


## CHAPTER 117

### Case Study: Forecast in ENADE for the Mathematics Course at the University of Pernambuco through simple linear regression analysis

 [10.56238/pacfdnsv1-117](https://doi.org/10.56238/pacfdnsv1-117)

**Leticia Karina Ramos de Lima**

ORCID: <https://orcid.org/0000-0001-8014-145X>  
University of Pernambuco, Brazil  
E-mail: [leticia.karina@upe.br](mailto:leticia.karina@upe.br)

**Damocles Aurélio Nascimento da Silva Alves**

ORCID: <https://orcid.org/0000-0002-7928-1276>  
University of Pernambuco, Brazil  
E-mail: [damocles.aurelio@upe.br](mailto:damocles.aurelio@upe.br)

**Lauane Raimundo Cordeiro**

ORCID: <https://orcid.org/0000-0002-1247-185X>  
University of Pernambuco, Brazil  
E-mail: [lauane.cordeiro@upe.br](mailto:lauane.cordeiro@upe.br)

**Mauricio Costa Goldfarb**

ORCID: <https://orcid.org/0000-0002-0909-1514>  
University of Pernambuco, Brazil  
E-mail: [mauricio.goldfarb@upe.br](mailto:mauricio.goldfarb@upe.br)

**Miriam Lecília Farias Ribeiro**

ORCID: <https://orcid.org/0000-0002-6439-2563>  
University of Pernambuco, Brazil  
E-mail: [miriam.ribeiro@upe.br](mailto:miriam.ribeiro@upe.br)

**Antonio Lopes Pessoa**

ORCID: <https://orcid.org/0000-0002-8057-9408>  
Department of Education of the State of Pernambuco, Brazil  
E-mail: [lopespessoaantonio@gmail.com](mailto:lopespessoaantonio@gmail.com)

**Danielle Loureiro Roges**

ORCID: <https://orcid.org/0000-0001-8502-1510>

Department of Education of the State of Pernambuco, Brazil  
E-mail: [danielle.lou.roges2022@gmail.com](mailto:danielle.lou.roges2022@gmail.com)

#### ABSTRACT

ENADE is a mandatory assessment composed of the weighted average of the General Training grades - FG and Specific Components - CE that, in graduates of all graduations, measures performance in characteristics such as skills, content and professional competences about their courses. In this sense, in this work, the objective was to carry out a study with the general purpose of carrying out a statistical modeling in relation to the ENADE grade of the Mathematics course at the University of Pernambuco, identifying the patterns of variation of the grade obtained by the course and, after that, through simple regression analysis, predict values of future exam scores of these courses. For this, the existing relationships between two variables (grades and years) were described using a simple linear regression analysis, taking into account the regression line, Pearson's Correlation Coefficient, the Determination and Analysis of Waste. From the observed results, it was concluded that the Enade grades tend to decrease over the years, since there is a negative linear relationship between the variables, this occurs even with all the CE grades and the Garanhuns grade in FG with a positive linear relationship.

**Keywords:** Enade, Simple Linear Regression, Teaching.

## 1 INTRODUCTION

In 1995, the National Course Examination (ENC) was created, called Provão, which, based on the evaluation of higher education courses and their results, had as main objective to obtain data that would influence the formulation of actions and improvements for Brazilian higher education. The mathematics course participated in the ENC from 1998 onwards. In 2004, a new National Higher Education Assessment System (SINAES) was implemented, with the same objectives as the Provão, but with the implementation of the National Student Performance Exam ( ENADE), being published for mathematics students only in 2005.

According to the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2005), ENADE is a mandatory assessment that, in graduates of all degrees, measures performance in

characteristics such as skills, content and professional competences about their courses and is carried out in periods not exceeding three years. It uses as instruments a test about the course and a socioeconomic questionnaire to obtain statistics on higher education institutions (HEIs) and students. These statistics are used to, from the leveling of performance, create public policies in favor of improvements to the quality and services in education.

The research in question has the general purpose of carrying out a statistical modeling in relation to the ENADE grade (composed of the weighted average of the General Training grades - FG and Specific Components - CE) of the Mathematics course at the University of Pernambuco - Campus Garanhuns, Nazaré da Mata e Petrolina, as specific objectives, aims to identify the patterns of variation of the grade obtained by the course and, after that, through simple regression analysis, predict values of the future grades of these courses in the exam.

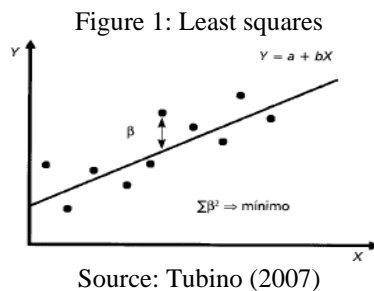
## 2 METHODOLOGY

In this work, the regression analysis aims to describe, through a mathematical model, the existing relationships between two or more variables from  $n$  observations of these variables and, in this way, predict future values of a (dependent) variable, that is, with on the basis of one or more variables explaining future potential.

