Journal of Namibian Studies, 33 S2(2023): 1661–1681 ISSN: 2197-5523 (online)

Sample Size in Multiple Regression Models: A simulation study

Dr. Essa J. H. Al-Harbi

Associate Professor of Measurement and Evaluation, Islamic University of Madinah, Saudi Arabia, Alhrbi909@gmail.com

Abstract

The aim of the research is to study the efficiency of the following indicators of the multiple regression model, i.e., the value of "F", R2, Adj., R2, Standard Beta B, Unstandard Beta B, and EMS, in the light of different sample sizes using three methods of multiple regression (standard – gradual – hierarchical). The descriptive method was used. The research community consisted of a virtual community that simulates reality, and it was obtained through the simulation method. It consists of 500 items, created using Mics. Excel program, which are observations that represent one dependent variable denoted by the symbol (Y), and five independent variables, coded (X1, X2, X3, X4, X5), with the aim of evaluating the efficiency of the indicators of multiple regression models (standard - progressive - hierarchical), in light of the different number of samples, range from ($10 \leq n \geq 500$). It was noticed that the conditions of multiple regression were not available in samples less than 50, so the indicators of regression models were studied with sample numbers ranging from (50 to 500). The results indicated that all the multiple regression models indicators increase and improve with increasing the number of the sample, and the best estimate for the regression models was when (n = 225), at a rate of (45) for each independent variable, as the value of R2 was (43%), and was equal to the value of the modified R2, and it was All indicators are good, and by increasing the sample from that, and until reaching the entire population (n=500), the percentage of improvement in the indicators was small, and the value of R2 was (45%). The results also indicated that there is no reliance on only one indicator to judge the quality of the multiple regression models, as well as the difference in the efficiency of the indicators according to the number of the sample, and the statistical method used, especially in the number of small samples. The study recommended using large samples in multiple regression studies, relying on more than one indicator to know the efficiency of regression models, and taking into account the regression method used according to the importance of the researcher's variables.

Keywords: Multiple regression models, sample size, simulation.

Introduction

Multivariate regression is an important tool in analyzing multivariate data when there is more than one response variable and more than one explanatory variable (Sperandei, 2014). It includes data analysis in many sciences such as educational, psychological, social, economic, engineering, agricultural and natural sciences. Regression analysis is a statistical analysis model commonly used to study many phenomena. It is a mathematical measure of the average relationship between two or more variables in terms of units of measurement for the explanatory variables in the relationship (Muhammad & Hussein, 2019). It can also predict the value of a random variable from a variable or group of variables, to help us perform analysis, judge future events, and obtain reasonable decision results. It is widely used in practical problems (Kang & Zhao, 2020).

The multiple regression model is sometimes called the general linear model. It is an extension of the simple model as it includes more than one independent variable. In the case of the simple model, the matter depends on two variables, a dependent variable and the other an independent variable, but in the case of the general model, it may include a number of variables, among which there may be one function and many independent variables (Sheikhi, 2012). Al-Salma (2020) defines multiple regression as a regression that is used to predict a future phenomenon through more than one independent variable.

Stamovlsis (2010) indicated the explanatory power of the stepwise multiple regression model for the change that occurs in the response variable, which is attributed to the explanatory variables. It is considered one of the most important and powerful statistical methods that are used in predictive studies that are conducted in order to predict a certain phenomenon through a group of factors that contribute to its occurrence. Al-Salma's (2020) study also confirmed that multiple regression analysis is significant and has the ability to predict, and statistical thinking can be predicted. Al-Jazzar (2021) confirmed that the polynomial regression model is free from all the standard problems of regression models. The results of the study of Qasim and Ismail (2013) were encouraging in identifying outliers in the data in a multivariate linear regression model.

Many researchers in the psychological, educational and social sciences use many statistical measures (Al-Salma, 2020; Stevens, 2012). The most important of which are prediction measures (Al-Salma, 2020; Dong & Taslimitehrani, 2015), including regression analysis, which is used in prediction studies where it is required to predict the relationship between variables and estimate their parameters. The selection of the research sample at this study is vital because it provides us with data that can be relied upon to generalize the results to the community. Therefore, the degree of accuracy in the data obtained by the researcher

from the sample depends on the appropriate number of the sample and the method of selection. The multiple regression model relies on a large number of data samples in the prediction process (Kang & Zhao, 2020). Regression models are greatly affected by the size of the sample, which affects the predictive ability of the regression model and the accuracy or credibility of the decision that is made (Midi et al., 2010). Therefore, when building a regression model to determine the form of the relationship, the optimal size of the independent variables is determined to obtain certain values of the dependent variable (Al-Ghamdi, 2013; Assas, 2019; Sabeel, 2015). Assas's study (2019) confirmed that any increase in the sample size is accompanied by an increase in the F value, taking into account that the largest change in the F value began when using a sample size of 150 cases.

