The Relationship Between the COVID-19 Cases and the Demand of Ventilators*

Can Case Numbers Predict the Medical Demand?

Siling Guo

October 29, 2025

This study explores the effect of new COVID-19 cases on ventilator demand during the COVID-19 pandemic. A simple linear regression model was established using data from the County of Santa Clara Public Health Department. In the model, we use a weighted index that combines daily and 7-day average case data as the independent variable, and daily ventilator demand as the dependent variable. The results show a positive correlation, with an expected increase of 1.2 units in ventilator demand for each point increase in the weighted index (p < 0.001, R^2 = 0.78). However, applying a simple regression model to time-series data has limitations, and this model may not have strong predictive power.

1 Introduction

One major symptom of a severe COVID-19 infection is a low blood oxygen saturation level, which is often treated by mechanical ventilation. During the 2020-2023 COVID-19 pandemic, the spread of the virus before the availability of effective treatments and vaccines caused an influx of demand for ventilators (Filip et al. 2022). Although the pandemic has ended, analyzing the relationship between the number of COVID-19 cases and ventilator demand may illustrate how medical equipment needs are affected by a surge in new patients during the beginning stages of a pandemic.

Our hypothesis is that an increase in the number of COVID-19 patients is followed by an increase in ventilator usage. The relationship is tested by using a linear regression model based on time-series data published by the County of Santa Clara Public Health Department (SCCPHD).

*Project repository available at: https://github.com/silingguo/MATH261A-project1.

We constructed a new independent variable, which we will call the Severity Score, which combines the daily new cases and the 7-day average new cases. We built this variable to smooth the data and show the general trend more clearly. The construction method will be shown in detail in the Data section. The number of patients on ventilators is the dependent variable. The results show a positive correlation, but the model's prediction is limited because the data are time-series, which may be affected by dependent error and unequal variance. We will discuss this issue later in the Discussion section.

In this paper, we will first discuss the research background and proposal, then describe the data usage and new variable construction and processing. Next, we explain the setup and validation of the model, as well as its limitations. After presenting the results with relevant tables and graphs, we present the findings and discuss the limitations

2 Data

In this study, we used the daily counts of new COVID-19 cases and the number of patients on ventilators from March 27, 2020 to May 23, 2021 ("COVID-19 Hospitalizations by Date | County of Santa Clara" 2025), provided by the SCCPHD and posted on the County of Santa Clara Open Data Portal. This time period represents the beginning of the COVID-19 pandemic. The World Health Organization (WHO) declared a pandemic in March 2020 due to the rising number of COVID-19 infections. By May 2021, several vaccines became available, which controlled the spread of the pandemic. The total sample size is 423 days with no missing values in the core variables. A small number of outliers were retained, as they likely reflect the peak pandemic conditions.

Overall, the data are clean, so the only processing performed was converting date strings into a datetime format as well as scaling and normalizing the key data. The original dataset consists of time-series data, the error terms are neither independent nor identically distributed, and may lead to misleading results due to fluctuations and time lag. This data does not meet the basic assumptions of regression analysis (Li 2025).

To reduce the 'noise' of time-series data and improve the predictive performance, we applied a Weighted Moving Average (WMA) to construct a scaled weighted index, Severity Score, as the independent variable.

$$Severity = (0.2 \times \text{daily new case count}) + (0.8 \times \text{7-day average percentage})$$

According to(Huang et al. 2020), the median time from illness onset to shortness of breath is 7 days. We assumed that the 7-day average case percentage reflects the overall trend better, so we gave it a higher weight (80%), while the daily new cases show short-term changes and were given a smaller weight (20%). Because the two variables use different units(counts and percentages), the Severity does not directly correspond to the magnitude of ventilator demand. For ease of understanding and presentation, the values were then scaled to a 0–100 range for statistical convenience. The score does not represent the actual number of cases; it only reflects the severity as defined by the index.

The dependent variable is the number of patients on ventilators, which is a continuous variable measured in terms of the number of people. According to the data summary (Table 1), the median composite severity score is 24.1 with a standard deviation of 25 and a maximum of 100, showing considerable variation. The daily average number of patients using ventilators is 179.7, with a standard deviation of 34.5. This data shows smaller fluctuations. Overall, the trend in the number of patients using ventilators is generally consistent with the trend in the number of new COVID-19 cases (Figure 1). Despite accounting for the time lag, the inherent limitations of time series analysis cannot be fully avoided. Limitations will be discussed in the discussion section.

