9th International scientific conference
Technics and Informatics in Education – TIE 2022
16-18 September 2022

Session: *IT Education and Practice*
Original scientific paper
DOI: 10.46793/TIE22.257J

Data analysis for COVID-19 using regression methods

Nemanja Jovanović [1*]
[1] University of Kragujevac, Faculty of Technical Sciences Čačak, Serbia
[*] neca.j.096@gmail.com

Abstract: *With the appearance of the first registered case of corona, as one of the world's most widespread and most dangerous viral infections, the need to monitor and predict the epidemiological situation is growing, both in the world and in our country. In this paper, the epidemiological data of the Republic of Serbia regarding the Corona virus in the period from 2020 to June 2021 are analyzed. Data were analyzed by regression methods, as one of the data mining techniques. Depending on the choice of regression method (simple, multiple and linear), a number of parameters were selected that include the number of persons (positive, tested, deceased, hospitalized and respirator) in relation to the time of the pandemic to make the most accurate prediction. As a result of the research using regression methods, it was found that the trend of development of the Corona virus epidemic is decreasing, i.e. (id est.) that preventive measures as well as the process of vaccination and revaccination have had an effect in the fight against Corona virus.*

Keywords: *regression; corona; data mining; analysis; the data*

1. INTRODUCTION

With the development of computer technology in the modern business world, there is a need for a variety of ways to research data on the Internet, documents and other content, in order to increase business efficiency, the need for new information and the contribution of scientific research. One of these ways is the application of data mining techniques, through which the extraction of useful information is performed.

There are several types of data mining techniques, and one that was chosen for the purpose of research is the regression technique, which contains a number of methods by which it is possible to perform the necessary analysis of the data set. Based on the data analysis using regression methods, new perspectives are opened for further prediction of observed phenomena and events, which as such the user is not able to notice and register in the future, and which are extremely important to the business world and the general public. One of such data is the most current topic related to COVID-19, i.e. (*id est.)* the corona virus that shook the world public at the end of 2019 and the beginning of 2020.

This topic is the goal of the research, and the data collected on this topic will be analyzed using regression methods, as one of the data mining techniques, with the help of software tools, in order to provide the best information, i.e. forecasts of further pandemic development. In other words, the importance of each regression method in data analysis will be emphasized, and their individual impact on the obtained results of the research itself will be compared with other related research.

The purpose of the research is to monitor the data in order to see the current state of health at the national level and based on that to create forecasts for the further development of the corona virus pandemic by using different regression methods.

In the next part of the paper it will be presented an overview of related research by other authors, theoretical overview of the regression, research methodology and results and discussion, which is the goal of this research.

2. A REVIEW OF RELATED RESEARCH

The beginning of the research is based on the study of regression methods, but also on the review of related research by other authors in the same field. For the field of data analysis using regression methods, the following studies were selected with the authors:

- Chauhan, P. & Kumar, A. & Jamdagni, P. (2020). Regression Analysis of COVID-19 Spread in India and its Different States.
- Tenenholtz, J. & George, F. & Gulati, S. (2020). Some Multiple Regression Models for the Number of COVID-19 Cases and Deaths in the United States. International Journal of Statistics and Probability.

In the first research “Regression Analysis of COVID-19 Spread in India and its Different States”, linear and polynomial regression model has been used to investigate the COVID-19 outbreak in India and its different states using time series epidemiological data up to 26th May 2020. Authors were analyzed number of deaths and recovered people by using simple linear regression. Also, the polynomial regression model is used to predict the number of patients in next three week. Both methods gives excellent and expected results in predicting COVID-19 in India and its Different States.

In the second research "Some multiple Regression Models for the Number of COVID-19 cases and deaths in the United States", epidemiological data are analyzed by using multiple linear regression model. Authors were analyzed number of COVID-19 cases by using more predictor variables. In this research, multiple linear regression models are here to identify the significant factors affecting the number of confirmed COVID-19 cases and the number of deaths per 100000 in the states of the US. Authors identified population density as the most influential factor in all the models.

