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/TIE22.251R

Determining the number of doctoral students in the Republic of Serbia using regression algorithm

Milica Radenković
University of Kragujevac, Faculty of Technical Sciences Čačak, Čačak, Serbia
* milicar298@gmail.com

Abstract: *The term data mining itself implies mining, i.e. the process of sorting, organizing or grouping a large amount of data, which enables the extraction of relevant information. More precisely, data mining leads to flexibility in data, discovery of relationships, regularity, legality and other structures where data can be organized into databases or can be textual, unstructured, derived from the Internet or data organized into time series. A significant change was made in the Bologna process with the introduction of doctoral studies, whose primary goal was to realize the link between education and research. As doctoral studies represent an important level of education, this paper is based on determining gender differences as determinants of the number of doctoral students in the Republic of Serbia. After downloading and installing the NCSS software tool, the downloading, transformation and preprocessing of data originating from the open data portal began. The result of this research is the analysis of data through regression methods, where the given regression mining technique with its set of methods made it possible to predict trends in gender differences as determinants of the number of doctors of science in the Republic of Serbia. The conducted research opens many possibilities for further research.*

Keywords: *mining; regression; doctoral students in the Republic of Serbia; gender.*

1. INTRODUCTION

In addition to the two already existing levels of higher education, undergraduate and graduate studies, the Bologna Process also introduced a third level-doctoral studies. It represented a significant change whose primary goal was to realize the link between education and research. In order to develop a society based on knowledge, the need for creative workers, who will meet in the future meet the changed requirements of all sectors of the economy and society as a whole, is becoming increasingly important in European countries. As the necessary skills are acquired primarily through experience in research, it was necessary to develop a new concept of doctoral studies, the basis of which would be the acquisition of professional experience through the management of original research projects in a high-quality scientific environment. In this way, this level of higher education is significantly and qualitatively different from the two firstly introduced levels. In this way, instead of training for research work, the university implements training through research work. Therefore, the new concept of doctoral studies basically views the doctorate as a professional experience gained on research projects, and the training during doctoral studies represents a whole.

2. APPLICATION OF THE REGRESSION ALGORITHM THROUGH RELATED RESEARCH

Namely, the idea of conducting research on the number of doctoral students in the Republic of Serbia, using the statistical regression method, arose from the very fact of the existence of a large number of conducted research on the same topic, as well as the relevance of the education system of the Republic of Serbia and the application of data mining. During the research by Vilotić and colleagues on the success of doctoral academic studies at the Faculty of Technical Sciences in Novi Sad, it was determined that up to year 2015 / 2016 (including that academic year), 1141 students were enrolled at doctoral academic studies, and only 146 candidates, i.e. 16, 33% gained the PhD degree. The best result in terms of the number of students, who completed doctoral studies, indicates that they were achieved in the interval of study length from 8 to 10 years. The assumption was that the duration of studies is affected by the workload, which refers to students of doctoral academic studies employed at the Faculty of Technical Sciences, which was denied pointing out the conclusion that the workload does not affect the duration of studies. Also, the assumption about the influence of the average grade of students of doctoral academic studies from previous studies is not significant for success [1].

Ojerinde et al.'s research proposes a framework for administering the prediction of student academic performance using learning analytic techniques. The research illustrates how this model is effectively used on secondary data collected from the Department of Computer Science, University of Jos, Nigeria. The study succeeded in achieving the goal of building a model to predict student academic performance. Analysis has shown that students who have good results in mathematic subjects have a higher chance of achieving excellence in other computer science subjects [2].

Mustapasha sought to conduct a more detailed investigation using web mining techniques to examine two well-known approaches to improving student success – the index of learning styles and learning strategies, in terms of self-confidence, learning aids, motivation, self-motivation, attitude, and the like. The results of linear regression indicated that the scales for selecting main topics, learning aids, self-motivation, self-confidence, goal, self-learning and information processing were significant with a relevance value of less than 0.05 [3].

Since doctoral studies represent an important level of education, this paper is based on gender differences as determinants of the number of doctoral students in the Republic of Serbia.

The development of modern technique and technology indicates an increasing use of digital tools, both in education and in other areas. In today's business, most data resides on the web. So all technologies related to data processing are very important for their success. Because of that, web mining represents an extraordinary technique for gathering useful information from the web, and precisely for the realization of research on gender differences as a determinant of the number of doctoral students in the Republic of Serbia.

