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

Simple Linear Model Project

Siling Guo

September 23, 2025

This study explores the relationship between the change in the number of new COVID-19 cases and daily ventilator demand during the COVID-19 pandemic. A simple linear regression model was established based on the data provided by the County of Santa Clara Public Health Department between 2020-03-27 and 2021-05-23. In the model, we use a weighted index of daily new COVID-19 cases (20%) and the 7-day average percent of COVID-19 cases (80%) as the independent variable, and ventilator demand as the dependent variable. The results shows a positive correlation, and there was an expected increase of 1.2 units of ventilator demand for each unit increase in the weighted index (β_1=1.2, p<0.001, R Squared=0.78). This model may not have strong predictive power because it is based on time series data. However, it helps illustrate the potential demand for medical equipment during pandemics.

1 Introduction

The 2020-2023 COVID-19 pandemic illustrated the challenges in balancing the limited availability of medical resources with the influx of demand caused by the surge in new COVID-19 cases, especially in resource-limited countries. Although the COVID-19 pandemic has ended, analyzing the relationship between the rate of new COVID-19 cases and the change in ventilator demand helps us understand how medical equipment needs are affected by the sudden surge of new patients expected in future pandemics. We will explore this relationship using a linear regression model based on data published by the County of Santa Clara Public Health Department from 2020-03-27 to 2021-05-23("COVID-19 Hospitalizations by Date | County of Santa Clara" n.d.).Because our study uses the time-series data, we constructed a weighted index to reflect time-lag effects. The index combines the daily number of new COVID-19

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

cases (20%) with the 7-day average COVID-19 case percentage (80%), scaled from 0 to 100, and is referred to as the 'severity' measure. This measure is the independent variable in our simple linear regression model, with the number of patients on ventilators as the dependent variable. First, we discuss the research background and proposal, then describe the data usage 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 draw our conclusions. Through analysis, we found a positive correlation between the COVID-19 severity index and the demand for ventilators.

2 Data

The observation units in this study are the daily count of new COVID-19 cases and the daily number of patients on ventilators from 2020-03-27 to 2021-05-23, as provided by the County of Santa Clara Public Health Department. The sample size is 423 days. The dataset is a time series, in which the error terms are neither independent nor identically distributed, which does not meet the basic assumptions of regression analysis (Li 2025). To fix this problem, we consider the time-lag effect. We defined a variable we will call the severity index to use as the independent variable for our analysis. This index combines the daily number of new COVID-19 cases (20% weight) with the 7-day average percentage of COVID-19 cases (80% weight).

$$Severity = (0.2 * dailynewcasecount) + (0.8 * 7 - dayaveragepercentage)$$

The values were then scaled to a 0–100 score range for statistical convenience. The score do not represent the actual number of cases; it only reflects the severity level 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. The data is generated by the County of Santa Clara Public Health Department and posted on the County of Santa Clara Open Data Portal. There are no missing values in the core variables. Additionally, 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. According to the data summary (Table 1), the average composite severity score was 24.1 with a standard deviation of 25 and a maximum of 100. This data has significant fluctuations. The daily average number of patients using ventilators is 179.7, with a standard deviation of 34.5. This data has 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 completely avoided. We can observe this time delay in the data.

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
sd	25.26331	34.50896
max	100.00000	283.00000
min	0.00000	123.00000