The linear regression model, according to Krajewski, Ritzman and Malhotra (2009), is one of the most known and used causal models, which consists of a variable called dependent being related to one or more independent variables by a linear equation. It can be said in technical language that the regression line minimizes squared deviations from the actual data. To obtain the equation of the line, simply apply the following equation:

$$y = a + bx$$

In the equation, “ $y$ ” refers to the dependent variable and “ $x$ ” to the independent variable. The “ $a$ ” represents the intersection of the line on the  $y$ -axis and the slope of the line. This formula establishes the equation that identifies the effect of the forecast variable (independent variable) on the demand for the product under analysis (dependent variable), this is because it seeks to forecast the demand for a certain item based on the forecast of another variable that is related to such item. In other words, it aims to find a linear prediction equation so that the sum of the squares of the prediction errors (beta) is the minimum possible. Below is a model where a regression line is formed in the Cartesian system, note that this line is calculated through the points constituted by the dependent and independent variables that are under analysis using the least squares method (Figure 1) :



In addition, the equations for obtaining the coefficients a and b that constitute the linear regression models, as described by Gaither and Frazier (2006), are described below (Figure 2):

Figure 2: Coefficients of the regression line

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}$$

Source: Authors (2019)

According to the figures presented above, the calculation of the coefficients a and b is intended to minimize the sum of squared deviations of the real data from the graph line. To exemplify what has been said, the linear regression model will be calculated for the data set analyzed during the results of this work, which is composed of the ENADE scores in the city of Garanhuns (Table 1 and Figure 3):

Table 1. Application example

Garanhuns (X)	Conceito Enade (Y)	X <sup>2</sup>	Y <sup>2</sup>	XY
2011	2,879	4044121	8,288641	5789,669
2014	2,715	4056196	7,371225	5468,01
2017	2,438	4068289	5,943844	4917,446
<b>Σ=6042</b>	<b>Σ=8,032</b>	<b>Σ=12168606</b>	<b>Σ=21,60371</b>	<b>Σ=16175,125</b>

Source: authors (2019).

Where:

Figure 3. Coefficients calculated from table 1

$$b = \frac{(3 * 16175,125) - (6042 * 8,032)}{3 * 12168606 - (6042)^2} = \frac{-3,969}{54} = -0,0735$$

$$a = \frac{(12168606 * 8,032) - (6042 * 16175,125)}{3 * 12168606 - (6042)^2} = \frac{8138,142}{54} \cong 150,7$$

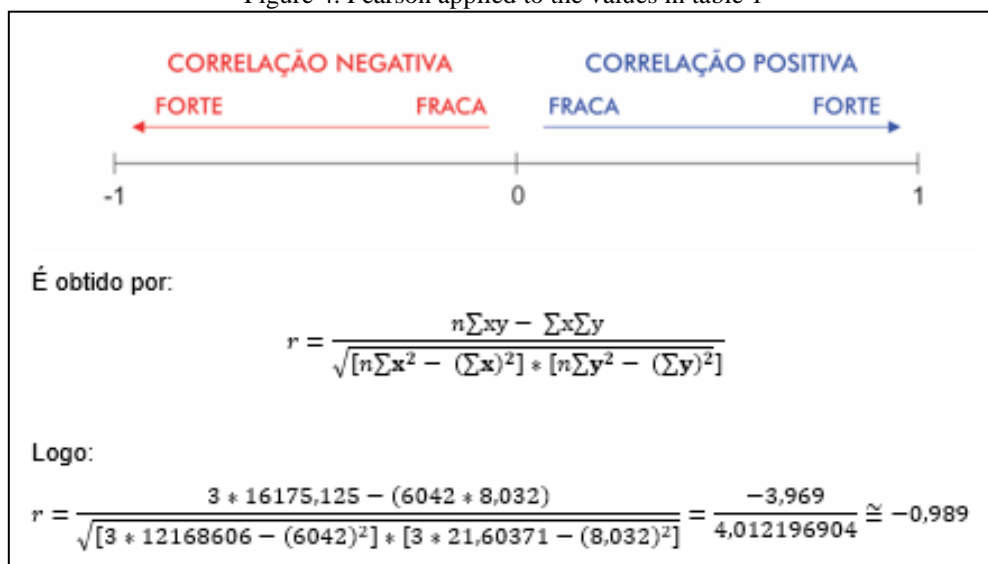
Source: authors (2019).

As  $y = a + bx$ , from the information obtained, we have that the equation of the line will be  $y = 150.7$

– 0.0735x, through which the value of the grades for the year 2020 was calculated, with 2020 being the variable x and the grade the variable y. In addition to the regression line, other deterministic factors that influence the parameters of the model were taken into account, in this case, we have the Pearson Correlation Coefficient, the Determination Coefficient and the Residual Analysis, see how each one was obtained:

**Pearson's Correlation Coefficient (r):** the study of correlation aims to measure and evaluate the degree of relationship existing between two or more variables in the hope that any relationship found can be used to make estimates or predictions of one of the variables that is, it is the number that summarizes the degree of relationship between the dependent variable and one or more independent variables. This degree of correlation varies on a scale from -1 to 1 (Figure 4) and is obtained through the dispositions of the points referring to the independent variables and the dependent variable in a straight line or a plane. 1 (Figure 4):

Figure 4. Pearson applied to the values in table 1



Source: authors (2019).

As the resulting r value was approximately -0.99, we assumed a very strong negative correlation, where as the values of the independent variable increase, the values of the dependent variable decrease.

