Journal of Mechanics Engineering and Automation 13 (2023) 93-99
doi: 10.17265/2159-5275/2023.04.001

# Predictive Modeling for Analysis of Coronavirus Symptoms Using Logistic Regression

Anatoli Nachev
*Business Information Systems, University of Galway, Galway H91 TK33, Ireland*

**Abstract:** This paper presents a case study on the IPUMS NHIS database, which provides data from censuses and surveys on the health of the U.S. population, including data related to COVID-19. By addressing gaps in previous studies, we propose a machine learning approach to train predictive models for identifying and measuring factors that affect the severity of COVID-19 symptoms. Our experiments focus on four groups of factors: demographic, socio-economic, health condition, and related to COVID-19 vaccination. By analysing the sensitivity of the variables used to train the models and the VEC (variable effect characteristics) analysis on the variable values, we identify and measure importance of various factors that influence the severity of COVID-19 symptoms.

**Key words:** COVID-19, supervised learning, models, classification, logistic regression.

## 1. Introduction

The COVID-19 pandemic, officially recognized in 2020, has dramatically changed all aspects of life around the world. The increasing mortality rate led to a crisis in the health care systems and subsequent national lockdowns, which in turn negatively affected economic and social welfare. Health care delivery was also adversely affected by outpatient medical facilities ceasing routine patient care. Medical risk factors influencing symptom severity and increased mortality rate have been well studied and reported in the literature. At the same time, additional factors of a social and demographic nature remain understudied in the literature. The purpose of this cross-sectional study was to use data from the US National Health Survey to estimate the potential impact of demographic, socio-economic, and general health factors on the severity of symptoms experienced by diagnosed COVID-19 cases.

The literature shows that studies related to the impact of COVID-19 on various aspects using the data we use include the study of socio-demographic and behavioural factors associated with receiving a seasonal flu vaccine [1]; utilization of medical care [2]; disruptions in cancer care and treatment [3]; racial and ethnic disparities in health outcomes [4]; differences in access to health care according to sexual orientation [5]. The literature also suggests that the most common approach to data analysis is statistical [1, 2, 3, 5].

In order to estimate the risk factors contributing to severity of COVID-19 symptoms, this study uses logistic regression predictive models trained by the data, combined with sensitivity and VEC (variable effect characteristic) analysis of variables.

The remainder of the paper is organized as follows: Section 2 provides an overview of the methods used to analyse data; Section 3 discusses the data used; Section 4 presents and discusses experimental results; and Section 5 gives conclusions.

## 2. Methods

### *2.1 CRISP-DM*

The CRISP-DM framework is a widely used approach to data analysis and mining, consisting of six main phases (Fig. 1) for successful transformation, development, testing, and deployment of machine learning solutions [6].

**Corresponding author:** Anatoli Nachev, Ph.D., assistant professor, research fields: data analytics and machine learning.

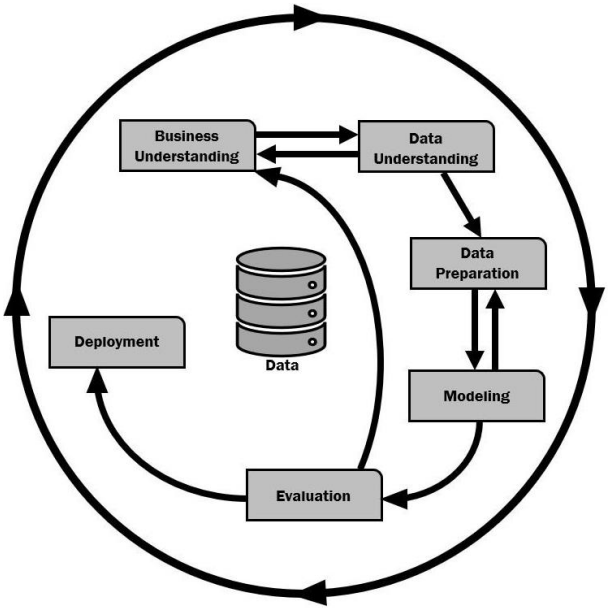


**Fig. 1 The CRISP-DM methodology and its phases.**

This study uses the CRISP-DM methodology as a research approach.

