

Logistic regression model to study the most important factors that affect students' academic commitment

A case study on Imam University students

Karema Ahmed Elsayed

Assistant Professor, Department of Business Administration
Faculty of Business & Administration

Princess Noura bint Abdulrahman University, Saudi Arabia

Lecturer, Department of Statistics, Faculty of Commerce

Al-Azhar University, Egypt

kaali@pnu.edu.sa, karemaelassasy@gmail.com

Abstract:

The research aimed at studying the factors affecting students' academic commitment using the method of logistic regression, Clarify their characteristics, and the maximum likelihood estimation MLE used to estimate the parameter of the model for a study of a sample of 1580 students. The characteristics of the logistic regression model, the method of constructing the regression model, and the method of estimating its parameters were highlighted.

The study showed the effectiveness of the logistic regression model to measure the factors that affect the academic commitment of Imam Muhammad University students, such as gender, culture, family environment, type of public or private high school, their location, academic advisor.

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

The most important factors affecting students' academic commitment (age, Admission point test at university, Social environment, and finally the students' follow-up to the academic advisor), also confirmed that the academic commitment of female students is better than the academic commitment of male students.

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

الملخص:

هدف البحث إلى دراسة العوامل المؤثرة على التزام الطلاب الأكاديمي باستخدام أسلوب الانحدار اللوجستي ، توضيح خصائص النموذج، وتقدير معالم النموذج باستخدام مقدر الامكان الاعظم لدراسة عينة من ١٥٨٠ طالب وطالبة من جامعه الإمام محمد بن سعود الإسلامية.

حيث تم تسليط الضوء على خصائص نموذج الانحدار اللوجستي وطريقة بناء وطريقة تقدير معاملاته.

أظهرت الدراسة فاعلية نموذج الانحدار اللوجستي لقياس العوامل المؤثرة في الالتزام الأكاديمي لدى طلاب جامعة الإمام ابن محمد (الجنس ، والثقافة ، والبيئة الأسرية ، ونوع المدرسة الثانوية العامة أو الخاصة ، وموقعهم الجغرافي ، دور المرشد الأكاديمي.. إلخ).

أهم العوامل التي تؤثر على التزام الطلاب الأكاديمي (العمر ، درجة اختبار القبول في الجامعة ، البيئة الاجتماعية ، وأخيراً متابعة الطلاب للمرشد الأكاديمي). كما أكدت أن الالتزام الأكاديمي للطالبات أفضل من الالتزام الأكاديمي لدى الطلاب.

الكلمات المفتاحية: الانحدار اللوجستي ، الالتزام الأكاديمي ، نسبة الأرجحية ، إحصاء والد ، الاحتمالية ، اختبارات النسبة ، أقصر فترة ثقة.

The first topic: methodology of research

1.1 Introduction:

The logistics regression models have been used in previous years in the field of medical sciences and have developed for recent use in the economic and applied field. It is widely used in life experiences, which are among the major concerns of the countries of the world for their relationship to human life and development by finding the best way to provide services for the best results. The method of analyzing the course of effective statistical methods in the analysis of data has enabled the researcher to identify and clarify the possible negative relationships of a combination of factors and determine the direct and indirect impact of the phenomenon of the macro study and thus help to draw logical explanations for this phenomenon and more efficient in data analysis. The importance of analyzing logistic regression is highlighted by the ability to study the impact of several factors on a particular phenomenon directly or indirectly, many studies deal with factors affecting the academic achievement of students but, the most important of which is the academic achievement at the secondary level of School, the educational level of parents and the financial level.[1],[3],[8].

In this research, the factors affecting the level of academic achievement were measured by the student's commitment to all credit hours for the first academic year at the university.

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

The number of credit hours for the first academic year is approximately thirty credit hours. According to the logistic regression model, the dependent variable takes a binary value which is zero for whom that's not passing the needed 30 credit hours and one for who passed it, the student passed the 30 hours answer with yes that equals one and who didn't pass the hours answer with no that equals zero.

1.2 Research problem:

A university student is the ultimate product and primary goal of all university activities. Therefore, the low level of student achievement, which depends on the student's commitment to credit hours each semester, is a failure and a challenge for the university. The research problem can be determined by the following question: What are the most important factors that affect the student's commitment to a record of credit hours according to the study plan from the viewpoint of Imam University students?

1.3 Research Importance:

The importance of the research lies in addressing one of the important issues, which is the extent of academic commitment to students, and knowledge of the factors that affect it to determine methods of raising and improving performance, as this issue has not been specifically addressed by Al-Imam University before.