Attia (2004) indicated that the literature in the field of psychological and educational measurement summarized three main methods for conducting linear regression analysis, and they differ among them in the method of arranging the introduction of independent variables in the different regression equations. The multiple regression model, which measures the effect of more than one variable on the dependent variable is the most widely used and has several methods (Assas, 2019). We also find that the multiple regression analysis method is used for three main purposes:

1. Description: The derivation of an equation that describes the form of the relationship between the explanatory variables and the dependent variables, (we try to choose the least number of explanatory variables that constitute the most essential part of the dependent variable variance) and that equation is used to find out the indicators only,

2. Estimation and prediction: The derivation of an equation that can be used to predict new values for the dependent variable (Y) through actual or expected values for a group of multiple independent variables, as this equation determines whether the independent variables contribute to explaining the change or difference in the values of the dependent variable (Y), or not,

3. Control: It is the interpretation of the variation or change in the values of the dependent variable in terms of the change in the values of the variable or the independent variables, on the grounds that the independent variables are a control, and the aim of building the model is to determine the size with which the independent variable must be modified to obtain certain values for the dependent variable (Chatterjee & Hadi, 2012).

The aim of regression analysis

Multiple regression models are among the most widely used statistical methods in research with non-experimental designs, and they are widely used in various research fields. The dependent, by means of one or more

independent variables, called the explanatory, explanatory, or independent variables. It also aims to evaluate the impact of the independent variables on the dependent variables (Henrik, 2010). It also aims to study the relationship between one or more dependent variables and an independent variable (Ayan & Garcia, 2008).

By mathematical formula:

$$Y_i = \beta_0 + \beta_{1xi1} + \beta_2 x_{2i} + \varepsilon_i$$

Where Y1 is the dependent variable and (Xi1; Xi2) are the independent variables, 1 is the random error, 0 is a constant value that expresses the value of y when the values of (Xi1; Xi2) are equal to zero (1, 2) represents regression coefficients for the independent variables.

Methods for performing linear regression analysis

The literature on educational measurement and evaluation indicated that there are three main methods for conducting linear regression analysis (Al-Akhdar, 2022; Qasim & Ismail, 2013), and they differ among themselves in the method of arranging the introduction of independent variables in the different regression equations, namely:

Hierarchical

Where predictive variables are chosen based on previous studies and the researcher's decision. As a general rule, the variables agreed upon in previous studies are added first based on their importance and ability to predict the dependent variable, then followed by the new variables that will be added from the researcher's point of view (Midi et al., 2010). The most important and then the important ones can be added.

Forced entry

Predictive variables are pushed exclusively into the model, which is similar to hierarchical regression, as it depends on previous studies in its selection of variables, but it differs from the hierarchical. As the researcher has no authority in the order of entering the variables. This method is the only suitable way to test theories.

Stepwise methods

In which the independent variables are included one after the other according to a mathematical criterion suggested by the same method. Three sub-methods are included in this method:

 Forward: In which the model contains only the regression constant, then a selection is made from among the independent variables most closely related to the dependent variable.

 Stepwise: It is similar to the forward entry method, except that in the multi-step method, each time a new variable is added to the equation, the regression analysis is re-worked to ensure the feasibility of the new

variable added to the predictive ability of the model, and to ensure that there are no redundant variables that do not add to the predictive ability of the model.

 Backward: It is the opposite of the forward method of input, where the program begins by adding all the predictive variables in the model and then calculates the contribution of each of them in predicting the dependent variable, and compares this value to the deletion criterion. seconds for the remaining predictive changes.

The posterior method is better than the forward method because of the suppressor effects, and the omitted effects mean that the predictor variable has a significant effect on stabilizing an independent variable only. The forward method is more likely than the back method to omit variables within the model that have omitted effects, so the forward method increases the probability of type II error (Almhairat & Al-Quraan, 2019).

Regression models

There are different types of regression analysis. Which allows us to evaluate the effect of the dependent variables on the independent variable. Regression models can be divided into the following three sections (Alexopoulos, 2010):

Standard regression

In this method, we enter the independent variables into the regression equation at once to obtain the equation that describes the relationship between all the independent variables and the dependent variable once without discussing whether all the independent variables should be included in the equation or not? We do not discuss whether the independent variables are related to each other or independent.

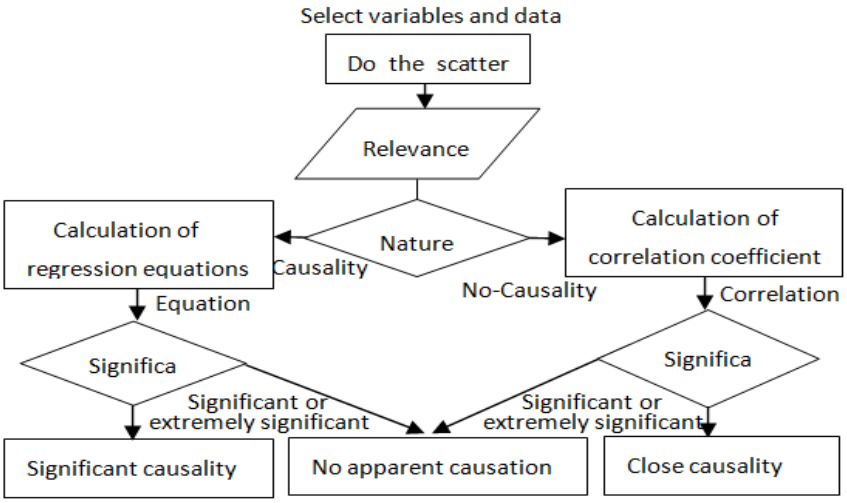
Hierarchical regression

In hierarchical regression, the independent variables are entered into the proposed equation successively, and we determine the order of entry of these variables into the proposed equation on a theoretical statistical basis.