In the following chapters, the theoretical part of the data mining technique, regression as well as its methods will be presented, on the basis of which the direction towards further research methodology has been developed.

3. DATA MINING TECHNIQUES

Data mining is the process of extracting out valid and unknown information from large databases and use it to make difficult decisions in business. Data mining or data analysis with complex and large datasets brings the wealth of research and knowledge in machine learning and statistics for the task of discovering new sets of knowledge in large databases. Over the past three decades, large amounts of difficult data's of business are stored electronically and this volume will continue to increase in future. In order to manage huge volumes of data, the techniques of data mining are also becoming sophisticated and advanced, day by day [1]. Fig. 1 shows the specific process of Data Mining, from the data itself, the selection of targeted data, processing them using some of the data mining techniques and presenting them according to a business template.

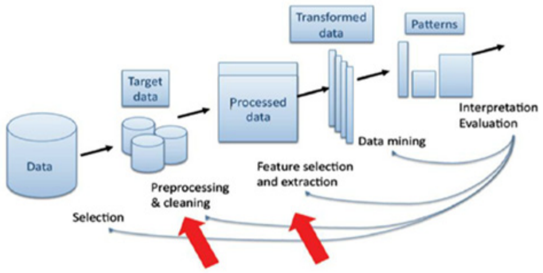


***Figure 1**. Data mining process [2]*

The difference between data analysis and research lies in the fact that data analysis is used to test statistical models and hypotheses on a set of data, e.g. (exempli gratia) when analyzing the effectiveness of a marketing campaign, regardless of the amount of data. In contrast, data research uses machine learning and statistical models to uncover secret or hidden patterns in large amounts of data [3]. It is important to emphasize that different data mining techniques are used for research, which can be seen in Fig. 2.

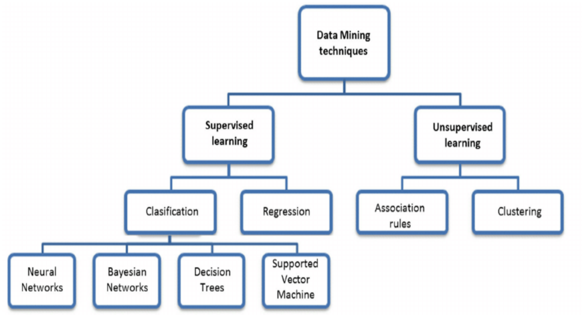


***Figure 2.** Data mining techniques [4]*

The Fig. 2 shows that Data mining techniques can be divided into supervised and unsupervised learning type. Supervised learning contains classification and regression. Unsupervised learning contains association rules and clustering.

In this research as it is mention before, regression methods will be used and the next chapter will describe technique and methods.

4. REGRESSION TECHNIQUE

Regression analysis is a collection of statistical techniques that serve as a basis for drawing conclusions on the relationships between interrelated variables. Since these techniques are applicable in almost all fields of study, including social, physical and biological sciences, business and engineering, regression analysis is now perhaps the most used of all methods of analysis data [5]. In order to determine whether and to what extent these phenomena are dependent, it is necessary make a regression model.

The regression technique contains a number of methods, on the basis of which the analysis and discussion of the achieved results over the data can be performed, and they are:

- Linear regression
    - Simple linear regression
    - Multiple (complex) linear regression
- Non-linear regression
- General linear model
    - Poisson model
    - Logistic model
- Log-linear models
- Regression tree and tree model

The following subchapters provide the theoretical basis for regression methods that will be used for further research.

4.1. Simple linear regression

The simplest form of regression is a simple linear regression containing one dependent and one independent variable, based on which a linear regression model can be created, which could be viewed as a line that minimizes the error rate between the actual prediction value and points on the line [6]. Most often, linear regression refers to a model in which the conditional mean value of Y, with a given value of X, is an affine function of X. The case with one independent variable is called simple, i.e. simple linear regression. When more than one independent variable is included, the process is called multiple (complex) linear regression [7].

