7639. It has been observed that a one unit change in the hardness index increases production by 7.8017. In our study, unlike these studies, the social factors on the number of cases was examined and Panel Data Analysis Fixed Effects Model was applied using R Studio. At the same time, the relationship between the measures taken by countries and the number of cases / death rates was also examined.

Keywords: COVID-19; Pandemic; Mathematical model; RStudio; Panel data analysis

2019 年最后几个月爆发并波及全球的 COVID-19 病毒, 被命名为“世界性传染病大爆发”。本文分析了病例

数、检测数量、人口比例、平均年龄、医院床位数、GDP 等环境因素对 COVID-19 病毒造成的病例/死亡人数的影响。首先，分析了 G20 国家的数据，并计算了描述性统计数据。在研究的第二阶段，土耳其、德国和英国两个不同的国家实施了由变量数据定义的策略，这些变量数据是为病例准备的，并调查了这些变量对患者的影响程度。

在本研究中，使用 R Studio 中的面板数据分析，分析了 10 个输入变量对病毒导致的个体死亡人数的影响。分析结果表明，由于 p 值（probably = 概率）小于 0.05，因此拒绝 H0 假设（零假设），接受 H1 假设，即变量与产出之间存在显著相关性。例如，在检验严格性指数对死亡人数的影响时，计算出的相关值为 0.7639。据观察，严格性指数每变化一个单位，产出就会增加 7.8017。

2019 年最后几个月出现并影响全球的 COVID-19 病毒迅速演变成全球流行病。本研究选取的 G20 国家占世界经济总量的 85%，投资流量的 80%，人口总数的三分之二。G20 包括美国、德国、阿根廷、澳大利亚、巴西、中国、印度尼西亚、法国、南非、韩国、印度、意大利、日本、加拿大、墨西哥、沙特阿拉伯，以及俄罗斯和土耳其。二十国集团峰会于 2008 年首次在华盛顿举行，已持续 10 年，已成为全球讨论经济政策的天然平台[1]。

例如，一些国家实行了隔离政策，例如土耳其和德国。对这三个国家的数据在 2020 年 3 月 12 日至 2020 年 5 月 29 日期间进行了分析。查阅文献发现，关于 COVID-19 的研究数量显著增加，尤其是在 2020 年的第一年。这些研究大多是从医学/生物学角度对冠状病毒进行研究。本文研究了在采取冠状病毒措施的框架内发生的疫情死亡人数。在这方面，这项研究具有与其他文章不同的特点。

我们将文献中关于 COVID 19 的研究分为不同的组：文献综述、数学模型和统计研究。Gulati 等人回顾了文献，Zhao 等人研究了该主题的案例研究[2]。此外，Mi 等人也回顾了文献，Zhao 等人研究了该主题的案例研究[2]。COVID-19 研究了肥胖率[2]。在该主题的数学建模研究中，Torrelba-Rodriguez 和 Marimuthu 等人、Adoke 等人、Briz-Redon 等人、Marimuthu 等人、Shie 等人和 Bonanad 等人进行了荟萃分析，并考察了该疾病与年龄因素的关系。此外，还使用了一些与疾病相关的预测技术、统计分类和机器学习技术。例如，Pathak 等人强调了诊断中要使用的分类技术[3]。Ceylan 研究了该疾病并做出了一些预测 [4]。一些统计研究已经检验了 COVID-19 的传播状况，并对未来的一些情况进行了预测。例如，世界各地的研究人员对伊朗、尼泊尔、非洲、沙特阿拉伯 和印度的 COVID-19 传播进行了统计分析。迄今为止的研究仅对基于国家的疾病传播数学模型进行了研究和分析。我们的研究对各国进行了分析和比较。

而我们在本研究中采用的面板数据分析研究通常只涉及时间因素（Aydin-a、Aydin-b、Williams 和 Zhao）。同时，尽管文献中关于面板数据分析的研究并不多，但时间因素已在一些领域得到应用，例如生物质能消耗、国际旅游对经济发展的影响[3]。由于本研究分析的是一定时间间隔内数据的投入产出关系，因此采用了 R Studio 的面板数据分析方法。由于本研究希望按一定时间间隔对数据进行分析，因此选择了面板数据分析方法。

一、方法

本节介绍了面板数据分析的步骤。面板数据分为平衡面板（每个单位始终可观测）和非平衡面板（某些单位在某些时期存在亏损）两种类型。将属于国家、个人、企业等单位或水平横截面观测值的混合数据在一定时期内进行组合，将同一横截面单位随时间变化的混合数据称为面板数据。

