Croatian Review of Economic, Business and Social Statistics (CREBSS)
UDK: 33;519,2; DOI: 10.1515/crebss; ISSN 1849-8531 (Print); ISSN 2459-5616 (Online)
Vol. 8, No. 2, 2022, pp. 66-83

sciendo

Conceptual understanding of linear regression among economics students at the university center of Tipaza, Algeria

Djouahra Idris
Centre Universitaire de Tipaza, Algeria, djouahra.idris@cu-tipaza.dz

Abstract
Solving problems related to econometrics requires a good knowledge of regression analysis concepts. The objective of this study is to evaluate students' difficulties resulting from the lack of knowledge of regression analysis concepts among economics students enrolled in the Master's cycle at the institute of economics at the university center of Tipaza (Algeria). In order to analyze students' answers, a typical correction was prepared based on professors' answers to this questionnaire. The procedure consists of comparing students' key answers with their corresponding typical answers to see how near or far it is from the right answer. In order to see whether the difficulties are originated from the same students, we analyzed the association between answers based on Multiple Correspondence Analysis (MCA) method. The principal results showed that difficulties resulting from the lack of knowledge of regression analysis concepts were prevalent among students. Their main causes were strongly related to misunderstanding, misconceptions and confusions. MCA analysis indicated that students can be categorized according to their answers into four groups: a very weak group, a weak group, an average group and a good group. We concluded that the difficulty of solving problems in the context of linear regression among students is the result of a lack of knowledge of regression concepts coupled with the inability to explain them.

Keywords: conceptual understanding, econometrics, economics, linear regression.

JEL classification: A20, C01.
DOI: 10.2478/crebss-2022-0011

Received: November 29, 2022
Accepted: December 5, 2022

Introduction
Teaching regression analysis is considered as a key introduction to teaching econometrics because studying modeling in all pedagogical phases requires an initial knowledge of statistical tools and mechanisms allowing the investigation of the relationship between the dependent variable and one or more independent variables (Angrist, Pischke, 2009). In this context, many empirical economists say that

students are not well prepared for serious problems they will encounter with real data (Swann, 2019). According to (Angrist, Pischke, 2017), econometric instruction remains mostly abstract; focusing on the search for technical concerns associated with estimation and assumptions rather than concentrating on the meaning of concepts (Angrist, Pischke, 2017). This has created a strong confusion between teaching econometrics and teaching regression analysis (Spanos, 1986).

Solving problems related to econometrics requires knowledge of the concepts of regression analysis coupled with the ability to use and explain the various assumptions and characteristics associated with estimation and estimators. It could be difficult initially to know whether students' errors are the result of a lack of knowledge of these concepts, or they are simply due to their inability to deal with and explain methods and steps of estimation. Several studies have shown that the inability to deal with these methods and steps is prevalent among students' difficulties in studying regression analysis. The research hypothesis states that the difficulty of solving problems in the context of linear regression among students is the result of a lack of knowledge of the concepts of regression in general coupled with the inability to explain them.

In this paper, we are interested in the evaluation of students' difficulties resulting from the lack of knowledge of regression analysis concepts, which are mainly related to model assumptions, model fitting, least square estimation, hypothesis testing, model selection and multicolinearity, in addition to their inability to explain them, which may complicate the educational process due to the difference in concepts in one hand and the multiplicity of ways to representing and explaining them in the other hand.

The chapters are structured as follows. The second chapter provides a literature review on the statistical and didactic considerations concerning the conceptual understanding of linear regression. The third chapter describes materials and methods as well as the methodology followed to analyse students' difficulties to answer questions related to regression analysis course. The fourth chapter presents the study results and discussion, while the fifth chapter concludes the study.

Literature review

Linear regression represents both an estimation technique and a computational device. Studying it in an economic context is concerned with the formulation and the use of models (Doran, Doran, 1989), and allows deriving many statistics arising in the context of econometrics (Davidson, MacKinnon, 1993).

In order to say that there is a conceptual attainment, the student must be able to define the concept then to position it in a node of the conceptual network in a coherent way that enables him to solve problems (Giordan, De Vecchi, 1987). To do this in a correct way, students, based on the information obtained from the various lessons, must achieve their "conceptual attainment" by linking the various existing concepts, and consequently they can build a self-knowledge structure that enables them to communicate knowledge in a correct manner (Taber, 2005).

Researches and experiences in this subject clearly show that learning and teaching regression analysis are mainly based on understanding the statistical meaning of different concepts, so that these latter can facilitate the understanding of econometrics (Gujarathi, 2004). It is known that statisticians use these methods with ease, unlike non-statisticians who often encounter difficulties in dealing with them (Madsen, 2016). Many studies have been reported in the literature dealt with misinterpretations in statistical thinking (Capraro et al., 2005; Pfannkuch, Ben-Zvi, 2011; Wild, Pfannkuch, 1999; Ben-Zvi, Garfield, 2004). Boels et al. (2019) and Cooper,

Shore (2008) reported the conceptual difficulties that students, teachers and researchers encounter when interpreting histograms and how to address the common misinterpretations more generally. Other studies explained the difficulties of choosing the kind of graph that intermediate grade students face and how they reason about graphical representations (Agro, 1977; Delmas et al., 2005).

Studies conducted by Birnbaum (1982) and Falk (1986) dealt with misconceptions of statistical significance when analyzing data. In the same context, Batanero et al. (1998), Batanero and Serrano(1999) and Bar-Hillel and Wagenaar (1991) highlighted challenges and difficulties in understanding randomness, its meanings, its perception and its implications. Misconceptions in statistical inference and hypothesis testing have also been reported in the literature through multiple studies. Their main subject concerned the investigation of the most common difficulties among university students and concentrated on the misconceptions that have not yet received much attention (Sotos et al., 2007, 2009; Krishnan, Idris, 2014; Reeves, Brewer, 1980). Studies conducted by Motulsky (2015) and Akobeng (2016) investigated the types of errors as well as the common misconceptions about data analysis and statistical interpretations. There are also studies that highlighted the common misconceptions concerning regression and linearity, with reference to some cautions and considerations to take into account as well as ways to dealing with them (Williams et al., 2013; Bossé et al., 2016; Tompkins, 1993; Hancock, 1965). Multicollinearity also took part among misconceptions faced by students. Authors, in this context, explored the insufficient and the inappropriate treatment of collinearity, and how collinearity issues that lead to spurious and unstable results can be avoided (Lindner et al., 2020).