1.4 Research Objective

1.5 Research is a major goal Determine the most important factors affecting the academic commitment of the university student from among the factors (economic, social, psychological and academic) and the ability of the logistic regression model to represent and measure this relationship and estimate the regression coefficients of the model.

1.6 Research Hypotheses

The research is based on the following hypotheses:

- 1- There is no statistically significant relationship between (economic - social - psychological and academic) factors and the academic committee of the university student from the viewpoint of the students of Imam University.
- 2- There is no difference in views among the sample members due to the two variables (gender - a decision in high school - type of school, and follow-up to the academic advisor).
- 3- The dependent variable takes (0,1), binary values.
- 4- There is a relationship between the dependent variable and explanatory variables take the following form:

$$P_i = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$

- 5- There is no correlation between random errors (independence of errors).

6- The random variable (Y_i) is assumed to be distributed by Bernoulli

$y \sim B(p)$ in the mean $E(Y_i) = p$

And variance $V(Y_i) = p(1-p)$.

7- The expected value of the random error is zero, since $Pr(x) = [1 - pr(x)]$

Also, the variance of the random error U_n expands based on the bar. [2],[6],[7].

1.7 Research methodology and tools

The study used the descriptive-analytical method, and the questionnaire used as a tool to know the factors affecting the academic achievement level of Imam University students.

Statistical software package program was used (Statistical Package for Social Sciences) SPSS.

1.8 Research sample

This paper deals with the presentation and analysis of the results of the examination of students' performance from Imam Mohamed Ibn Saud Islamic University - Riyadh, where the resolution was used as a data collection tool to verify the impact of the factors specified in the model on 1850 students.

1.9 Research limits

Spatial boundaries: Imam University in Riyadh for some of its scientific colleges

Temporal boundaries: the academic year 1439-1440.

The objective limits a study of the factors affecting the academic achievement level of the university student.

The second topic: Theoretical side.

2.1 Logistic Regression

It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one, more nominal, ordinal, interval, or ratio-level independent variables.

2.2 Characteristics of the Logistic Regression Model

1. This model does not put any preconditions on the explanatory variables
2. The model does not specify which vectors belong to the new observations but also determines the probability of this affiliation. It can also be used to analyze the binary and multivariate descriptive variables.
3. The Maximum Likelihood Estimation (MLE) is used to estimate its parameters and therefore the quality conditions are met in these variables
4. Due to the ease of calculations used in the Model, logistic regression becomes one of the most appropriate models for the analysis of binary and multivariate descriptive variables.

2.3 Method of Building the Regression Model

1. Achieving the relationship between the binary nominal variable and the nominal independent variable through a single analysis using the Chi-square and the correlation test
2. To achieve the correlation between the variable logarithm of the binary nominal and the independent quantitative constant variable by the scheme of dispersion between the two variables where the relationship must be positive
3. Analysis of the linear relationship between independent variables.[9]

2.4 Analysis of Simple and Multiple Logistic Regression

The analysis of logistic regression is used in epidemiological and medical studies, in which the quantitative and qualitative independent variables that affect the probability of the resulting variable are determined when the logistic regression is applied.

2.4.1 Required For Simple Logistic Regression

Quantitative or nominal independent variables such as weight, height, marital status, and gender. It is a binary variable, such as a disease (yes, no), gender (male, female), and others.

2.4.2 Required For multiple logistic regressions

Two or more independent quantitative or nominal variables such as weight, height, marital status, and gender. One variable my name is a dual-type follower such as an illness (yes, no) or gender (male, female) and others.

2.5 Logistic Regression Model

The logistic regression model is defined as one of the regression models where the relationship between the y variable and the explanatory variables $(x_1, x_2, x_3, \dots, x_K)$ is nonlinear where the y variable of the binary response takes the values (0, 1) and the success is the probability of (p)or failure to respond to the probability of $[q = (1-p)]$ so the dependent variable y follows the Bernoulli distribution and the probability density function will be as follow:

$$P(Y = Y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \dots\dots\dots (2)$$

Sine: $y_i = (0,1)$

y_i : Binary variable dependent response

p_i :The probability of a response when $y_i = 1$

Therefore, signing the dependent variable represents the probability of a response

$$E(y_i) = p(y = 1) = p_i \dots\dots\dots (3)$$

The variance of the dependent variable by Bernoulli distribution is

$$V(y_i) = p_i(1 - p_i) \dots\dots\dots (4)$$

Let $(x_0, x_1, x_2, x_3, \dots, x_n)$ be a set of explanatory variables and let n have many observations of these variables that are matrix X.

