ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **10** (2021), no.3, 1285–1299
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.10.3.16

USE OF REGRESSION ANALYSIS METHODS TO DETERMINE FACTORS RELATED TO SCHOOL DROPOUT (AN APPLIED STUDY OF A SAMPLE OF PRIMARY SCHOOLS IN BAGHDAD)

Sabah Haseeb Hasan[1] and Narjis Hadi Irhaif

ABSTRACT. The study dealt with the use of one of the important statistical techniques, namely regression analysis in determining the factors related to the school dropout. Two variable selecting methods were used, which are the method of all possible regressions and the method of stepwise regression. The two methods were applied to the data collected about the dropout of students of primary schools in the capital, Baghdad. The study reached statistically significant results that the frequent failure of study, poor living, changing family housing, poor security, and bad relationship between teachers and students are factors that increase school dropout. A multiple linear regression model was constructed that quantifies the effect of each factor on the number of students who drop out, where all parameter values in the estimated model were positive. The most important result is that the use of the regression analysis technique has given results that are closely compatible with the results of many descriptive studies that dealt with the causes and factors that contribute to the spread of this phenomenon in society.

[1]*corresponding author*

2020 *Mathematics Subject Classification.* 62-08.

Key words and phrases. regression analysis, multiple linear regression, variable selection, stepwise regression, all possible regressions, school dropout.

Submitted: 04.02.2021; *Accepted:* 18.02.2021; *Published:* 16.03.2021.

1. Introduction

School dropout is a phenomenon that affects most educational systems at the global and local levels, and it is one of difficult problems which floats on the educational scene and is as old as the emergence of the school as an educational institution. But, it did not appear as an educational obstacle sociological and cognitive, except in the modern era that the dropout has become unable to be adapted to the developments brought about by technological progress resulted from the rapid development of knowledge. The dropout student has become a wasteful energy that impedes the progress sought by societies. [4].

This study aims to use some statistical techniques such as regression analysis and some methods of selecting variables to determine the most important factors that affect the number of students dropping out of primary schools in some areas of the capital, Baghdad, by analyzing real data on students who leave school for several reasons. It is an attempt to arrive at a statistical model that is used in estimating and predicting relationships between student dropout and the factors causing this phenomenon. Accordingly, we address the concept of regression analysis, the multiple linear regression model, and the most popular methods used in selecting variables to determine the best regression model, depending on some statistical measures used for this purpose. Then some concepts about school dropout and its causes will be identified. Likewise, regression analysis methods will be applied to real data on the dropout students from some primary schools in the capital, Baghdad. The results of the statistical analysis will be discussed in detail. Finally, the study ends with some conclusions that contribute to the analysis and interpretation of the phenomenon of school dropout.

2. Concept of Regression Analysis

In statistics, regression analysis is a statistical technique based on modeling the relationship between a dependent variable (also called a response variable) and one or more independent variables (also known as explanatory variables), where the dependent variable is modeled as a function of the independent variables, the corresponding regression coefficients (coefficients), and the random error term that represents the unexplained variation in the dependent variable through the function. In linear regression, the dependent variable is modeled as

a linear function of a set of regression parameters and random error. The parameters must be estimated so that the model gives the "best fit" of the data. Parameters are estimated by the most commonly used method called the least squares method. If the regression model reflects the real relationship between the dependent variable and the independent variables adequately, then this model can be used to predict the dependent variable, identify the important independent variables, and build a desired causal relationship between the response variable and the independent variables [10].

2.1. **Multiple Linear Regression Model.** Multiple linear regression model is one of the most used statistical analysis tools. Multiple regression is concerned with estimating the relationship between the dependent variable and several independent variables. When there is k of independent variables, the multiple linear regression model can be formulated as follows [8]:

$$y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik} + u_1 \qquad (i = 1, 2, 3 \ldots, n).$$