Table 1: Data Summary of Composite Covid-19 Cases and Number of Patients on Ventilators

Data Summary of composite Covid-19 cases and Number of patients on ventilators		
Statistic	**COVID-19 Composite Severity Score**	**Patients on Ventilators**
Mean	24.18056	179.74194
Standard Deviation	25.26331	34.50896
Maximum	100.00000	283.00000
Minimum	0.00000	123.00000

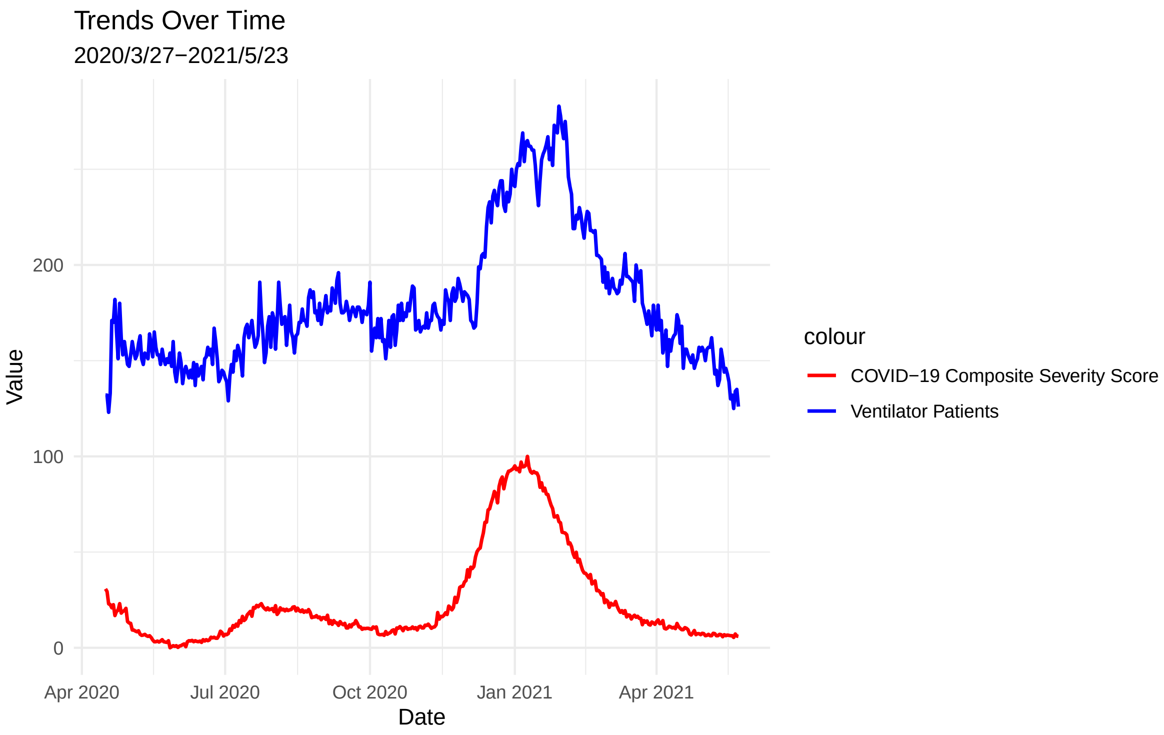


Figure 1: Trends Over Time

3 Methods

The score of COVID-19 severity and the number of patients on ventilators are continuous variables. The trend chart of the data suggests a potential linear correlation between two variables. We apply a simple linear regression model.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The dependent variable, Y_i, represents the number of patients on ventilators on a given day. The independent variable, X_i, represents the COVID-19 composite severity score on a given day. The intercept, β_0, represents the average number of patients on ventilators when the composite severity score is 0. The slope, β_1, represents the expected change in the number of patients using ventilators for each additional unit of COVID-19 composite severity. The error, ε_i, represents the error term. In this model, the weighted method and scale method was used to clean up the data, the ordinary least squares (OLS) method was used to estimate the parameters. To ensure valid inference, we need to meet these four assumptions:

Linearity: The scatter plot (Figure 2) shows that two variables are linearly correlated; the number of ventilators increases as the Severity Score increases in general.