$y_i = (y_1, y_2, y_3, \dots, y_n)$ is a random sample of the binary response variable and $y_i \in \{0,1\}$

This leads to a regression model given as follows:

$$y_i = p_i + \varepsilon_i \dots\dots\dots (5)$$

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

Represents the regression function (probability of response)

$$p_i = p(y = 1) = \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \dots\dots\dots (6)$$

$$0 \leq \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \leq 1$$

$$0 \leq odd \leq 1$$

β : Vector of information dimensions ($p * 1$).

$x_i = (x_{i0}, x_{i1}, \dots, x_{ip})$ A class vector of explanatory variables ($1 * p$).

ε_i : The error limit that will have an average of zero as in the following formula:

$$\varepsilon_i = y_i - p_i$$

$$E(\varepsilon_i) = E(y_i) - E(p_i) = p_i - p_i = 0 \dots\dots\dots (7)$$

Either the error limit variance is equal to the variance of the adopted binary response variable.

$$V(\varepsilon_i) = V(y_i) = p_i(1 - p_i) \dots\dots\dots (8)$$

The error threshold follows the Bernoulli distribution with an average of (0) and the variance $p_i(1 - p_i)$, it is noted that the variance of the error limit depends on the values of the response probability (p_i).

i.e. on vector values x_i and there is the variance of error non-homogeneous. [3],[6].

2.6 Methods for Estimating the Parameters of the Logistic Regression Model

To estimate the parameters of the logistic regression model, it was based on the maximum possible method (MLE), the maximum possible method requires repetitive methods for its calculation. Therefore, the initial values of β_1 and β_0 are required.

Before analyzing this method, it is important to know that binary data such as success and failure appear in most areas of study. The analysis of logistic regression is most often used to examine the relationship between intermittent responses and total explanatory variables. Some discussed logistic regression. [1].

Since in binary data the distribution of random error U_i is discrete, it is an abnormal distribution, but a binomial distribution is distributed. If the error is distributed naturally, the variance is not equal in all aggregates. In the case of discrete response models, the model is as follows:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + U_i \quad \dots\dots\dots (9)$$

Y_i : Dependent Variable.

x_i : Independent Variable.

U_i : is a random error

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$ Regression equation parameters.

Since the variable (Y_i) a distribution of any Bernoulli there are only two values, so the random distribution of error (U_i) distribution is discrete and does not continue any:

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

$$Y_i = 1; \quad Y_i = 0; \quad E(Y_i) = p_i; \quad V(Y_i) = 1 - p_i = Q_i$$

While:

p_i : Response Ratio When $Y_i = 1$

Q_i : Response Ratio When $Y_i = 0$

2.6.1 Maximum Likelihood Estimation (MLE)

This method is based on the finding of β values, which are estimates of vector β , which makes the function at its end, and assuming that we have n of independent random variables) ;

$(y_1, y_2, y_3 \dots \dots y_r)$; n_i, p_i, y_i that represents the sum of successes at each end of the function at n_i the end of and that (K) of the independent variables in each set of totals, the probability density function of (Yi):

$$P_i(Y_i = y_i) = C_{p_i}^{n_i} p_i^{n_i} (1 - p_i)^{n_i - y_i} \dots \dots \dots (10)$$

$$i = 1, 2, \dots, r; \quad y_i = 0, 1, 2, \dots, r$$

$$E(y_i) = n_i p_i \dots \dots \dots (11)$$

$$V(y_i) = n_i p_i (1 - p_i) \dots \dots \dots (12)$$

The p_i response rate is estimated as follows: $p_i = \frac{y_i}{n_i}$; $Q_i =$

$$1 - p_i = \frac{1 - y_i}{n_i} = \frac{n_i - y_i}{n_i}$$

Maximum Likelihood Estimation (MLE) functions for the distribution of data (Yi) is by the formula:

$$L(P) = \prod_{i=1} C_{p_i}^{n_i} p_i^{n_i} (1 - p_i)^{n_i - y_i} \dots \dots \dots (13)$$

$$L(P) = \prod_{i=1} C_{p_i}^{n_i} p_i^{n_i} \left[\frac{p_i}{1 - p_i} \right]^{y_i} (1 - p_i)^{n_i} \dots \dots \dots (14)$$