**Determination Coefficient (R<sup>2</sup>):** this coefficient measures the relationship between the dependent variable and the independent variables. Indicating how many percent the variation explained by the regression represents of the total variation (population). When:

- R<sup>2</sup> = 1: All observed points lie exactly on the regression line (perfect fit), that is, the y variations are 100% explained by the variation of x's through the specified function, with no deviations around the estimated function.
- R<sup>2</sup> = 0: It is concluded that the variations of y are exclusively random and the introduction of the variables x's in the model will not incorporate any information about the variations of y. To exemplify the calculation of R<sup>2</sup> we apply the values recorded in table 1 (Figure 5):

Figure 5. Coefficient of determination ( $R^2$ ) of table 1

É obtido por:

$$R^2 = \frac{(\sum xy - \sum x * \sum \frac{y}{n})^2}{(\sum x^2 - \frac{(\sum x)^2}{n}) * (\sum y^2 - \frac{(\sum y)^2}{n})}$$

Logo:

$$R^2 = \frac{(16175,125 - 6042 * \frac{8,032}{3})^2}{(12168606 - \frac{(6042)^2}{3}) * (21,60371 - \frac{(8,032)^2}{3})} = \frac{-1,750329}{1,78863606} \cong -0,9786$$

Source: authors (2019).

The coefficient of determination obtained indicates that approximately 98% of the dependent variables are explained by the generated regression model.

**Residual Analysis:** Residual analysis plays a fundamental role in evaluating the fit of a simple linear regression model, investigating the model's adequacy regarding basic assumptions, as well as normality, error independence, homoscedasticity, linear relationship of X and Y and lack of fit. In addition to the significance and adequacy tests, residual analysis complements the list of procedures that must be performed after adjusting any model (Messeti, 2013). The residuals of the values visualized in table 1 were described below (table 2):

Table 2: Analysis of residues from table 1

Garanhuns		
Observação	Y previsto (através da reta de regressão)	Resíduos (diferença entre o y calculado e o y observado)
1	2,897833	-0,01883
2	2,677333	0,037667
3	2,456833	-0,01883

Source: authors (2019).

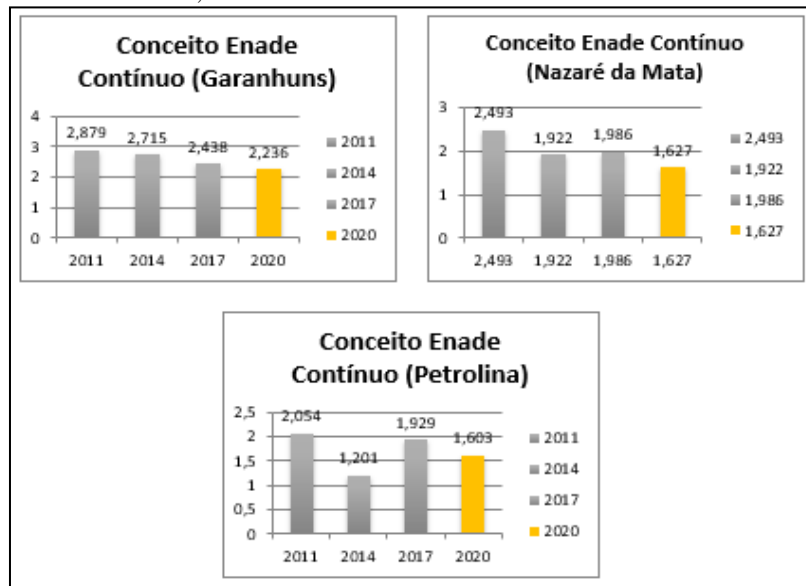
Note that the residual values are small, which helps to understand the quality of the model, in addition to having constant variance, without outliers or influential points.

### 3 RESULTS AND DISCUSSION

The National Student Performance Examination (Enade), as is already known, is composed of the weighted average of the General Training (FG) and Specific Components (CE) grades. The observed grades were organized by city (Garanhuns, Nazaré da Mata and Petrolina), Categories (Enade, General Formation or Specific Components) and the years in which the grades were obtained, seeking, through simple linear regression, to predict the grades that will be obtained by the Mathematics Degree courses at the University of Pernambuco in the year 2020 (not yet disclosed).

### 3.1 ENADE: THE FOLLOWING GRAPHS SHOW THE ENADE SCORES OBTAINED BY STUDENTS IN 2011, 2014 AND 2017, IN ADDITION TO THE EXPECTED SCORES FOR 2020:

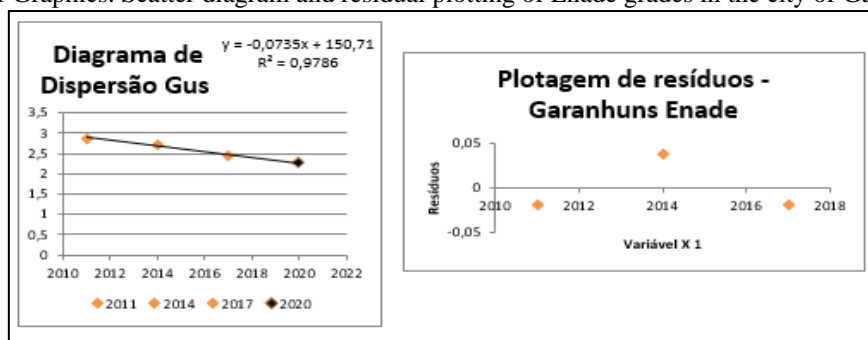
**Set 1 of Graphs:** Enade grades in the years 2011, 2014 and 2017 and expected grades for the year 2020 according to the regression model in the cities of Garanhuns, Nazaré da Mata and Petrolina .



Source: Authors (2019).

The information described in this set 1 of graphs was used to calculate the linear regression models for each city in the general notes of ENADE. In set 2 of graphs, information about the city of Garanhuns is described:

Set 2 of Graphics. Scatter diagram and residual plotting of Enade grades in the city of Garanhuns.