固定效应模型：本质上是控制或部分扣除那些不随时间变化的变量及其对其他变量的影响和响应。

该检验通过考虑横截面依赖性实现了变量之间的协整关系。

二、应用

本研究考察了 COVID-19 疫情传播函数之间的关系，该函数的输入参数为病例数、检测数量、年龄、住院床位比例和死亡率。疾病传播频率与人口密度之间存在关联。

进行的检测、隔离过程、治疗过程与死亡率之间存在关联。如果在面板数据模型中未将横截面数据作为划分单元处理，则变量之间的关系可能会产生误导。本研究使用了相同的 G20 国家 COVID-19 数据集。

数据分析过程包括数据分析、清理、转换和建模。本研究对 G20 国家进行了面板数据分析，数据收集时间为 2020 年 3 月 12 日至 2020 年 5 月 29 日。稳固性指数是指各国在政策领域采取的措施的 1-100 分之间的分数，该分数来自公开数据来源。GDP 以美元计算。人口密度以每平方米人口数量计算。

结果显示，美国的病例数（10048,60）和死亡人数（929,60）的偏差最大，而英国的病例数在指定日期开始时势头强劲，第一个月就增加了 8719 例，增幅达 99% [31,32,33]。表 2 计算了标准差、平均值和变异系数等描述性统计数据。

本研究采用了面板数据分析。最终，面板数据分析可以监测数据行为。通过基于 COVID-19 动态传播大流行模型和时间序列分析建立统计模型，旨在建立后期时间段的预测模型。由于疫情爆发时可用的数据样本相对较大，在 2019-nCoV 传播的现阶段，所建立的模型能够显示疫情的发展趋势、峰值大小等。基于时间序列分析的统计建模能够更准确地预测疫情的短期走势。

三、结果与讨论

CDBP、CDLM 和 CD 等横截面相关性和斜率齐性检验结果，我们拒绝了横截面独立性零假设，即变量之间不存在相关性。该假设的结果表明，所有被考察的变量都存在横截面相关性。使用 R Studio、foreign、readxl、car、apsrtable、plm 和 gplots 等程序包进行数据面板分析。

英国的死亡人数在 4 月至 5 月期间居高不下。虽然 5 月之后有所下降，但最新的死亡人数仍在 400 人左右。德国的死亡人数继续在这个分布中波动较小，而土耳其的死亡人数则保持在 0-200 人左右，较为稳定。常规 OLS（普通最小二乘法）回归未考虑组间或时间间的异质性。图 3 和图 4 显示了不同国家和不同日期间的异质性。

分析结果表明，该模型显著，因为 p 值 = $2e-16$ 小于 0.05。相关系数 = 0.7639 接近于 1，表明紧密度指数与测量值之间存在很强的相关性。紧密度指数每变化一个单位，死亡人数就会增加 7.8017 个单位。德国的僵硬指数每变化一个单位，土耳其为 399.4421 个单位，英国为 514.5541 个单位，英国为 31.7702 个单位，这些国家的死亡人数都会减少。

在研究疫情期间，研究人员发现，各国经济措施的有效性，尤其是紧缩指数的差异，可能对疫情的有效性产生显著影响。

分析中使用的变量会影响病例数和死亡人数。例如，紧缩指数与疫情死亡人数之间存在显著关系。然而，紧缩指数值是采取必要措施后得出的。

可以对病毒感染者、该群体中曾患病的人或未受影响人群的物理环境或生物学特征进行回顾性研究。

参考文献

- [1] Adekola HA, Adekunle IA, Egberongbe HO, et al. Mathematical modeling for infectious viral disease: The COVID-19 perspective. *Journal of Public Affairs*. DOI: 10.1002/pa.2306.
- [2] Briz-Redón A, Serrano-Aroca A. The effect of climate on the spread of the COVID-19 pandemic: A review of findings, and statistical and modelling techniques. *Progress in Physical Geography: Earth and Environment*, 2020: 1-14. DOI: 10.1177/0309133320946302.
- [3] Orfali R, Shafiq ul Azam M, Aati HY, et al. COVID-19: A pandemic disease and the Saudi precautions. *Saudi Pharmaceutical Journal*, 2020, 28(7): 888-897.

- [4] Williams R. Panel Data 4: Fixed Effects vs Random Effects Models. *Sociology*, 2021, 73994: 1–6.