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

$$\ln L(P) = \sum_{i=1}^r \left[\ln C_{p_i}^{n_i} + y_i \ln \left[\frac{p_i}{1-p_i} \right] + n_i \ln(1 - p_i) \right] \dots (15)$$

Since p_i the probability of success is affected by the independent explanatory variables according to the logistic model (in the form of matrices):

$$p_i = \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \dots (16)$$

$$1 - p_i = \frac{1}{1 + \exp(x_i' \beta)} \dots (17)$$

When compensating the formulas (16) and (17) in (15) we get the following formula:

$$L(P) = \sum_{i=1}^r \left\{ \ln C_{p_i}^{n_i} + y_i X_i' \beta + n_i \ln \left[\frac{1}{1 + \exp(x_i' \beta)} \right] \right\} \dots (18)$$

We use The Newton-Raphson method to find the *Maximum Likelihood* estimates according to the Formula:

$$t_{s+1} = t_s - G^{-1} g_{(s)} \dots (19)$$

t_{s+1} : represents the vector of β parameters to be estimated

t_s : The vector represents the initial values of the parameters

$g_{(s)}$: The first derivative vector of the logarithm represents the function Likelihood

$$g_{(s)} = \frac{\partial \ln L}{\partial \hat{\beta}} \begin{bmatrix} \frac{\partial \ln L}{\partial \beta_0} \\ \cdot \\ \cdot \\ \frac{\partial \ln L}{\partial \beta_K} \end{bmatrix}$$

G: Matrix The expected value of the negative value of the second derivative of the logarithm is the Maximum Likelihood function

$$G = \left[\frac{\partial^2 \ln l}{\partial \hat{\beta}_i \partial \hat{\beta}_l} \right] = \begin{bmatrix} -E \frac{\partial^2 \ln l}{\partial \beta_0^2} & -E \frac{\partial^2 \ln l}{\partial \beta_0 \partial \beta_1} & -E \frac{\partial^2 \ln l}{\partial \beta_0 \partial \beta_K} \\ \cdot & -E \frac{\partial^2 \ln l}{\partial \beta_1^2} & -E \frac{\partial^2 \ln l}{\partial \beta_1 \partial \beta_K} \\ \cdot & \cdot & -E \frac{\partial^2 \ln l}{\partial \beta_K^2} \end{bmatrix}$$

Thus, the formula for finding estimates is as follows:

$$\hat{\beta}_{s+1} = \hat{\beta}_s + \left[\frac{\partial^2 \ln l}{\partial \hat{\beta}_j \partial \hat{\beta}_j} \right]^{-1} \frac{\partial \ln L}{\partial \hat{\beta}_s} \dots\dots\dots (20)$$

The original Newton-Raphson formula in the case of the logistic model to find the Maximum Likelihood estimates is:

$$\hat{\beta}_{s+1} = \hat{\beta}_s + (X'VX)^{-1} X'(y - \hat{y}_s) \dots\dots\dots (21)$$

$\hat{\beta}_{s+1}$: Vertical direction of the estimate values in the cycle (s+1) at grade ((k+1)*1) ^ Vertical direction of the estimate values in the cycle (s) at grade ((k+1)*1)

X: matrix of independent variables in the cycle (r*) (k + 1)

V: a diagonal matrix of the rank (r*r)

$$V = \begin{bmatrix} n_1 \hat{p}_1 \hat{q}_1 & 0 & 0 & 0 & 0 \\ 0 & n_2 \hat{p}_2 \hat{q}_2 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & . & . \\ 0 & 0 & 0 & . & n_r \hat{p}_r \hat{q}_r \end{bmatrix}$$

Estimation of the vector of parameters (β) is discontinued when the difference from the previous cycle and the subsequent cycle is very small and approaches zero.[4]

The third topic: Applied side

3.1 Description of data

Students differ in terms of gender, culture, family environment, the financial status of parents, etc., while schools differ in the number of students, the quality of teachers, infrastructure, location, the assistance provided by the government, etc. Naturally, student performance measured in terms of credit points obtained in the first year of university is a dependent variable of the above parameters.

Variability in results is a social function, which should be studied and analyzed scientifically as a study of students' performance is as old as the history of education.

The analysis has started around the thirties of the 20th century. Performance measurement corresponds to different independent

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

variables that can be analyzed using logistic regression analysis as logistic regression has been successfully used in the social sciences.