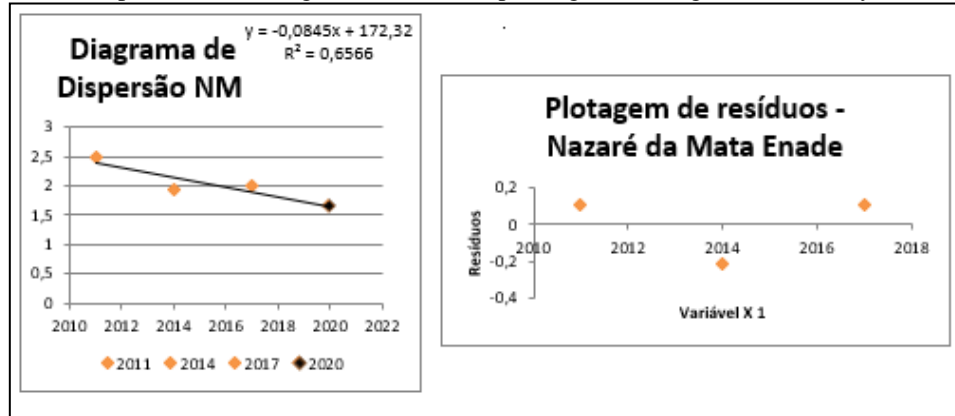


Source: authors (2019).

We note that in the city of Garanhuns in all variables as X increases Y decreases, therefore, there is a negative linear relationship where Pearson's Correlation Coefficient (r) is equal to -0.98923 (Set 2 of Graphs), characterizing a correlation very strong. The regression line is given by the equation  $y = -0.0735x + 150.71$ , where the Coefficient of Determination (R<sup>2</sup>) is equal to 0.9786, that is, the line explains almost 100% of the observed values. The residuals between the observed and predicted scores are almost equal, ranging between 0.05 and -0.05. The data above suggest that the relationship between the variables analyzed is strong, as a consequence, it is expected that the Enade score in Garanhuns in 2020 will be even lower

than in previous years. In set 3 of graphs, the information for the city of Nazaré da Mata is described:

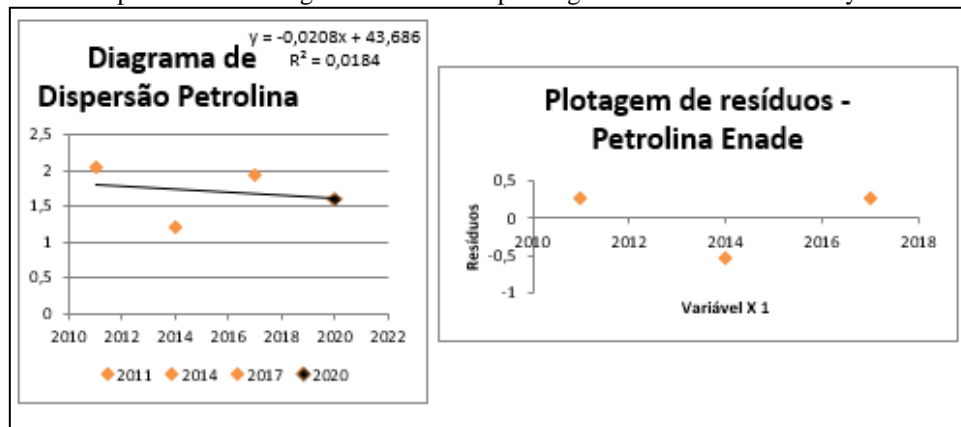
Set 3 of Graphics. Scatter diagram and residual plotting of Enade grades in the city of Nazaré



Source: authors (2019)

In Nazaré da Mata we have that, on average, X increases and Y decreases, r is equal to -0.81034, indicating a strong correlation, the equation of the regression line is  $y = -0.0845x + 172.32$  and  $R^2 = 0.6566$ , being then approximately 65% how much the regression model can explain the observed values. As for residues, they do not vary much, they are not greater than 0.2 or less than 0.3, so the predicted grade for 2020 can be taken into account, that is, the data suggest that in the year 2020 the grade will be again the smallest in the data set. In set 4 of graphs, information for the city of Petrolina is described:

Set 4 of Graphics. Scatter diagram and residual plotting of Enade notes in the city of Petrolina.