Gradual regression

In the stepwise regression model, the number of independent variables entered into the model, as well as the order of their entry, is determined by a statistical criterion that is reached by the stepwise regression procedure (Al-Akhdar, 2022). These models differ in two respects: the first in dealing with overlapping differences due to the correlation of the independent variables, and the second in the order of entering the independent variables into the equation.

Figure 1. Regression models



Research objectives

The current research aims to:

1. Identify the degree of influence of the sample size on the efficiency of the standard multiple regression model.

2. Identify the degree of effect of sample size on the efficiency of the stepwise multiple regression model.

3. Identify the degree of effect of sample size on the efficiency of the hierarchical multiple regression model.

Research problems

The progressive multiple regression method is commonly used in the field of educational and psychological sciences, as it aims to reduce the number of the many independent variables that affect the phenomenon, to the lowest possible number so that it has a large explanatory ratio for the variance from the values of the dependent variable, and thus increase the predictive ability of the dependent variable through a model Regression. Which contains some independent variables, and the current research came to answer the following questions:

1. What is the effect of sample size on the efficiency of the standard multiple regression model?

2. What is the effect of the sample size on the efficiency of the stepwise multiple regression model?

3. What is the effect of the sample size on the efficiency of the hierarchical multiple regression model?

Significance of study

The study at hand dealt with a statistical method commonly used in educational and psychological research, and it contributes to indicating the efficiency of the multiple regression model (standard - gradual - hierarchical) in determining the proportion of explained variance, as well as the effect of sample size on the proportion of variance explained by the multiple regression model.

Methodology

Research design

In light of the problem and objectives of the research, the descriptive approach was used to describe the extent of change in the indicators of multiple regression models (standard - gradual - hierarchical) in light of the different sample sizes.

Population

The research community consisted of a virtual community that simulates reality and was obtained through the simulation method, and it consisted of (500) items, which were generated using the Excel program, and they are observations that represent (1) a dependent variable symbolized by the symbol (Y), (5) independent variables, were coded as follows: (X1, X2, X3, X4, X5), in order to evaluate the efficiency of multiple regression models (standard - stepwise - hierarchical) in light of the different sample sizes, ranging from ($10 \leq n \leq 500$).

To ensure the suitability of the study population for the multiple regression models, the availability of multiple regression conditions was tested as follows:

1. The linearity of the relationship between the dependent variable and the independent variables.

Table 1 Pearson correlation coefficients between the dependent variable and the independent variables (n = 500)

	y	**x_1**	**x_2**	**x_3**	**x_4**	**x_5**
Y	1	0.517	0.438	0.407	0.375	0.523
x_1		1	0.024	0.109	0.104	0.117
x_2			1	0.114	0.068	0.069
x_3				1	0.119	0.048
x_4					1	0.065
x_5						1

Table 1 indicates that there is a statistically significant correlation at a level of significance less than (0.05) between the dependent variable and each of the independent variables, and all the values of the Pearson correlation coefficients are greater than (0.3), where the values of the Pearson correlation coefficients ranged from (0.375) for the correlation between the dependent variable (Y) and the independent variable (x4) to (0.523) for the correlation between the dependent variable (Y) and the independent variable (x5).

2. No autocorrelation between the independent variables and some of them:

Table 1 indicates that there is no strong correlation with more than (0.7), as the highest correlation coefficient value between the independent variables was equal to (0.119) between the two variables (x3, x4), which is less than (0.7), so we will keep all generated variables.

3. Tolerance

It expresses the amount of variation of the specified independent variable that is not explained by other independent variables in the model and should not be less than the allowable limit (0,1), as the lack of permittivity of (0,1) means that the multiple correlation with other variables is high, which increases the probability of achieving multiple linear accompaniment, and the results were as follows:

Table 2 Tolerance values to ensure non-linearity

Independent variables	Tolerance
x_1	0.973
x2	0.922
x3	0.895
x4	0.966
x5	0.901

Table 2 shows that all tolerance values are greater than (0.1), and therefore this assumption is available in the generated data.

4. Variance inflation factor (VIF)

The variance inflation coefficient is an indicator of the presence of multiple collinearity and should not exceed the allowable limit (10), as an increase in VIF above (10) means an increased probability of achieving the multicollinearity, and the results were as follows:

Table 3. Variance inflation coefficient values to ensure non-linearity

Independent variables	VIF
x1	1.322
x2	1.230
x3	1.029

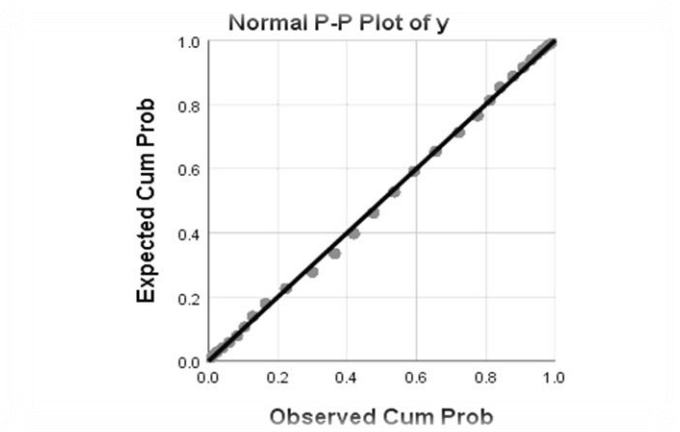
x4	1.187
x5	1.209

Table 3 and Figure indicate that all values of the variance inflation coefficient are less than (10), and therefore this assumption is available in the generated data.

5. Standardized residual

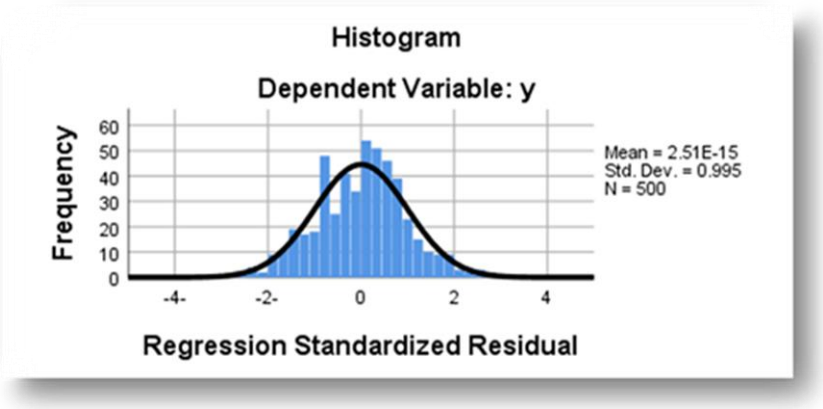
The availability of this assumption has been confirmed through the (Normal P-P Plot) and the standard distribution of the residuals. The following figures show the availability of this condition.

Figure 2. The line of moderation of the values of the dependent variable



It is noted from the previous figure that all points lie on or close to the equinox line

Figure 3. The normal distribution of the residuals



It is noted from the previous figure that the residuals are distributed in a "normal" standard.

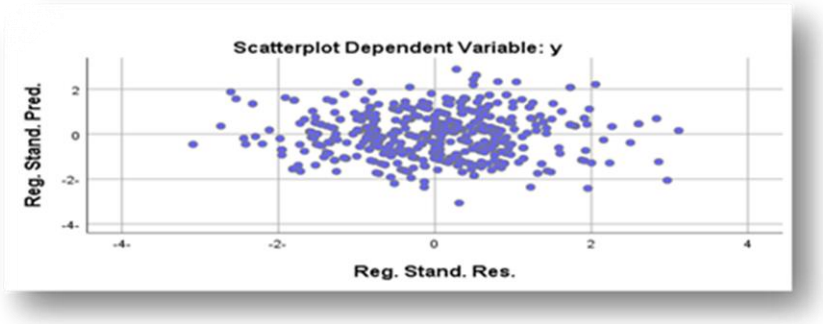
5. No outliers

This assumption has been verified in two ways as follows:

1. Using scatterplot,

Extreme points can be defined as cases with standard residuals of not more than 3.3 or less than -3.3, bearing in mind that in large samples it is okay to have a few of them and it is not necessary to take any measures. Also, the residuals must be distributed in a rectangular shape so that most of the values are concentrated in the middle (along the zero point) and do not take the curved or inclined pattern in which one side is significantly higher than the other, as the deviation from the rectangular shape indicates a violation of the outlier values hypothesis.

Figure 4. Diffusion diagram of the residuals



2. Mahalanobis distance

The results of this method indicated that none of the Mahalanobis values exceeded the critical value (which is determined by the value of chi squared at the degrees of freedom corresponding to the number of independent variables) and since the number of independent variables is equal to (5), so the critical value of the Mahalanobis values is equal to (11.07) and the following table Descriptive statistics show the Mahalanobis values.

Table 4 Descriptive statistics of Mahalanobis values

pointer	**Mahalanobis**
lower value	1.002
the largest value	1.1420
Arithmetic mean	1.028

standard deviation	0.161
number of values	500
number of independent variables	5
critical value	11.07

From all the previous results, it can be said that there is availability of all the conditions for the multiple regression models, the appropriateness of the statistical vocabulary that was generated for use in the current study, and the reliability of the results.

Sample

Random samples were selected from the statistical observations, which were generated using different sample sizes, ranging from ($10 \leq n \leq 500$). It consists of observations of one dependent variable, which is symbolized by the symbol (Y), and five independent variables, which are symbolized by symbols (X1, X2, X3, X4, X5). with the aim of studying the efficiency of multiple regression models (standard - gradual - hierarchical) in light of the different sizes of samples, ranging from ($10 \leq n \leq 500$), in which the assumptions of the regression models are available.

Instrument

The study tool is an analysis of a table of contents, consisting of a set of rows representing cases (individuals), numbering (500), and a set of columns representing independent variables, numbering 5.