Materials and methods

The research consists of an exploratory study, through which we sought to know the students' ability to retrieve the concepts of regression analysis course starting from the research hypothesis. The questions aim to assess the students' ability to define the concepts of simple linear regression and multiple linear regression as well as their ability to explain them clearly.

Data collection was done from April 1st 2022 to April 30th 2022 based on a written questionnaire including questions revolving around concepts and topics taught during the third year undergraduate degree which are namely related to simple linear regression (SLR), multiple linear regression (MLR), estimators, hypotheses testing and multicolinearity. The questionnaire was assessed by three professors specialized in statistics and econometrics in order to make sure that the asked questions are clear and not ambiguous to students. It should be noted that Arabic is the teaching language at the institute of economics and it was this language that was used for data collection (questionnaire). Answers to the questionnaire were then translated into English.

The study sample consists of 180 students (59% females and 41% males) enrolled in the Master's cycle at the institute of economics at the university center of Tipaza (Algeria), who have taken the same econometrics course in their third year undergraduate degree. The author has explained to students the answers' conditions (student anonymity, answers are not subject to marks, individual answers), as well as the maximum time of answering that was fixed at 45 minutes. In order to analyze students' answers; a typical correction (typical answers) was prepared relying on professors' answers to this questionnaire. In fact, for each question, the procedure consists of recording students' key answers so that we can classify them under the same category (Robert, Bouillaguet, 2002). Each answer category was then compared with its corresponding typical answer to see how near or far it is from

the right answer. In this context, we distinguished between five categories of answers: successful answers (SA) that are very close to the typical answer, approximate answers (AA) that are considered as right answers but lack some important detail, other answers (OA) that seem to be successful but don't really respond to the asked question and wrong answers (WA) that are completely far from the typical answer. If the student doesn't provide any answer, he will be oriented to the category of non-answers (NA).

In order to see whether the various difficulties are originated from the same students or they are dispersed between students, we analyzed the association between answers based on Multiple Correspondence Analysis (MCA). This method is a data analysis technique that allows the exploration of large set of nominal data by representing them as points in a low dimensional Euclidean space. In our case, MCA allows identifying groups of students with similar profiles in terms of the studied variables (answer categories). Measuring the association between categories of answers is done through the analysis of proximities. To do this, we relied on "test values" in order to measure proximities between the different answer categories. These latter are calculated based on Chi-square distance in MCA. The interpretation of this measure says that when test values corresponding to different variables (answer categories) are very close, this indicates a strong resemblance between the corresponding individuals (students) (Escofier, Pagès, 1998). Graphically, a short distance between two variables is interpreted as a resemblance between the corresponding individuals in terms of the studied variables (Lebart et al., 1995).

Results and Discussion

Didactical results

Results corresponding to students' answers according to each topic are presented in tables 1 to 12.

Q1. The scatter plot is a useful tool before any statistical analysis. Explain!

The typical answer to this question is: "The scatter plot is used to investigate the possible relationship between the independent variable and the response variable. i.e.: if there is a linear relationship or a non-linear relationship between the independent variable and the response variable". The classification of students' answers to question Q1 is presented in table 1.

Table 1 Students' key answers records related to Q1

Answer category		Number	%
AA1	The scatter plot is used to check for a possible relationship between the independent variable and the response variable.	70	38.89
SA1	The scatter plot is used to investigate the possible linear relationship between the independent variable and the response variable.	47	26.11
OA1_1	The scatter plot gives an idea about how variables are correlated	36	20
OA1_2	The scatter plot gives an idea about how variables are associated	20	11.11
NA1	Non-answer	7	3.89

Source: author calculations based on students answers to the questionnaire.

Direct observation of answers shows that most of students (38.89%) could explain the usefulness of using the scatter plot as "a tool that allows checking the possible relationship between the independent variable and the response variable" (AA1), while 26.11% of students improved the answer by adding "linear relationship" to the previous definition (SA1), since it consists of the course of regression analysis that

deals only with linear regression. The other answers show that some students (20%) referred to scatter plot as a tool allowing having an idea about how variables are correlated (OA1_1), while some others (11.11%) evoked the notion of "association" (OA1_2). These are considered as less correct answers because notions of "correlation" and "association" are used in the context of measures (coefficient of correlation of Pearson for example), while the scatter plot is a primary graphical tool that gives an idea about the general tendency of the studied phenomenon. The proportion of non-answers to question Q1 is equal to 3.89% (NA1).

Q2. In the simple linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the error ε_i is random. Explain!

The typical answer is: "This is the first assumption of the simple linear model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Given that the left hand side of this equation $y_i$is always random because we cannot control the dependent variable, the independent variable $x_i$in not always random (sometimes it is a controlled variable). As a result, the quantity $\beta_1 x_i$ is not always random. This means that ε_i is the quantity of the right hand side of the equation that is always random (a random quantity = a random quantity)". The classification of students' answers to question Q2 is presented in table 2.

Table 2 Students' key answers records related to Q2

Answer category		Number	%
AA2	ε_i is a random variable because its values vary according to variations of Y_i	87	48.33
SA2	This is one of the assumptions of the linear regression. ε_i is a random variable because is the error associated to the dependent variable Y_i.	58	32.22
WA2	ε_i is not random because it represents the complement of the information brought by the independent random variable x_i	27	15
NA2	Non-answer	8	4.44

Source: author calculations based on students answers to the questionnaire.

From table 2, we notice that about one half of students (48.33%) refer the randomness of the error $\varepsilon_i$to the randomness of the dependent variable Y_i(AA2). This is an incomplete answer especially that the students didn't mention the mechanisms of this fact as it is shown in the typical answer related to Q2. About one third of students (32.22%) improved the answer by showing that the question Q2 concerns one of the assumptions of simple linear regression (SA2). Concerning wrong answers, 15% of students said that the error ε_i is not random (WA2) and think that the fact of being the complement of the information brought by the independent variable x_i makes form the error ε_i dependent to the independent variable x_i in terms of randomness. In this context, these students think that the independent variable is always random; but the reality says that this latter might be sometimes random and sometimes not random (a controlled variable). The proportion of non-answers to question Q2 is equal to 4.44% (NA2).

Q3. In the simple linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $Cov(\varepsilon_i, \varepsilon_j) = 0$ for every $i \neq j$. Explain!

The typical answer is: "This is the second assumption of the simple linear model. ε_i is the error associated with y_i and ε_j is the error associated with y_j. Since the response y_i is independent (uncorrelated) from the response y_j for every $i \neq j$, the corresponding errors ε_i and ε_j are uncorrelated as well". The classification of students' answers to question Q3 is presented in table 3.