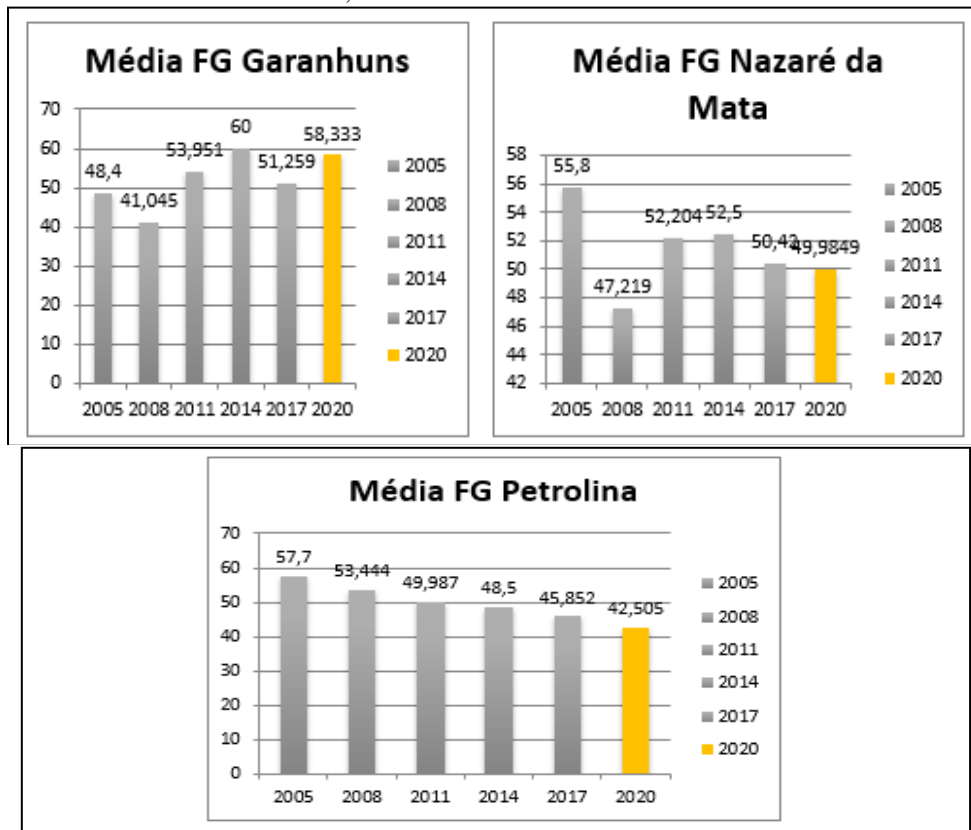


Source: authors (2019).

In the city of Petrolina, on average, while X grows Y decreases, unlike other campuses, we have that the correlation between the variables that were analyzed is almost negligible, this is because r is equal to -0.13568. The Regression Line is given by the equation  $y = -0.0208x + 43.686$  and the Determination Coefficient is 0.0184, also very low, meaning that the line does not explain even 1% of the observed values, this situation shows us that practically there is no relationship between the variables and, therefore, in the predicted grade for the year 2020, there is a high chance that there will be a greater residue of the observed grade for the calculated grade.

3.2 GENERAL TRAINING (FG): FROM THE GENERAL TRAINING GRADES, THE YEARS 2005, 2008, 2011, 2015 AND 2017 WERE TAKEN, THE GRAPHS SHOW THE GRADES OBSERVED IN THESE YEARS PLUS THE PREDICTED GRADE FOR THE YEAR 2020:

Set 5 of Graphs: General Training Grades in the years 2005, 2008, 2011, 2015 and 2017 and expected grades for 2020 according to the regression model in the cities of Garanhuns, Nazaré da Mata and Petrolina .



Source: Authors (2019).

The information described in this set 5 of graphs was used to perform the calculations of linear regression models for each city in the General Training notes, the data that are described in set 6 of graphs below are also summarized in table 3:

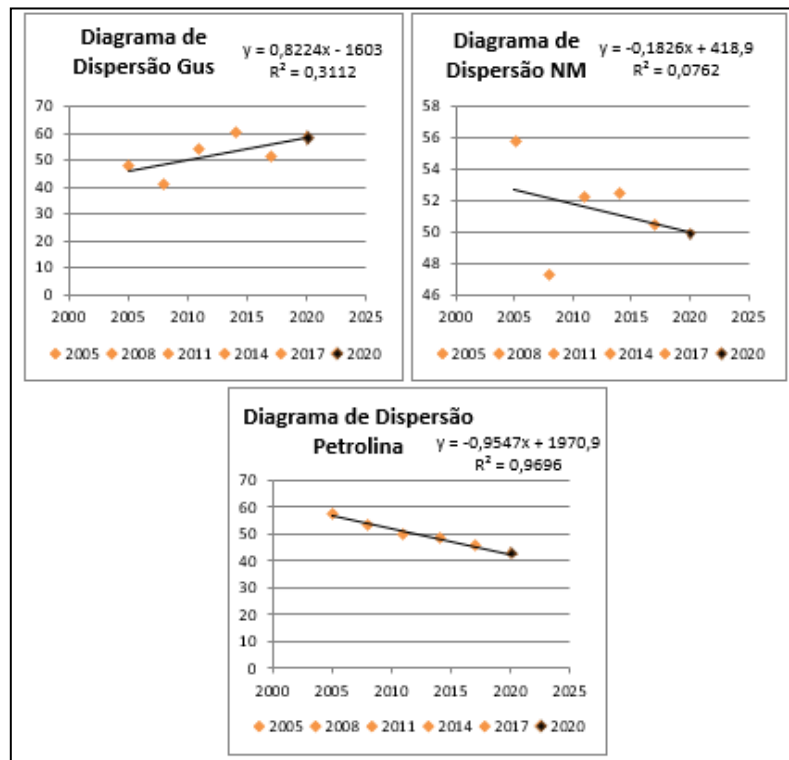
Table 3. Regression data for cities in General Training notes

	<b>Garanhuns:</b>	<b>Nazaré da Mata</b>	<b>Petrolina</b>
r	0,557856	-0,27604	-0,9847
Reta	$y = 0,8224x - 1603$	$y = -0,1826x + 418,9$	$y = -0,9547x + 1970,9$
R <sup>2</sup>	0,3112	0,0762	0,9696

Source: authors (2019)