Validity and reliability of the tool

In simulation studies, the study tool does not need to calculate the validity and reliability indicators, when it is not a tool in the form of a questionnaire or test, but rather the nature of the data and its suitability to the statistical methods in the current study, which gives credibility to the results obtained.

Statistical tests

To answer the study questions, the following statistical methods were used:

1. Pearson correlation coefficient.
2. permittivity.
3. Contrast inflation coefficient.
4. Equinox line.
5. Distribution of leftovers.
6. Diffusion drawing of the residue.

7. Mahalanops haciendas.

8. F-value

9. R2

10. Adj. R2

11. βStand

12. β Unstand

13. EMS

Results and discussion

RQ1: What is the effect of sample size on the efficiency of the standard multiple regression model?

Sample sizes were chosen from 10 individuals to 500 individuals, and each time the efficiency of the standard multiple regression model was studied through five indicators (the "F" value of the significance of the regression model "F", the estimation coefficient "R2", the estimation coefficient Average "Adj. R2", standard partial regression coefficients "Standard Beta B", partial regression coefficients "B Unstandard Beta", error mean squares "EMS"). Before starting the study of the regression model, it was confirmed that its conditions were met in the size of the selected sample, and it was noted that the conditions for its use were not met when the sample size was small (10, 20, 30, 40).

The effect of sample size on the efficiency of the standard multiple regression model was studied starting from the sample size (50), then (75, 100, 125, 150, 175, 200, 225). It was noted that with an increase in the sample sizes from (225), the rate of change of the improvement of the regression model indicators was small. So (300, 400, 500) were chosen.

Table 5 The effect of sample size on the efficiency of the standard multiple regression model

Sample	F	Sig	R^2	ADJ. R^2	β values (shaded values mean that they are statistically significant)						EMS
					non-normative	X_1	X_2	X_3	X_4	X_5	
					normative	X_1	X_2	X_3	X_4	X_5	
50	1.47	0.22	0.14	0.05	non-normative	0.07	0.94	0.68	0.18	0.33	5.99
					normative	0.02	0.31	0.23	0.06	0.12	
75	3.55	0.01	0.20	0.15	non-normative	0.31	0.91	0.88	0.54	0.45	5.48

					normative	0.10	0.29	0.23	0.18	0.16	
100	6.75	0.001	0.26	0.23	non-normative	0.56	0.86	0.87	0.50	0.54	5.03
					normative	0.20	0.29	0.31	0.17	0.19	
125	10.22	0.001	0.30	0.27	non-normative	0.70	0.86	0.78	0.55	0.58	4.97
					normative	0.25	0.39	0.25	0.19	0.20	
150	13.91	0.001	0.33	0.30	non-normative	0.82	0.79	0.64	0.47	0.62	4.95
					normative	0.29	0.26	0.29	0.23	0.20	
175	18.89	0.001	0.36	0.34	non-normative	0.82	0.72	0.93	0.83	0.77	4.84
					normative	0.29	0.24	0.32	0.27	0.25	
200	24.29	0.001	0.39	0.37	non-normative	0.75	0.87	0.98	0.83	0.80	4.79
					normative	0.26	0.29	0.34	0.27	0.26	
225	34.78	0.001	0.43	0.43	non-normative	0.77	0.90	1.04	0.83	0.87	4.62
					normative	0.27	0.30	0.37	0.27	0.29	
300	45.19	0.001	0.43	0.43	non-normative	0.90	0.95	10.01	0.87	0.82	4.61
					normative	0.31	0.32	0.36	0.27	0.26	
400	67.89	0.001	0.44	0.44	non-normative	0.86	1.03	0.96	0.92	1.00	4.57
					normative	0.29	0.33	0.33	0.29	0.31	
500	91.72	0.001	0.45	0.45	non-normative	0.84	1.02	1.01	0.95	1.04	4.42
					normative	0.29	0.33	0.35	0.30	0.33	

Table 5 indicates that when the sample size was 50, that is, an average of (10) individuals for each independent variable, the regression model was not good, as the value F was (1.47), which is not statistically significant, and the ratio of explained variance R2 (14%), and there is a significant difference between it. The modified R2 value (5%), and it was noted that there was one independent variable with a statistically significant effect (x2), and the highest value of the error contrast ratio (EMS), which is (5.99).

By increasing the sample size to 75, that is, at a rate of (15) for each independent variable, the regression model improved relatively, as the F value reached (3.55), which is statistically significant, and the proportion of explained variance R2 rose to (20%), and the difference between it and the value of R2 decreased. modified (15%), and it was noted that there were two independent variables with a statistically significant

effect (x2, x3), and the value of the error variance ratio (EMS) decreased, and it became (5.48).

The same trend was observed in the improvement of all indicators of the quality of the standard multiple regression model with increasing sample size.

The researcher believes that the results obtained in the sample (225), i.e. at a rate of (45) for each independent variable, were good and sufficient, as the R2 value reached (43%), and became equal to the modified R2 value (43%), and the percentage of improvement in the indicators of The standard regression model is slightly more than the sample (225), and this can be seen in the sample (500), that is, the entire population, the value of R2 is (45%), that is, an improvement rate of (3%) only compared to the sample size (225).