3. BASICS OF WEB MINING

The term mining itself means mining. Data mining is the process of sorting, organizing or grouping large amounts of data and extracting relevant information. We could also define data mining as finding patterns in data. These data can be organized into databases, but they can also be textual data, unstructured data from the Internet, or data organized into time series. Data mining leads to logicality in the data, that is, the discovery of relationships, regularities and other structures among the data.

Web mining is also the use of techniques for extracting useful information from the web. Web data refers to web content (text, images, etc.), web structure (links), and web usage (http logs, server logs, etc.).

3.1. Categories of web mining

As already stated in the introductory part, there are three categories of web mining and they are as follows:

- Web content mining – This is the process of handling useful information from the content of web pages and web documents, which are mainly text, images and audio / video files. The techniques used in this discipline were drawn heavily from natural language processing (NLP) and information retrieval. It is mainly performed through a web browser, with the help of a web crawler and is used for the purpose of indexing websites.
- Web structure mining – This is the process of analyzing the nodes and structure of a web page through the use of graph theory. There are two things that can be gained from this: the structure of the website in terms of how it is linked to other sites, while the second is the structure of the website itself, that is, how each page within the site is linked.
- Web usage mining – This is the process of removing form and information from server logs to see user activities, including where users are from, how many clicks there are on which items on the site, and the types of activities performed on the site, [4].

4. STATISTICAL REGRESSION MODEL

In a large number of researches, one of the goals is to describe the connections between the phenomena that surround us. This can be achieved by finding a formula or equation that relates the quantities we observe. In statistics, regression analysis deals with finding statistical connections between phenomena. Regression is of great importance, both in economics and business, and in other natural sciences, such as: chemistry, physics, biology, pharmacology, toxicology, biochemistry and forensic medicine and so on, [5].

Thus, regression analyzes the relationships between variables. Regression is a data mining technique used to predict a range of numerical values (also called continuous values), given a particular set of data. For example, regression can be used to predict the cost of products or services, given other variables. Regression is used in many industries for business and marketing planning, financial forecasting, environmental modeling, and trend analysis.

Regression is one of the methods within the methods that make up statistical learning - a large set of methods, techniques, statistics tools for modeling and understanding complex data sets. Regression is one of the ways to build a model for predicting and evaluating one or more dependent variables based on one or more independent variables. In regression, there is an output, unlike other statistical techniques that deal with problems in which there is no dependent variable [6].

5. RESEARCH METHODOLOGY

5.1. Data analysis software tool "NCSS"

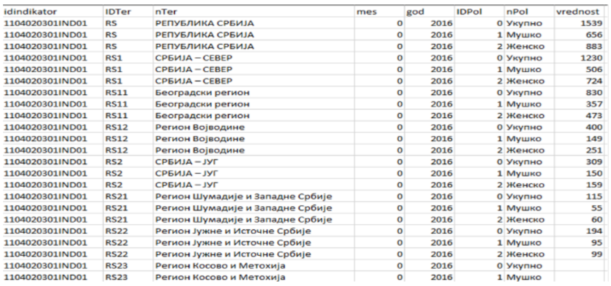
The software provides a complete and easy-to-use collection of hundreds of statistical and graphical tools for analyzing and visualizing your data. This data analysis software comes in complete with integrated documentation, free training videos, and full phone and email support from a team of PhD statisticians. With a few simple steps, meaningful numerical results and clean, clear graphics can be obtained. In order to perform the necessary data analysis, it is necessary to download the software tool from the web address: Free Trial | NCSS Statistical Software | NCSS.com. This is followed by program installation and data download.

5.2. Data for analysis

The data that will be analyzed in this paper were taken from the open data portal, provided by the government of the Republic of Serbia, from the following web address: Број доктора наука по полу - студије III степена - Отворени подаци (data.gov.rs).

5.3. Data transformation and preprocessing

After the successful collection of data in .xlsx format, the next step related to data preprocessing and transformation was approached. Therefore, the data in its original form looks like in Figure 1. Data preprocessing represents one of the most important tasks of data mining and includes the preparation and transformation of data into a suitable template, which is suitable for research methods. The mentioned preprocessing activity tends to reducing the amount of data, finding connections between data, normalization, as well as eliminating redundancies and extracting new data. Techniques such as cleaning, integration, transformation and removal of redundant data are included in the preprocess After preprocessing the data, the downloaded set must correspond to a template that is suitable for research, so that the appropriate method, the regression method, can be applied, which is the case in this paper.

