ANALYSIS OF STUDENTS' PERFORMANCES ON MATHEMATICS EXAM USING MULTIPLE LINEAR REGRESSION

Ana Anokić[1] Svjetlana Janković-Šoja[2] Tatjana Keča[3]

Abstract: The aim of this study is to explore the impact of different factors, such as: students' activity, colloquium result and time until the exam is passed, on students' success on Mathematics exam. Multiple regression and Chi-square test of independence are applied to the real-life data based on teacher's notes. Two groups of students, different in terms of: group size, trimester in which they attend classes, their performances and average success in the subject, are considered. Additional statistical tests are performed to confirm assumptions for correct application of the proposed methodology. The linear regression models are obtained and adequate conclusions are drawn.

Key words: Education, Mathematics, Statistics, Multiple linear regression, Chi-square test of independence

1. INTRODUCTION

The majority of students that have attended higher educational institutions (HEIs) in the last three years were in high schools during the pandemic of Covid 19 virus. In the Republic of Serbia, state of emergency due to the pandemic was declared on March, 15th 2020. Almost immediately, traditional classes were replaced by online lessons. A lot of effort and significant time were required to provide acceptable models of teaching with adequate tools, materials, estimation of students' workload and examination methods for each subject. In addition, different models were tested and changed several times, depending on the current epidemiological situation. The same challenges in education appeared worldwide at the same time. Study [1], that includes 438 university members in Bangladesh, with different perspectives depending on their age, experience and academic discipline, pointed out the most important challenges of online teaching: difficulty in practical work, monitoring students and insufficient feedback. In addition, the analysis showed that during the pandemic 75% of the teachers preferred online lessons, while only 10% of participants preferred this teaching model after pandemic.

Online learning had many negative psycho-educational impacts on students' motivation, stress level and well-being [2]. However, many high school students developed solid skills for self-learning using technology with a small participation or even without direct interaction and collaborative work with immediate feedback, that are essential benefits of traditional teaching in classrooms. Some of these students have preserved the same habits at their HEIs. They demonstrate poor visible interest and activity in-person, their communication is very brief and infrequent and, despite of constant encouragement, they rarely use help of teaching stuff in the form of consultation or insight in the exam papers.

Many different scenarios of students' performances in their activity and success in completing their pre-exam and final-exam obligations can be noticed in large groups, particularly in the first year of study. Students can be visibly inactive in class but with good exam results, active in presence but with poor result on the final exam, entering studies with modest background knowledge but with excellent willpower, dropping out their studies at the very beginning, etc. We believe that one of the main goals in contemporary education process is to create stimulating environment, that will encourage students to experience the benefits of in-person communication in classroom. The focus of this study is to explore the impact of different factors to the success in the final exam. The factors that can be relevant are: students' activity in class, results on colloquiums and the time until the exam is passed. Appropriate statistical methods are applied to the real-life data, based on teacher's detailed records during trimester on *Mathematics*, the mandatory subject for the first-year students of the two

[1]Senior lecturer, Department School of Information and Communication Technologies, Academy of Technical and Art Applied Studies Belgrade, Zdravka Čelara St, 16, Belgrade, email: ana.anokic@ict.edu.rs
[2]Associate professor, Faculty of Agriculture, University of Belgrade, Nemanjina St 6, Belgrade, email: svjetlanajs@agrif.bg.ac.rs
[3]Professor of applied studies, Department School of Information and Communication Technologies, Academy of Technical and Art Applied Studies Belgrade, Zdravka Čelara St, 16, Belgrade, email: tatjana.keca@ict.edu.rs

largest study programs of a HEI in Belgrade. The observed period is from January 2022. to October 2024.

The remainder of the paper is organized as follows. The applied statistical methodology is described in Section 2, while computational results with the corresponding interpretations are presented in Section 3. The conclusion remarks are given in Section 4.

2. STATISTICAL METODOLOGY

Two main approaches in statistical analysis are descriptive and inferential statistics. The former is used for summarizing and describing the dataset in terms of its central tendencies, variability and shape, and the latter addresses to drawing conclusions based on the observed data. Measures of descriptive statistics used in this study are well known and frequently applied. Therefore, their description is omitted. We outline the main principles of the inferential statistics methods used in this data analysis.