It can be expressed in the shortened form:

$$y_i = B_0 + \sum_{j=1}^{k} B_i X_{ij} + u_i \qquad (i = 1, 2, 3 \ldots, n),$$

while y_i is the dependent variable and X_1, X_2,..., X_k are the independent variables, B_0 is a constant value expressing the value of y_i when the values of the independent variables are equal to zero. B_1, B_2, ..., B_k are the regression coefficients associated with the independent variables. Therefore, the number of parameters in the model is equal to p, $(p = k + 1)$. As for n the number of observation for the variables. Using the matrix notation, the multiple linear regression model can be expressed as follows:

$$Y = X\beta + u$$

While Y is a vector (n $\times$ 1) of the observed responses, X is a matrix with a (n $\times$ p) dimension of constant values, β is a vector (p $\times$1) of the unknown fixed parameters and u is a vector (n $\times$1) of (unobserved) random errors with a zero mean. In this model the quantities u and y are random vectors (ie, vectors of random variables). The error vector u is considered random according to our assumption. Since y depends on u, it is also random. And when the assumptions

of the regression model are fulfilled, using the method of least squares, it yields the parameter vector estimator with the following formula:

$$b = (X'X)^{-1}X'Y,$$

where b is the vector (p $\times$1) of the estimated parameters. i.e. : b_0, b_1, b_2, b_2, ..., b_k.

2.2. **Variable Selection Methods.** One of the important topics in regression analysis is reducing the number of explanatory variables that are used in the final model, which leads to less effort, time and cost, as well as ensuring ease of analysis and understanding. Therefore, there must be a balance between the process of reducing the explanatory variables and increasing their number to obtain accurate predictive results. It is better to choose the model with the least number of explanatory variables, so that these variables are important and have an impact on the dependent variable. There are several methods of selecting the variables in regression analysis. The following are the most commonly used methods for this purpose: [5].

(1) The forward regression method.
(2) Backward regression method.
(3) Stepwise regression method.
(4) All possible regressions method.

In the forward selection method, the backward deletion and the stepwise regression method, the comparison will be made by F-test of the hypotheses about the total and partial sources of variance, while in the method of all possible regressions the comparison between the models is done through some statistical criteria. In this study, two of the aforementioned methods will be used, which are the all possible regression method and the stepwise regression method.

- **All possible regression method:**

It depends on finding all possible models (their number equals 2^p -1). Which contain one explanatory variable and even a number (k) of explanatory variables. Models that contain the explanatory variables are placed in groups, so the number of groups is k, and some of those models may be found according to what the researcher deems necessary, which gives the largest amount of information related to the nature of the relationship between the dependent variable and the set of explanatory variables. In order to select the variables, this method

is characterized by relying on experience and the use of relevant analytical results and some statistical comparison criteria, such as the coefficient of determination (R_P^2) and the adjusted coefficient of determination (Adj.R_P^2) and the mean square error (MSE_P) for each model. The comparison between models is done through the values of these criteria, that the model with the highest value for the coefficient of determination and the adjusted coefficient of determination is the best, and the model with the lowest value for the mean squares of error is the best [7].

- **Stepwise regression method:**

In this method, all the independent variables that entered the equation are calculated for a partial F-statistic in each step and evaluated on its basis again because our early selection of one of the independent variables sometimes may give a partial F-statistic less than the tabular value of F in the later stages, due to the presence of a strong relationship between it and one of the other independent variables, that was chosen in the equation, and this method needs two tabular F values, namely F_{in} and F_{out}. This method is also called the sequential steps method, combining the forward selection method for entering the variables into the model and the backward deletion method that works to delete variables that have no statistical significance, and for this reason, it has two tabular values, one is F_{in} and the other is F_{out} [6].

3. The Concept of School Dropout

School dropout has been defined by many authors. Abd al-Dayem [1] was defined it that a student left school for some reason before the end of the last year of the educational stage in which he was registered, this student is not considered a dropout if he leaves a certain educational stage after its end and does not belong to the next stage. Whereas, Baftoom [3] defined it as the student's cessation of studies at a certain stage of study before completing the study there. As for the Arab Organization for Education, Culture and Science, [9], the dropout was defined as the discontinuation of school from a certain stage before completing studies there. Accordingly, from all this, school dropout is a student who stops school before completing it for any reason (except for death) and does not enroll in any other school.