\

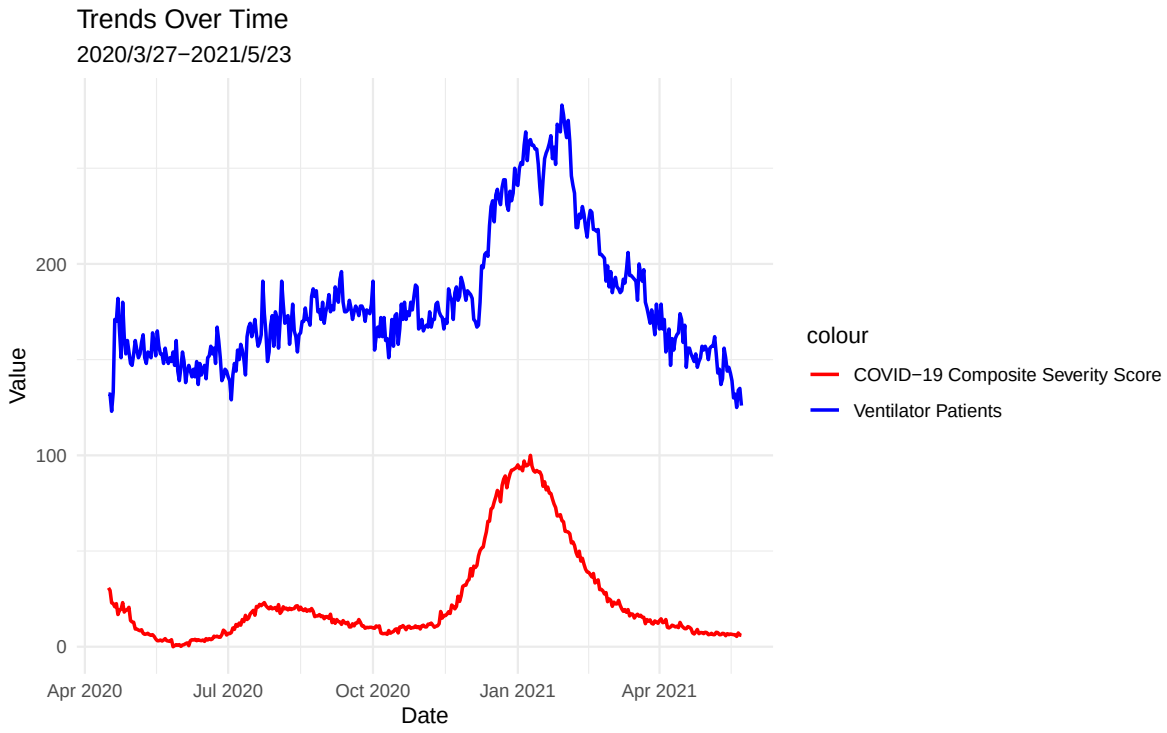

Figure 1: Trends Over Time

3 Methods

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Here,

y_i = The number of patients on ventilators on a given day
X_i = The score of COVID-19 composite severity on a given day β_0 = The average number of patients on ventilators when composite severity score = 0(intercept) β_1 = The expected change in the number of patients using ventilators for each additional unit of COVID-19 composite severity (slope)
ε_i = Error term

In this model, we use these methods to process the data: Using the weighted method and scale method to clean up the data, using the ordinary least squares (OLS) method to estimate the parameters, and applying the lm() function in R to fit the model. The OLS method assumes that the error term -ε_i follows a normal distribution. The QQ Plot (Figure 2)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. The Residual histogram(Figure 3) shows that the residual distribution is approximately normal, but a higher value on the right tail makes the right side exceed the normal curve. Together with the QQ plot, this indicates the normality assumption is not fully satisfied. The Residual and Projector plot (Figure 4) shows that as the predictor increases, the residuals form a cone shape and are clustered. This is because we use time-series data as the key variables, where the error terms are not independent and the variance is not constant. In addition to the limitations of error, the definition of the severity scale and the choice of weighting ratio also present limitations. Because we are using time series data, we must take the time-lag effect into consideration. In the raw data, the units of measurement for the number of new cases (number) and the 7-day average case percentage (percentage) don't match. Even after normalizing the two and calculating a composite value using a 2:8 weighting method, the result still does not correspond to the magnitude of ventilator demand. For ease of understanding and presentation, we used a 0–100 scale to measure severity. However, in reality, COVID-19 composite severity scores are inherently difficult to define, and a single score cannot accurately reflect the number of new cases. Furthermore, I have consulted numerous sources but 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, drawing on relevant COVID-19 disease progression data(Huang et al. 2020), shows that the median time from illness onset to dyspnoea is 7 days. It shows that the 7-day average case percentage should account for a larger proportion. Therefore, we established a 2:8 weighting ratio. These issues

limit the value of the measurement of key variables in guiding, predicting, and referencing practical applications.