Multiple regression is a powerful method for estimation the relation between two or more independent and one dependent variable. When this relation is linear, then dependent variable y and the independent variables $x_1, x_2, \dots, x_k$ are connected by formula:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon, \quad (1)$$

which is known as multiple linear regression model [3], [4] with a constant term β_0, slope coefficients β_i for each independent variable x_i $(i \in \{1,2,\dots,k\})$, and model error's term ε, known as residual or random error. The constant term β_0 represents the value of the dependent variable y when all independent variables are equal to 0, while β_i for $i \in \{1,2,\dots,k\}$ are equal to the average change in y-value when the corresponding variable x_i increases for one unit, leaving the remaining independent variables unchanged. Coefficients β_i for $i \in \{0,1,2,\dots,k\}$ are known as regression parameters.

There are certain assumptions related to random error and independent variables that must be satisfied in every linear regression model [4], [5]. Otherwise, estimates of regression parameters will not have desirable properties. The random error needs to follow the normal distribution and it should not be auto correlated. Independent variables must not be mutually correlated. If these conditions are not fulfilled, the corresponding model faces the following problems: heteroscedasticity (the variance of random error is not constant, it is different for each observation), autocorrelation (there is a correlation between random errors of different observations) and multicollinearity (independent variables are linearly dependent). A regression model, that meets above assumptions, can be used for approximation of the relation between variables and prediction of dependent variable based on the values of independent ones. A useful guide on multiple regression analysis with applications can be found in [6].

Another inferential statistical method is *Chi-square test of independence*. This test is used to determine the independence between two variables, X and Y, that are considered to be independent if their joined probability is equal to the product of the marginal probabilities:

$P(X = i, Y = j) = P(X = i)P(Y = j)$, for each pair of $i \in \{1,2,\dots,r\}$ and $j \in \{1,2,\dots,c\}$.

This means that behavior of these variables does not influence each other. The Chi-square test is used to confirm or reject the null hypothesis H_0 about independence between two variables [3], which is tested against the alternative hypothesis H_A. More precisely, H_0 and H_A are defined as:

H_0: the independence model is true i.e. $P(X = i, Y = j) = P(X = i)P(Y = j)$ for all pairs of (i, j),

H_A: $P(X = i, Y = j) \neq P(X = i)P(Y = j)$ for at least one pair of (i, j).

In particular, this test is applied to determine the independence of the arrangement of observation units within a set according to two variables (X and Y) with two or more modalities [7]. Computationally, the first step consists of calculating the expected frequencies, denoted by n_{ij}^*, of the observed variables X and Y under the assumption that they are independent. The test statistics, the Chi-square value, is

obtained based on the differences between n_{ij}^* and the corresponding empirical frequencies n_{ij} using the following formula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*} \quad (2)$$

The obtained value is then compared by critical values of the Chi-square distribution $\chi^2_{\alpha,\nu}$.

Analysis of variance (ANOVA) is a statistical test usually applied to compare the means of more than two data sets in basic statistics. It is based on measuring the difference between datasets using their mean values and variances, resulting in F-ratio, that is compared by *Fisher* distribution [8]. In addition, this method can be used to test the significance of regression coefficients and multiple correlation coefficient in multiple regression analysis.

The *Durbin-Watson* statistic is a test for detecting the presence of autocorrelation of residuals from a regression analysis [9]. The value of this statistic ranges between 0 and 4, where 2 indicates there is no autocorrelation, less than 2 means positive and greater than 2 indicates negative autocorrelation.

3. COMPUTATIONAL RESULTS

The described statistical methods are applied on the data related to students' performances on subject *Mathematics* in the *Department School of Information and Communication Technologies* within the *Academy of Technical and Art Applied Studies Belgrade* from January 2022. to October 2024. The IBM software *Statistical Package for the Social Sciences* (SPSS) version 20 is used for analysis. SPSS is a comprehensive system used for descriptive statistics and variety of complex statistical tests.

3.1. Data

The two study programs, *Internet technologies* (IT) and *Communication technologies* (CT), include *Mathematics* as a mandatory subject in the first year of study. As these groups are large, CT and IT students attend the subject separately in different trimesters following the same curriculum and the same evaluation methods. Statistical analysis is performed on the detailed data from teacher's notes about students' activity during trimesters, their colloquium results, the order number of examination period in which the exam is passed and the exam result. Only data related to students who passed the exam during the observed period (by October 2024), resulting in total of $n_1 = 271$ IT students from the last three generations (2021-2023) and $n_2 = 75$ CT students from the last two generations (2022 and 2023), are considered. The independent variables used in the regression model are:

- x_1: the students total activity calculated based on teacher's notes on presence, demonstrating interest and participating in class discussion, expressed in percentage, i.e. $x_1 \in [0, 100]\%$,
- x_2: the total number of points achieved through two colloquiums, $x_2 \in [0, 40]$,
- x_3: the order number of examination period in which student passes the exam, $x_3 \in \{1,2, \dots ,15\}$.