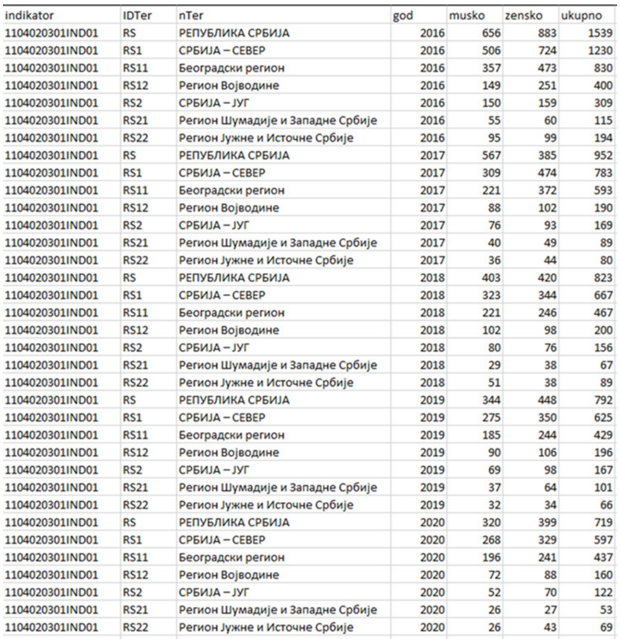


idindikator	IDTer	nTer	mes	god	IDPol	nPol	vrednost
1104020301IND01	RS	РЕПУБЛИКА СРБИЈА	0	2016	0	Укупно	1539
1104020301IND01	RS	РЕПУБЛИКА СРБИЈА	0	2016	1	Мушко	656
1104020301IND01	RS	РЕПУБЛИКА СРБИЈА	0	2016	2	Женско	883
1104020301IND01	RS1	СРБИЈА – СЕВЕР	0	2016	0	Укупно	1230
1104020301IND01	RS1	СРБИЈА – СЕВЕР	0	2016	1	Мушко	506
1104020301IND01	RS1	СРБИЈА – СЕВЕР	0	2016	2	Женско	724
1104020301IND01	RS11	Београдски регион	0	2016	0	Укупно	830
1104020301IND01	RS11	Београдски регион	0	2016	1	Мушко	357
1104020301IND01	RS11	Београдски регион	0	2016	2	Женско	473
1104020301IND01	RS12	Регион Војводине	0	2016	0	Укупно	400
1104020301IND01	RS12	Регион Војводине	0	2016	1	Мушко	149
1104020301IND01	RS12	Регион Војводине	0	2016	2	Женско	251
1104020301IND01	RS2	СРБИЈА – ЈУГ	0	2016	0	Укупно	309
1104020301IND01	RS2	СРБИЈА – ЈУГ	0	2016	1	Мушко	150
1104020301IND01	RS2	СРБИЈА – ЈУГ	0	2016	2	Женско	159
1104020301IND01	RS21	Регион Шумадије и Западне Србије	0	2016	0	Укупно	115
1104020301IND01	RS21	Регион Шумадије и Западне Србије	0	2016	1	Мушко	55
1104020301IND01	RS21	Регион Шумадије и Западне Србије	0	2016	2	Женско	60
1104020301IND01	RS22	Регион Јужне и Источне Србије	0	2016	0	Укупно	194
1104020301IND01	RS22	Регион Јужне и Источне Србије	0	2016	1	Мушко	95
1104020301IND01	RS22	Регион Јужне и Источне Србије	0	2016	2	Женско	99
1104020301IND01	RS23	Регион Косово и Метохија	0	2016	0	Укупно	
1104020301IND01	RS23	Регион Косово и Метохија	0	2016	1	Мушко	

Figure 1. *Unprocessed Data*

The original data set contained many rows, there were empty cells, which did not ensure accuracy during data analysis, and they were transformed. Certain columns from the above data set have been removed because they are not relevant for this type of analysis. Also, special columns have been created that enable better quality results because they leave the possibility of data analysis according to more dependent and independent variables (x, y).