Table 3 Students' key answers records related to Q3

Answer category		Number	%
AA3_1	This is one of the assumptions of the simple linear model that says that errors of the theoretical model are uncorrelated.	72	40
AA3_2	Errors are uncorrelated, because the different observations y_j's are independent.	55	30.56
SA3	This is an assumption of the simple linear model. For a representative sample of data, the errors must be uncorrelated because error correlation leads to the correlation of the corresponding observations y_j's. As a result, we will have an information redundancy and a less representative sample.	38	21.11
NA3	Non-answer	15	8.33

Source: author calculations based on students answers to the questionnaire.

From table 3, we distinguish between two kinds of approximate answers. The first kind of answers (40% of students) explain error uncorrelation just as an assumption of the simple linear model (AA3_1), while the second kind of answers (30.56% of students) are better since they focus on the uncorrelation of observations that leads to the uncorrelation of the corresponding errors (AA3_2). Concerning successful answers, 21.11% of students gave a more innovative explanation to the reason behind error uncorrelation. In fact, in addition to the information brought by the typical answer, students relate the assumption of error uncorrelation to good sampling (SA3) which is really true. The proportion of non-answers to question Q3 is equal to 8.33% (NA3), which is quite considerable proportion.

Q4. In the simple linear model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, is it possible to calculate the expected value of an observation y_i. Explain!

The typical answer is: "It is possible to calculate $E(y_i)$ because of sampling fluctuations, which makes from y_i itself a random variable. As a result, we can calculate its expected value $E(y_i)$". The classification of students' answers to question Q4 is presented in table 4.

Table 4 Students' key answers records related to Q4

Answer category		Number	%
OA4	Yes, mathematically we can calculate $E(y_i)$.	69	38.33
AA4	Yes, the observation y_i fluctuates from a sample to a sample. These fluctuations are summarized by their mean $E(y_i)$.	50	27.78
SA4	Yes, the randomness characterizing y_i causes value fluctuations from a sample to another. This means that y_i is itself a variable and its fluctuations can then be summarized by their mean $E(y_i)$.	38	21.11
WA4	It is not possible, because y_i represents just one value (value of the response variable corresponding to the ith observation).	18	10
NA4	Non-answer	5	2.78

Source: author calculations based on students answers to the questionnaire.

From table 4, we notice that 38.33% of students explain the possibility of calculating the expected value of the observation y_i to mathematical feasibility (OA4). This perception is not really true because being able to apply the mathematical tool on observations must be coupled with the possibility to interpret the resulting outcome statistically (statistical signification). On the other hand, 10% of students were completely wrong because they think that it is not possible to calculate the expected value of y_i and perceive it like just one value (WA4). Concerning approximate answers, 27.78% approach the right answer by evoking the

notion of sampling fluctuation, but they didn't mention notions of 'variable" and "randomness" that are the basis of the calculation of the expected value (AA4). Only 21.11% of students who really satisfy the conditions of the typical answer related to the question Q4 (sampling fluctuations, randomness and variability). This is considered as a category of successful answers (SA4). The proportion of non-answers to question Q4 is equal to 2.78% (NA4).

Q5. Explain the meaning of fitting a simple linear regression model.

The typical answer is: "fitting a simple linear regression model consists of estimating regression parameters of the theoretical model in such a way that minimizes the sum of squares of residuals". The classification of students' answers to question Q5 is presented in table 5.

Table 5 Students' key answers records related to Q5

Answer category		Number	%
OA5_1	Fitting a simple linear regression model consists of estimating the slope and the intercept of the model.	50	27.78
OA5_2	It consists of finding the line that best represents the scatter plot of data.	41	22.78
AA5	Fitting a simple linear regression model consists of estimating regression coefficients in such a way that minimizes the error.	40	22.22
SA5	Fitting a simple linear regression model consists of estimating regression coefficients in such a way that minimizes the sum of squares of residuals.	36	20
NA5	Non-answer	13	7.22

Source: author calculations based on students answers to the questionnaire.

From table 5, we notice that the majority of students tend to limit the definition of "fitting a model" to "parameter estimation" (27.78%) in one side (OA5_1) or to finding "the best line of the scatter plot" (22.78%) in the other side (OA5_2), without showing the mechanism of doing that (minimization of the residual sum of squares). We can see that 22.22% of students provided a good answer by evoking the concept of "minimization", but there is still confusion between "the error" and "the residual" (AA5). About one fifth of students (20%) improved the answer by evoking all the necessary keywords associated to the typical answer of the question Q5 (estimation, minimization, sum of squares of residuals). This latter is considered as the category of successful answers (SA5). The proportion of non-answers to question Q5 is equal to 7.22% (NA5) which is also a considerable proportion.

Q6. Least square estimators $\widehat{\beta_0}$, $\widehat{\beta_1}$ are unbiased estimators. Explain!

The typical answer is: "Least square estimators $\widehat{\beta_0}$, $\widehat{\beta_1}$ are unbiased estimators means that the expected values corresponding to these estimators are equal to the real values of the parameters $\beta_0 \; and \; \beta_1$ respectively". The classification of students' answers to question Q6 is presented in table 6.

Table 6 Students' key answers records related to Q6

Answer category		Number	%
WA6_1	Unbiased estimators mean that the estimators converge to the real parameters.	60	33.33
WA6_2	Unbiased estimators are estimators without error.	55	30.55
AA6	Because the estimates of the unknown parameters are, on average, around the real parameters.	52	28.89
NA6	Non-answer	13	7.22

Source: author calculations based on students answers to the questionnaire.

Direct observation of table 6 shows that a large proportion of students (33.33%) perceives "unbiased estimator" as a "convergent estimator" (WA6_1), while other proportion (30.55%) confuses between "bias" and "error" (WA6_2). In this context, we recall that "convergence" means that the chances of observing a difference between the value of the estimator and that of the true parameter are as lower as the number of observations is large. Concerning the "error", it is an unknown quantity associated with the real model, while bias is associated with the estimated model. Concerning approximate answers, 28.89% of students approach the typical answer related to Q6 (AA6). To improve the answer, students should focus on the concept of "expected value" and how to use it in order to define the concept of "unbiased estimator". The proportion of non-answers to question Q6 is equal to 7.22% (NA6) which is a high proportion.

Q7. Least square (LS) estimators $\widehat{\beta_0}$, $\widehat{\beta_1}$ are called linear estimators of β_0 and β_1 respectively. Explain!

The typical answer is: "LS estimators $\widehat{\beta_0}$, $\widehat{\beta_1}$ are called linear estimators because they are linear combinations of the response variables $(y_i's)$".