The least squares method is a statistical way of evaluating regression analysis in order to predict a solution using a certain system that contains several unknown variables in a set of equations. The least squares method can be used to denote the reduction of the sum of the final solution, i.e. the squares that represent the residues made in the results of each equation [8].

4.2. Multiple linear regression

Ordinary least square method which is widely used in case of simple regression is also most widely used in case of predicting the value of dependent variable from the values of two or more independent variables. Regression equation in which dependent variable is estimated by using two or more independent variables is known as multiple regression [9].

In developing a multiple regression equation, one needs to know the efficiency in estimating the dependent variable on the basis of the identified independent variables in the model. The efficiency of estimation is measured by the coefficient of determination (R2) which is the square of multiple correlation. The coefficient of determination explains the percentage of variance in the dependent variable by the identified independent variables in the model. The multiple correlation explains the relationship between the group of independent variables and dependent variable. Thus, high multiple correlation ensures greater accuracy in estimating the value of dependent variable on the basis of independent variables. Usually multiple correlation, R is computed during regression analysis to indicate the validity of regression model. It is necessary to show the value of R2 along with regression equation for having an idea about the efficiency in prediction [9].

4.3. Nonlinear regression

In statistics, nonlinear regression is a form of regression analysis in which experimental data are modeled by a function that is a nonlinear combination of model parameters, and depends on one or more independent variables. Data were processed by the method of successive approximations. The data consists of independent variables that do not contain errors (explanatory variables), k, and related experimental dependent variables (responsive variables) y. Each value of y is modeled as a random variable with the average given in the form of a nonlinear function f (k, β). Systematic errors may be present, but their treatment is beyond the scope of regression analysis. If the independent variables contain errors, variable error models can be used [10].

5. RESEARCH METHODOLOGY

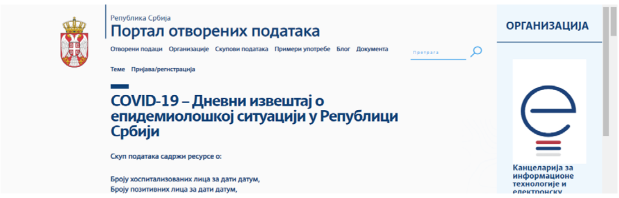
This research methodology at first starts by choosing tool for analyzing the data. After choosing the tool, it has to get the data from valid resource. After getting data, data need to be preformatted in model that is common for regression analysis. In the next chapters, it will be presented whole methodology implementation.

5.1. NCSS Data Analysis Tool

"NCSS" is a statistical package produced and distributed by NCSS, a limited liability company. It was created in 1981 by Jerry L. Hintze, as NCSS, a limited liability company, specializing in providing software services in the form of statistical analysis intended for researchers, companies and academic institutions [11]. After the tool is installed, it is necessary to download the data for analysis.

5.2. Data collected for analysis

Data for analysis were collected from the open data portal, which is provided by the Government of the Republic of Serbia. The selection of data for analysis refers to the daily report on the epidemiological situation for Covid 19 in the Republic of Serbia, which can be seen in Fig. 3.



COVID-19 – Дневни извештај о епидемиолошкој ситуацији у Републици Србији

Скуп података садржи ресурсе о:

Броју хоспитализованих лица за дати датум,
Броју позитивних лица за дати датум,

***Figure 3**. Open Data Portal – Covid 19 – Daily report of the epidemiological situation in the Republic of Serbia [12]*

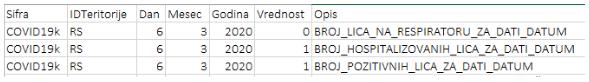
5.3. Preprocessing and data transformation

After successfully collecting the data in .xlsx format, the next step is related to data preprocessing and transformation.

