Application of Logistic Regression in the Study of Students' Performance Level (Case Study of Vlora University)

PhD. Miftar Ramosacaj[1]

Prof. Dr. Vjollca Hasani[2]

Prof. Dr. Alba Dumi[3]

[1] Department of Math and Statistic" University of Vlore, Albania; Email: mramosaco@yahoo.it
[2]Dean of Economy Faculty, AAB Pristina University Kosovo; Email: v.hasani@gmail.com
[3]Management Department" Economy Faculty University of Vlore, Albania; Email: alba.besi12@gmail.com

Doi:10.5901/jesr.2015.v5n3p239

Abstract

The problem with the students' performance level in the first year of studies has been subject of many other similar studies. We have used the opportunity that binary logistic regression model offers. In this study, took part around 240 freshmen. The data were obtained from the analysis of the questionnaires. The analysis was conducted based on variables such as gender, environment where they live, type of private or public school, school location, points taken in high school, the mode of perception of the social environment. This paper deals with the presentation and analysis of the results of the examination of students' performance from University of Vlore by also establishing a more general assessment. It is assumed that students' results are affected by gender, type of private or public school, their location etc. The approach via logistic regression is done to study the results of the students after the first semester according to the variables mentioned above. The analysis consists of an examination of the impact of these independent factors in the performance of students in the first year of study.

Keywords: *Logistic regression, Odds Ratio, Wald Statistic, Likelihood, Ratio Tests, Shortest Confidence Interval.*

1. Introduction

Students differ in terms of gender, culture, family environment, financial status of parents, etc., while schools differ in number of students, the quality of teachers, infrastructure, location, assistance provided by the government, etc. Naturally, student performance measured in terms of credit points obtained in the first semester of university is a dependent variable of the above parameters. Variability in results is a social function, which should be studied and analyzed scientifically.

Study of students' performance is as old as the history of education. The analysis has started around thirties of the 20th century. Performance measurement corresponds to different independent variables that can be analyzed using logistic regression analysis. Logistic regression has been successfully used in the social sciences.

This paper deals with the presentation and analysis of the results of the examination of students' performance from University of Vlore by also establishing a more general assessment. It is assumed that students' results are affected by gender, type of private or public school, their location etc.

2. Literature Review And Hypotheses

The approach via logistic regression is done to study the results of the students after the first semester according to the variables mentioned above. Results of the students are divided into two groups: less than 30 credits, which mean that they have not taken all the provided courses, and more than 30 credits, for those who have. About 29% of students have less than 30 credits, while 61% of them have more than 30 credits. Students are classified into two different categories. The idea of binary logistic regression analysis seems to be appropriate when the results are functions of independent variables mentioned above.

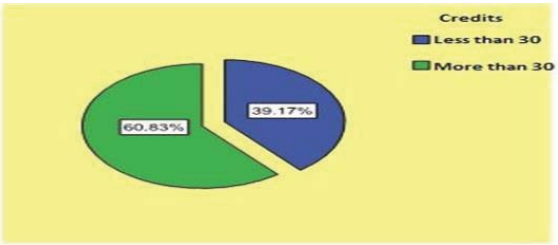

Figure 1: The two categories of students in base of first year CFU

Students come from high school with relatively high points. 75% of them have 4000-5000 points, while about 21% have 3000-4000 points. The important thing is to look at how students perform based on the high school points. Referring to the graphic, we note that: from the students that have 2000-3000 points, 45% of them have poor performance, while from students with 4000-5000 points, 34% results in poor performance. We think that the problem is related to the students with 3000-4000 point, from which 58% results in poor performance.

Summing up on this issue, we can say that there is an inflation of the high school points, and consequently the performance in the first year is as described. Greater inflation has the students with 3000-4000 points, based on the performance of the first semester. For this reason, we think that should be taken more into consideration the points calculated on the basis of merit-preference system, perhaps their calculation should be done in terms of criteria which lead to a more accurate assessment of merit.

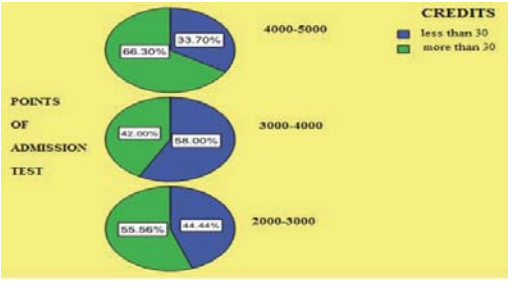

Figure 2: % of CFU first year and admission point test at university

3. Research Goal

Environment where the students study has a major impact in their performance, in terms of how stressful is for them. From the students who claim that the environment is less stressful, only 34% of them results in poor performance, while among those who claim that the social environment is stressful 52% of them results in lower performance.