The school dropout is due to many reasons, including economic, educational, family and personal reasons. There are several studies and sources that dealt in detail with these reasons. Here we deal with the most prominent of them as follows [2]:

(1) Economic reasons such as: the low living standards of students' families, which drives some of them to leave school in search of low-wage jobs to support their families.
(2) Educational reasons, including what is attributed to the teacher in failing to take into account individual differences among the pupil, being overburdened with many homework, using physical and moral punishment, and using traditional methods of teaching and evaluation. One of the reasons is related to the school, such as the inability of some school administrations to provide affordable services to students at an appropriate cost, the failure to introduce modern technological educational methods to the school, the absence of supervision over the performance and behavior of the teacher, the overcrowding of classrooms with students, and negligence in following the absence of students who have dropped out of school.
(3) Reasons related to the student himself such as: health and physical factors, family factors such as family disintegration, illiteracy of one or both parents, lack of family follow-up for children, cruelty and corporal punishment of children in the family, some customs and traditions of some families in depriving girls from education and far from family housing for school.

4. Application to the Dropout Students

'Data were collected from 25 primary schools from different regions of the capital, Baghdad, in cooperation with the Directorate of Education to obtain some data necessary for the study. Due to the lack of data on the number of students dropping out of schools and its reasons, data were collected in this study by means of investigation, observation and personal interviews, in order to reach the data required for the study.

As for the number of dropout students, the number of students registered in the school and the number of students graduating for the same group were monitored. Where the accreditation was the year 2012-2013 school entry and the completion of primary school in 2017-2018. With regard to the reasons for students leaving their schools, an interview method was used with a group of teachers, social researchers and school assistants. After recording all the information required for the study, the variables were defined as follows:

Y: Dependent variable representing the number of students who drop out of school.

As for the explanatory variables, they were defined as follows:

X_1: Repeated failure of the student.
X_2: Immigration (relocation of family housing).
X_3: Poverty (low standard of living for the family).
X_4: Lack of safety (kidnapping children, explosions, etc.).
X_5: Bad relationship between student and teacher.

After defining the variables, depending on the SPSS 25 package, the regression analysis technique was applied using all the possible regression method and the stepwise regression method to obtain the results of the statistical analysis of the study.

5. Results and Discussion

Table 1 displays the results of applying the method of all possible regressions. The table shows all models, the variables included in each model, the number of parameters in each model and the estimated values of the parameters for each model. Five explanatory variables were used (k = 5), so the number of parameters would be six (p = 6), and therefore the number of models estimated would be 31 models, (2^p-1 = 2^5-1 =31). We see in Table 1 that five models contain one explanatory variable, ten models contain two explanatory variables, ten other models contain three explanatory variables, five models contain four explanatory variables, and finally one model contains all the five explanatory variables included in the study. The table also contains the estimated parameter values for all models.

Table 1: All possible models with entered variables and coefficient estimates

Model no.	Variables in the Model	Estimates of Coefficients					
		b_0	b_1	b_2	b_3	b_4	b_5
1	X_1	25.403	1.170				
2	X_2	17.740		2.173			
3	X_3	25.028			2.364		
4	X_4	27.907				4.184	
5	X_5	28.788					2.152
6	X_1X_2	8.263	.908	2.055			
7	X_1X_3	10.165	1.239		2.422		
8	X_1X_4	22.824	.529			3.766	
9	X_1X_5	11.767	1.359				2.380
10	X_2X_3	15.289		1.651	1.272		
11	X_2X_4	17.802		1.562		2.220	
12	X_2X_5	14.205		1.892			1.312
13	X_3X_4	19.330			1.818	3.342	
14	X_3X_5	21.770			1.904		1.250
15	X_4X_5	23.008				3.522	1.390
16	$X_1X_2X_3$	4.202	1.027	1.442	1.458		
17	$X_1X_2X_4$	10.537	.694	1.655		1.555	
18	$X_1X_2X_5$	2.214	1.079	1.695			1.580
19	$X_1X_3X_4$	11.219	.772		1.964	2.664	
20	$X_1X_3X_5$	5.061	1.342		1.879		1.486
21	$X_1X_4X_5$	13.759	.846			2.702	1.709
22	$X_2X_3X_4$	15.241		.987	1.331	2.326	
23	$X_2X_3X_5$	13.189		1.565	.960		1.002
24	$X_2X_4X_5$	14.780		1.420		1.866	1.118
25	$X_3X_4X_5$	17.843			1.589	3.101	.729
26	$X_1X_2X_3X_4$	6.477	.813	1.041	1.459	1.557	
27	$X_1X_2X_3X_5$	.472	1.132	1.313	1.091		1.241
28	$X_1X_2X_4X_5$	3.916	.949	1.504		.848	1.460
29	$X_1X_3X_4X_5$	7.432	.929		1.665	2.180	1.047