The first stage is *business understanding*, defining the project goals from a business perspective and turning that knowledge into a data mining problem definition.

From this point of view, the main objective of this study is to train binary classification models that can predict severity of COVID-19 symptoms based on various features from available data. The *data understanding* stage includes initial data collection and activities to understand the meaning of the data, identify data quality issues, and obtain an initial understanding of the data. The *data preparation* stage includes all activities to build the final dataset from the initial raw data. The *modelling* stage is focused on training classification models and optimizing their parameters for maximum performance. Finally, the *evaluation* stage involves a thorough evaluation of the models and a review of all CRISP-DM stages. The *deployment* stage is not applicable to this study.

### *2.2 Logistic Regression*

This research uses binary logistic regression as machine learning technique for binary classification. It measures relationship between one or more independent variables $X_i$ and a categorical dependent variable $Y$ by estimating probabilities using the logistic function (1).

$$Y(X_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_{1n})}} \quad (1)$$

also known as odds ratio, where $\beta_i$ are regression coefficients. The logistic function will always produce an S-shaped curve as shown in Fig. 2, so values of Y close to 0 indicate low probability of belonging to the success class 1 and values close to 1 indicate high probability of belonging to that class. The coefficients $\beta_i$ are estimated using the maximum likelihood technique on the training data.

## 3. Dataset

This study uses cross-sectional microdata collected from the IPUMS NHIS, providing information on U.S. population health [7]. The extracted dataset initially contained 68 variables and 75,101 records from years 2021-2022. With reference to the CRISP-DM data understanding and preparation stages, variables were reduced to 33 to avoid multicollinearity. Records were also reduced to 4,340 representing only diagnosed COVID-19 cases.

Variables were grouped by meaning into four categories, namely:

- *Demographic*: year, region of residence, urban-rural, age, sex, sex orientation, marital status, race, Hispanic ethnicity, US born, and citizenship.