It is assumed that students' results are affected by gender, type of private or public school, their location, etc. Literature Review and Hypotheses the approach via logistic regression is done to study the results of the students after the first year according to the variables mentioned above. The results of the students are divided into two groups: less than 30 credits, which means that they have not taken all the provided courses, and more than 30 credits, for those who have. About 35% of students have less than 30 credits, while 65% 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 the independent variables mentioned above.[5]

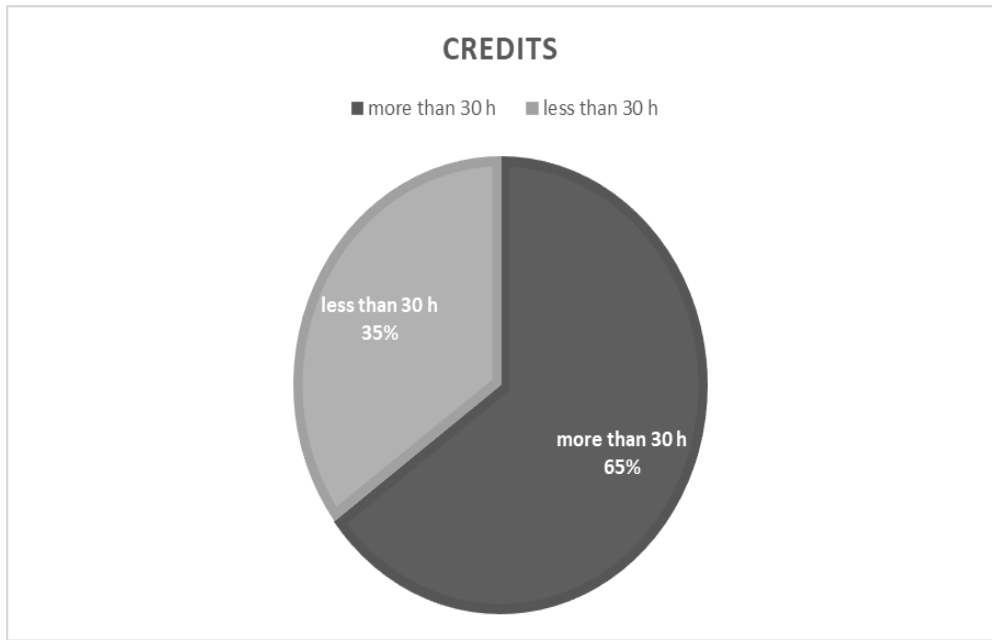


Figure 1: The two categories of students in the base of first-year CFU.

Students who came from high school with relatively high points: 66% of them have more than 400 points, while about 34% have 300-400 points. Therefore, the points that students achieved during high school should be considered when assessing the performance of the students.

Referring to the graphic, we note that: the students with poor performance had 200-300 points, 42%, while the students who had more than 400 points were 34%. The students with 300-400 points during high school whose performance was poor were

55% and 45% had a good performance. We may conclude that the higher points of students in high school (300-400), so we think this criterion should be taken into consideration based on the merit-preference system, which results in a more accurate assessment of merit.

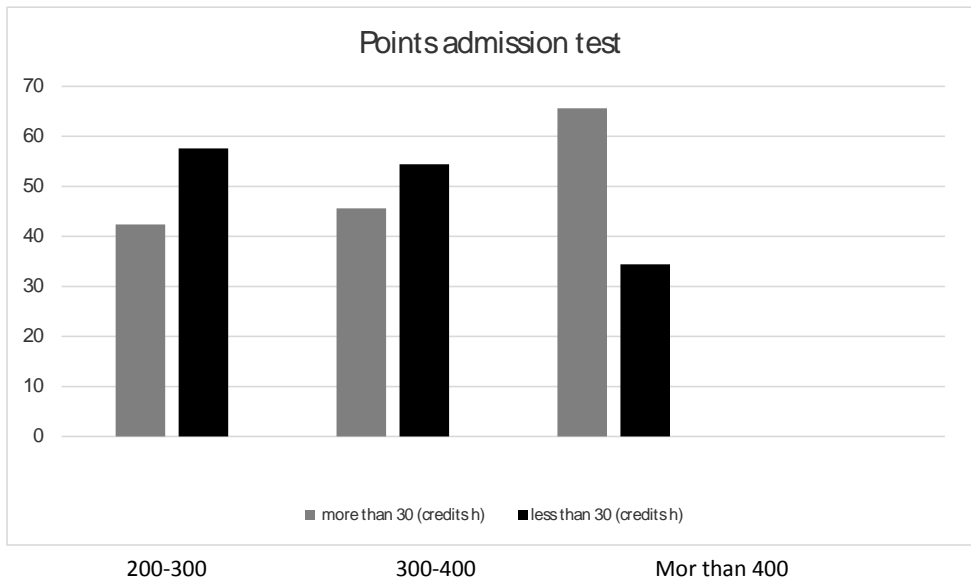


Figure 2: % of CFU first year and admission point test at university.