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})w_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} c_i \times w_i \quad (1)$$

where: $c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and $w_i = y_i - \bar{y}$, $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \widehat{\beta_1}\bar{x}$. Since $\widehat{\beta_1}$ is a linear combination of $y_i's$, $\widehat{\beta_0} = \frac{\sum_{i=1}^{n} y_i}{n} - \widehat{\beta_1}\bar{x}$ is also a linear combination of $y_i's$. The classification of students' answers to question Q7 is presented in table 7.

Table 7 Students' key answers records related to Q7

Answer category		Number	%
WA7	LS estimators $\widehat{\beta_0}$, $\widehat{\beta_1}$ are linear because they are the estimators of the linear parameters β_0 and β_1 respectively.	45	25
OA7	LS estimators $\widehat{\beta_0}$, $\widehat{\beta_1}$ are linear because they are the intercept and the slope of the estimated linear line: $y_i = \widehat{\beta_0} + \widehat{\beta_1}x_i$.	41	22.78
AA7	If we use the mathematical tool, we can prove that LS estimators are linear combinations of the independent variable values.	40	22.22
SA7	$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ *and* $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$ If we look to LS estimates, we can see that $\widehat{\beta_1}$ is linear combination of x_i and y_i. Since $\widehat{\beta_0}$ is directly connected to $\widehat{\beta_1}$, it is linear as well	36	20
NA7	Non-answer	18	10

Source: author calculations based on students answers to the questionnaire.

From table 7, we can detect a lot of misconceptions related to the concept of "linearity". In fact, 25% of students refer the linearity of LS estimators to the fact that the real parameters are linear (WA7). This is not true because the real model is unknown and we cannot attribute certain characteristics to it. All features and properties are attributed to the estimated model and the estimators. We also notice that 22.78% of students refer the linearity of LS estimators to the linearity of the estimated linear model, without giving a detailed explanation about that (OA7). Concerning approximate answers, 22.22% of students said that it is possible to prove the linearity of LS estimators based on the mathematical tool, but they didn't mention the details of calculations (AA7). In the other side, 20% of students improved

the answer by showing that the linearity of the LS estimators is explained by the possibility of representing them as linear combination of the response variable y_i (SA7). The proportion of non-answers to question Q7 is equal to 10% (NA7) and it is the highest among all other answers.

Q8. What are the name and the objective of following hypothesis test? $\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$

The typical answer is: "This is called the test of slope coefficient. It aims to show if there is a significant linear relationship between the independent variable and the dependent variable". The classification of students' answers to question Q8 is presented in table 8.

Table 8 Students' key answers records related to Q8

Answer category		Number	%
OA8_1	This is "the test of effect". It aims to study the effect of the independent variable on the dependent variable.	55	30.55
OA8_2	This is "the test of variation". It aims to study the variation of the dependent variable with respect to variations in the independent variable.	50	27.77
AA8	The objective of the test is to test the linear relationship between the independent variable and the response variable.	59	32.78
NA8	Non-answer	16	8.89

Source: author calculations based on students answers to the questionnaire.

From table 8, we notice a lot of difficulties and misconceptions in both the definition and the interpretation of the hypothesis test evoked in question Q8. In fact, although the above mentioned test allows showing the effect of the independent variable on the response variable, it is not called "the test of effect" as answered by 30.55% of students (OA8_1) or "the test of variation" as answered by 27.77% of students (OA8_2). About one third of students (32.78%) who could really explain the objective of the test "testing the linear relationship" (AA8), but their answers lack the name of the test (test of slope coefficient). The proportion of non-answers to question Q8 is equal to 8.89% (NA8) which is also a considerable proportion.

Q9. In multiple regression model with k regressors, adding an irrelevant regressor variable to the model has an impact on the residual sum of squares (RSS). Explain!

The typical answer is: "The RSS decreases as we add a new regressor variable (either relevant or not). However, if the new added variable is very relevant for the model, the RSS decreases more, but if it is not so much relevant for the model, then RSS decreases less". The classification of students' answers to question Q9 is presented in table 9.

Table 9 Students' key answers records related to Q9

Answer category		Number	%
AA9	The RSS decreases as we add a new irrelevant regressor variable.	92	51.11
OA9_1	The RSS decreases if a relevant regressor variable is added to the model. Otherwise, the RSS stays stable.	35	19.44
OA9_2	The RSS decreases if a relevant regressor variable is added to the model and increases if an irrelevant regressor variable is added to the model.	32	17.77
WA9	Adding an irrelevant regressor variable to the model doesn't have any impact on the RSS.	14	7.78
NA9	Non-answer	7	3.89

Source: author calculations based on students answers to the questionnaire.

According to table 9, we notice that more than one half of students (51.11%) were very close to the typical answer related to the question Q9 (AA9), but without giving more details about the different situations regarding the new added regressor variable (relevant or not relevant). Some other students (19.44%) were right in one part of the answer (when saying that "the RSS decreases if a relevant regressor variable is added to the model") and wrong in the other part (when saying that "the RSS stays stable" when we and an irrelevant variable) (OA9_1). We can also notice that 17.77% of students think that the RSS increases if an irrelevant regressor variable is added to the model. These perceptions are probably caused by misunderstandings of proofs related to hypotheses testing as well as tests related to variable significance in multiple regression modeling. A few proportion of students (7.78%) think that adding an irrelevant regressor variable to the model doesn't have any impact on the RSS (WA9). This is completely a wrong answer and it seems that these students suffer from a major misunderstanding. The proportion of non-answers to question Q9 is equal to 3.89% (NA9).

Q10. To select the best multiple linear regression model, we must check some important criteria. Explain!

The typical answer is: "The model is evaluated according to four criteria: the coefficient of multiple determination (R^2), the adjusted R^2, the mean square of residuals (MSR) and Mallow's statistics". The classification of students' answers to question Q10 is presented in table 10.

Table 10 Students' key answers records related to Q10

Answer category		Number	%
AA10	We can check the model based on the coefficient of multiple determination (R^2).	66	36.67
SA10	The model is evaluated based on the coefficient of multiple determination (R^2) and the adjusted R^2.	56	31.11
WA10	The model is evaluated based on forward selection or backward selection methods.	51	28.33
NA10	Non-answer	7	3.89

Source: author calculations based on students answers to the questionnaire.