To finish up, it can be said that the size of small samples does not fit standard multiple regression models, whether in terms of the availability of its conditions, or the efficiency of the model in explaining the phenomenon under study. And the most appropriate sample is (45). These results are consistent with the findings of Al-Ghamdi (2013).

RQ2: What is the effect of the sample size on the efficiency of the stepwise multiple regression model?

Sample sizes were chosen from 10 to 500, and each time the efficiency of the standard multiple regression model was studied through five indicators (the “F” value of the significance of the regression model “F”, the estimation coefficient “R2”, the modified estimation coefficient " Adj. R2", Standard Partial Regression Coefficients "Standard Beta B", Partial Regression Factors Nonstandard "B Unstandard Beta", Error Mean Squares "EMS"). Before starting to study the regression model, it was confirmed that its conditions were met in the size of the selected sample, and it was noted that the conditions for its use were not met when the sample was small (10, 20, 30, 40). Therefore, the effect of the sample on the efficiency of the standard multiple regression model was studied, starting with the sample (50) and then (75, 100, 125, 150, 175, 200, 225). So, samples (300, 400, 500) were selected. The results were as follows:

Table 6 Indicators of the efficiency of the stepwise multiple regression model according to the size of the samples

Sample No	F	Sig	R^2	ADJ. R^2	β values (shaded values mean that they are statistically significant)						EMS
					within the form	X_1	X_2	X_3	X_4	X_5	
					out of form	X_1	X_2	X_3	X_4	X_5	
50	4.47	0.04	0.09	0.03	within the form		0.89				5.93
					out of form	0.03		0.21	0.02	0.08	
75	5.78	0.01	0.14	0,08	within the form		0.80	0.87			5.59
					out of form	0.09			0.18	0.16	
100	7.30	0.001	0.24	0,20	within the form	0.51	0.87	0.85		0.54	5.03
					out of form				0.17		
125	10.22	0.001	0.30	0.27	within the form	0,70	0.85	0.78	0.55	0.85	4.97
					out of form						
150	13.91	0.001	0.33	0.30	within the form	0.82	0.79	0.64	0.74	0.62	4.95
					out of form						
175	18.89	0.001	0.36	0.34	within the form	0.82	0.72	0.93	0.83	0.77	4.84
					out of form						
200	24.29	0.001	0.39	0.37	within the form	0.75	0.87	0.89	0.83	0.80	4.79
					out of form						
225	34.78	0.001	0.43	0.43	within the form	0.77	0.90	1.04	0.83	0.87	4.62
					out of form						
300	34.19	0.001	0.43	0.43	within the form	0.90	0.95	1.01	0.87	0.82	4.61
					out of form						
400	67.89	0.001	0.44	0.44	within the form	0.86	1.03	0.96	0.92	1.00	4.57
					out of form						
500	91.72	0.001	0.45	0.45	within the form	0.84	1.01	1.01	0.95	1.04	4.42
					out of form						

Table 6 indicates that when the sample was 50, that is, at a rate of (10) for each independent variable, the regression model was appropriate, only according to the standard of the F value (4.47), which is statistically significant at (0.05), while the rest of the indicators The other indicated that the regression model is not good, as the rate of explained variance was R2 (9%) only, and there is a significant difference between it and the modified R2 value (3%), and it was noted that there was only one independent variable in the model with a statistically significant effect (x2), and it was excluded Four independent variables from the model (x1, x3, x4, x5), and the highest value of the error variance ratio (EMS), which is (5.93). Accordingly, the researcher believes that in the case of small samples, relying on the value of F only as an indicator of the efficiency of the regression model is wrong and inaccurate, and other

indicators must be considered for accurate judgment on the efficiency of the multiple regression model.

By increasing the sample to (75), that is, at a rate of (15) for each independent variable, the regression model improved relatively, as the F value reached (5.78), which is statistically significant at (0.01), while the rest of the indicators are still low, as the ratio of explained variance became (R2). 14%, and the difference between it and the modified R2 value decreased (11%), and it was noted that there were two independent variables with a statistically significant effect (x2, x3), and the value of the error variance ratio (EMS) decreased and became (5.59).

In light of this, the researcher believes that despite the improvement in the values of the model's indicators, the reliance on the value of F only as an indicator of the efficiency of the regression model is wrong and inaccurate, and other indicators must be considered for accurate judgment on the efficiency of the multiple regression model.

The same trend was observed in the improvement of all indicators of the quality of the gradual multiple regression model by increasing the sample, and the researcher noted that starting from the sample (125), i.e. at a rate of (25) for each independent variable, the five independent variables (x1, x2, x3, x4, x5) were introduced. In the model because it is statistically significant, none of the variables were excluded, and therefore the results of all indicators in the progressive multiple regression model were similar to the results of the standard multiple regression model.

In light of this, the results obtained for the sample (225), i.e. an average of (45) for each independent variable, were good and sufficient, as the R2 value reached (43%) and became equal to the modified R2 value (43%), and the improvement rate became In the indicators of the standard regression model are few, with an increase in the sample from (225) individuals, and this can be seen in the sample (500), I, e. the entire population, the value of R2 is (45%), i.e. an improvement rate of (3%) only compared to the sample (225).