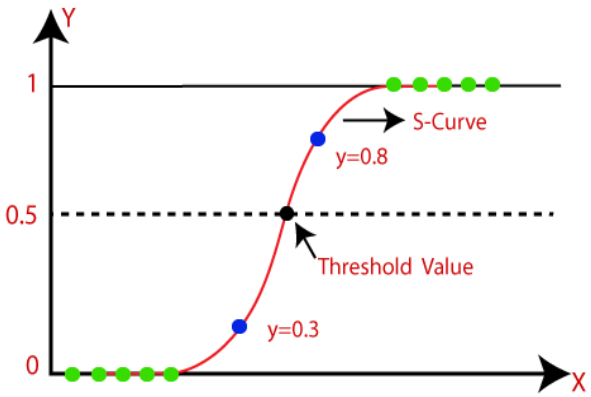


**Fig. 2 Logistic regression function.**

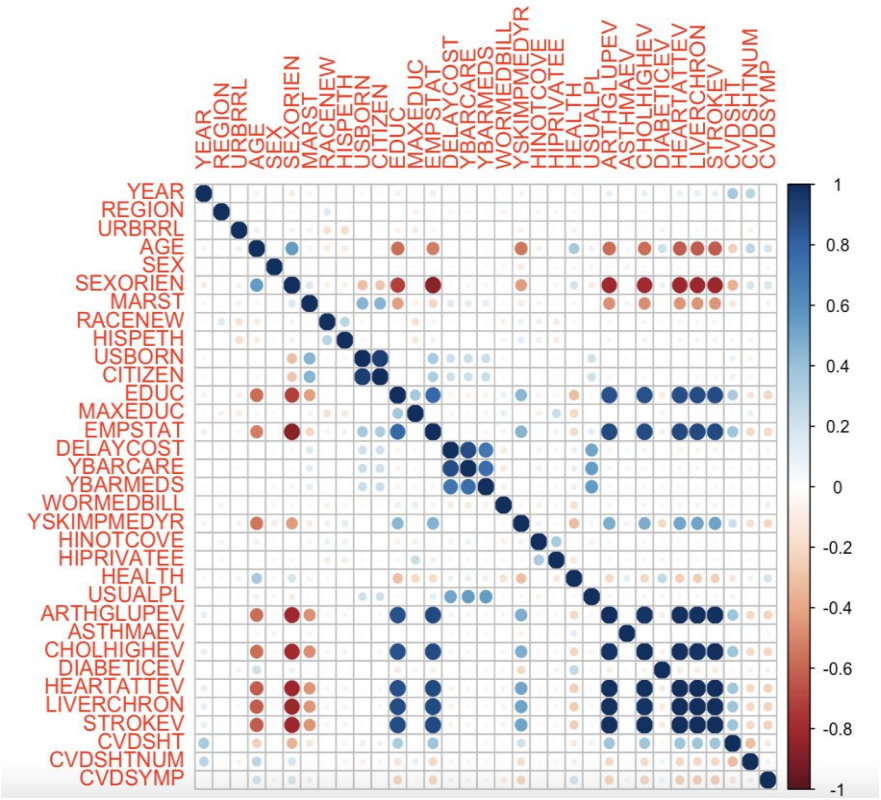


**Fig. 3 Correlation matrix plot. Positive and negative correlations are represented by colored circles.**

- *Socio-economic*: education, max family education, employment, delayed medical care due to cost, unaffordable medical care, unaffordable medicines, worried about paying medical bills, less medications to save money, unable to pay medical bills, health insurance, and private health insurance.
- *Health condition*: health status, usual place for medical care, arthritis, asthma, high cholesterol, diabetes, heart attack, liver condition, and stroke.
- *COVID-19 related*: covid vaccination, number of vaccinations, and covid severity symptoms (dependent variable).

The correlation analysis of variables illustrated in Fig. 3 showed very strong positive and negative correlations between variables with a value above 0.8, which led to the elimination of four more variables: SEXORIENTATION, YBARCARE, USUALPLACE, and HISPET.

We further analysed these factors by training predictive models based on logistic regression, which is discussed in the next section.

## 4. Experiments and Analysis

The focus of our experiments was to investigate the relationship between different factors and the severity of symptoms of diagnosed COVID-19 cases. For that purpose, in R [8] we trained four logistic regression classifiers with data from each of the four categories to examine in detail the factors within those categories. With reference to the modelling stage of CRISP-DM, we set aside 20% of the data for testing purposes only and the rest of 80% were split for training and validation purposes in 2:1 ratio. To avoid training bias caused by composing “lucky sets” we used 5-fold cross-validation and averaged results from 10 instances of each classifier. Metrics used to evaluate performance were prediction accuracy (Acc), sensitivity (TPR), specificity (TPR), precision, and F1. We also performed Receiver Operating Characteristics (ROC) analysis [9] of the classifiers and calculated the Area Under the

**Table 1 Performance metrics of logistic regression models trained by four categories of variables.**

| | Demographic | Socio-economic | Health | COVID-19 |
|---|---|---|---|---|
| Acc | 68.29 | 71.06 | 72.33 | 72.33 |
| AUC | 0.70 | 0.76 | 0.76 | 0.76 |
| TPR | 56.25 | 74.05 | 76.84 | 76.84 |
| TNR | 60.64 | 47.27 | 46.78 | 46.78 |
| Precision | 62.14 | 59.86 | 60.74 | 60.74 |
| F1 | 59.05 | 66.20 | 67.85 | 67.85 |

Curve (AUC) metric to evaluate the overall performance regardless of the choice of threshold operating point for mapping predicted probability to a class label.

Table 1 shows the performance metrics for each classifier. According to accuracy, as the primary performance estimator for a classifier, factors related to overall health and COVID-19-related to vaccination categories play an equally important and most significant role in COVID-19 symptoms. This is confirmed by the equal and highest values of all other metrics. The next in importance is the category of socio-economic factors. It shows slightly lower accuracy in predictions but the same AUC value of 76%, almost equalizing the importance of this category with the previous two. However, the demographic factors generally play the least significant role in COVID-19 symptoms, with some exceptions. This is confirmed by the lowest values of nearly all metrics in the table. To rank variable significance, we used the sensitivity analysis method [10], which varies each input variable through its range from min to max value and use the gradient measure to rank it. To analyse how individual variable values contribute to the output, we performed a VEC analysis [11] on some significant non-binary variables.

The following sections show those details for each category of factors.

