Understanding and interpreting regression analysis

Parveen Ali [1,2], Ahtisham Younas [3,4]

10.1136/ebnurs-2021-103425

[1]School of Nursing and Midwifery, University of Sheffield, Sheffield, UK
[2]Sheffiled University Interpersonal Violence Research Group, The University of Sheffiled SEAS, Sheffield, UK
[3]Faculty of Nursing, Memorial University of Newfoundland, St. John's, Newfoundland and Labrador, Canada
[4]Swat College of Nursing, Mingora, Swat, Pakistan

Correspondence to:
Ahtisham Younas, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada; ay6133@mun.ca

Introduction

A nurse educator is interested in finding out the academic and non-academic predictors of success in nursing students. Given the complexity of educational and clinical learning environments, demographic, clinical and academic factors (age, gender, previous educational training, personal stressors, learning demands, motivation, assignment workload, etc) influencing nursing students' success, she was able to list various potential factors contributing towards success relatively easily. Nevertheless, not all of the identified factors will be plausible predictors of increased success. Therefore, she could use a powerful statistical procedure called *regression analysis* to identify whether the likelihood of increased success is influenced by factors such as age, stressors, learning demands, motivation and education.

What is regression?

Regression analysis allows for investigating the relationship between variables.[1] Usually, the variables are labelled as dependent or independent. An independent variable is an input, driver or factor that has an impact on a dependent variable (which can also be called an outcome). For example, if we were to say age affects academic performance of students, what will be the independent and dependent variables here? Well here age is an independent variable, and it has the potential to impact on outcome/dependent variable—in this case, academic performance. Similarly, in the nurse educator's example, critical thinking is a dependent variable and age, experience and training are independent variables.

Purposes of regression analysis

Regression analysis has four primary purposes: description, estimation, prediction and control.[12] By description, regression can explain the relationship between dependent and independent variables. Estimation means that by using the observed values of independent variables, the value of dependent variable can be estimated.[2] Regression analysis can be useful for predicting the outcomes and changes in dependent variables based on the relationships of dependent and independent variables. Finally, regression enables in controlling the effect of one or more independent variables while investigating the relationship of one independent variable with the dependent variable.[1]

Types of regression analyses

There are commonly three types of regression analyses, namely, linear, logistic and multiple regression. The differences among these types are outlined in table 1 in terms of their purpose, nature of dependent and independent variables, underlying assumptions, and nature of curve.[1 3] However, more detailed discussion for linear regression is presented as follows.

Linear regression and interpretation

Linear regression analysis involves examining the relationship between one independent and dependent variable. Statistically, the relationship between one independent variable (x) and a dependent variable (y) is expressed as: $y = \beta_0 + \beta_1 x + \varepsilon$. In this equation, β_0 is the y intercept and refers to the estimated value of y when x is equal to 0. The coefficient β_1 is the regression coefficient and denotes that the estimated increase in the dependent variable for every unit increase in the independent variable. The symbol ε is a random error component and signifies imprecision of regression indicating that, in actual practice, the independent variables are cannot perfectly predict the change in any dependent variable.[1] Multiple linear regression follows the same logic as univariate linear regression except (a) multiple regression, there are more than one independent variable and (b) there should be non-collinearity among the independent variables.

Factors affecting regression

Linear and multiple regression analyses are affected by factors, namely, sample size, missing data and the nature of sample.[2]

- Small sample size may only demonstrate connections among variables with strong relationship. Therefore, sample size must be chosen based on the number of independent variables and expect strength of relationship.
- Many missing values in the data set may affect the sample size. Therefore, all the missing values should be adequately dealt with before conducting regression analyses.
- The subsamples within the larger sample may mask the actual effect of independent and dependent variables. Therefore, if subsamples are predefined, a regression within the sample could be used to detect true relationships. Otherwise, the analysis should be undertaken on the whole sample.

Example

Building on her research interest mentioned in the beginning, let us consider a study by Ali and Naylor.[4] They were interested in identifying the academic and non-academic factors which predict the academic success of nursing diploma students. This purpose is consistent with one of the above-mentioned purposes of regression analysis (ie, prediction). Ali and Naylor's chosen academic independent variables were preadmission qualification, previous academic performance and school type and the non-academic variables were age, gender, marital status and time gap. To achieve their purpose, they collected data from 628 nursing students between the age range of 15–34 years. They used both linear and multiple regression analyses to identify the predictors of student success. For analysis, they examined the relationship of academic and non-academic

Table 1 Comparison of linear, logistic and multiple regression

Linear	Logistic	Multiple
Purpose		
Examines the relationship between one independent variables with one dependent continuous variable	Calculates the likelihood of event with binary outcome (ie, yes or no)	It is an extension of simple linear regression and examines the relationship between one or more independent and dependent variables simultaneously
Nature of dependent and independent variables		
1. Dependent variable should be continuous 2. Independent variables could be at any level of measurement	1. Dependent variable should be categorial 2. Independent variables could be at any level of measurement	1. Dependent variables should be continuous 2. Independent variables could be at any level of measurement
Assumptions		
1. Assumes that the distribution of dependent data is normal or Gaussian 2. Requires a linear relationship between dependent and independent variables	1. Assumes that the distribution of dependent data is binomial. 2. It does not require a linear relationship between dependent and independent variables 3. The independent variables should not be correlated	1. Assumes that the distribution of dependent data is normal or Gaussian 2. Requires a linear relationship between dependent and independent variables 3. The independent variables should not be correlated. Higher correlation among the independent variables may affect the relationship between independent and dependent variable
Nature of curve		
It uses a straight line	It uses an S-curve	It uses a straight line
Example		
Examining the relationship between hours of training and levels of patient self-care and predict how long training should last for every unit increase in self-care levels	Estimating the likelihood of development of pressure ulcers (dichotomous outcome: yes or no) due to longer hospital stay, number of times of positioning, BMI (Body Mass Index) and age	Examining the relationship between hours of training and patient self-care levels while controlling for other variables (eg, family support, duration of disease) that may affect the relationship

variables across different years of study and noted that academic factors accounted for 36.6%, 44.3% and 50.4% variability in academic success of students in year 1, year 2 and year 3, respectively.[4]