From table 10, 36.67% of students tend to limit model checking to the coefficient of multiple determination (R^2) (AA10). Some other students (31.11%) improved this answer by adding the adjusted R^2 instead of R^2 only (SA10). Concerning unsuccessful answers, 28.33% of students demonstrated a strong confusion between criteria evoked in the typical answer related to the question Q10 that are part of the "all possible approach method" and the approach called "the sequential selection" that includes forward selection, backward selection and stepwise selection methods (WA10). This confusion is explained by a deficit of understanding of the course "selection of the best regression model". The proportion of non-answers to question Q10 is equal to 3.89% (NA10).

Q11. What does the coefficient of multiple determination (R^2) measure?

The typical answer is: "the coefficient of multiple determination (R^2) measures the proportion of variability in the response variable which is explained by the regressor variables". The classification of students' answers to question Q11 is presented in table 11.

Table 11 Students' key answers records related to Q11

Answer category		Number	%
SA11	The coefficient of multiple determination (R^2) measures the degree of variation of the response variable with respect to independent variables.	94	52.22
WA11_1	It measures the degree of representation of reality that is done by the regressor variables.	50	27.78
WA11_2	It measures how much regressor variables could represent the studied phenomenon.	27	15
NA11	Non-answer	9	5

Source: author calculations based on students answers to the questionnaire.

From table 11, we notice that more than one half of students (52.22%) succeeded to a large extent in giving a definition to the coefficient of multiple determination by evoking the notion of "variability" (SA11). However, a lot of misconceptions are recorded in this context. In fact, some students (27.78%) perceive the coefficient of multiple determination as a measure allowing the representation of a particular "reality" (WA11_1), while some others (15%) perceive it as a measure allowing the representation of a particular "phenomenon" (WA11_2). These confusions are explained by misunderstandings of the course "measurement of goodness of fit". The proportion of non-answers to question Q11 is equal to 5% (NA11).

Q12. Explain the concept of multicolinearity?

The typical answer is: "In multiple regressions, multicolinearity is a problem that occurs when two or more regressor variables are strongly correlated or linearly dependent". The classification of students' answers to question Q12 is presented in table 12.

Table 12 Students' key answers records related to Q12

Answer category		Number	%
OA12_1	Multicolinearity is a phenomenon that occurs when two or more regressor variables are strongly correlated.	51	28.33
OA12_2	Multicolinearity happens when there is a correlation between two regressor variables.	36	20
WA12_1	Multicolinearity occurs when model errors are dependent	27	15
WA12_2	Multicolinearity is a phenomenon that occurs when the independent variable and the dependent variable are strongly correlated.	16	9
SA12	Multicolinearity is a problem that happens when there is a dependence between two or more regressor variables.	34	18.89
NA12	Non-answer	16	8.89

Source: author calculations based on students answers to the questionnaire.

From table 12, we notice some misconceptions related to the concept of "multicolinearity". In fact, 28.33% of students perceive it as a "phenomenon" (OA12_1), some others (20%) perceive it as a "correlation" between regressor variable (OA12_2). Concerning answers that are completely wrong, 15% of students see that multicolinearity occurs when "model errors" are dependent (WA12_1), while some others (9%) see it as a "strong correlation" between the "independent variable" and the "dependent variable" (WA12_2). These are the result of a misunderstanding related to the course of multicolinearity. We notice that only a few proportion of students (18.89%) that really succeeded to provide a write definition to multicolinearity without any confusion. The proportion of non-answers to question Q12 is equal to 8.89% which is also a high proportion.

Statistical results

In table 13, we present a summary of statistics related to different answer categories.

Table 13 Summary statistics (values in %)

Answer category	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Average percentage
SA	26.11	32.22	21.11	21.11	20.00	0.00	20.00	0.00	0.00	31.11	52.22	18.89	20.23
AA	38.89	48.33	70.56	27.78	22.22	28.89	22.22	32.78	51.11	36.67	0.00	0.00	31.62
OA	31.11	0.00	0.00	38.33	50.56	0.00	22.78	58.33	37.22	0.00	0.00	48.33	23.89
WA	0.00	15.00	0.00	10.00	0.00	63.89	25.00	0.00	7.78	28.33	42.78	23.89	18.06
NA	3.89	4.44	8.33	2.78	7.22	7.22	10.00	8.89	3.89	3.89	5.00	8.89	6.20

Source: author calculations based on students answers to the questionnaire.

As we can see from table 13, the average percentage of approximate answers (AA) is the highest among all answer categories (31.62%) followed by other answer (OA) category with an average percentage of 23.89%. The percentage of successful answers (SA) represents only one fifth of the total answers, which is comparable to the percentage of wrong answers (WA) that is around 18.06%. The average percentage of non-answers to the questionnaire is equal to 6.20%. This relatively high percentage of non-answers is much affected by the considerable percentage of non-answers recorded in questions Q3, Q5, Q6, Q7, Q8 and Q12. Results of table 13 are visualized in figure1.

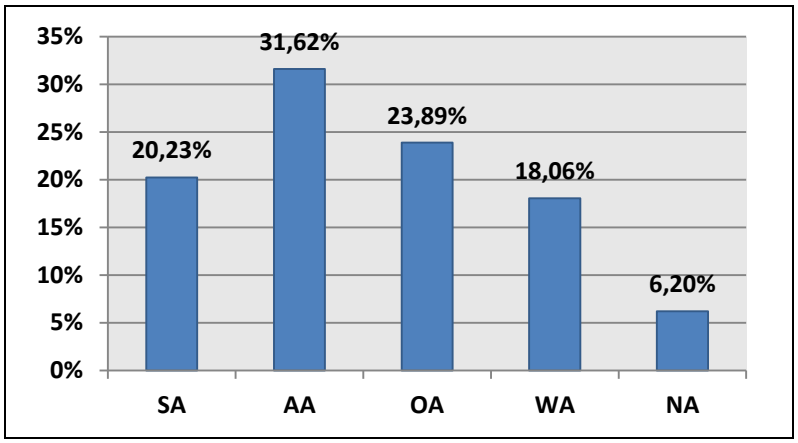

Figure 1 Average proportions corresponding to answer categories

Source: Done by the author based on average proportions presented in table 13.

The output of Multiple Correspondence Analysis (MCA) indicate that the two first factorial axes (F1, F2) contains 80.79% of the total variability of the studied phenomenon. Because of the considerable variation explained by the two first factorial axes, we could neglect the information brought by the rest of the axes. The "test values" of the different answer categories corresponding to factorial axes (F1, F2) are presented in table 14 (results include only values that are significant at 5% level).