Independence of error: The Residual vs Time plot (Figure 3) shows that the residuals are clustering and exhibit a cyclical pattern. This pattern indicates the errors are dependent, which may lead to underestimated standard errors

Equal Variance of error: The Residual vs Predictor plot (Figure 4) shows that as the predictor increases, the residuals form a cone shape and are clustered. The variance is not constant, so the standard t-test might not give accurate results.

Normality of error: If the residuals approximately along the reference line in QQ plot, the residuals are approximately normally distributed.The QQ plot (Figure 5) indicates that the majority of the residuals closely align with the reference line. A few data points fall below the expected quantiles on the left, while a few data points fall above the expected quantiles. This suggests that there might be outliers. In the Residual histogram (Figure 6), the density line intuitively shows that the residual distribution is approximately normal, but a higher value on the right tail makes the right side exceed the curve. Together with the QQ plot, the pattern indicates that the normality assumption is not fully satisfied, so the validity of confidence intervals and t-tests may be affected.

The value of measurements of key variables for guiding, predicting, and informing practical applications is limited because time-series data violate several model assumptions.

We applied the R's lm() function to fit the linear model(R Core Team 2025), and we used several R packages to generate a table, manipulate and visualize the data, including dplyr (Wickham et al. 2023), lubridate (Grolemund and Wickham 2011), ggplot2(Wickham 2016), broom(Robinson et al. 2023), flextable (Gohel and Skintzos 2024), and scales(Wickham and Seidel 2022).

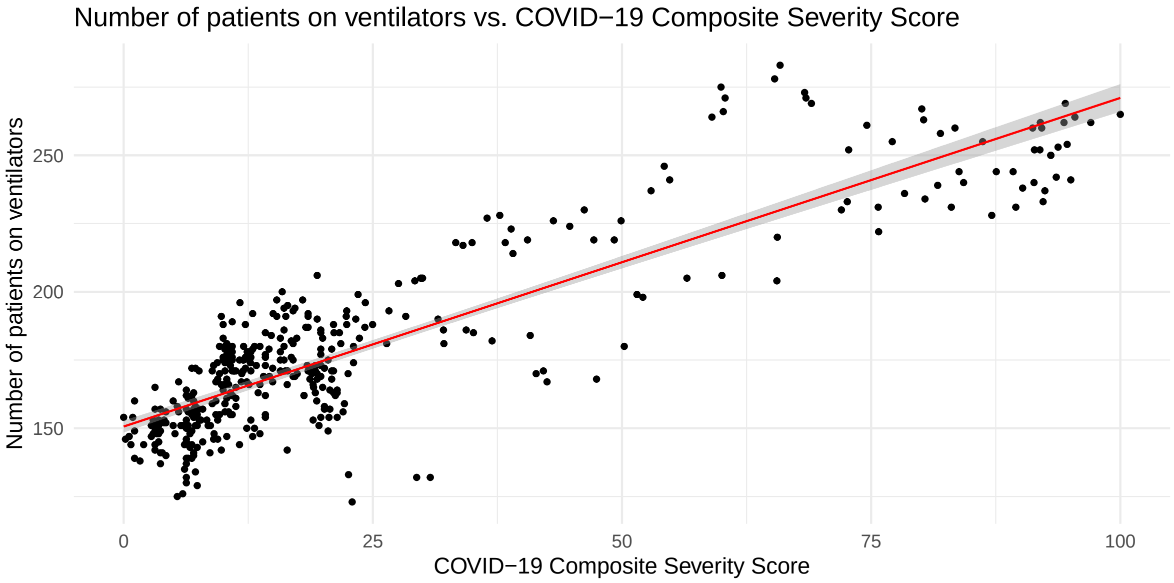


Figure 2: Number of patients on ventilators vs. COVID-19 Composite Severity Score