The dependent variable is defined as:

- y: the number of points achieved in the exam $y \in [11, 60]$.

It can be expected that the final exam result is a product of students' activity during trimesters as well as their colloquium results, representing effort and engagement in the study process from the beginning and finally, the third factor important for preparing the exam is time, expressed as the order number of examination period in which the exam is passed.

3.2. Measures of decriptive statistics

To summarize the data for every defined variable, measures of descriptive statistics: minimum (MIN), average (MEAN) and maximum (MAX) value are used to describe central tendencies, while standard deviation (Std. Dev.) and coefficient of variation (CV) are calculated to measure the variability. Their values for both study programs, IT and CT, are shown in Table 1.

Table 1 – Measures of descriptive statistics for IT and CT

Variable	Measures (IT | CT)				
	MIN	MEAN	MAX	Std. Dev.	CV (%)
Activity x_1 (%)	0 | 0	41.18 | 55.33	100 | 100	29.63 | 23.95	71.95 | 43.29
Colloquium result x_2	0 | 0	30.82 | 29.47	40 | 40	7.33 | 7.12	23.78 | 24.16
The examination period x_3	1 | 1	1.47 | 2.28	9 | 11	0.94 | 2.12	63.94 | 92.98
Exam result y	13 | 12	31.31 | 26.03	60 | 54	10.32 | 8.26	32.96 | 31.73

It can be concluded from the results given in Table 1 that IT students perform better on the final exam compared to CT students as on average IT students achieve 5 points more and some of them, unlike CT students, reach the maximum result of 60 points. However, the difference in colloquium results is very small between IT students and CT students. This could be a consequence of lower activity of IT students in class during trimesters, which is on average 14% less than the activity of CT students. The reasonable explanation lies in the fact that IT students attend the subject during the last trimester of the first year when they already managed to find all information and materials, while CT students study the subject in the first trimester, so their activity in class is mostly constant. In addition, due to the popularity of the study program, IT students have a better background knowledge on the subject, as their average results on the entrance exam in Mathematics is significantly better compared to CT students. Therefore, it is natural that they invest more time in activities on other subjects they consider as more challenging.

The largest variability of 92.98% can be noticed for variable x_3 corresponding to the examination period for CT students that varies mostly due to the poor prior knowledge of certain number of students. This value is followed by the variability measure of activity of IT students (71.95%), while colloquium results show the highest consistency with CV values less than 25% for both study programs.

3.3. Multiple linear regression

For the correct application of linear regression analysis, it is necessary to verify the assumptions described in Section 2. Therefore, correlations between variables of both study programs, IT and CT, are tested. The corresponding *Pearson* coefficients (r), p-value and squared *Pearson* coefficients (r^2) are given in Table 2. Significant correlation at the 0.01 level is marked with **.

Table 2 – Correlation of variables for IT and CT

Variable	Coefficients	Variables for IT			Variables for CT		
		Exam result y	Activity x_1	Colloquium result x_2	Exam result y	Activity x_1	Colloquium result x_2
Activity x_1	r (p-value)	0,260** (0,000)	1	-	-0,085 (0,466)	1	-
	r^2	0,068	1	-	0,007	1	-
Colloquium result x_2	r (p-value)	0,251** (0,000)	0,301** (0,000)	1	-0,300** (0,009)	0,411** (0,000)	1
	r^2	0,063	0,091	1	0,090	0,169	1
The examination period x_3	r (p-value)	-0,052 (0,391)	-0,163** (0,007)	-0,504** (0,000)	0,089 (0,447)	-0,226 (0,051)	-0,461** (0,000)
	r^2	0,003	0,027	0,254	0,008	0,051	0,213