### *4.1 Demographic Factors*

Demographic factors in the data include REGION—region of residence; URBRRL—urban-rural residence; AGE—age; SEX—gender; MARST—current marital status; RACENEW—self-reported race; USBORN—U.S. born; and CITIZEN—citizenship.

The sensitivity analysis performed on the trained model by these variables ranks their importance, as seen in Fig. 4. AGE dominates over the others with a score of 60%, while the remaining 40% is shared among CITIZEN, MARST, and USBORN. The rest of the variables have, in practice, no significance towards the severity of symptoms. When examining how values of AGE affect the symptoms through VEC (Fig. 5), an almost linear dependence between increasing age and worsening of symptoms can be observed.

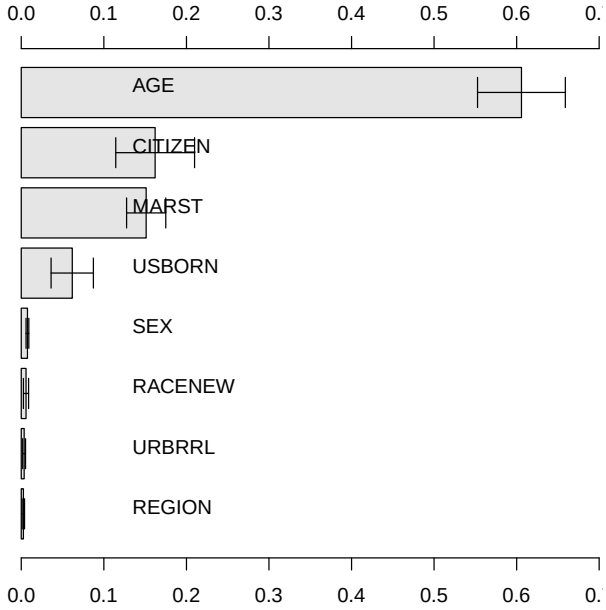


**Fig. 4 Significance of demographic variables.**

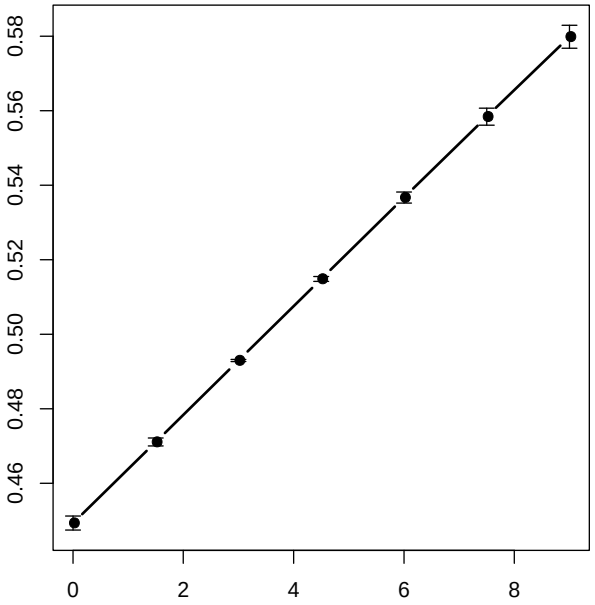


**Fig. 5 VEC analysis of AGE.**

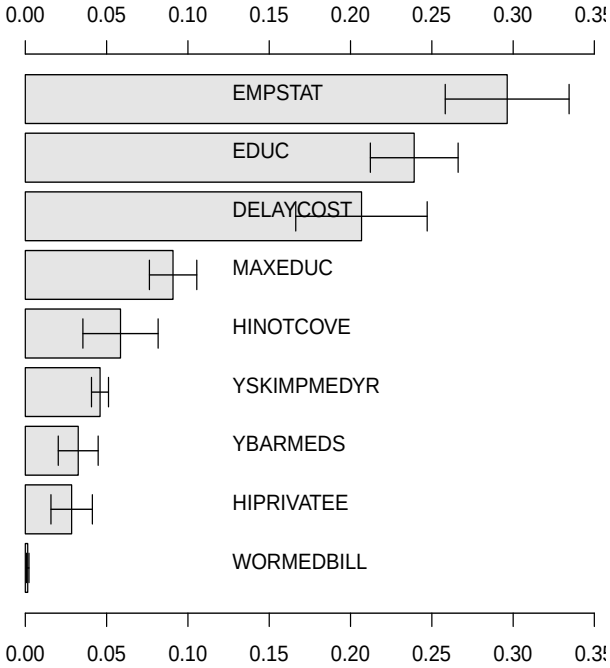


**Fig. 6 Significance of socio-economic variables.**

*4.2 Socio-Economic Factors*

Socio-economic factors in the data include EMPSTAT—employment status; EDUC—educational attainment; MAXEDUC—highest level of education of all the adults in the family; DELAYCOST—delayed medical care due to cost; YBARMEDS—could not afford prescription medicines; WORMEDBILL—worried about paying medical bills; YSKIMPMEDYR—took less medication to save money; HINOTCOVE—health insurance coverage status; and HIPRIVATEE—private health insurance.