From the foregoing, it can be said that small samples are not suitable for progressive multiple regression models, whether in terms of the availability of its conditions, or the efficiency of the model in explaining the phenomenon under study, and also the necessity of not relying on the F index only to judge the quality of the model, and when using gradual multiple regression one must take into account Considering the ratio between the number of individuals and the number of independent variables under study, and that the most appropriate sample size is (45) individuals/independent variable. These results are consistent with the findings of Assas (2019) that any increase in the sample size is accompanied by an increase in the F value, taking into

account that the largest change in the F value began when using a sample size of 150 cases.

RQ3: What is the degree of influence of the sample size on the efficiency of the hierarchical multiple regression model?

Samples were selected from 10 to 500 individuals, and each time the efficiency of the standard multiple regression model was studied through five indicators, which are the value of "F", "R2", "Adj.R2", "Standard Beta B", "B Unstandard Beta", "EMS".

Before starting to study the regression model, it was confirmed that its conditions were met in the selected sample, and it was noted that the conditions for its use were not met when the sample was small (10, 20, 30, 40).

Therefore, the effect of the sample on the efficiency of the standard multiple regression model was studied, starting with the sample (50), then (75, 100, 125, 150, 175, 200, 225). The slopes were small, so sizes (300, 400, 500) individuals were chosen. The results were as follows:

Table 7 Indicators of the efficiency of the hierarchical multiple regression model according to the size of the samples

Sample No	F	Sig	R^2	ADJ. R^2	β values (shaded values mean that they are statistically significant)						EMS
					within the form	X_1	X_2	X_3	X_4	X_5	
					out of form	X_1	X_2	X_3	X_4	X_5	
50	2.24	0.09	0.07	0.03	within the form		0.97				5.91
					out of form	0.01		0.61	0.05	0.12	
75	3.87	0.01	0.18	0.13	within the form		0.89	0.81			5.52
					out of form	0.30			0.56	0.16	
100	8.12	0.001	0.19	0.20	within the form	0.51	0.87	0.85			5.18
					out of form				0.49	0.19	
125	10.22	0.001	0.30	0.27	within the form	0.70	0.85	0.78	0.55	0.58	4.97
					out of form						
150	13.91	0.001	0.33	0.30	within the form	0.82	0.79	0.64	0.74	0.62	4.95
					out of form						
175	18.89	0.001	0.36	0.34	within the form	0.82	0.72	0.93	0.83	0.77	4.84
					out of form						
200	24.29	0.001	0.39	0.37	within the form	0.75	0.78	0.98	0.83	0.80	4.79
					out of form						
225	34.78	0.001	0.43	0.43	within the form	0.77	0.90	0.1.04	0.83	0.87	4.62
					out of form						
300	45.19	0.001	0.43	0.43	within the form	0.90	0.95	1.01	0.87	0.82	4.61
					out of form						

400	67.89	0.001	0.44	0.44	within the form	0.86	1.03	0.96	0.92	1.00	4.57
					out of form						
500	91.72	0.001	0.45	0.45	within the form	0.84	1.02	1.01	0.95	1.04	4.42
					out of form						

Table 7 shows that when the sample was 50, with a rate of (10) for each independent variable, the regression model was not appropriate according to all indicators, as the value of the F indicator was (2.24), which is not statistically significant at (0.05), and it was The ratio of the explained variance is R2 (9%) only, and there is a difference between it and the modified R2 value (7%). It was noted that there was only one independent variable in the model with a statistically significant effect (x2), and four independent variables were excluded from the model (x1, x3, x4, x5), and the highest value of EMS, which is (5.91).

Accordingly, the researcher believes that in the hierarchical regression model, small samples are not good for accurate judgment on the efficiency of the hierarchical multiple regression model.

By increasing the sample to (75), at a rate of (15) for each independent variable, the regression model improved relatively, as the F value reached (3.78), which is statistically significant at (0.01), while the rest of the indicators are still low, as the rate of explained variance became R2 (18%).) and there is a difference between it and the modified R2 value (13%).

It was noted that there were two independent variables with a statistically significant effect (x2, x3), and the value of the error variance ratio (EMS) decreased, and became (5.52).

In light of this, the researcher believes that despite the improvement in the values of the model's indicators, the reliance on small samples for the efficiency of the regression model is still wrong and inaccurate.

The same trend was observed in the improvement of all indicators of the quality of the hierarchical multiple regression model by increasing the sample, and the researcher noted that starting from the sample (125) at a rate of (25) individuals for each independent variable, the five independent variables (x1, x2, x3, x4, x5) were entered into the model was statistically significant, and variables were not excluded.

Thus, the results of all indicators in the hierarchical multiple regression model were similar to the results of the standard multiple regression and progressive multiple regression models. The results obtained in the sample (225), with an average of (45) for each independent variable, were good and sufficient, as the R2 value reached (43%), and became equal to the modified R2 value (43%). The sample is about (225), and this can be seen in the sample (500), the entire community, the value of R2 is (45%), an improvement rate of only (3%) compared to the sample (225).