Data preprocessing is one of the most important tasks of Data Mining and involves the preparation and transformation of data into an appropriate template, which is suitable for research methods. The preprocessing activity strives to reduce the amount of data, find connections between data, normalize, as well as eliminate surpluses and extract new data. The process itself involves several techniques such as [13]:

- cleaning,
- integrations,
- transformations and
- removal of redundant data.

The downloaded data set must correspond to a template that is suitable for research in order to be able to apply the appropriate method, in this case the regression method. The original data set contained a large number of rows, then columns that did not fit the appropriate analytical template, as well as empty cells that had to be formatted in a way that ensures accuracy in data analysis. The original data set can be seen in Fig. 4.

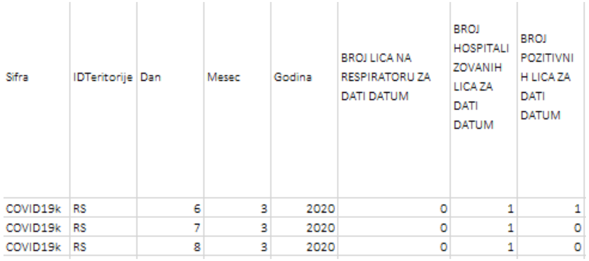


Sifra	IDTeritorije	Dan	Mesec	Godina	Vrednost	Opis
COVID19k	RS	6	3	2020	0	BROJ_LICA_NA_RESPIRATORU_ZA_DATI_DATUM
COVID19k	RS	6	3	2020	1	BROJ_HOSPITALIZOVANIH_LICA_ZA_DATI_DATUM
COVID19k	RS	6	3	2020	1	BROJ_POZITIVNIH_LICA_ZA_DATI_DATUM

Figure 4*. Original data set*

Fig. 5 clearly shows the repetition of the number of rows caused by the last column, which is a description of different categories of the epidemiological situation. Also, the penultimate two columns can be seen to be unfilled, i.e. not assigned values, where such a set of data later gives incorrect forecasts. In order to reduce the number of rows in the data set, it is necessary to create separate columns for the values from the description column.

By creating new columns, the number of rows will be reduced, because when entering data for a given date, only one row will be needed for all the listed columns. Numeric values have been added to the blank fields to reduce the possibility of computational problems. Also, separate columns are created, which enable better results, because they leave the possibility of analyzing data according to several dependent and independent variables (x and y). In this way, a successful transformation of the data itself was performed, which can be seen in Fig. 5.



Sifra	IDTeritorije	Dan	Mesec	Godina	BROJ LICA NA RESPIRATORU ZA DATI DATUM	BROJ HOSPITALIZOVANIH LICA ZA DATI DATUM	BROJ POZITIVNIH LICA ZA DATI DATUM
COVID19k	RS	6	3	2020	0	1	1
COVID19k	RS	7	3	2020	0	1	0
COVID19k	RS	8	3	2020	0	1	0

Figure 5*. Data for analysis after transformation*

Data transformation of data favors analysis using regression methods, because the data are adapted to the appropriate template. The further course of data transformation would be reflected in the logical division of data, by column of the year, where two separate sets of data for 2020. and 2021. are created on the basis of the division.

In the next chapter, after successfully installed software, collected, pre-processed and transformed data from the open data portal, the analysis and discussion based on the achieved results using certain regression methods is approached.

6. RESULTS AND DISCUSSION

The following subsections will present the results and discussion using the above regression methods on the collected data.

6.1. Results and discussion using simple linear regression

Based on the imported data as well as the applied filters, a simple linear regression is applied by selecting the option Analysis, Regression, Simple Linear Regression. In order to perform a simple linear regression, it is necessary to choose one dependent and one independent variable, i.e. x and y.

For the analysis of results in 2020. and 2021., the following were selected:

- For x (independent variable) - months of the year
- For y (dependent variable) - number of people on respirator

Based on the data for 2020., i.e. defined columns for x and y, using simple linear regression, a graphical representation was created showing the flow of the number of persons on the respirator in relation to the months of the year, which can be seen in Fig. 6.