The sensitivity analysis conducted on this model ranks the importance of the variables, as seen in Fig. 6. The group of three variables, EMPSTAT, EDUC, and DELAYCOST, dominate over the others with similar scores, showing a combined significance of 75%, while the remaining six variables share a 25% significance, making them less significant, overall. According to the results, the main socio-economic factors that influence severity of symptoms are primarily economic, reflecting financial ability to bear the costs of treatment, as well as educational, perhaps related to the assumption that more educated people take more adequate and timely decisions about risks and necessary actions for treatment. Also educated people are generally better paid and secure financially.

The study on how EDUC values affect symptoms through VEC (Fig. 7) shows that the relationship between education attainment and severe symptoms is almost linear, with an increase in patients' education levels leading to a decrease in symptom severity.

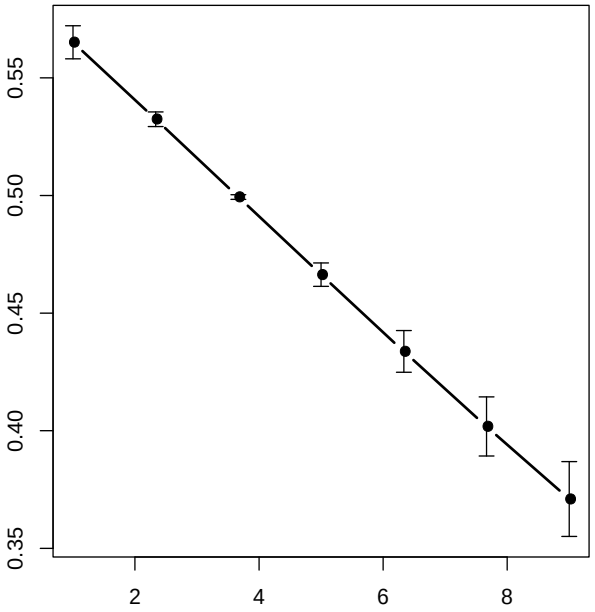


**Fig. 7 VEC analysis of EDUC.**

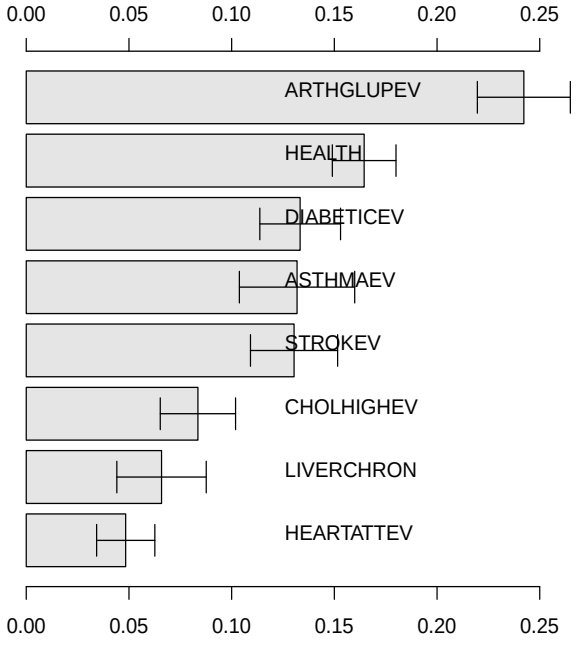


**Fig. 8 Significance of health variables.**

*4.3 Health Factors*

The third group of factors are related to health and include HEALTH—a healthy state; ARTHGLUPEV—autoimmune conditions; ASTHMAEV—asthma; CHOLHIGHEV—high cholesterol; DIABETICEV—diabetes; HEARTATTEV—had a heart attack; and LIVERCHRON—chronic liver disease.