After data preprocessing, a form of data is obtained that is suitable for analysis using regression methods, which can be seen in Figure 2.



indikator	IDTer	nTer	god	musko	zensko	ukupno
1104020301IND01	RS	РЕПУБЛИКА СРБИЈА	2016	656	883	1539
1104020301IND01	RS1	СРБИЈА – СЕВЕР	2016	506	724	1230
1104020301IND01	RS11	Београдски регион	2016	357	473	830
1104020301IND01	RS12	Регион Војводине	2016	149	251	400
1104020301IND01	RS2	СРБИЈА – ЈУГ	2016	150	159	309
1104020301IND01	RS21	Регион Шумадије и Западне Србије	2016	55	60	115
1104020301IND01	RS22	Регион Јужне и Источне Србије	2016	95	99	194
1104020301IND01	RS	РЕПУБЛИКА СРБИЈА	2017	567	385	952
1104020301IND01	RS1	СРБИЈА – СЕВЕР	2017	309	474	783
1104020301IND01	RS11	Београдски регион	2017	221	372	593
1104020301IND01	RS12	Регион Војводине	2017	88	102	190
1104020301IND01	RS2	СРБИЈА – ЈУГ	2017	76	93	169
1104020301IND01	RS21	Регион Шумадије и Западне Србије	2017	40	49	89
1104020301IND01	RS22	Регион Јужне и Источне Србије	2017	36	44	80
1104020301IND01	RS	РЕПУБЛИКА СРБИЈА	2018	403	420	823
1104020301IND01	RS1	СРБИЈА – СЕВЕР	2018	323	344	667
1104020301IND01	RS11	Београдски регион	2018	221	246	467
1104020301IND01	RS12	Регион Војводине	2018	102	98	200
1104020301IND01	RS2	СРБИЈА – ЈУГ	2018	80	76	156
1104020301IND01	RS21	Регион Шумадије и Западне Србије	2018	29	38	67
1104020301IND01	RS22	Регион Јужне и Источне Србије	2018	51	38	89
1104020301IND01	RS	РЕПУБЛИКА СРБИЈА	2019	344	448	792
1104020301IND01	RS1	СРБИЈА – СЕВЕР	2019	275	350	625
1104020301IND01	RS11	Београдски регион	2019	185	244	429
1104020301IND01	RS12	Регион Војводине	2019	90	106	196
1104020301IND01	RS2	СРБИЈА – ЈУГ	2019	69	98	167
1104020301IND01	RS21	Регион Шумадије и Западне Србије	2019	37	64	101
1104020301IND01	RS22	Регион Јужне и Источне Србије	2019	32	34	66
1104020301IND01	RS	РЕПУБЛИКА СРБИЈА	2020	320	399	719
1104020301IND01	RS1	СРБИЈА – СЕВЕР	2020	268	329	597
1104020301IND01	RS11	Београдски регион	2020	196	241	437
1104020301IND01	RS12	Регион Војводине	2020	72	88	160
1104020301IND01	RS2	СРБИЈА – ЈУГ	2020	52	70	122
1104020301IND01	RS21	Регион Шумадије и Западне Србије	2020	26	27	53
1104020301IND01	RS22	Регион Јужне и Источне Србије	2020	26	43	69

Figure 2. *Transformed data*

6. RESULTS

6.1. Results and discussion using simple linear regression

After the imported data and applied filters, a simple linear regression was applied. In order to conduct a simple linear regression, it was necessary to choose two dependent and one independent variable, namely h(year) and u(male and female).

Based on the data on the number of completed doctoral studies, taking into account only the female gender, i.e. the defined columns for X and Y, using simple linear regression, a graphic representation was created that shows the flow of completed doctoral studies of female individuals in relation to the years from 2016 to 2020, which can be seen in Figure 3. The graph clearly indicates that a simple linear regression predicts a lower number of female PhD graduates starting in 2016 and ending in 2020.