Ali and Naylor presented the relationship among these variables using scatter plots, which are commonly used graphs for data display in regression analysis—see examples of various scatter plots in figure 1.[4] In a scatter plot, the clustering of the dots denoted the strength of relationship, whereas the direction indicates the nature of relationships among variables as positive (ie, increase in one variable results in an increase in the other) and negative (ie, increase in one variable results in decrease in the other).

Table 2 presents the results of regression analysis for academic and non-academic variables for year 4 students' success. The significant predictors of student success are denoted with a significant p value. For every, significant predictor, the beta value indicates the percentage increase in students' academic success with one unit increase in the variable.

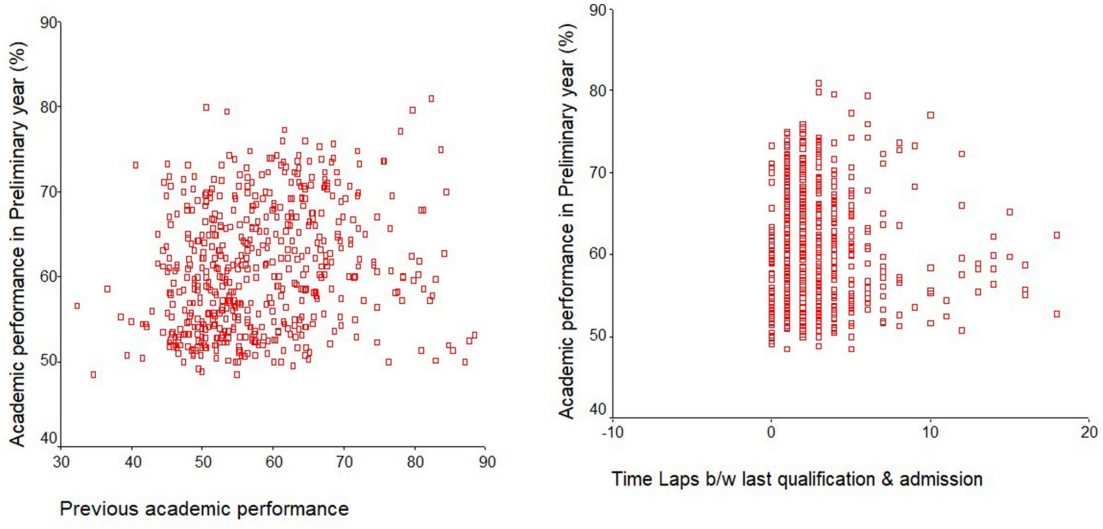

Figure 1 An Example of Scatter Plot for Regression.

Table 2 Regression model for the final year students (N=343)

Variables	β	SE (β)	P value
Academic variables			
Entry qualification			
Intermediate science	5.47	1.054	‹0.001**
Intermediate arts/ commerce	4.436	1.214	‹0.001**
Matric science (ref=matric arts)	2.041	0.81	0.004**
Previous academic performance (%)	0.344	0.057	‹0.001**
Schools type			
Private (ref=public)	5.086	0.667	‹0.001**
Non-academic variables			
Gender			
Male (ref=female)	1.398	4.006	0.727
Place of domicile			
Urban (ref=rural area)	4.268	3.675	0.246
Interaction: gender and previous academic performance	−0.0953	0.065	0.146
Interaction: place of domicile and previous academic performance	−0.0512	0.062	0.401

Interaction between gender and previous academic performance=0.146.

Interaction between place of domicile and previous academic performance=0.401

**p‹0.01.

Conclusions

Regression analysis is a powerful and useful statistical procedure with many implications for nursing research. It enables researchers to describe, predict and estimate the relationships and draw plausible conclusions about the interrelated variables in relation to any studied phenomena. Regression also allows for controlling one or more variables when researchers are interested in examining the relationship among specific variables. Some of the key considerations are presented that may be useful for researchers undertaking regression analysis. While planning and conducting regression analysis, researchers should consider the type and number of dependent and independent variables as well as the nature and size of sample. Choosing a wrong type of regression analysis with small sample may result in erroneous conclusions about the studied phenomenon.

Twitter Parveen Ali @parveenazamali and Ahtisham Younas @@Ahtisham04

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iDs
Parveen Ali http://orcid.org/0000-0002-7839-8130
Ahtisham Younas http://orcid.org/0000-0003-0157-5319

References

1 **Montgomery DC**, Peck EA, Vining GG. *Introduction to linear regression analysis*. John Wiley & Sons, 2012.
2 **Schneider A**, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Dtsch Arztebl Int* 2010;**107**:776.
3 **Hilbe JM**. *Practical guide to logistic regression*. CRC Press, 2016.
4 **Ali PA**, Naylor PB. Association between academic and non-academic variables and academic success of diploma nursing students in Pakistan. *Nurse Educ Today* 2010;**30**:157–62.