30	$X_2X_3X_4X_5$	13.687		.996	1.093	2.071	.744
31	$X_1X_2X_3X_4X_5$	2.469	.976	1.064	1.139	1.034	1.080

Table 2 shows the values of the three criteria corresponding to each model. It is noticed in the table that if one variable is entered into the model, the three criteria indicate that the best model that contains the variable X_2 has the highest value for the two criteria coefficient of determination, the adjusted coefficient of determination, and the lowest value for the mean square error. In the case of entering two variables, the best model that contains two variables X_3 and X_4, is because this model has the highest value for the coefficient of determination, the adjusted coefficient of determination and the lowest value for the mean square error. If three variables are entered, the best model that contains the variables X_1, X_2 and X_5 will have the highest value of the coefficient of determination, the adjusted coefficient of determination and the lowest value for the mean squares error. In the case of entering four variables, the best model contains the variables X_2, X_1, X_3 and X_5, according to the values of the three criteria. Finally, if all variables are used, the values of the three criteria indicate the preference for this model over the rest of the estimated models.

Table 2: Criteria values of all possible models with entered variables

Model no.	Variables in the Model	P	Criteria		
			R_p^2	Adj. R_p^2	MSE_p
1	X_1	2	.189	.154	65.675
2	X_2	2	.642	.627	28.983
3	X_3	2	.471	.448	42.833
4	X_4	2	.521	.500	38.788
5	X_5	2	.303	.273	56.409
6	X_1X_2	3	.754	.732	20.837
7	X_1X_3	3	.683	.654	26.868
8	X_1X_4	3	.554	.514	37.719
9	X_1X_5	3	.555	.515	37.673
10	X_2X_3	3	.741	.718	21.888
11	X_2X_4	3	.738	.714	22.180

12	X_2X_5	3	.744	.721	21.663
13	X_3X_4	3	.778	.758	18.766
14	X_3X_5	3	.556	.515	37.622
15	X_4X_5	3	.635	.601	30.936
16	$X_1X_2X_3$	4	.883	.866	10.417
17	$X_1X_2X_4$	4	.795	.765	18.202
18	$X_1X_2X_5$	4	.898	.883	9.059
19	$X_1X_3X_4$	4	.848	.826	13.487
20	$X_1X_3X_5$	4	.801	.772	17.679
21	$X_1X_4X_5$	4	.714	.673	25.356
22	$X_2X_3X_4$	4	.847	.825	13.608
23	$X_2X_3X_5$	4	.795	.766	18.177
24	$X_2X_4X_5$	4	.810	.782	16.883
25	$X_3X_4X_5$	4	.805	.778	17.255
26	$X_1X_2X_3X_4$	5	.924	.908	7.119
27	$X_1X_2X_3X_5$	5	.963	.956	3.419
28	$X_1X_2X_4X_5$	5	.909	.891	8.456
29	$X_1X_3X_4X_5$	5	.901	.881	.215
30	$X_2X_3X_4X_5$	5	.875	.850	11.656
31	$X_1X_2X_3X_4X_5$	6	.980	.975	1.962

Table 3 summarizes the results of the best estimated models by entering the variables in each model. When comparing the best models, we notice that the criteria values are in favor of the model that contains all the variables. Since it is noticed that the highest values of the two criteria, coefficient of determination adjusted coefficient of determination are with model no. 5 in the table, the same is the case for the criterion of mean squares error, as the lowest value for this criterion was in model no. 5, which contains all the explanatory variables.