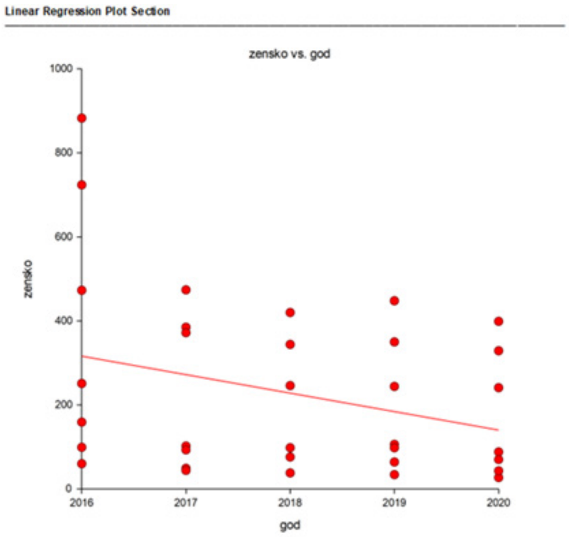


Figure 3. *The results of the analysis achieved by applying a simple linear regression for the females*

Then, a further analysis for the male sex was undertaken. The graph clearly indicates that a simple linear regression predicts a lower number of male PhD graduates starting in 2016 and ending in 2020, which can be seen in Figure 4.

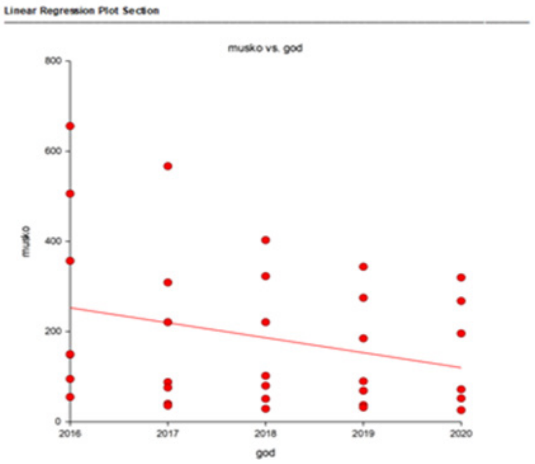


Figure 4. *The results of the analysis achieved by applying a simple linear regression for male*

6.2. Results and discussion using multiple linear regression

By applying multiple linear regression, a graphic representation was created that shows the flow of completed doctoral studies of individuals of both sexes as well as the total number (Y) in relation to the years from 2016 to 2020 (X), which can be seen in Figure 4. The graphic clearly indicates that the multiple linear regression indicates a decrease in the number of male and female individuals, as well as the total number, who end up doctoral studies, starting from 2016 until 2020.

Furthermore, the analysis was started only for individuals of the female sex, i.e. defined columns X and Y, using multiple linear regression, based on which a graphic display was created that shows the flow of the mentioned data in given period, which can be seen in Figure number 5.

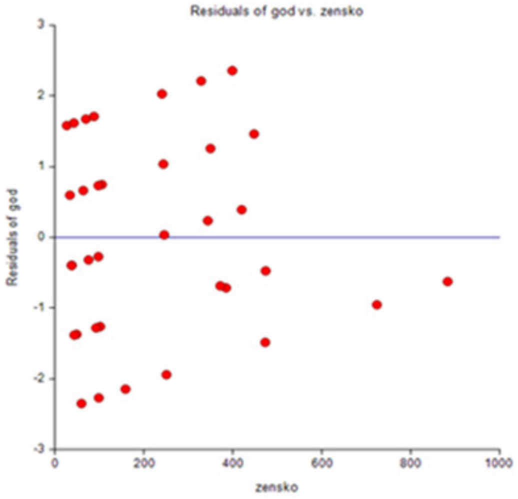


Figure 5. *The result of the analysis achieved by applying multiple linear regression for females*

After the analyzed data for the female persons, further analysis is done only for the male persons, that is, the defined columns X and Y, by applying multiple linear regression, on the basis of which a graphic representation was created that shows the flow of the mentioned data in a given period, which can be seen in Figure number 6.

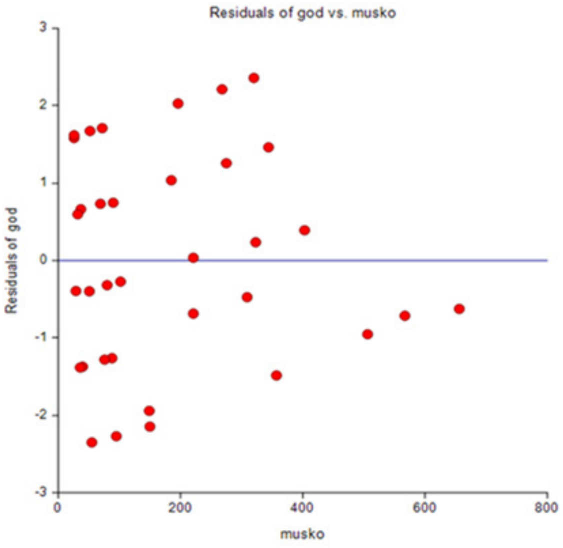


Figure 6. *The result of the analysis achieved by applying multiple linear regression for male*