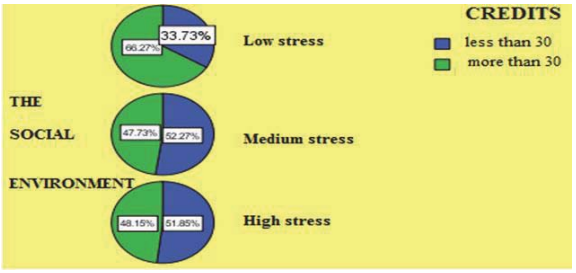

Figure 3: The descriptive link between environment level of stress and first year CFU

However, based on the data, we can say that 71% of students claim that social environment is less stressful.

Referring to the analysis of the below graphic, 59% of male students results in low performance, while 66% of female students in high performance.

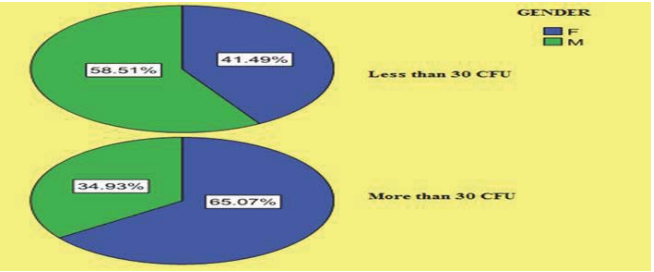

Figure 3: The descriptive link between student gender and first year CFU

The type of school from where they come from, isn´t significant for their performance after the first semester

4. Sample and Data Collections

Let's take into consideration the indicator variable: y_1=1, if y_1 has less than 30 credits and y_1=0, if y_1 has more than 30 credits. Also, we assume that:

P(Y=1) = θ=1-P(Y=0), where θ can be written as

$$\theta = x = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_{n)}}}$$

, where $\boldsymbol{\alpha}$ is a constant and $\boldsymbol{\beta}_i$ are the regression coefficients for each variable X_i for i=1,2,…, n. An alternative form of the regression equation is:

$$logit[\theta(x)] = log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

The purpose of logistic regression is to correctly predict the outcome category for special occasions using logistic regression model. To achieve this goal, it has been created a model that includes all predictor variables that are useful to predict the outcome.

5. Logistic Regression Calculates the Probability of Success Over the Probability of Failure

Results of the analyses are presented as a report of the opportunities. The odds ratio is a good indicator that shows the chances of success against failure under specific conditions of the data. It is used as a descriptive statistics and plays an important role in "logistic regression".

The first assumption or the null hypothesis (H0) is that the predictive coefficient is zero among the population.

2. Test: The hypothesis arises if there are sufficient evidence in the sample to refuse the null hypothesis and therefore to accept the alternative one, that the coefficients are different from zero. Confidence intervals can be used for hypothesis testing and for regression coefficients as well.

In this paper, logistic regression analysis was performed using as dependent variable the amount of credits, assuming that the student with less than 30 credits are failing in one or more subjects, and as dependent variables: gender, type of public or private school, the location, the level of stress in the social environment.

By applying the method of correlation analysis, we noted that there is not a significant correlation between the type of school and results of exams (around 0.02), as well as between other variables, which shows that it is less likely to have (collations) between independent variables.

DATA		DATA TYPE	VALUES
	Dependent	Variable	
Taken credits	Passing Failing	Binary	0- Less than 30 credits 1- More than 30 credits
	Independent	Variable	
P1 Gender	Independent	Binary	1- F 0- M
P2 Type of school	Independent	Binary	0- Public 1- private
P5 High school points	Independent	Non- binary	1- 5000-4000 2- 4000-3000 3- 3000-2000
P4 Taken credits	Credits	Binary	0- Less than 30 credits 1- More than 30 credits
P3 School location	Independent	Binary	0- Village 1- Urban area
P6 Social environment	Independent	Non- binary	1- stressful 2- not stressful 3- very stressful

Discussion: Referring to logistic regression equation, taken from the SPSS program, we can give the general form of the regression equation:

$$\theta = \frac{e^{(0.75-0.88\,x_1+0.12x_2-0.28x_3-0.46\,x_5+0.61x_6-0.44x_7)}}{1+e^{(0.75-0.88\,x_1+0.12x_2-0.28x_3-0.46\,x_5+0.61x_6-0.44x_7)}}$$

Based on the equation analysis, we can see which variables are significant and which are not. If the significance of the respective coefficient is less than 0.05, then this coefficient is statistically significant. Thus, the significance for P7 (Is the social environment stressfully?) affects the level of student performance, it has a negative impact on it.

H1: A student who considers the social environment as not stressful has 4.6 times more chance of being successful than a student who has different perception on the social environment.

H2: The environment where the student lives has another impact on his performance, this impact is positive.

Significance for this coefficient is 0.04. Those who live in their homes have 8 times more chance for a higher performance compared to students who live in the dormitory or rented house.