From the foregoing, it can be said that the size of small samples does not fit the hierarchical multiple regression models, whether in terms of the availability of its conditions, or the efficiency of the model in explaining the phenomenon under study (Alexopoulos, 2010).

Conclusion

Multiple regression models are among the most widely used statistical methods in research with non-experimental designs, and they are widely used in various fields of research. This paper conducted a review of the efficiency of multiple regression models (standard - stepwise - hierarchical) in light of the different number of samples ranging from (10 ≤ n≤500) in which the assumptions of the regression models are available, and random samples were selected from the statistical observations that were generated using the number of samples It consists of observations of one dependent variable, which is designated by the symbol (Y), and five independent variables, which are symbolized by (X1, X2, X3, X4, X5).

The results indicated that in the standard regression when the sample was (50) with an average of (10) for each independent variable, the regression model was not good, by increasing the number of the sample to (75) with an average of (15) for each independent variable, the regression model improved relatively, and the same trend was observed in All indicators of the quality of the standard multiple regression model improved by increasing the sample. As for the gradual regression, when the sample was (50) with an average of (10) for each independent variable, the regression model was appropriate, while the rest of the other indicators indicated that the regression model was not good. By increasing the sample to (75) with an average of (15) for each independent variable, an improvement The regression model is relatively, while the rest of the indicators are still low. In light of this, the results obtained from the sample (225) at a rate of (45) for each independent variable were good and sufficient. As for the hierarchical regression, when the sample was (50) with an average of (10) for each independent variable, the regression model was not appropriate according to all indicators. By increasing the sample to (75) with an average of (15) for each independent variable, the regression model improved relatively. The results obtained when The sample (225), with an average of (45) for each independent variable, was good and sufficient. It can be said that the small samples do not fit standard, gradual, or hierarchical multiple regression models, either in terms of the availability of its conditions, or the efficiency of the model in explaining the phenomenon under study.

The study recommends researchers to use large samples in multiple regression studies in the future. It is also recommended to rely on more

than one indicator to know the efficiency of regression models, and take into account the regression method used according to the importance of the variables in this study.

Bibliography

Al-Akhdar, N. (2022). Multiple linear regression. Algeria: University of Kasdi Merbah Ouargla.

Alexopoulos. E. (2010). Introduction to multivariate regression analysis. Hippokratia. 14(1), 23-28.

Al-Ghamdi. A. (2013). The effect of sample size on the predictive power of the standard multiple regression model, (Unpublished Master Thesis). Umm Al-Qura University.

Al-Jazzar, F. (2021). Multiple regression model as a treatment for standard problems: an empirical study on the relationship between the economic growth rate and the inflation rate in the Egyptian Economy. Journal of Financial and Business Research, 4, 1-32.

Almhairat, L. & Al-Quraan, M. (2019). The effectiveness of using multiple regression models in predicting the variables contributing to success in the corresponding (regular) courses of Yarmouk University students. Journal of Educational Science Studies, 46(2), 68-79.

Al-Salma, A. S. (2020). Multiple linear regression model for predicting statistical thinking in light of some variables. Al-Manara Journal for Research and Studies. 26(1), 171-195.

Assas, F. (2019). Studying the explained variance ratio in the stepwise multiple regression model in the light of different sample sizes. Journal of Scientific Research in Education, 9(20), 317-380.

Attia, A. (2005). Modern econometrics between theory and practice. Alexandria: University House.

Ayan, M. & Garcia, M. (2008). Prediction of university students' academic achievement by linear and logistic models. The Spanish Journal of Psychology, 1, 275.288.

Chatterjee, S & Hadi, A. (2012). Regression analysis by example, (5th Edition). Cnada: WILEY.

Dong G. & Taslimitehrani V. (2015). Pattern-aided regression modeling and prediction model analysis. IEEE Transactions on Knowledge and Data Engineering, 27(9), 2452-2465.

Henrik, M. (2010). Introduction to general and generalized linear models. UK, Chapman and Hall/CRC Press.

Kang, H. & Zhao. H. (2020, September). Description and application research of multiple regression model optimization algorithm based on data set denoising. Journal of Physics: Conference Series 1631(1), p. 012063.

Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. Journal of interdisciplinary mathematics, 13(3), 253-267.

Muhammad, M. & Hussein, H., (2019). Addressing the effects of linear duality in the multiple regression model. Fayoum J. Agric. Res. & Dev, 33(1) 77-91

Qasim. M., & Ismail. Y. (2013). Detection of outliers in a multivariate linear regression model using JPS sampling. Journal of Education and Science, 26(1) 149-161

Sabeel, A. (2015). The effect of auxiliary variables and sample size on probability sampling estimates, (Unpublished PhD thesis). Sudan University of Science and Technology.

Sheikhi M. (2012). Econometric methods, lectures and applications, (1 st Edition). Dar Al-Hamid for Publishing and Distribution.

Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia medical, 24(1), 12-18.

Stamovlsis, D. (2010). Methodological and epistemological issues on liner regression applied to psychometric variable in problem solving. Chemistry Education Research and Practice, 1, 59-68.

Stevens, J. P. (2012). Applied multivariate statistics for the social sciences. Routledge