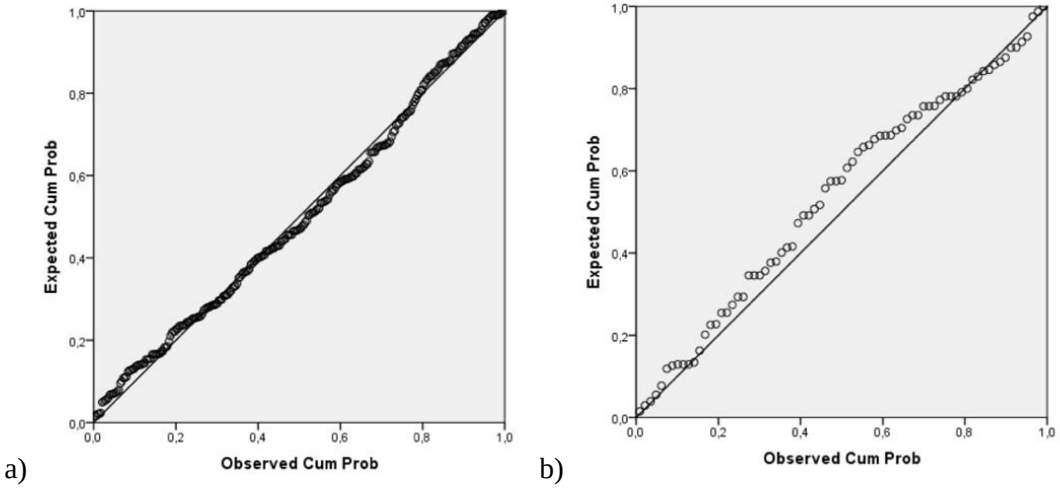
Based on values from Table 2, it can be concluded that for the study program IT, the variable y very significantly depends on variables x_1 and x_2, while for CT students it depends significantly only on x_2. Independent variables are mutually correlated, which can cause the problem of multicollinearity. However, according to [4], multicollinearity problem is present if $|r| > 0.7$ stands, and that is not the case here. In addition, the authors of [4] state that the multicollinearity problem also exists when the squared *Pearson* coefficient r^2 is greater than the coefficient of determination related

to multi-regression model R^2, that is equal to 0.906 and 0.854 for IT and CT, respectively (see below).

The estimated linear regression model for data related to IT is: $y = 0.074x_1 + 0.739x_2 + 3.433x_3$ with coefficient of determination $R^2 = 0.906$. The adjusted value of coefficient of determination $R^2 = 0.905$ indicates that 90.5% of variations of random variable y are the consequence of the independent variables x_1, x_2 and x_3, while 9.5% of them depend on some other factors that are not included in the model. The interpretation of the obtained regression coefficient $\beta_1 = 0.074$ is that the exam result increases on average for 0.074 points if students invest 1% more in class activities, $\beta_2 = 0.739$ indicates on average 0.739 points more in exam result when the colloquium result increases for 1 point, and $\beta_3 = 3.433$ is the average additional number of exam points if a student passes the exam in the next exam period. The regression coefficients are tested and they appear to be statistically significant as their p-values are less than 0.01. Based on their standardized values: $\beta_1' = 0.114$, $\beta_2' = 0.710$ and $\beta_3' = 0.182$, it can be concluded that the colloquium result (x_2) has the major impact on the exam result, while activity (x_1) has the lowest impact. In addition, ANOVA is applied to the regression model to test the hypothesis of the simultaneous equality of the regression parameters to zero, but also the hypothesis that the multiple correlation coefficient is statistically significant. All parameters and the multiple correlation coefficient are significant with p-value of the corresponding F-statistics less than 0.01.

The linear regression model for the data related to CT students is: $y = 0.681x_2 + 2.053x_3$ with $R^2 = 0.854$. Note that this model does not include the variable x_1 as the corresponding coefficient β_1 turned out to be not statistically significant, unlike the remaining coefficients β_2 and β_3, whose p-values are less than 0.01 in the applied tests. These coefficients have similar interpretation as in the previous model. Their standardized values $\beta_2' = 0.757$ and $\beta_3' = 0.234$ show that colloquium result has more impact than the period when student passes the exam. The adjusted value of the coefficient of determination $R^2 = 0.850$ implies that 85% of variations in exam result can be explained by colloquium results and the exam period when a student passes the exam, while 15% represents the impact of other factors that are not included in the model. ANOVA test confirms the significance of regression coefficients and the multiple correlation coefficients as well.