After analyzing the data for men, a further analysis is made only for men and women, i.e. defined columns X and Y, using multiple linear regression, on the basis of which a graphic display was created that shows the flow of the mentioned data in a given period, i.e. forecasts an decrease in the number of male and female individuals who have completed doctoral studies, which can be seen in Figure number 7.

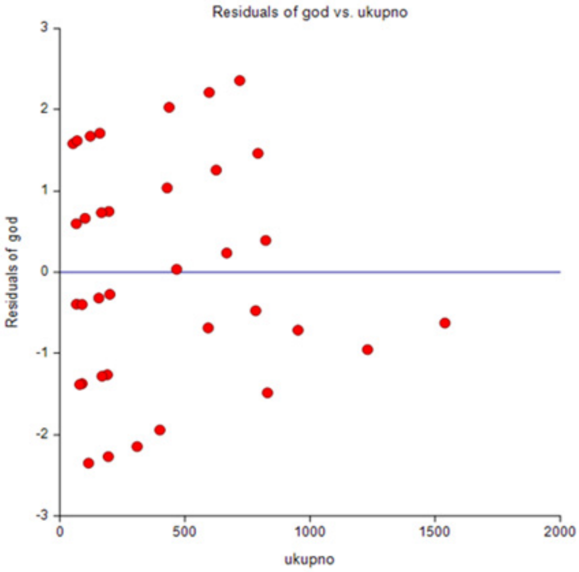


Figure 7. *The result of the analysis achieved by applying multiple linear regression for male and females*

6.3. Results and discussion using non-linear regression

Based on the data for the given period from 2016 to 2020, i.e. defined columns for X (year) and Y (female), using non-linear regression, a graphic display was created that shows the trend of the number of female individuals who completed doctoral studies in relation to the already mentioned period. Non-linear regression predicts an decrease in the number of female individuals completing doctoral studies, with 99% certainty, which is shown in Figure 8.

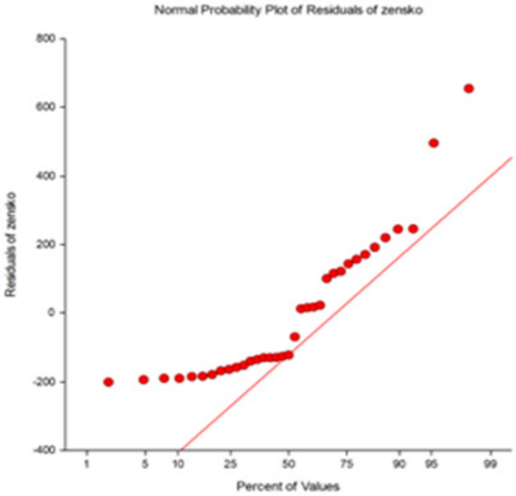


Figure 8. *The result of the analysis achieved by the application of non-linear regression for female*

While, the non-linear regression predicts a decrease in the number of male individuals completing doctoral studies, with 99% certainty, which can be seen in Figure 9.

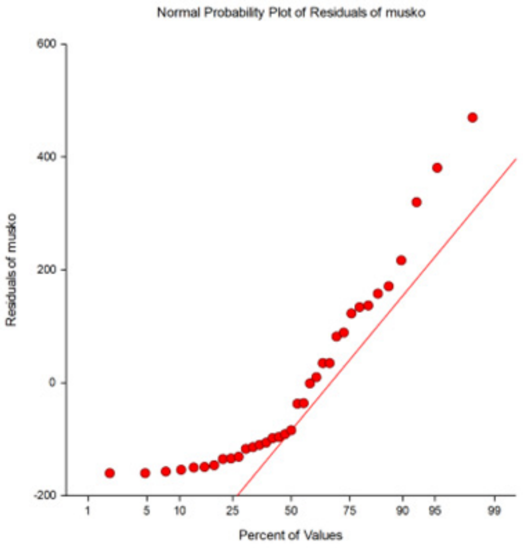


Figure 9. *The result of the analysis achieved by applying non-linear regression for male*

6.4. Correlation of results with related research

The results of the first related research paper indicate that by the end of the 2015/2016 academic year, only 16.33% of the total number of students enrolled in the third degree of academic studies at the Faculty of Technical Sciences in Novi Sad had completed their doctoral studies. It can be said that the number of doctoral students at the level of the Republic of Serbia is decreasing in the incoming period, as we can see in this paper, which talks about gender differences as a determinant of the number of doctoral students, but also about their total number. Analysis of the results indicates a decrease in the number of male and female persons, i.e. of the total number of PhDs from 2016 to 2020.