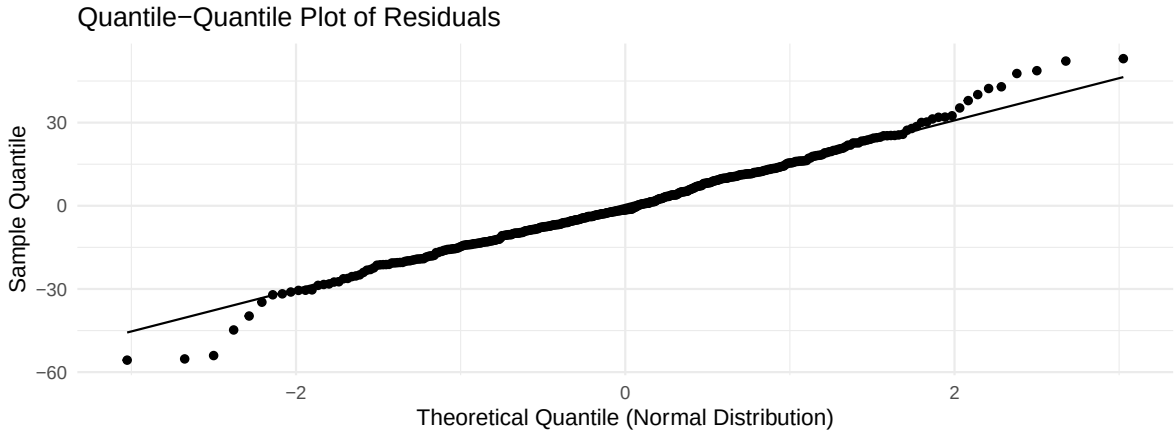

Figure 2: Quantile-quantile plot of residuals

\

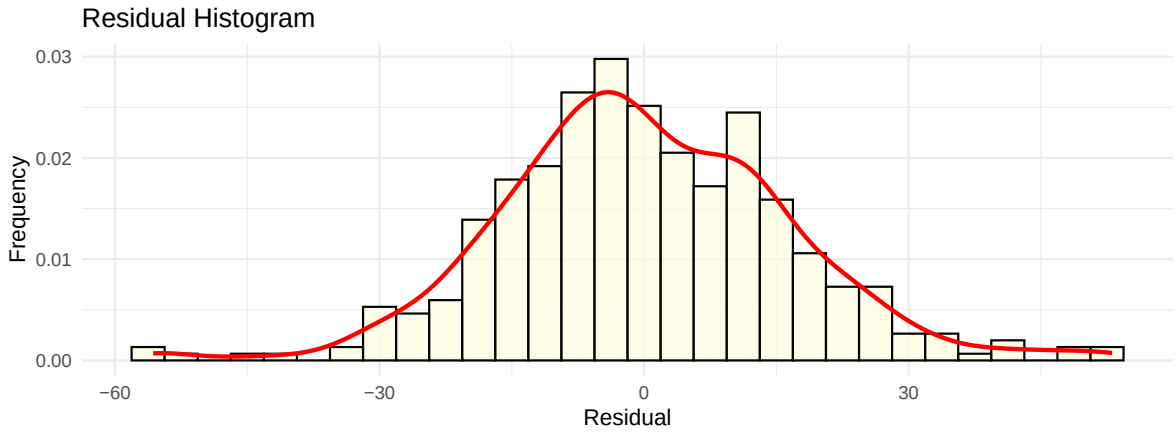

Figure 3: Residual Histogram

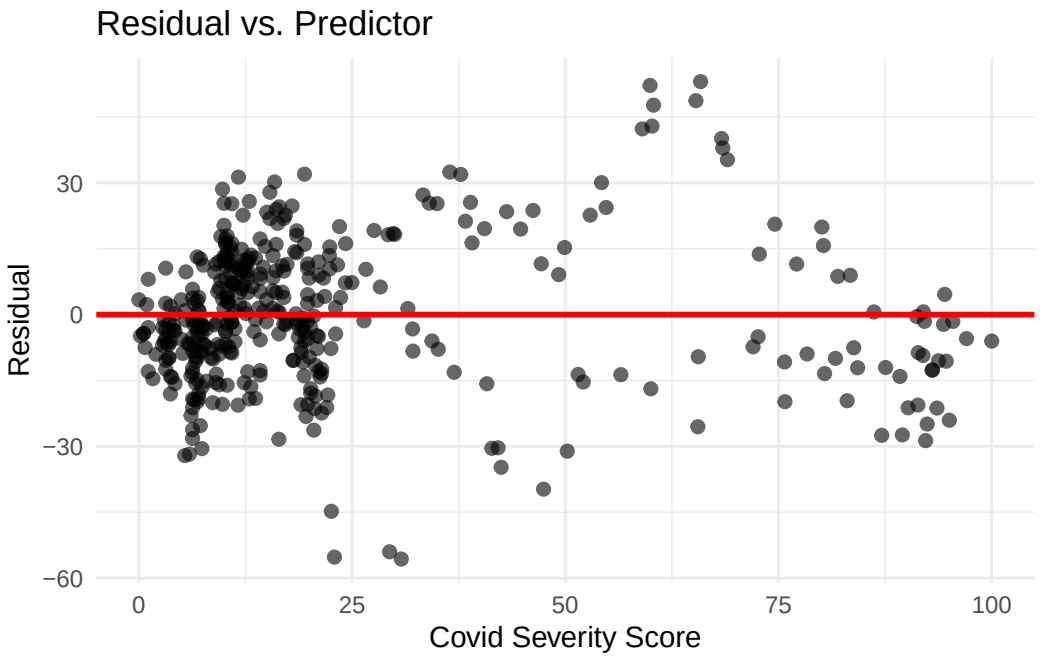

Figure 4: Residual vs. Predictor

4 Results

Figure 5 indicates that the number of patients on ventilators becomes increasingly dispersed as the score of COVID-19 composite severity rises, and the model demonstrates a moderate fit.From the data summary((Table 2)), we find F = 1396.39, p-value $< .001$, and estimate of covid_score = 1.20, indicating that the linear regression model is highly significant, and the predictor variable (x) can explain the variation in the dependent variable (y). Multiple R-squared: 0.777 and Adjusted R-squared: 0.776 indicate a model fit of approximately 70%.Besides, the