As a result of all possible regressions method, it is not possible to neglect any of the variables that were identified in this study as they are related to the phenomenon of school dropout. Thus, the best model for representing the phenomenon of school dropout out of 31 models that was estimated is the following model:

$$y = 2.469 + .976X_1 + 1.064X_2 + 1.139X_3 + 1.034X_4 + 1.080X_5$$

TABLE 3. Best models with entered variables according to the criteria values

Model no.	Variables in model	P	R^2_P	Adj.R^2_P	MSE_p
1	X_2	2	.642	.627	28.983
2	X_3, X_4	3	.778	.758	18.766
3	X_1, X_2, X_5	4	.898	.883	9.059
4	X_1, X_2, X_3 , X_5	5	0.963	.956	3.419
5	X_1, X_2, X_3 , X_4, X_5	6	0.980	.975	1.962

In order to confirm this results, the stepwise regression method was applied to the same variables, and the results were as follows:

Table 4 shows the values of the correlation coefficient between the dependent variable y and the five explanatory variables, where it is noticed that all the correlation values are positive and significant. The strongest correlation was with the variable X_2, equal to 0.801, with a significant level less than 0.01. Therefore, this variable will be the first candidate variable to enter the model.

Table 5 shows the variables included and excluded from the model according to the stepwise regression method. The sequence of entering the variables in the model is evident. It is noticed from the results that the first variable entered the model is the variable X_2, followed by the variable X_1, then the variable X_5, after the variable X_3 and finally the variable X_4. Therefore, all the variables entered the model and none of the variables were excluded, as is evident in the removed variables column in the table. In other words, the variables were entered into the model according to their importance, leading to the final model, where all the variables were entered. This result is consistent with the result of all possible regressions method that chose the model that contains all five variables.

Therefore, the final estimated model in this way was the same model obtained by the previous method as follows:

$$y = 2.469 + .976X_1 + 1.064X_2 + 1.139X_3 + 1.034X_4 + 1.080X_5$$

By analyzing the results of the regression analysis technique for the study variables using the two methods, it is clear from the above estimated multiple regression model that the number of students dropouts (the dependent variable y) is affected by the explanatory variables used in this study as follows: The

TABLE 4. Pearson correlation between variables and its significances

Variables		Y	X_1	X_2	X_3	X_4	X_5
Pearson Correlation	Y	1.000	.435	.801	.686	.722	.551
	X_1	.435	1.000	.129	-.036-	.367	-.115-
	X_2	.801	.129	1.000	.521	.588	.308
	X_3	.686	-.036-	.521	1.000	.275	.418
	X_4	.722	.367	.588	.275	1.000	.321
	X_5	.551	-.115-	.308	.418	.321	1.000
Sig.(1-Tailed)	Y	.	.015	.000	.000	.000	.002
	X_1	.015	.	.270	.432	.036	.292
	X_2	.000	.270	.	.004	.001	.067
	X_3	.000	.432	.004	.	.091	.019
	X_4	.000	.036	.001	.091	.	.059
	X_5	.002	.292	.067	.019	.059	.

increase in the repeated failure of the student X_1 increases his dropout by 0.976 units of the number of dropouts, the increase in the change of family housing X_2 increases by 1.062 units of dropouts, increases in family poverty X_3 increases 1.139 units of dropouts, increases insecurity X_4 increases by 1.034 units of dropouts and finally increases bad relationship between students and teachers X_5 increases by 1.080 units of student dropouts.

The following are some other statistical indicators about the quality of the final model. Table 6 represents the ANOVA table for the final model. The value of F statistic notes the significance of the final model at the level of 0.01, where

TABLE 5. Variables entered \removed in the models

Model no.	Variables Entered	Variables Removed	Method
1	X_2	-	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	X_1	-	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	X_5	-	Stepwise (Criteria: Probability-of-F-to-enter <=.050, Probability-of-F-to-remove >= .100).
4	X_3	-	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
5	X_4	-	Stepwise (Criteria: Probability-of-F-to-enter <=.050, Probability-of-F-to-remove >= .100).

the value of F was equal to 186.103, which is greater than the tabular value at the level of 0.01. Which indicates the significant linear relation between the dependent variable and the five explanatory variables.