The analysis of first factorial axis (F1) indicates that students are categorized according to their answer categories into two different groups; a good group and a weak group. The first group of answers (colored in yellow) includes successful answer (SA) categories and approximate answer (AA) categories where the second group (colored in green) includes wrong answer (WA) categories and other answer (OA) categories (except for questions Q2 and Q3). This means that students, in the first group of answers (the good group), tend to answer correctly questions related to linear regression analysis and, at worst, they tend to approach the right answers

(approximate answers). For the second group (the weak group), students tend to wrongly answer the questions and, at best, they tend to give other responses that don't really answer the asked questions related to linear regression.

Table 13 Test value corresponding to factorial axes F1 and F2

F1		F2	
Answer category	Test value	Answer category	Test value
Q10-SA	11.1896	Q6-NA	10.4549
Q8-AA	11.0517	Q11-NA	9.4667
Q11-SA	10.9823	Q10-NA	9.1372
Q9-AA	10.9100	Q5-NA	8.8795
Q6-AA	10.8635	Q7-NA	8.5064
Q2-SA	9.9818	Q12-NA	8.3012
Q3-SA	9.6752	Q3-NA	6.9755
Q7-SA	9.4385	Q8-NA	6.4975
Q4-SA	9.2366	Q9-NA	6.2530
Q12-SA	9.1343	Q4-NA	5.3447
Q5-AA	7.3783	Q1-NA	3.9228
Q1-SA	6.9947	Q1-SA	3.9013
Q5-SA	6.1643	Q4-SA	3.7848
Q1-AA	3.9420	Q4-WA	3.6870
Q4-AA	3.1816	Q12-SA	3.5459
Q7-AA	2.5371	Q2-WA	2.0590
Q3-AA	-6.8307	Q8-OA	-2.2787
Q2-AA	-7.3863	Q5-AA	-2.5907
Q4-OA	-7.9579	Q7-AA	-2.9640
Q12-WA	-8.5216	Q9-AA	-3.1264
Q7-WA	-9.0223	Q11-SA	-4.5382
Q9-OA	-9.3348	Q3-AA	-5.7256
Q10-WA	-9.3888	Q7-OA	-5.7611
Q1-OA	-9.7335	Q10-AA	-6.0076
Q6-WA	-9.8130	Q1-AA	-6.2564
Q11-WA	-10.9880	Q4-AA	-6.5199
Q8-OA	-11.4858	Q6-WA	-6.5241
Q5-OA	-11.5485	Q12-OA	-8.3973

Source: author calculations bases on ACM analysis.

The analysis of the second factorial axis (F2) indicates two other groups of students according to answer categories; an average group and a very weak group. The first group of answers (colored in brown) includes mostly approximate answer (AA) categories and other answer (OA) categories where the second group (colored in blue) includes mostly non answer (NA) categories and wrong answer (WA) categories (except for questions Q1, Q4 and Q12). This means that students of the first group (the average group) tend, in general, to approximately answer questions related to linear regression analysis. For the second group, students tend to not responding with any answer. This is considered as a very weak group of students. The resulting four groups of students according to their answer categories (SA, AA, OA, WR and NA) are represented graphically in figure 2.

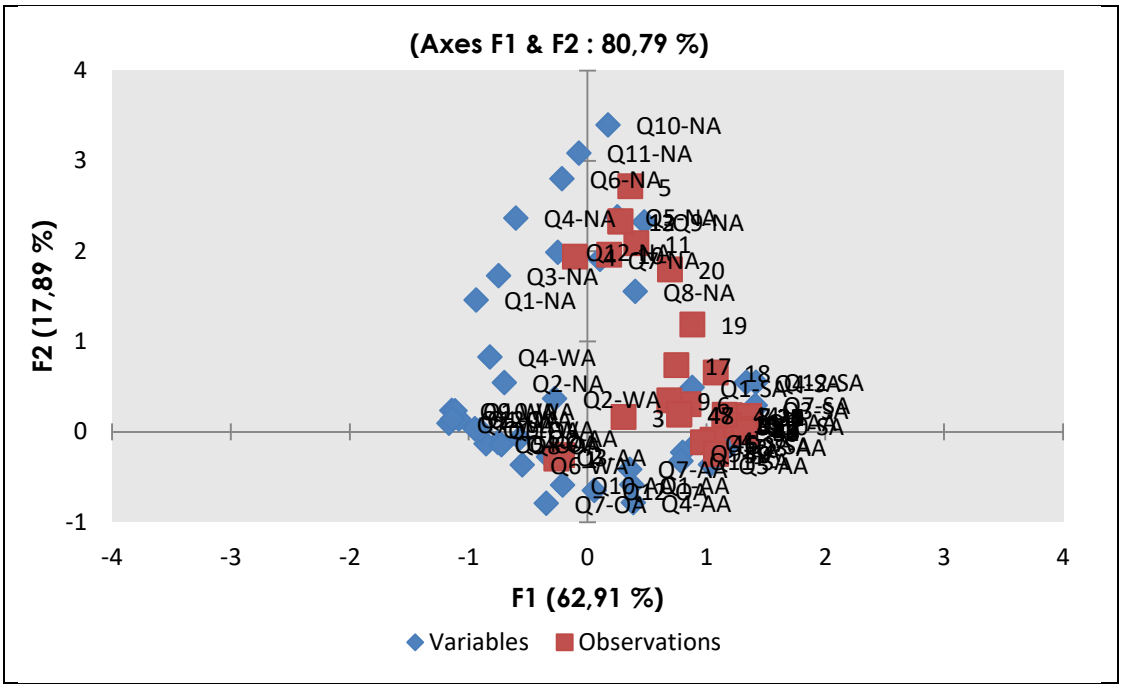

Figure 2 Graph of variables-observations
Source: author calculations bases on ACM analysis.

From figure 2, we clearly notice that the good group of students corresponding to SA and AA answer categories is concentrated in the right hand side of the axis F1, where the weak group of students corresponding to WA and OA categories is mostly concentrated in the left hand side of the axis F1. Concerning the average group of students corresponding to OA and AA categories tend to be concentrated in the lower side of axis F2, where the very weak group of students corresponding mostly to NA categories is concentrated in the upper side of axis F2.

Conclusion

After the presentation of the findings, we can confirm the research hypothesis stating that the difficulty of solving problems in the context of linear regression among students is the result of a lack of knowledge of regression concepts in general coupled with the inability to explain them. According to the difficulty of questions answering, we conclude that the studied students can be classified into four different groups; a good group which tends to answer successfully and doesn't have any difficulty, an average group which tends to answer approximately by facing some difficulties, a weak group which tends to answer wrongly and suffer from a lot of difficulties and a very weak group whose students tend to not answer.