The environment had a major impact on student's performance as the students with high stress having poor performance (34%) while those with low stress having good performance (52%).

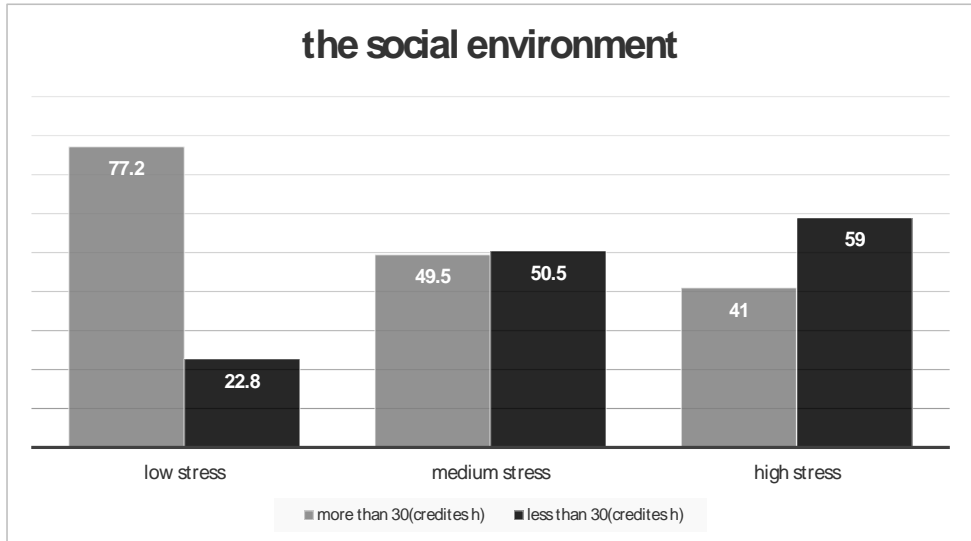


Figure 3: The descriptive link between the environment level of stress and first-year CFU.

However, the study showed that 66% of the female students had higher performance though they had social stress while 58% of male students had low performance though they did not suffer social stress.

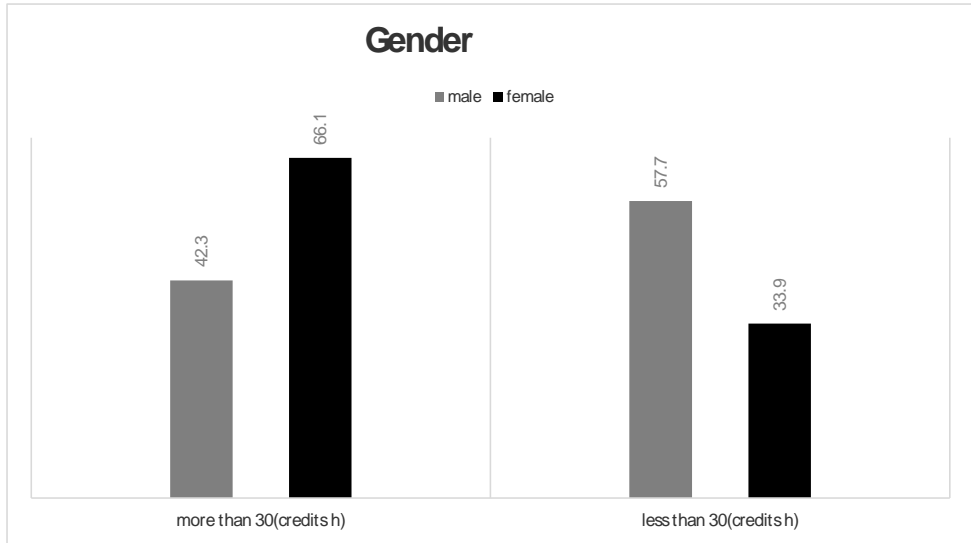


Figure4: The descriptive link between student gender and first-year CFU.

The study concluded that the type of school was not a major criterion to assess the performance of the students during the first year therefore; it can be excluded from the independent variables. Figure (5) shows that 92% of the female students followed the academic advisor, while for the students, the percentage was 73%.



Figure5: Follow-up students to the academic advisor

Test Data Of Dependent Variable.