TABLE 6. ANOVA table of the final model

Score of variance	Sum of Squares	d.f.	Mean Square	F	Sig.
Regression	1825.368	5	365.074	186.103	0.000
Residual	37.272	19	1.962		
Total	1862.640	24			

Table 7 represents the estimated parameters in the final model, along with some statistical indicators related to the evaluation of the estimates. It is noted from the table that all the estimates are highly significant, as all t-test values are significant at a level of 0.01 for all variables in the model, which indicates the importance and influence of the variables in the dependent variable. It is also noticed that the values of all the estimates are positive, which indicates their existence positively affecting the increase in students' dropout from schools.

TABLE 7. Coefficients Estimates and Some Statistical Indicators

Variables	Coefficients	Std. Error	t	Sig.
Constant	2.469	1.497	1.649	.116
X_1	.976	.098	9.992	.000
X_2	1.064	.123	8.657	.000
X_3	1.139	.139	8.198	.000
X_4	1.034	.260	3.982	.001
X_5	1.080	.147	7.320	.000

6. CONCLUSIONS

In this study we dealt with the use of regression analysis technique and some methods of selecting variables in determining the factors that affect students' dropout from primary schools. After applying all possible regression method and the method of stepwise regressions to the variables that were entered into the regression model which represents the number of students dropped out of primary schools in Baghdad, it was found that they affect positively and significantly in the model. The results showed that the factors of repeated failure, poor living of families, change in family housing, lack safety and the bad relationship between the student and the teacher are factors that have a significant impact on the phenomenon of school dropout, and any increase in these factors increases the number of students who drop out of schools. All these results were closely compatible with the results of many descriptive studies, which indicated that all these factors are important reasons for students' dropout from schools.

REFERENCES

[1] A. ABD AL-DAYEM: *Educational Planning, Dar Al-Alam Al-Malayn*, 2nd Edition, Beirut, 1972.

[2] A.S. AL-NASSER: *School dropout, the open path to child labor*, The National Library, Amman, **56** (2014), 87–92.

[3] S.A. BAFTOOM: *Reasons for students dropping out of the English language department in the College of Education in Al-Mahra from the students' own point of view*, Journal of Anbar University for Human Sciences, **3** (2012), 788–767.

[4] S. BRAHIMI, D. BURHLI : *Effectiveness of a counseling program (cognitive-behavioral) in reducing school dropouts among fourth year students*, Psychological and Educational Studies, **11**(2) (2018), 60–82.

[5] S. CHATTERJEE, A. HADI: *Regression Analysis by Example*, Fourth Edition, John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

[6] D.C. MONTGOMERY, A. PECK, G.G. VINING: *Introduction to Linear Regression Analysis*, John Wiley & Sons, Inc., 2012.

[7] J.O. RAWLINGS, S.G. PANTULA, D.A. DICKEY: *Applied Regression Analysis: A Research Tool*, Second Edition, Springer-Verlag New York, Inc., 1998.

[8] A.C. RENCHER, G.B. SCHAALJE: *Linear Models in Statistics*, 2nd ed., John Wiley & Sons, Inc., 2008.

[9] THE ARAB ORGANIZATION FOR EDUCATION, CULTURE AND SCIENCE: *the circle of students dropping out in the primary education stage held in Algeria in the year (1972)*, the General Authority for the Amiri Press, Cairo, 1973.

[10] X. YAN, X.G. SU: *Linear Regression Analysis: Theory and Computing*, World Scientific Publishing Co. Pte. Ltd., 2009.

COLLEGE OF ADMINISTRATION AND ECONOMICS
UNIVERSITY OF KIRKUK. IRAQ.
Email address: sabahsaqi@uokirkuk.edu.iq

COLLEGE OF MEDIA
UNIVERSITY OF BAGHDAD, IRAQ.
Email address: Narjishadi@comc.uobaghdad.edu.iq