The difficulties resulting from the lack of knowledge of regression analysis concepts were prevalent among students in the different axes of the questionnaire (graphical representation, model assumptions, model fitting, least square estimation, hypothesis testing, model selection and multicolinearity). These difficulties were strongly caused by confusions, misconceptions and misunderstandings, which engendered high proportions of wrong answers (WA), other answers (OA) and non-answers (NA) when answering the questionnaire. The results corresponding to the various axes of the questionnaire are not the first of their kind, as many researches confirm the lack of knowledge that students suffer with regard to the definition of concepts of linear regression analysis as well as the inability to explain them correctly.

The difficulty to give a correct interpretation to the scatter plot is classified among graphical issues. Research indicates that these latter are often misinterpreted by students, teachers and researchers in different fields. In this context, Boels et al. (2019) revealed that the common conceptual difficulties related to constructing and interpreting graphs are dependent to many concepts of statistics such as data and distribution. In addition to that, certain skills and facts must be recognized by students in order to better choose the kind of graph to make (Agro, 1977). Also, when discussing misconceptions that students encounter in analyzing center and variability measures for graphically represented data, Cooper and Shore (2008) concluded that students' misconceptions mostly stem from the difficulty to maintain the acquired understanding as long as possible of the concepts being analyzed.

The misconceptions recorded when answering questions related to model assumptions and hypothesis testing strongly stem from a consistent misunderstanding of statistical inference. A solid understanding of this latter is of crucial importance for fitting an econometric model and interpreting its results (Sotos et al., 2007). Also, when compared with earlier studies (Vallecillos, 2000; Rossman et al., 2004), our results confirm that there is still a considerable number of economics students holding misconceptions about the concepts of bias, error, estimation and multicolinearity. In order to address them, Rossman et al. (2004), suggests initiatives of struggling and wrestling with wrong ideas as well as encouraging students to confront their misconceptions.

Analyzing the problem of non-answers (NA) in the results of a query has proven to be of immense importance (Lee et al., 2018). In fact, providing people querying the collected data with explanations for the non-answers is of great usefulness (Huang et al., 2008). However, giving general explanations to non-answers is not an easy task since these latter can be very large and, thus, may not be very helpful to the desired analysis (Glavic et al., 2015). In our case, the proportion of non-answers (total average of 6,2%) is probably explained by two factors: the first one is that the concerned students have not acquired the knowledge related to regression analysis course because of a severe misunderstanding. The second one is that the concerned students did not attend the courses because of the non-obligation to attend classes in the Algerian university system, which led to a lot of absenteeism and a lack of interest among students.

At the end, we conclude that the passage from teaching regression analysis to teaching econometrics represents a big challenge. Improving the level of understanding and eliminating confusions and misconceptions related to econometrics must start first by eliminating confusions and misconceptions related to regression analysis course. To do this, we first suggest enhancing statistical thinking regarding the use of techniques as well as theory understanding in addition to teaching formulas and proving them. Also, it is important to learn how to understand students' thinking in order to address their confusions. Second, we recommend the development of understanding of concepts related to statistical inference and the necessary prerequisites, and the alignment of regression analysis courses with teaching approaches that are based on interpretations of concepts more than calculations. Last but not the least, adopting strategies based on confronting students with their misconceptions and educating them that holding correct ideas does not necessarily lead to right answers if they are not used in the appropriate contexts (the case of other answers (OA)). Regarding the problem of absenteeism, we suggest developing mechanisms that oblige students to attend lectures in order to ensure the minimum level of cognitive attainment.

Due to time limitations, the questions of the survey were mostly based on the course of simple linear regression. To further investigate students' conceptual difficulties, future research should give priority to questions related to multiple linear regression. The study could also be extended to other courses related to nonlinear regression.

References

1. Agro, S. (1977). *Graphing. USMES Intermediate "How to" Set*. Available at https://files.eric.ed.gov/fulltext/ED220330.pdf [01 August 2022].
2. Akobeng, A. K. (2016). Understanding type I and type II errors, statistical power and sample size. *Acta Paediatrica*, Vol. 105, No. 6, pp. 605-609. DOI: 10.1111/apa.13384
3. Angrist, J. D., Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press. DOI: 10.2307/j.ctvcm4j72
4. Angrist, J. D., Pischke, J. S. (2017). Undergraduate Econometrics Instruction: Through Our Classes, Darkly. *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 125-144. DOI: 10.1257/jep.31.2.125
5. Bar-Hillel, M., Wagenaar, W. A. (1991). The perception of randomness. *Advances in applied mathematics*, Vol. 12, No. 4, pp. 428-454. DOI: 10.1016/0196-8858(91)90029-I
6. Batanero, C., Green, D. R., Serrano, L. R. (1998). Randomness, its meanings and educational implications. *International Journal of Mathematical Education in Science and Technology*, Vol. 29, No. 1, pp. 113-123. DOI: 10.1080/0020739980290111
7. Batanero, C., Serrano, L. (1999). The meaning of randomness for secondary school students. *Journal for Research in Mathematics Education*, Vol. 30, No. 5, pp. 558-567. DOI: 10.2307/749774
8. Ben-Zvi, D., Garfield, J. B. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking*. Dordrecht, The Netherlands: Kluwer academic publishers. DOI: 10.1007/1-4020-2278-6
9. Birnbaum, I. (1982). Interpreting statistical significance. *Teaching Statistics*, Vol. 4, No. 1, pp. 24-26. DOI: 10.1111/j.1467-9639.1982.tb00451.x
10. Boels, L., Bakker, A., Van Dooren, W., Drijvers, P. (2019). Conceptual difficulties when interpreting histograms: A review. *Educational Research Review*, Vol. 28, pp. 1-23. DOI: 10.1016/j.edurev.2019.100291
11. Bossé, M., Marland, E., Rhoads, G., Rudziewicz, M. (2016). Searching for the Black Box: Misconceptions of Linearity. *Chance*, Vol. 29, No. 4, pp. 14-23. DOI: 10.1080/09332480.2016.1263094
12. Capraro, M. M., Kulm, G., Capraro, R. M. (2005). Middle grades: Misconceptions in statistical thinking. *School Science and Mathematics*, Vol. 105, No. 4, pp. 165-174. DOI: 10.1111/j.1949-8594.2005.tb18156.x
13. Cooper, L. L., Shore, F. S. (2008). Students' misconceptions in interpreting center and variability of data represented via histograms and stem-and-leaf plots. *Journal of Statistics Education*, Vol. 16, No. 2, pp. 1-13. DOI: 10.1080/10691898.2008.11889559
14. Davidson, R., MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. Available at https://russell-davidson.arts.mcgill.ca/textbooks/EIE-davidson-mackinnon-2021.pdf [13 May 2022].
15. Delmas, R., Garfield, J., Ooms, A. (2005, July). *Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions*. Available at https://www.causeweb.org/cause/archive/artist/articles/SRTL4_ARTIST.pdf [21 May 2022].
16. Doran, H. E., Doran, H. (1989). *Applied regression analysis in econometrics*. CRC Press.
17. Escofier, B., Pagès, J. (1998). *Analyses factorielles simples et multiples*. Available at https://cdn-cms.f-static.com/uploads/1460418/normal_5b9ba5dc15394.pdf [10 September 2022].
18. Falk, R. (1986). Misconceptions of statistical significance. *Journal of Structural Learning*, Vol. 9, No. 1, pp. 83-96.
19. Giordan, A., De Vecchi, G. (1987). *Les origines du savoir. Des conceptions des apprenants aux concepts scientifiques*. Delachaux et Nestlé, Neuchâtel-Paris.