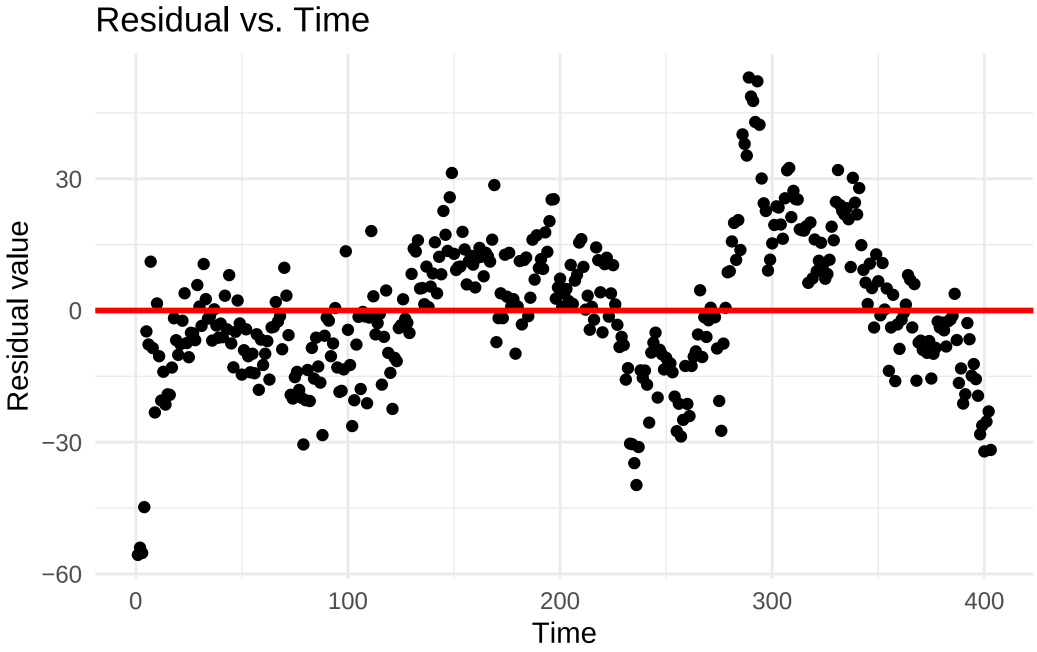


Figure 3: Residual vs. Time

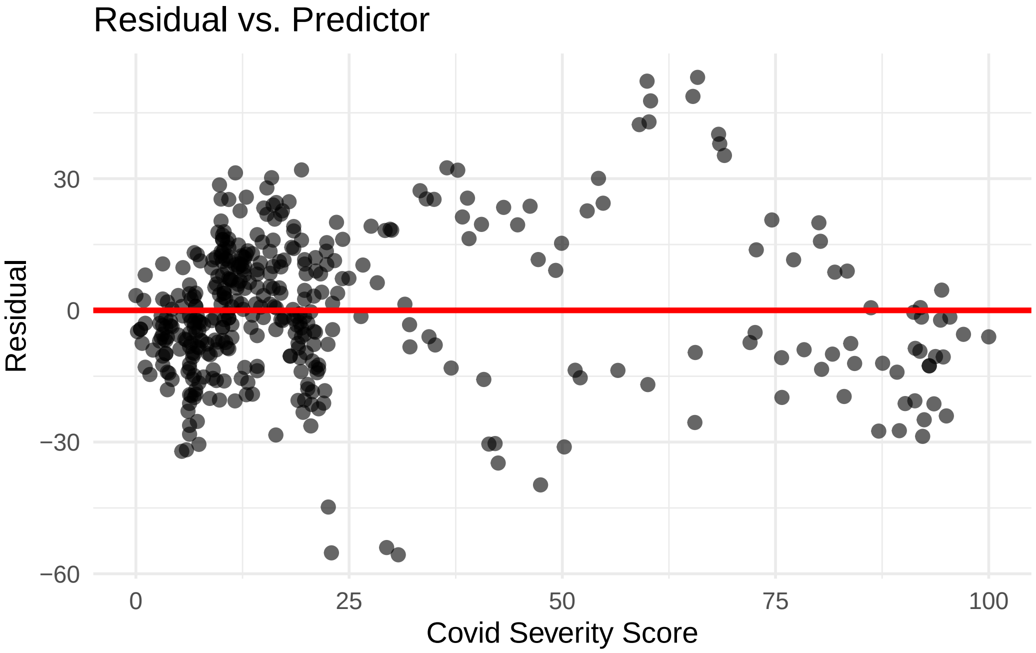


Figure 4: Residual vs. Predictor

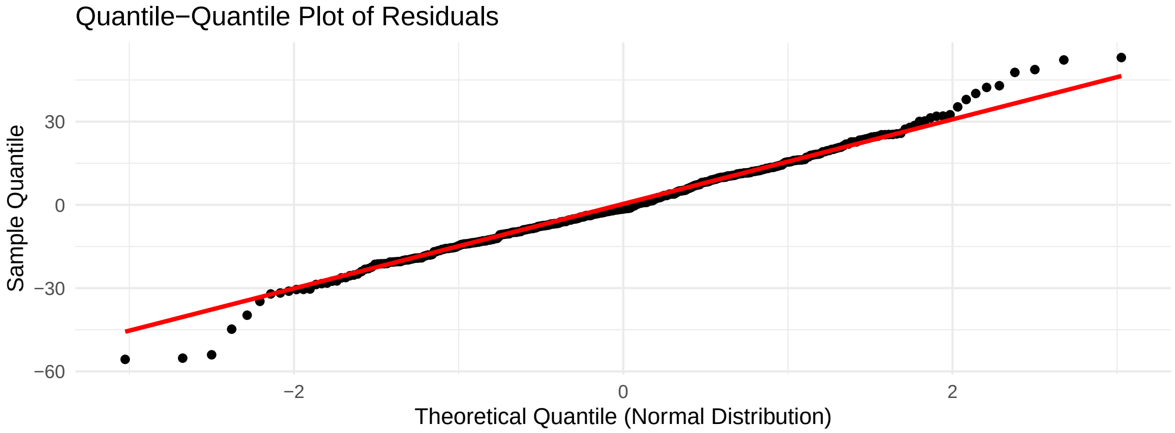


Figure 5: Quantile-quantile plot of residuals

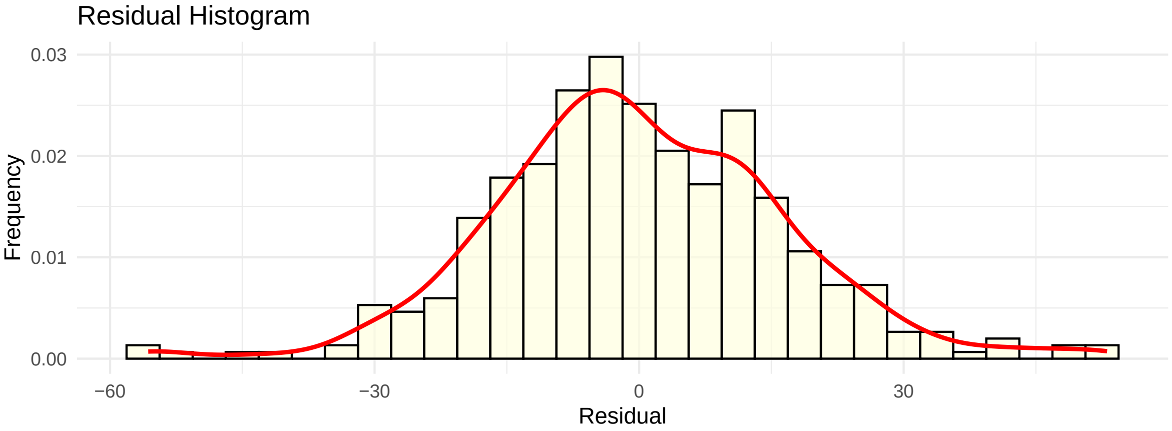


Figure 6: Residual Histogram

4 Results

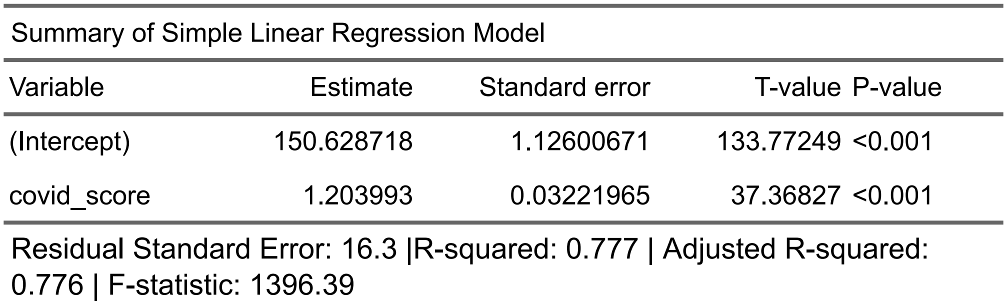
We conducted a t-test to determine whether the Severity Score influences ventilator demand. The setup is as follows: The null hypothesis, H_0, represents $\beta_1 = 0$, meaning there is no relationship between the Severity Score and the ventilator demand.