Similarly, the sensitivity analysis performed on this model ranks the variable significance, as seen in Fig. 8. Results show that the leading factor is presence of autoimmune disease (24%), followed by the overall health state (17%), and a group of three factors for the presence of asthma, diabetes, and stroke (13% each). The least significant factors are those related to high cholesterol, liver condition, and heart attack (combined 20%).

The study of how overall health outcomes affect symptoms through VEC analysis (Fig. 9) shows that while the self-reported overall health status varies from excellent to poor, the severity of symptoms changes from low to high.

*4.4 COVID-19-Related Factors*

The fourth group of COVID-19-related factors contains two variables: CVDSHT—at least one covid vaccination; and CVDSHTNUM—number of covid vaccinations.

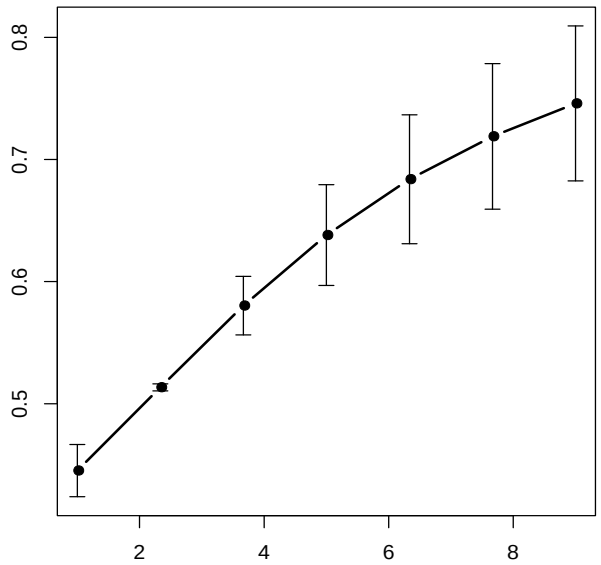


**Fig. 9 VEC analysis of HEALTH.**

Sensitivity analysis performed on this model ranks the presence of vaccination as most significant with 73%, as seen in Fig. 10. The second significant factor with 27% is the number of vaccinations. Studying how the number of vaccinations affects symptoms through VEC analysis (Fig. 11) shows that the relationship between that number and severe symptom is inversely proportional.