COVID-19 composite severity is a key factor influencing the demand in ventilator usage. Specifically, for every composite severity score during this time period, there is an expected increase of 1.2 ventilator patients. Although the model seems to fit the data well, its true explanatory and predictive capabilities are limited when combined with the aforementioned limitations.

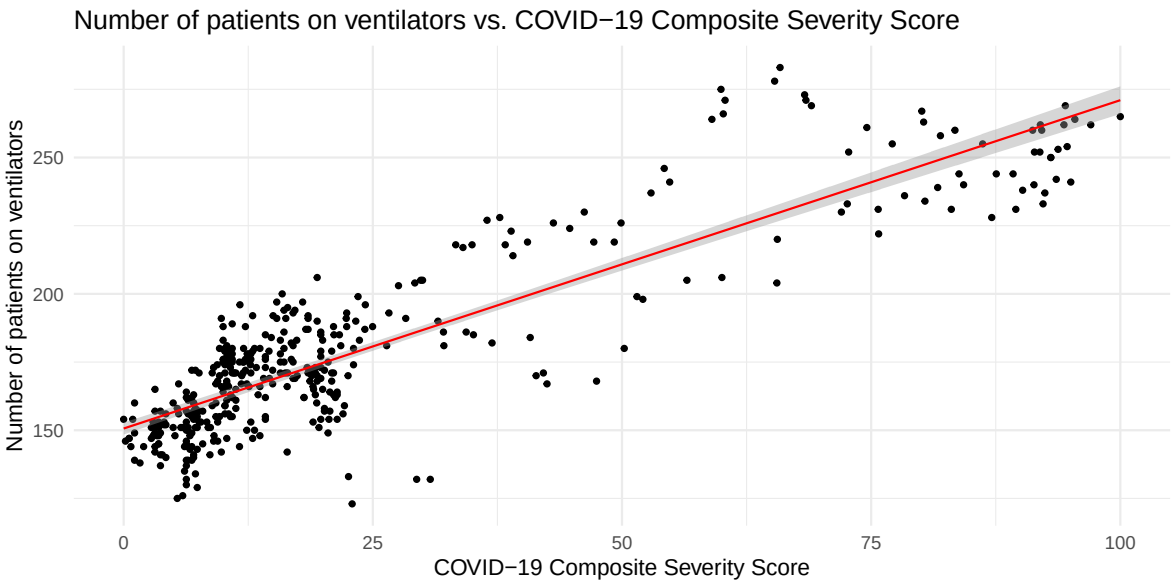

Figure 5: Number of patients on ventilators vs. COVID-19 Composite Severity Score

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

References

"COVID-19 Hospitalizations by Date | County of Santa Clara." n.d. Accessed September 23, 2025. https://data.sccgov.org/COVID-19/COVID-19-hospitalizations-by-date/5xkz-6esm/about_data.

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.