The alternative hypothesis, H_1, represents $\beta_1 \neq 0$, meaning the relationship between Severity Score and the ventilator demand is significant.

The results (Table 2) show that the coefficient for Severity Score is positive (t = 37.37, p < 0.001). Therefore, we reject H_0 and conclude that there is a significant relationship between the Severity Score and ventilator demand.

As we discussed above, the time-series data violate the assumptions of equal variance, normality, and independence of the error terms are not fully satisfied. These violations will reduce the t-test's accuracy. The estimate of covid_score is 1.20, indicating that the linear regression model is highly significant.For every composite severity score point increase during this time period, there is an expected increase of 1.2 ventilator patients. The intercept is 150.62, it is positive because there is ventilator usage for reasons other than COVID-19. Multiple R-squared: 0.777 and Adjusted R-squared: 0.776 indicate a model fit of approximately 77%.

Table 2: Summary of Simple Linear Regression Model



Summary of Simple Linear Regression Model

Variable	Estimate	Standard error	T-value	P-value
(Intercept)	150.628718	1.12600671	133.77249	<0.001
covid_score	1.203993	0.03221965	37.36827	<0.001

Residual Standard Error: 16.3 |R-squared: 0.777 | Adjusted R-squared: 0.776 | F-statistic: 1396.39

5 Discussion

Overall, according to the results, the model seems to fit the data well; however, its true explanatory and predictive capabilities is limited. The time-series data we used here violate the assumptions of the simple linear regression model, so the result should be interpreted with caution.

We applied a smoothing method to construct a Severity Score variable to reduce fluctuations in the data, but it does not actually address the temporal dependence in the dataset. The smoothing method is a way to enhance the model's stability, reduce lag effect; however, it does not fundamentally change characteristics of the time-series data, such as dependence of error and temporal dependence.

The error term is the basis of statistical inference because it determines how variance and uncertainty are estimated in the model. When its assumptions are not met, the standard errors, t-tests, and confidence intervals become unreliable, weakening both the model's inference and its predictions.

If error terms are highly related, the model may misinterpret the relationship between the two variables, giving an underestimated p-value and inflated t-value. In our study, this issue makes it difficult to interpret the real causal relation between new COVID-19 cases and ventilator demand. In reality, ventilator demand is likely influenced by other factors beyond infection severity, such as population age structure, vaccination rates, and seasonal factors.

In practice, the COVID-19 composite Severity Score is inherently difficult to define, and a single score cannot accurately reflect the number of new cases, giving it limited practical value. Moreover, despite consulting multiple sources, we have not found a clear definition of the "appropriate ratio" between the number of new cases and the 7-day average in the weighting method.

This study intends to explore the effect of COVID-19 new cases on ventilator demand. However, we cannot use this model to examine the relationship due to limitations in the dataset and the regression model. Nevertheless, while the simple linear regression model is limited and cannot fit complex datasets, it is a good starting point to identify the data pattern and gain insight.

References

"COVID-19 Hospitalizations by Date | County of Santa Clara." 2025. https://data.sccgov.org/COVID-19/COVID-19-hospitalizations-by-date/5xkz-6esm/about_data.

Filip, Roxana, Roxana Gheorghita Puscaselu, Liliana Anchidin-Norocel, Mihai Dimian, and Wesley K. Savage. 2022. "Global Challenges to Public Health Care Systems During the COVID-19 Pandemic: A Review of Pandemic Measures and Problems." *Journal of Personalized Medicine* 12 (8): 1295. https://doi.org/10.3390/jpm12081295.

Gohel, David, and Panos Skintzos. 2024. *Flextable: Functions for Tabular Reporting*. https://CRAN.R-project.org/package=flextable.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Huang, Chaolin, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Liangfang Zhang, et al. 2020. "Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China." *The Lancet* 395 (10223): 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.

Li, Dongfeng. 2025. "Cointegration and Vector Error Correction Model." https://www.math.pku.edu.cn/teachers/lidf/course/fts/ftsnotes/html/_ftsnotes/fts-coint.html.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, Simon Couch, and Max Kuhn. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Dana Seidel. 2022. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.