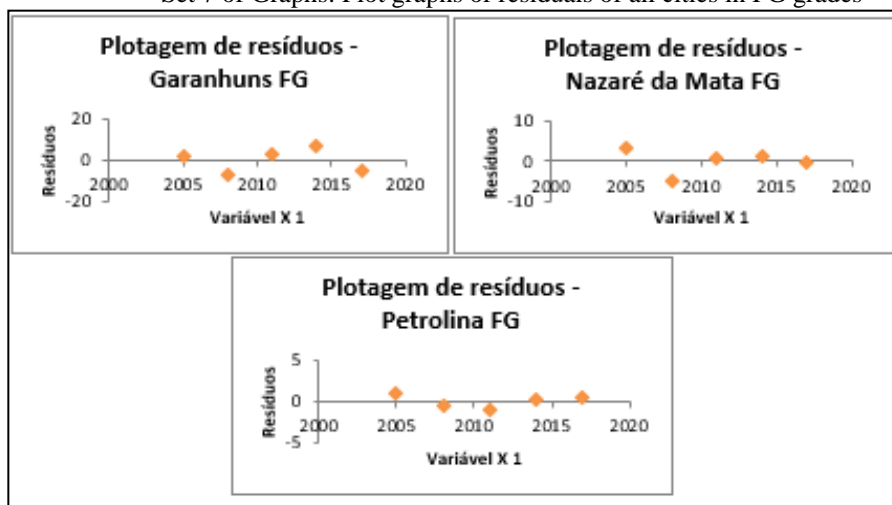
Set 6 of Graphs: Scatterplots of all cities in General Training notes:



Source: Authors (2019).

Observe the information, in the General Formation (FG) notes we noticed that where there is a stronger correlation is in Petrolina, classified as very strong, followed by Garanhuns, with a moderate correlation and Nazaré da Mata, which has a negligible correlation, this reflects in the Coefficients of Determination, where the regression line, in Nazaré da Mata, explains less than 1% of the observed data, followed by Garanhuns, with 31% and Petrolina with 96%. As for residual analysis (set 7 of graphs):

Set 7 of Graphs: Plot graphs of residuals of all cities in FG grades



Source: Authors (2019).

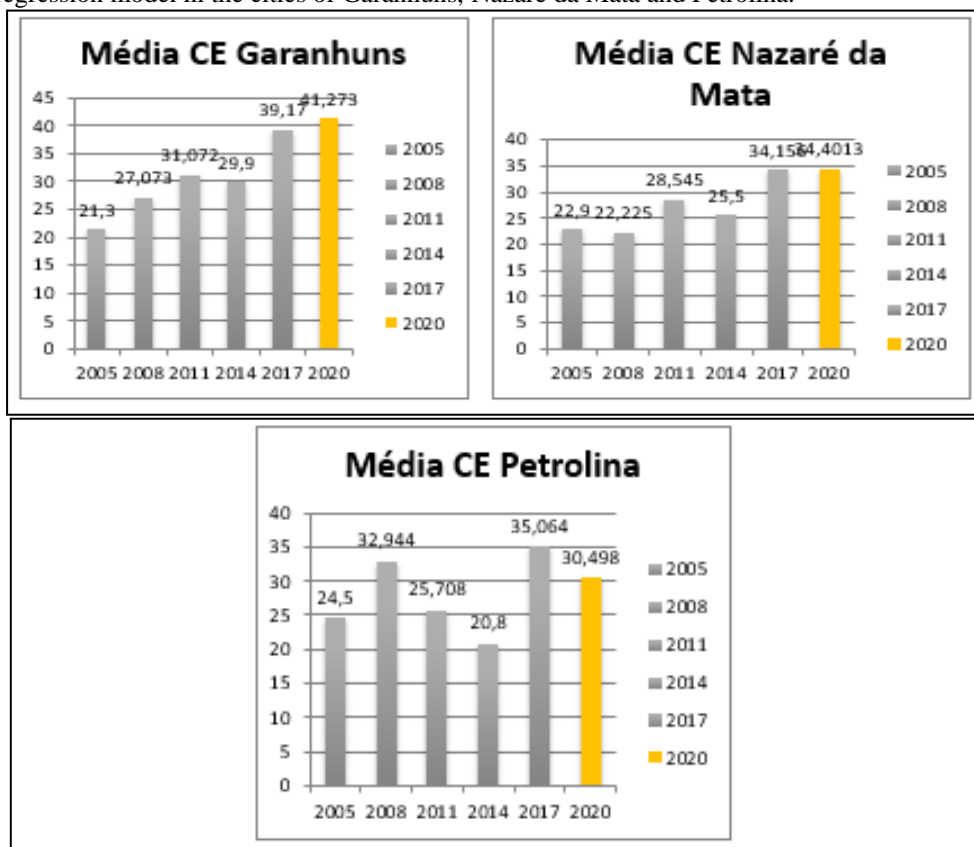
Note that the residues, in Garanhuns, vary considerably from each other. Nazaré da Mata, in turn, although it has a negligible correlation between the variables, does not have very distinct residues, whereas

Petrolina naturally has the smallest residues, with the majority reaching almost zero.

Based on the observations, it can be seen that the analysis of the linear relationship between the two variables (year and grade FG) in the three cities suggests that Petrolina has a direct relationship of cause and effect between the variables, therefore, it is expected that its lowest grade in General Training or in 2020, in Nazaré there is no considerable linear relationship, with the predicted grade being more uncertain, while Garanhuns, if the value obtained is close to what was predicted, will reach its second highest grade compared to previous years.

**3.3 SPECIFIC COMPONENTS (EC): FINALLY, FROM THE GRADES OF SPECIFIC COMPONENTS, THE YEARS 2005, 2008, 2011, 2015 AND 2017 WERE ALSO TAKEN, THE GRAPHS SHOW THE GRADES OBSERVED IN THESE YEARS PLUS THE PREDICTED GRADE FOR THE YEAR 2020:**

Set 8 of Graphs: Specific Component Grades in the years 2005, 2008, 2011, 2015 and 2017 and forecasted grades for 2020 according to the regression model in the cities of Garanhuns, Nazaré da Mata and Petrolina.