To determine the distribution of variable y , variable data (the binary response variable) we will adopt the ready-made program (EASY-FIT) test (Anderson-Darling), which shows that these data are distributed according to Bernoulli's distribution with their estimation parameter ($p = 0.178$) as shown in Table (1).

	Distribution	Parameters
1	Bernoulli	$n=1, p = 0.178$
2	Binomial	$n=1850, p = 0.178$
3	D. Uniform	$a=0, b=1$
4	Geometric	$P = 0.749$
5	Poisson	$\theta = 0.178$
6	Hypergeometric	No fit
7	Logarithmic	No fit (data min < 1)
8	Neg. Binomial	No fit

Where:

Y: Probability of Success over the Probability of Failure.

X_1 : Gender.

X_2 : Type of school

X_3 : School Location

X_4 : Taken Credits

X_5 : Admission point test at university

X_6 : Social environment

X_7 : following the academic advisor

Logistic Regression Calculates the Probability of Success over the Probability of Failure Let's take into consideration the indicator variable: $y_1 = 1$, it y_1 has less than 30 credits and $y_1 = 0$, it y_1 has more than 30 credits.

3.2 Statistical analysis

These data are analyzed by the statistical program SPSS to determine the importance of variables and their impact on Student Performance, results of the analyses are presented as a report of the opportunities in the table (2). 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 descriptive statistics and plays an important role in "logistic regression".

The first assumption or the null hypothesis:

H_0 : Is that the predictive coefficient is zero among the population.

The second assumption:

H_0 : The hypothesis arises if there is 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 regression coefficients as well. In this paper, logistic regression analysis was performed using as dependent variable the number 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 and 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 the results of exams (around 0.02), as well as between the other.

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

Table (2)			
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	0 – M 1 - F
P2 Type of school	independent	Binary	0 –Public 1 –Private
P3 School Location	independent	Binary	0 – Village 1 –Urban area
P4 Taken Credits	independent	Binary	0 - Less than 30credits 1- More than 30 credits
P5 Admission point test at university	independent	Non-Binary	200-300 300-400 Mor than 400
P6 Social environment	independent	Non-Binary	Stressful Not stressful Very stressful
P7 the academic advisor	independent	Binary	0: Not following the academic advisor 1: following the academic advisor

Discussion: Referring to the logistic regression equation, taken from the SPSS program, we can give the general form of the regression equation (6):

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

$$p_i = p(y = 1)$$

$$= \frac{e^{\beta x_i}}{1 + e^{\beta x_i}}$$

Logistic regression model for all factors:

$$\frac{e^{0.73-0.85x_1+0.11x_2-0.25x_3-0.43x_4-0.60x_5-0.43x_6-0.42x_7}}{1 + e^{0.73-0.85x_1+0.11x_2-0.25x_3-0.43x_4-0.60x_5-0.43x_6-0.42x_7}}$$

Based on this 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 P6 (Is the social environment stressful?) affects the level of student performance, has a negative impact on it.

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

H2: The environment where the students live has another impact on their performance, this impact is positive. The significance of this coefficient is 0.04. Those who live in their homes have 8 times a chance for higher performance compared to students who live in the dormitory or rented houses.

H3: It has probably to do with the fact that most of the students are from the University of El-Imam Mohamed. In terms of the type of school and its location, we can say that they do not have a

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

significant impact; since these two variables do not have an influential role model, they should be removed from the equation see table (3).

Logistic regression model form would be:

$$\frac{e^{0.757-0.889x_1+0.127x_2-0.284x_3-0.468x_5-0.616x_6-0.441x_7}}{1 + e^{0.757-0.889x_1+0.127x_2-0.284x_3-0.468x_5-0.616x_6-0.441x_7}}$$

Table (3): Variables in the Equation

Step 1	B	S.E.	Wald	Df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
P7	-.441	0.203	4.693	1	0.030	0.644	0.432	0.959
P6	0.616	.216	8.143	1	.004	1.852	1.213	2.828
P5	-.468	.265	3.118	1	.077	.626	.372	1.053
P3	-.284	.412	.477	1	.0490	.752	.336	1.686
P1	-.889	.288	9.514	1	.002	.411	.234	.723
P2	.127	.372	.116	1	.733	.1135	.548	2.353
Constant	.757	.952	.633	1	.426	2.132		

a. Variable(s) entered on step 1: P7, P6, P5, P3, P1, P2.