The diagrams in Figure 1 illustrate the normality of residuals of the obtained models using the cumulative distribution function (CDF). The observed CDF of the standardized residuals, represented by points, is compared to the expected CDF of the normal distribution (line) to confirm the normality. of residuals.



a) b)

Figure 1: P-P diagram of standardized residuals of estimated models for a) IT and b) CT

The *Durbin-Watson* test is applied to both models, resulting in values of the corresponding coefficients, 1.734 for IT and 1.881 for CT, which are between 1.7 and 2.3. According to [10] the sufficient condition for eliminating the possibility of autocorrelation is satisfied.

3.4. Chi-square test of independance

From descriptive statistic measures given in Table 1, it can be noticed that variable x_3 (the examination period when student passes the exam), shows the largest difference between IT and CT students. We examine the independence of distributions of students according to variable x_3 values, and the study program that students attend. To satisfy the main assumptions of the Chi-square test of independence, the three modalities of variable x_3 are determined. They classify students according to the period in which exam is passed into: the first examination period (I), the second (II), and more than two examination periods have passed before student passes the exam (III). The contingency table is represented as Table 3.

Table 3 – Contingency table

Study program	The examination period x_3			Total
	I	II	III	
IT	197	39	35	271
CT	46	11	18	75
Total	243	50	53	346

The obtained χ^2 value is equal to 5,80 which is less than the corresponding critical values $\chi^2_{0,05;2} = 5{,}99$ and $\chi^2_{0,01;2} = 9{,}21$, leading to the conclusion that null hypothesis should not be rejected, i.e., there is no statistical dependency between the order number of examination period in which student passes the exam and the study program that student attends.

4. CONCLUSION

Adequate statistical methods are applied to the real-life data on the students' performances within the mandatory subject *Mathematics* in a higher educational institution in Belgrade. Students from the two study programs, IT and CT, that completed the final exam from January 2022 till October 2024, are considered, resulting in total of 271 IT students and 75 CT students. The data are summarized using descriptive statistic measures. Impacts of the three independent variables: activity in class, colloquium results and the order number of examination period in which student completes the exam to the exam result, are tested using multiple linear regression analysis. The Chi-square test proved the independence between study program and the order number of examination period when a student passes the exam. The regression analysis has shown that colloquium results have the major impact on the final exam for both study programs. In the case of the larger and more successful group of IT students, the next factor is time until the exam is passed, and the factor with the lowest impact is activity in class. For the smaller group of CT students, consisting of students that on average achieve lower results in the subject, the activity in class is omitted as a statistically insignificant factor. The analysis confirmed the teacher's assumptions on superficial or less visible activity in class, as well as the fact that more intensive activity is demonstrated by students of lower pre-knowledge and lower current results, and finally the presence of self-learning using materials and technology to the larger extent.

5. REFERENCES

[1] Saha, S. M.; Pranty, S. A.; Rana, M. J.; Islam, M. J.; Hossain, M. E.: *Teaching during a pandemic: do university teachers prefer online teaching?*, Heliyon, 8 (1), 2022.
[2] King, R.B.; Chai, C.S.; Korpershoek, H.: *Learning and teaching during Covid-19 and beyond:educational psyhology perspectives,* Educational Psyhology, 42 (10), pp. 1199-1203, 2022.
[3] Maričić, M.; Ignjatović, M.; Jeremić, V.: *Modeli statističkog učenja,* Akademska misao, Beograd, 2022.
[4] Mladenović, Z; Petrović, P.: *Uvod u ekonometriju*, Ekonomski fakultet, Univerzitet u Beogradu, Beograd, 2007.
[5] Jovičić, M.; Dragutinović-Mitrović, R.: *Ekonometrijski metodi i modeli*, Ekonomski fakultet, Univerzitet u Beogradu, Beograd, 2018.

[6] Tranmer, M.; Murphy, J.; Elliot, M.; Pampaka, M.: *Multiple Linear Regression* (2nd Edition); Cathie Marsh Institute Working Paper 2020-01, 2020.
[7] Lakić, N.; Janković Šoja, S: *Statistika*, Poljoprivredni fakultet, Univerzitet u Beogradu, Beograd, 2021.
[8] Gamst, G., Meyers, L. S., Guarino, A.: *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*, New York, NY: Cambridge University Press, 2008.
[9] Durbin, J., Watson, G. S., Testing for serial correlation in least squares regression. Parts I and II. *Biometrika*, *37*, 409-428, 1950-1951.
[10] Soldić Aleksić, J.: *Primenjena analiza podataka*, Ekonomski fakultet, Univerzitet u Beogradu, Beograd, 2018.