It can be clearly seen that simple linear regression predicts an increase in the number of people on respirators, starting from the beginning of the pandemic in March until December 2020.

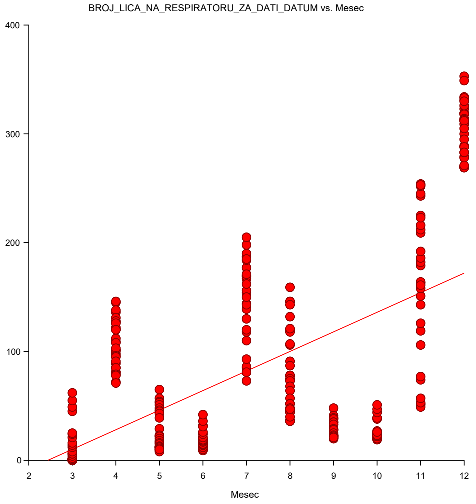


Figure 6*. The result of the analysis was achieved by applying a simple linear regression for 2020.*

After analyzing the data for 2020, we proceed to further analysis for 2021, i.e. defined columns for x and y, using simple linear regression, based on which a graphical representation is created showing the flow of persons on the respirator in relation to the months of the year, which can be seen in Fig. 7. From the graph it can be clearly seen that a simple linear regression predicts a decrease in the number of people on the respirator, starting from January until June 2021.

The result of the reduction in the number of people on respirators can be linked to clearly defined measures in the fight against Covid 19 virus issued by the Government of the Republic of Serbia, as well as voluntary vaccination conducted in early 2021.

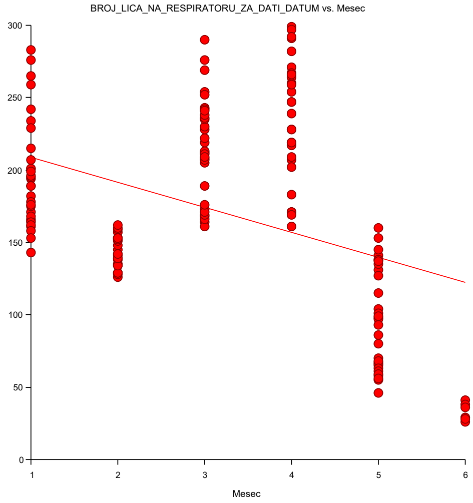


***Figure 7**. The result of the analysis was achieved by applying a simple linear regression for 2021.*

6.2. Result and discussion using multiple (complex) linear regression

Based on the imported data as well as the applied filters, multiple (linear) regression is applied by selecting the option Analysis, Regression, Multiple Linear Regression. In order to perform multiple linear regression, it is necessary to choose one dependent and one independent variable, i.e. x and y. For the analysis of results in 2020. and 2021., the following were selected:

- For x (independent variables) - number of persons on respirator, number of hospitalized persons, number of positive persons, number of tested persons as well as number of deceased persons for a given date
- For y (dependent variable) - month of the year

Based on the data for 2020., i.e. defined columns for x and y, using multiple (complex) linear regression, a graphical representation was created showing the flow of the above parameters in relation to the months of the year, which can be seen in Fig. 8. the graph clearly shows that multiple (complex) linear regression predicts a general increase in all independent variables starting from x_1 to x_5, relative to the dependent variable y, starting from the beginning of the pandemic, more precisely from March until December 2020.