H3: This probably has to do with the fact that most of the students are from city of Vlore.

In terms of the type of school and its location, we can say that they don't have a significant impact; since these two variables don't have an influential role model, they should be removed from the equation.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	P7	-.441	.203	4.693	1	.030	.644	.432	.959
	P6	.616	.216	8.143	1	.004	1.852	1.213	2.828
	P5	-.468	.265	3.118	1	.077	.626	.372	1.053
	P31	-.284	.412	.477	1	.490	.752	.336	1.686
	P1	-.889	.288	9.514	1	.002	.411	.234	.723
	P2	.127	.372	.116	1	.733	1.135	.548	2.353
	Constant	.757	.952	.633	1	.426	2.132		

a. Variable(s) entered on step 1: P7, P6, P5, P31, P1, P2.

Regression equation form would be:

$$\theta = \frac{e^{(0.54-0.91\,x_1-0.47\,x_5+0.62x_6-0.45x_7)}}{1+e^{(0.54-0.91\,x_1-0.47\,x_5+0.62x_6-0.45x_7)}}$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1[a]	P1(femaleMale)	-.919	.284	10.462	1	.001	.399	.229	.696
	P5	-.477	.264	3.257	1	.071	.621	.370	1.042
	P6	.626	.210	8.903	1	.003	1.870	1.240	2.821
	P7	-.450	.203	4.928	1	.026	.638	.429	.949
	Constant	.547	.721	.575	1	.448	1.728		

a. Variable(s) entered on step 1: P1, P5, P6, P7.

Considering Exp (B), we can say that the gender has an impact on the model; female performance is 0.4 times higher than male performance. The influence of the social environment of the school is significant in the model; 0.6 times higher is the performance of the students who consider it less stressful.

Wald for parameter P1 is 10.5 which show that when other parameters are unchanged, the chances for a higher performance among females are 0.4 times better than among males. Also, we can see the confidence interval for Exp (B) for each coefficient. Independent variable, P2 has a significance of 0.733 therefor; it is not statistically significant if students come from private or public schools in assessing their performance during the first semester. Also P31, the school location isn't a statistically significant variable; significance is 0.490 <0.05 and therefore should not be included in the model.

Independent variables that are significant in the model according Forward Stepwise method are P1, P6 and P7, respectively, gender, environment where you live, and the level of stress in the social environment. Also, it is noted that the performance of females and males are 0.669 and 1.979 times better than co-educational schools respectively. Similarly, manifestations of urban students are 1.409 times better than those of students from rural areas.

By applying the method of correlation analysis, it is noted that there is not a significant correlation between the variables included in the model.

Independent variables, school location, are irrelevant in the model; significance is 0.733> 0.05, as well as P31 and P5, while the student gender and type of schools have significantly affected the results of students in the first semester. Regarding the type of private or public school, it can be seen that the performance of private high schools is not significant comparing to public schools.

This is shown by Wald 0.116. A similar conclusion is reached related to their location, even though 86% of students come from high schools in urban areas and only 14% from schools in rural areas. Performance of urban students is 0.750 times better than rural students. Considering the length of confidence interval and the estimated risk, we find that the school location has a narrow 95% confidence interval. Comparison of the odds ratio is considered to set the independent variables.

6. Conclusions

From the obtained data and statistical analysis we can conclude that: Study the performance of students is a known problem which has its significance in establishing teaching and research policies in order to increase the level of quality.

The identification of such phenomena is an advantage to the society. Referred to the data, we can tell that the level of female students' performance is better than that of male students, in other words, it would be appropriate to recommend in these degrees female students to attend, considering the extent of their performances.

The environment where they live is far from being appropriate for the research. This requires radical measures to be taken to increase not only economic conditions but also standards for a research environment in the university.

We conclude from the study that the level of student performance is affected from high school results and outperform those who have a higher valuation. So we should pay specific attention to increasing the level of performance since in high school. Continuous improvement of socio-economic conditions of students, and the creation of non-stressful conditions are important contributing factors in increasing student performance.

References

Agresti, A. (1996). An Introduction to Categorical Data Analysis, John Wiley and Sons, Inc.
Dumi A Ramosacaj M, JERM 2013, "The method of math analyses in hotel reservation", Roma Italy.

Dumi A Ramosacaj M, MSCER 2014, "The method of math in economy", Roma Italy.
Danny A, JOLIS Journal 2008, Logistic Regression Analysis USA
Hosmer, D. and Stanley, L. (1989). Applied Logistic Regression, John Wiley and Sons, Inc.
Menard, S. (1995). Applied logistic regression analysis (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–106), Thousand Oaks, CA: Sage.
Menard, S. (1995). Applied Logistic Regression Analysis. Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 106.
Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis, The American Statistician, 54(1), p. 17–24.