Source: Authors (2019).

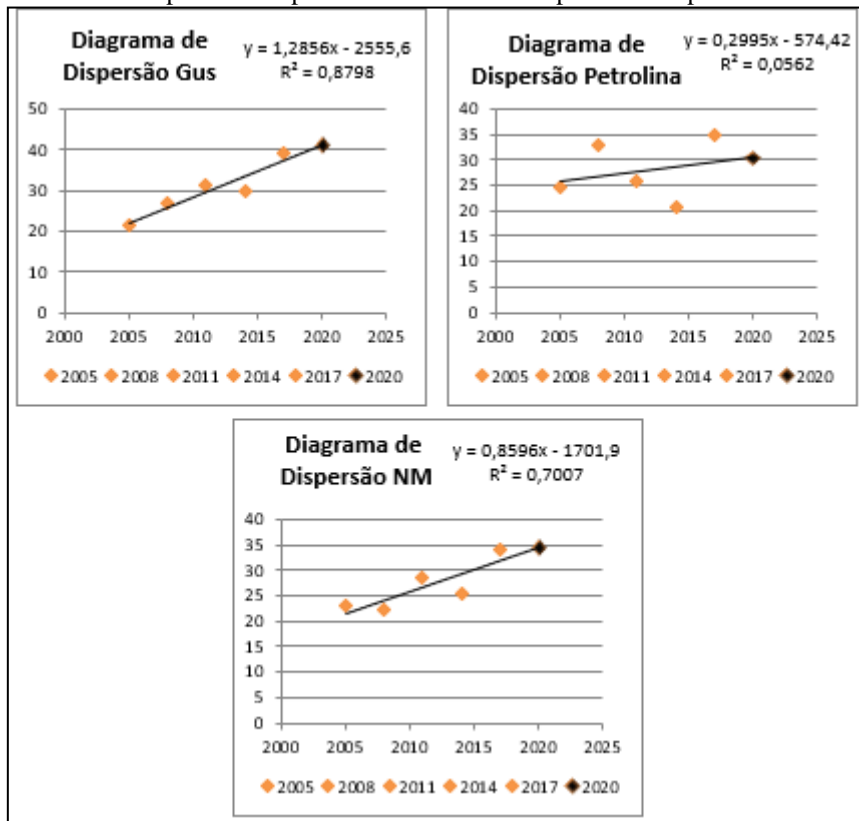
The information described in this set 8 of graphs was used to perform the linear regression model calculations for each city in the Specific Components notes, the data that are described in the sets 9 of graphs below are also summarized in table 4:

Table 4. Regression data for cities in the Specific Components scores

	<b>Garanhuns:</b>	<b>Nazaré da Mata</b>	<b>Petrolina</b>
r	0,937972	0,837101	0,237167
Reta	$y = 1,2856x - 2555,6$	$y = 0,8596x - 1701,9$	$y = 0,2995x - 574,42$
R <sup>2</sup>	0,8798	0,7007	0,0562

Source: authors (2019).

Set 9 of Graphs: Scatterplots of all cities in the Specific Components notes:

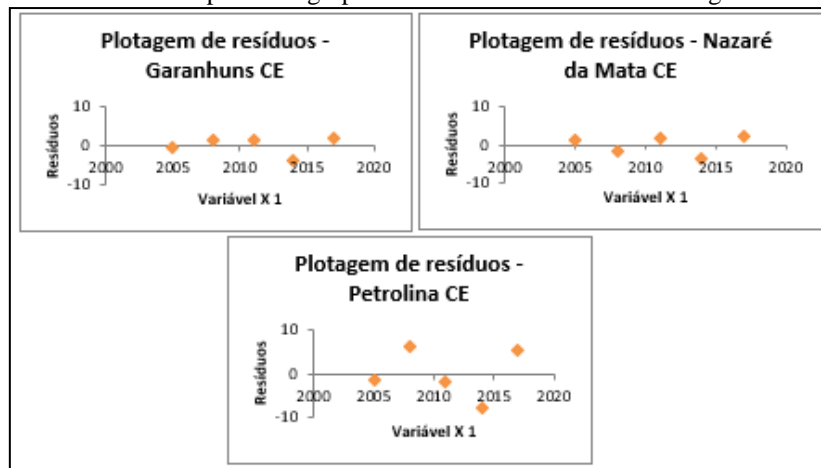


Source: Authors (2019).

In the city of Garanhuns, when analyzing the grades, we noticed that, on average, if X grows Y grows, with the exception of one occasion between 2011 and 2014, this leaves us with a positive linear correlation where the Pearson correlation coefficient has a value close to one, which leads to a strong correlation, a similar situation occurs with Nazaré da Mata, which obtains a strong correlation, whereas Petrolina presents a negligible correlation. As for the Determination Coefficient (R<sup>2</sup>), from its analysis, it is noted that the regression line explains less than 1% of the values observed in the city of Petrolina, in Nazaré da Mata this value is approximately 70% and in Garanhuns 87%.

Also by analyzing the set of graphs 9, we saw that the scores predicted for the year 2020 in Garanhuns and Nazaré have a strong cause and effect relationship, with the highest ever achieved by both, which definitely does not occur in Petrolina, which, in addition to its values have an almost null linear relationship, it would only reach its third highest score if what was predicted happened. Now look at the residuals (set 10 of graphs):

Set 10 of Graphs: Plot graphs of residuals of all cities in EC grades



Source: Authors (2019).