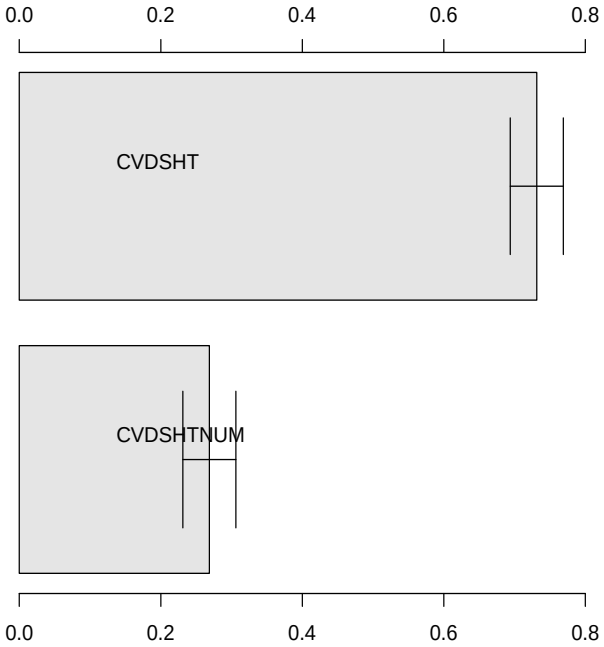


**Fig. 10 Significance of COVID-19-related variables.**

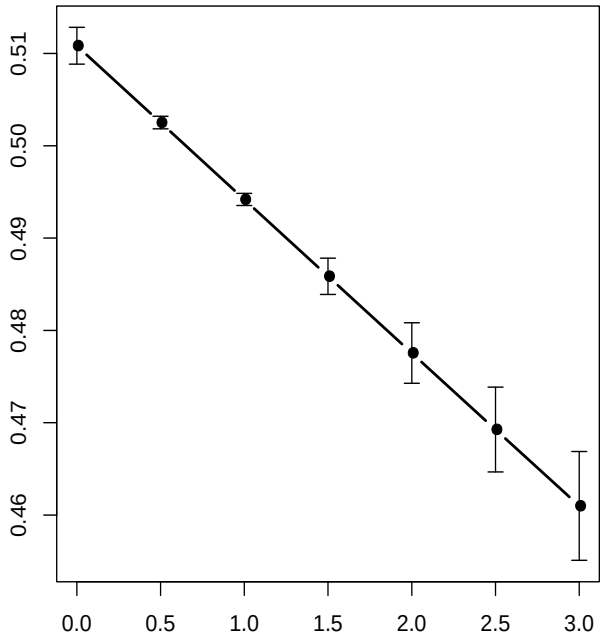


**Fig. 11 VEC analysis of CVDSHNUM.**

## 5. Conclusions

The aim of this study is to investigate factors affecting the severity of symptoms in COVID-19 disease, using data from surveys conducted between 2021-2022 in the U.S. The data contain four sets of variables related to demo-graphic, socio-economic, health, and COVID-19 vaccination indicators. Although the specialized medical literature has studied extensively severity of symptoms from medical perspective, factors of demographic, social, economic, and health-related nature still deserve a better estimation. We explore those factors by training predictive models with available data and estimate their significance through sensitivity and VEC analyses. The results discussed in the paper show that in addition to the importance of some health factors and the role of COVID-19 vaccination, socio-economic factors such as financial status and education level of patients also play an important role. Another outcome from this approach is measuring of the factor importance.

Those findings and discussion can be seen as a basis for decision making and policy development in the domain.

## References

[1] Vashist, K., Choi, D., and Patel, S. A. 2022. "Identification of Groups at High Risk for Under-Coverage of Seasonal Influenza Vaccination: A National Study to Inform Vaccination Priorities during the COVID-19 Pandemic." *Annals of Epidemiology* 68: 16-23.

[2] Weissman, J. D., Pinder, N., Jay, M., and Taylor, J. 2023. "The Impact of Health Coverage, Race and Ethnicity on Utilization of Preventive Medical Care during the First Year of the Covid-19 Pandemic: Racial and Ethnic Health Disparities." *Journal of Racial and Ethnic Health Disparities* 10: 1-9.

[3] Lang, J. J., Narendrula, A., Iyer, S., Zanotti, K., Sindhwani, P., Mossialos, E., and Ekwenna, O. 2022. "Patient-Reported Disruptions to Cancer Care during the COVID-19 Pandemic: A National Cross-Sectional Study." *Cancer Medicine* 12 (4): 4773-85.

[4] Hoang, M. T. 2022. "The Evolution of Racial and Ethnic Disparities in Health Outcomes." Bachelor of Science thesis, University of Central Florida. https://stars.library.ucf.edu/honorstheses/1147.

[5] Gonzales, G., and de Mola, E. L. 2021. "Potential COVID-19 Vulnerabilities in Employment and Healthcare Access by Sexual Orientation." *Annals of LGBTQ Public and Population Health* 2 (2): LGBTQ-2020-0052.

[6] Chapman, J. C., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. 2000. "CRISP-DM 1.0—Step-by-Step Data Mining Guide." CRISP-DM Consortium.

[7] Blewett, L. A., Drew, J. A. R., King, M. L., and Williams, K. C. W. 2022. *Natalie Del Ponte and Pat Convey. IPUMS Health Surveys: National Health Interview Survey, Version 7.2 [dataset]*. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D070.V7.2.

[8] R Development Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

[9] Fawcett, T. 2005. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27 (8): 861-74.

[10] Kewley, R., Embrechts, M., and Breneman, C. 2000. "Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks." *IEEE Transactions on Neural Networks* 11 (3): 668-79.

[11] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. 2009. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." *Decision Support Systems* 47 (4): 547-553.