20. Glavic, B., Köhler, S., Riddle, S., Ludäscher, B. (2015). *Towards Constraint-based Explanations for Answers and {Non-Answers}*. Available at https://www.usenix.org/system/files/conference/tapp15/tapp15-glavic-revised.pdf [3 August 2022].
21. Gujarathi, D. M. (2004). *Gujarati: Basic Econometrics*. Available at http://portal.belesparadisecollege.edu.et:8080/library/bitstream/123456789/3407/1/10.Gujarat.PDF [25 June 2022].
22. Hancock, C. K. (1965). Some misconceptions of regression analysis in physical organic chemistry. *Journal of Chemical Education*, Vol. 42, No. 11, pp. 608-609. DOI: 10.1021/ed042p608
23. Huang, J., Chen, T., Doan, A., Naughton, J. F. (2008). *On the provenance of non-answers to queries over extracted data*. Available at https://pages.cs.wisc.edu/~jhuang/case.pdf [17 May 2022].
24. Krishnan, S., Idris, N. (2014). Students' misconceptions about hypothesis test. *Redimat*, Vol. 3, No. 3, pp. 276-293. DOI: 10.4471/redimat.2014.54
25. Lebart, L., Morineau, A., Piron, M. (1995). *Statistique exploratoire multidimensionnelle*. Available at https://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-10/010007837.pdf [7 June 2022].
26. Lee, S., Ludäscher, B., Glavic, B. (2018). Provenance summaries for answers and non-answers. *Proceedings of the VLDB Endowment*, Vol. 11, No. 12, pp. 1954-1957. DOI: 10.14778/3229863.3236233
27. Lindner, T., Puck, J., Verbeke, A. (2020). Misconceptions about multicollinearity in international business research: Identification, consequences, and remedies. *Journal of International Business Studies*, Vol. 51, No. 3, pp. 283-298. DOI: 10.1057/s41267-019-00257-1
28. Madsen, B. S. (2016). Data collection. In *Statistics for Non-Statisticians*, Springer, Berlin, Heidelberg, pp. 1-13. DOI: 10.1007/978-3-662-49349-6_1
29. Motulsky, H. J. (2015). Common misconceptions about data analysis and statistics. *Pharmacology research perspectives*, Vol. 3, No. 1, pp. 1-8. DOI: 10.1002/prp2.93
30. Pfannkuch, M., Ben-Zvi, D. (2011). Developing teachers' statistical thinking. In *Teaching statistics in school mathematics-challenges for teaching and teacher education*, Springer, Dordrecht, pp. 323-333. DOI: 10.1007/978-94-007-1131-0_31
31. Reeves, C. A., Brewer, J. K. (1980). Hypothesis testing and proof by contradiction: an analogy. *Teaching Statistics*, Vol. 2, No. 2, pp. 57-59. DOI: 10.1111/j.1467-9639.1980.tb00387.x
32. Robert, A. D., Bouillaguet, A. (2002). *L'analyse de contenu*. Presses Universitaires de France, Paris.
33. Rossman, A. J., Chance, B., Obispo, C. P. S. L. (2004). *Anticipating and addressing student misconceptions*. Available at https://www.rossmanchance.com/artist/proceedings/rossman.pdf [17 June 2022].
34. Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational research review*, Vol. 2, No. 2, pp. 98-113. DOI: 10.1016/j.edurev.2007.04.001
35. Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., Onghena, P. (2009). How confident are students in their misconceptions about hypothesis tests?. *Journal of Statistics Education*, Vol. 17, No. 2, pp. 1-21. DOI: 10.1080/10691898.2009.11889514
36. Spanos, A. (1986). *Statistical foundations of econometric modeling*. Cambridge University Press. DOI: 10.1017/CBO9780511599293
37. Swann, G. P. (2019). Is precise econometrics an illusion?. *The Journal of Economic Education*, Vol. 50, No. 4, pp. 343-355. DOI: 10.1080/00220485.2019.1654956
38. Taber, K. S. (2005). Learning quanta: Barriers to stimulating transitions in student understanding of orbital ideas. *Science Education*, Vol. 89, No. 1, pp. 94-116. DOI: 10.1002/sce.20038
39. Tompkins, C. A. (1993). *Using and interpreting linear regression and correlation analyses: Some cautions and considerations*. Available at http://aphasiology.pitt.edu/1435/1/21-04.pdf [10 April 2022].

40. Vallecillos, A. (2000). Understanding of the logic of hypothesis testing amongst university students. *Journal für Mathematik-Didaktik*, Vol. 21, No. 2, pp. 101-123. DOI:10.1007/BF03338912
41. Wild, C. J., Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International statistical review*, Vol. 67, No. 3, pp. 223-248. DOI: 10.1111/j.1751-5823.1999.tb00442.x
42. Williams, M. N., Grajales, C. A. G., Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research, and Evaluation*, Vol. 18, No. 1, pp. 1-14. DOI: 10.7275/55hn-wk47

About the author

Djouahra Idris obtained a PhD in Economics and Applied Statistics at the National Higher School of Statistics and Applied Economics (Kolea, Algeria). He is currently a Lecturer of Statistics and Econometrics at the Institute of Economics at the University Center of Tipaza (Algeria). He is also the president of the pedagogical accompaniment cell for newly recruited teachers since 2019 and member of the National Committee for Distance Education starting from September 2022. The author can be contacted at: *djouahra.idris@cu-tipaza.dz*.