Both the Garanhuns and Nazaré residuals do not have a large variance between them, in both cities the lowest value is no more than -4 and the highest value is no more than 3, having, in addition, several residues with almost equal values. Petrolina, in turn, has a much greater variation, with residuals between -8 and 7 and quite distinct from each other.

#### 4 FINAL CONSIDERATIONS

With the accomplishment of this work, it was noticed that the data obtained from the application of the proposed method revealed the value of the predicted grades for the years 2020, in addition, through the Linear Correlation Coefficient it was possible to quantify the correlations as strong or weak, through the Determination Coefficient, the proportion of the variability of the predicted values for pata y was explained, which is explained by the simple linear regression model in each case, and, finally, through the Residual Analysis, the basic assumptions of each were confirmed. model.

At the end of the study, it was found, regarding the Enade scores, that the cities that had a good relationship between the variables studied have as supposed scores for the year 2020 a lower value than all the previous ones, that is, the score tends to decrease each year, the same happens with the General Education grade in Petrolina. The General Training grade in Garanhuns, which has a moderate correlation, tends to grow and expects to reach its second highest grade in 2020.

As for the scores in Specific Components, the two cities that have good correlation achieve their highest scores in 2020, that is, the CE score tends to grow over the years. Finally, of the nine analyzes carried out, the three that obtained a weak or insignificant correlation (Enade de Petrolina, FG de Nazaré da Mata and CE de Petrolina) naturally have a large variation in their scores, without a specific standard, consequently, the Predicted notes for the year 2020 have the same characteristic.

For future work, it is suggested to use simple linear regression for an analysis of ENADE in a specific municipality in the state in other undergraduate or Bachelor's degrees at the University of Pernambuco.

## REFERENCES

- Barros, JLDC, Campos, MZD, Teixeira, DDC, & Cabral, BGDAT (2020). Reflections on the level of specific knowledge of undergraduate students in Physical Education at Enade 2014. *Revista Brasileira de Estudos Pedagógicos* , 101, 99-119.
- Beltrão, KI, & Mandarino, MCF (2014). Evidence from ENADE-Changes in the Profile of Graduate Mathematicians. *Essay: Evaluation and Public Policies in Education* , 22, 733-753.
- Da Silva Dias, J., de Magalhães Porto, C., & Nunes, AKF (2016). General training and specific knowledge in the Enade test. *International Teacher Training Meeting and Permanent Forum on Educational Innovation* , 9(9).
- De Almeida, DA, Almeida, SPN de C. e, & Amorim, MMT (2021). Profile of graduates in mathematics degree: an analysis from enade's data (2005-2017). In *SciELO Preprints* . <https://doi.org/10.1590/SciELOPreprints.2561>.
- DeMelo, AM (2009). Correlation Analysis and Simple Linear Regression: Accounting Applied to Economic-Financial Indicators of 2009 of Publicly Traded Companies of the Civil Construction Segment that are Members of the Novo Mercado. *UFSC Congress of Controllershship and Finance and Scientific Initiation in Accounting* , Brazil.
- DE OLIVEIRA FILHO, ML (2002). The use of linear regression as a strategic tool for projecting production costs. In *Annals of Brazilian Congress on Costs-ABC* .
- Do Amaral, LS, dos Santos, ALP, de Figueiredo, MPS, de Almeida Ferreira, DS, Silva, JE, dos Santos, HCT, ... & Moreira, GR (2020). Internalization of Covid-19: An analysis of the evolution of cases/10 thousand inhabitants in municipalities of the Microregion of Garanhuns in the State of Pernambuco, through non-linear regression models. *Research, Society and Development* , 9(9), e293996582-e293996582.
- Morettin, PA, & Bussab, WO (2017). *Basic Statistics* . Saraiva Education SA.
- Gaither, N., & Frazier, G. (2001). *Production and operations management* . Pioneer Thomson Learning.
- INEP - National Institute of Educational Studies and Research Anísio Teixeira. (2020). *Enade Concept* . Retrieved from <http://portal.inep.gov.br/conceito-enade> .
- INEP - National Institute of Educational Studies and Research Anísio Teixeira (2020). *National Student Performance Exam (Enade)* . Retrieved from <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade> .
- Krajewski, LJ; Ritzman, LP; Malhotra, M. (2009). *Production and operations management* . 8. ed. Sao Paulo: Pearson Prentice Hall.
- Lima, BAT, & dos Santos, MBH (2021). Students' perception of education at the university through the ENADE student questionnaire. *Research, Society and Development* , 10(3), e19510313150-e19510313150.
- Malhotra, NK, Rocha, I., Laudisio, MC, Altheman, É., Borges, FM, & Taylor, RB (2005). *Introduction to marketing research*.
- Medeiros, FSB, & Bianchi, RC (2009). The application of the simple linear regression method in the demand for seasonal products: a case study. *Disciplinarum Scientia| Applied Social* , 5(1), 35-53.

MESSETI, Ana Verginia Libos. (2013). Correlation and regression analysis. Specialization course “Lato Sensu” in statistics. *State University of Londrina, Londrina*.

Ramos, PNF, da Silveira, OR, & de Souza Maia, JC (2022). Determination of simple linear regression analysis to explain the influence of soil physical attributes on cotton production . *Research, Society and Development* , 11(8), e28411830591-e28411830591.

Smith, MM *ENADE Commented: Component: General Training: 2006, 2007, 2008 editions* . EDIPUCRS.  
Triola, MF (2008). Introduction to statistics. In *Introduction to Statistics* (pp. xxvi-310).

TUBINO, DF (2007). *Production planning and control: theory and practice* . Sao Paulo: Atlas.