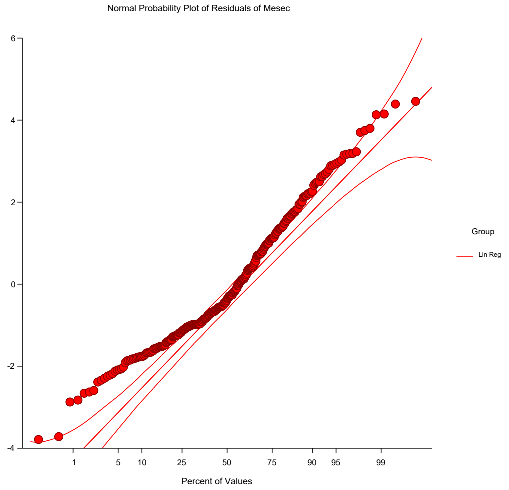
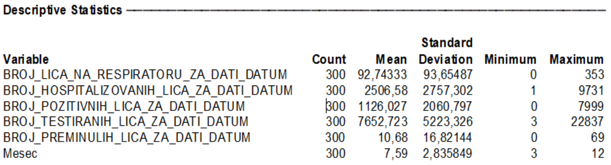


***Figure 8**. The result of the analysis was achieved by applying multiple (complex) linear regression for 2020.*

Also, from the descriptive analysis itself, it can be concluded that the variable related to the number of tested persons for a given date has the greatest significance in the prediction by applying multiple (complex) linear regression. This result shows that the prediction largely depends on the number of tested persons, because based on that number there is a possibility for a better and more accurate forecast of the further course of the pandemic, which can be seen in Fig. 9.



Descriptive Statistics

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
BROJ_LICA_NA_RESPIRATORU_ZA_DATI_DATUM	300	92,74333	93,65487	0	353
BROJ_HOSPITALIZOVANIH_LICA_ZA_DATI_DATUM	300	2506,58	2757,302	1	9731
BROJ_POZITIVNIH_LICA_ZA_DATI_DATUM	300	1126,027	2060,797	0	7999
BROJ_TESTIRANIH_LICA_ZA_DATI_DATUM	300	7652,723	5223,326	3	22837
BROJ_PREMINULIH_LICA_ZA_DATI_DATUM	300	10,68	16,82144	0	69
Mesec	300	7,59	2,835849	3	12

***Figure 9**. Descriptive statistics achieved by applying multiple (complex) linear regression for 2020.*

After analyzing the data for 2020., we proceed to further analysis for 2021., i.e. defined columns for x and y, using multiple (complex) linear regression, based on which a graph is created showing the flow above the above parameters in relation to the months in years, which can be seen in Fig. 10.

From the graph it can be clearly seen that multiple (complex) linear regression predicts general stagnation and a slight decrease in all independent variables starting from x_1 to x_5, relative to the dependent variable y, in the time interval of January until June 2021.

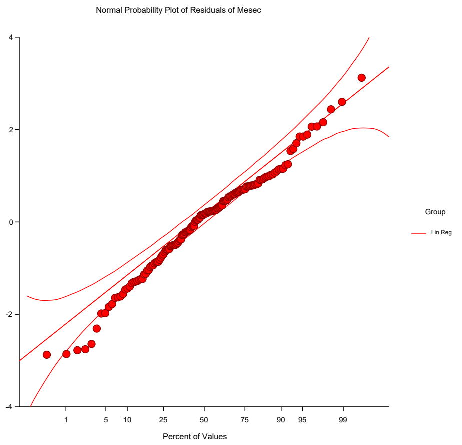
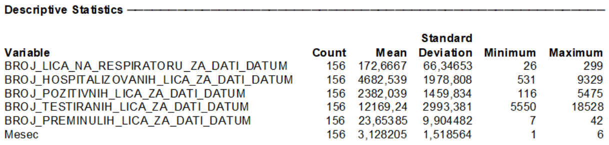


***Figure 10**. The result of the analysis was achieved by applying multiple (complex) linear regression for 2021.*

Also, from the descriptive analysis itself, it can be concluded that the variable related to the number of tested persons for a given date has the greatest significance in the prediction by applying multiple (complex) linear regression. This result shows that the prediction largely depends on the number of tested persons, because based on that number there is a possibility for a better and more accurate forecast of the further course of the pandemic, which can be seen in Fig. 11.