As in the research work of Ojerdindej and his associates, the multiple linear regression technique was used in this work, but not in combination with the SPSS software tool, but with the NCSS software tool. Their research illustrates how the multiple linear regression model is effectively used on secondary data and the study was able to achieve the objective of building a model for predicting students' academic performance. While in this paper the application of the multiple linear regression clearly predicted the decrease in the number of male and female doctoral students, that is, their performance.

Reviewing the research, it can be seen that the students had ten different combinations of good aspects, eight of which are on the strategy scale which is good. This can be correlated with the results of this research, which using simple and complex linear regression, as well as non-linear regression, shows an decrease in the number of PhDs from 2016 to 2020, so students should apply better aspects for learning.

7. DISCUSSION

In the modern conditions of life and work, education represents an important stage in the life of every individual, therefore it is important to promote its advancement through constant research. Advances in technique and technology have contributed to a more comprehensive application of various research opportunities and the use of various techniques, such as data mining techniques. The advantage of the research paper "Determining the number of doctoral students in the Republic of Serbia by regression algorithm", in relation to the aforementioned related research, is the application of web mining techniques, more precisely statistical methods of simple and multiple linear regression, as well as non-linear regression, to data collected from the portal, which it was not applied only to the total number of persons who completed doctoral studies. The mentioned techniques were applied especially for men, especially for women, as well as for the total number of persons who became doctors of science in the given period. The obtained results unequivocally showed that more effective modernization and more intensive application of regression methods and more realistic planning and programming of possible research can be achieved only if a sufficient amount of objective information is available on the basis of which the current situation can be diagnosed and procedures for further work can be determined. Of course, this research should initiate further, more complex and broader research, with a larger number of respondents, in a wider area, which will lead to more efficient data transformations.

8. CONCLUSION

The result of this research is data analysis through regression methods with the use of the NCSS software tool, in order to collect and predict significant information related to gender differences as determinants of the number of PhDs in the Republic of Serbia.

The data mining regression technique with its set of methods made it possible to predict in time the trends of gender differences as determinants of the number of PhDs in the Republic of Serbia, in relation to male and female gender, as well as the total number of individuals completing doctoral studies. From the research itself, an decrease in finishing doctoral studies was established when it comes to the female gender, while it is seen that the male gender is not lagging behind either. The further development of the research would be reflected in the continuation of monitoring gender differences in the upcoming period and the continuous collection of data and comparison with the previous ones, so that during a certain period of time it would be predicted which segments of information are circular for education in the Republic of Serbia.

REFERENCES

[1] D. Vilotic, R. Doroslovacki, I. Kovacevic, V. Katic, D. Seslija, S. Kolakovic, Z. Konjovic (2017). *Analysis of the success of doctoral academic studies at the Faculty of Technical Sciences, Novi Sad.* Faculty of Technical Sciences.

[2] O.D. Oyerinde, P.A. Chia, Predicting Students' Academic Performances. *A Learning Analytics Approach using Multiple Linear Regression, Nigeria*. University of Jos.

[3] O. Mustapaúa, A. Karahocaa, D. Karahoca, H.Uzunboylu (2010). *"Hello World", Web Mining for E-Learning, Istanbul-Turkey*. University of Bahçeúehir.

[4] Milica Radenkovic, unpublished master's thesis*. Application of web mining techniques in the analysis of links, logos, text and opinions*. Faculty of Technical Sciences in Cacak

[5] K. A. F. Copeland (1997). *Applied Linear Statistical Models. New York.* McGraw-Hill/Irwin.

[6] D. H. Barlow, V. M. Durand, S. G. Hofmann (2018*). Essentials of Modern Business Statistics with Microsoft Office Excel, Boston*. Cengage Learning.