Considering Exp (B), we can say that 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 was significant in the model as the performance of the students who consider it was less stressful was 0.6 times higher. Wald for parameter P1 is 10.5 which showed that when other

Logistic regression model to study the most important factors that affect ...

Dr/ Karema Ahmed Elsayed

parameters are unchanged, the chances for a higher performance among females are 0.4 times higher 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 therefore; it is not statistically significant if students come from private or public schools in assessing their performance during the first semester. Also P3, the school location isn't a statistically significant variable; significance is 0.490 0.05, as well as P3 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 compared to public schools. This is shown by Wald 0.116. A similar conclusion is concluded related to their location, even though 86% of students come from high schools in urban areas and only 14% from schools in rural areas. The performance of urban students is 0.750 times better than rural students. Considering the length of a confidence interval and the estimated risk, we find that the school location has a narrow confidence interval of 95%. A comparison of the odds ratio is considered to set the independent variables to see table (4).

Logistic regression model form would be:

$$\frac{e^{0.547-0.919x_1-0.477x_5-0.626x_6-0.450x_7}}{1 + e^{0.547-0.919x_1-0.477x_5-0.626x_6-0.450x_7}}$$

Step 1a	B	S.E.	Wald	Df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
P1	-.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		

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

It is clear from the table (4) that the most important factors affecting the degree of students' academic commitment (age, Admission point test at university, Social environment, and finally the students' follow-up to the academic advisor).

The fourth topic: Conclusions and Recommendations

The Main Conclusions and Recommendations reached will be discussed in this section.

4.1 Conclusions:

Based on the data obtained and statistical analysis, we can conclude the following: It is very important to review the academic performance of university students during the first year of study (one year) to be able to improve the factors affecting their academic commitment. Improving these factors creates an advantage for the student, the university, and society as a whole. It can also be concluded that: The

level of academic commitment of female students is better than that of male students. As well as follow-up of the students to the academic advisor higher. It has been proven that the environment in which they live affects the extent of their academic commitment. Therefore, it requires improving the economic conditions and the social and psychological aspects, urging students and making them aware of the importance of following the academic advisor and creating the appropriate atmosphere for study at the university. The level of academic commitment is affected by the results of the secondary school; Therefore, we must pay special attention to increasing the performance of students in high schools. In conclusion, the continuous improvement of students' social, economic, and psychological conditions during all stages is an important contributing factor in raising the level of academic commitment of students.

4.2 Recommendations

According to the findings of the research, it is recommended to:

1. Other variables that are not mentioned in this study may be applied and to examine how significant they are in the quality of education.
2. Use the Cox model to analyze the impact of the factors that have been studied, comparing with the present results by the logistic model.
3. The use of other methods for estimating logistic regression such as the BIZI methods in statistical analysis for the regression model.
4. Using nonlinear regression models of various types to analyze Social and economic phenomena.
5. Use the multiple logistic regression model.

Reference:

- [1] Agresti, A. (2007). An introduction to categorical data analysis. 2nd. Hoboken NJ: Wiley-Interscience.
- [2] Guo, P., Xiao, B., & Li, J. (2012). Unconstraining methods in revenue management systems: Research overview and prospects. *Advances in Operations Research*, 2012.
- [3] Hilbe, J. M. (2009). *Logistic regression models*. Chapman and hall/CRC.
- [4] Nemes, S., Jonasson, J. M., Genell, A., & Steineck, G. (2009). Bias in odds ratios by logistic regression modelling and sample size. *BMC Medical Research Methodology*, 9(1), 56.
- [5] OLANIRAN, Y. O. A. (n.d.). *The Application of Logistic Regression*

Analysis to the Cumulative Grade Point Average of Graduating Students: A Case Study of Students' of Applied Science, Federal Polytechnic, Ilaro.

- [6] Ramosacaj, M., Hasani, V., & Dumi, A. (2015). Application of logistic regression in the study of students' performance level (Case Study of Vlora University). *Journal of Educational and Social Research*, 5(3), 239.
- [7] Shuai, J.-J., & Wu, W.-W. (2011). Evaluating the influence of E-marketing on hotel performance by DEA and grey entropy. *Expert Systems with Applications*, 38(7), 8763–8769.
- [8] Sloane, D., & Morgan, S. P. (1996). An introduction to categorical data analysis. *Annual Review of Sociology*, 22(1), 351–375.
- [9] Talib, H. R., Abd, A. R., & Ahmed, A. H. (2018). Use of logistic regression to study the most important factors affecting the incidence of tuberculosis. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, 10(2), 1–9.