Descriptive Statistics

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
BROJ_LICA_NA_RESPIRATORU_ZA_DATI_DATUM	156	172,6667	66,34653	26	299
BROJ_HOSPITALIZOVANIH_LICA_ZA_DATI_DATUM	156	4682,539	1978,808	531	9329
BROJ_POZITIVNIH_LICA_ZA_DATI_DATUM	156	2382,039	1459,834	116	5475
BROJ_TESTIRANIH_LICA_ZA_DATI_DATUM	156	12169,24	2993,381	5550	18528
BROJ_PREMINULIH_LICA_ZA_DATI_DATUM	156	23,65385	9,904482	7	42
Mesec	156	3,128205	1,518564	1	6

***Figure 11**. Descriptive statistics achieved by applying multiple (complex) linear regression for 2021.*

6.3. Result and discussion using nonlinear regression

Based on the imported data as well as the applied filters, nonlinear regression is applied by selecting the option Analysis, Regression, Nonlinear Regression. In order to perform nonlinear regression, it is necessary to choose one dependent variable y and one independent variable x that fits into the nonlinear expression of the function A + B * X, where A and B represent the set of all positive and negative numbers.

For the analysis of results in 2020. and 2021., the following were selected:

- For x (independent variable) - number of deaths for a given date
- For y (dependent variable) - month of the year

Based on the data for 2020., i.e. defined columns for x and y, using nonlinear regression, a graphical representation was created showing the flow of deaths in relation to the months of the year, which can be seen in Fig. 12. The graph clearly shows that nonlinear regression predicts an increase in deaths with over 90% certainty, up to 40 people per day, if the same growth trend continues until the end of 2020.

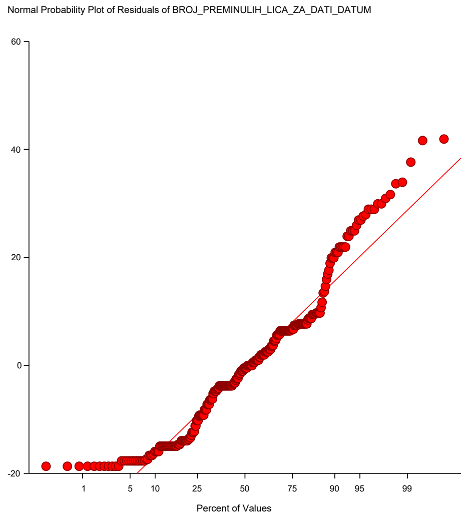


***Figure 12.** The result of the analysis achieved by applying nonlinear regression for 2020.*

Based on the data for 2021., i.e. defined columns for x and y, using nonlinear regression, a graphical representation was created showing the flow of deaths in relation to the months of the year, which can be seen in Fig. 13. From the graph it is clear see that nonlinear regression predicts a slight increase in deaths with a certainty of over 90%, in values of up to 20 people per day, if the same growth trend continues until the end of 2021. The reduced number of deaths can also be related to the fact that the Government of the Republic of Serbia has issued measures and a vaccination plan to reduce the spread of Covid 19, which will cause fewer deaths in the future.

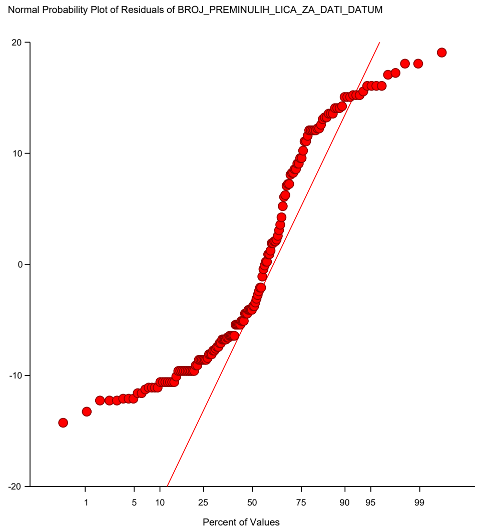


***Figure 13**. The result of the analysis achieved by applying nonlinear regression for 2021.*

7. CONCLUSION

The result of this research is the analysis of data using regression methods with the NCSS software tool, in order to collect and predict significant information related to the current epidemiological situation in the Republic of Serbia regarding the Covid virus 19.

In comparison with other related research, presented in this paper, the research results of this paper are partially similar, but there are also certain differences in which the scientific contribution itself is reflected. When it comes to linear regression, it is basically the same for all cases and the only difference would be that in this paper the data set on which the analysis was performed was wider over time and thus got a broader picture of further forecasts of virus development. When it comes to multiple linear regression, related research gives the impression that the parameters for analysis were chosen randomly, while in this paper it was taken into account that the parameters (predictor variables) are strictly correlated to make the accuracy of the multiple linear model much better. By selecting the best correlated predictor variables, the accuracy of the multiple linear model is achieved. Also, this research has a nonlinear regression model that can be identified with a polynomial regression model. In related research, the polynomial regression model was applied for a significantly shorter period of time, while in this paper, the time period of the data set is significantly longer and more objective for the forecast.

The data mining technique of regression, with its set of methods, enabled the timely prediction of trends in the further course of the development of the epidemiological situation in the Republic of Serbia. The research itself showed that the trend of the epidemic is decreasing, i.e. that preventive measures as well as the process of vaccination and revaccination have had an effect in the fight against the Covid 19 virus.

Further development of the research would be reflected in the continuation of monitoring the current epidemiological situation, and continuous data collection and comparison with the previous one, in order to predict for a certain period of time which segments of information are crucial for health in the Republic of Serbia.

REFERENCES

[1] Agha, S. & Haider, A. (2014). An Introduction to Data Mining Technique. *International Journal of Advancement in Engineering Technology, Management & Applied Science*, 1(3), 66.

[2] Principal Component Analysis (PCA). (2020, October 24). Towards AI. Retrieved April 5, 2020, from https://towardsai.net/p/data-mining/principal-component-analysis-pca

[3] Olson, D. L. (2007). Data mining in business services*. Service Business*, 1(3), 181-193, doi:10.1007/s11628-006-0014-7

[4] Researchgate.net Retrieved April 10, 2022, from https://www.researchgate.net/figure/Main-data-mining-techniques_fig2_270552309

[5] Golberg, M., Cho, H. (2010). *Introduction to Regression Analysis*. Department of Mathematical Sciences, University of Nevada, Las Vegas, USA, 1.

[6] Halili, F., Rustemi, A. (2016). *Predictive Modeling: Data Mining Regression Technique Applied in a Prototype*, 212-213.

[7] Freedman, D.A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press, 26.

[8] Abazid, M. (2018). Least squares methods to forecast sales for a company. *International Journal of Scientific and Engineering Research*, 864.

[9] Verma, J.P. (2013). Data Analysis in Management with SPPS Software. *Springer*, 145-146, doi: 10.1007/978-81-322-0786-3.

[10] Bethea, R. M., Duran, B. S. and Boullion, T. L. (1985) Statistical methods for engineers and scientists: Second edition, revised and expanded. New York, NY: Marcel Dekker.*, 48(9), 351.*

[11] Lawson, K., Buncher, D. (2014, December 2). *Statistical software*. Ncss.com. Retrieved April 22, 2022, from https://www.ncss.com/

[12] Office for Information Technologies and Electronic Administration (2020) *Covid-19 Daily report on the epidemiological situation in the Republic of Serbia* [Data set]. Retrieved May 5, 2022, from https://data.gov.rs/sr/datasets/covid-19-dnevni-izveshtaj-o-epidemioloshkoj-situatsiji-u-republitsi-srbiji

[13] Bhaya, W.S., S.A. Alasadi (2017). Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, 12(16), 4102-4107, doi:10.36478/jeasci.2017.4102.4107