# ANALYSIS OF LINEAR REGRESSION IN SPANISH BACCALAUREATE TEXTBOOKS

P. Díaz and Luis J. Rodríguez-Muñiz
Department of Statistics and Operations Research, and Mathematics Education
University of Oviedo (Spain)
patrycyadd@gmail.com, luisj@uniovi.es

*In this paper we study the presentation of linear regression in Spanish textbooks, from different publishers, written for the first year of* Bachillerato *(16 years old), Spanish post-compulsory Secondary Education (Baccalaureate). We perform a three-way analysis, based on conceptual, educational-cognitive, and phenomenological points of view. Furthermore, we perform a study of the type of problems and exercises proposed in the textbooks. Finally, we analyze correspondences between contents in textbooks and Spanish official curriculum for this year. Our goal is to check whether textbooks include innovative key points recently introduced in the official curriculum, because, in Spain, textbooks are still very important tools in the classroom.*

## INTRODUCTION

Linear regression is one of the topics in Statistics in post-compulsory Secondary School (called *Bachillerato* in Spanish Educational system). It is related to other concepts as: variation, scatter plots, functional dependence, estimation, etc. In Spain, it is studied in the first year of *Bachillerato* (16 years old) in two courses: Applied Mathematics to Social Sciences I, and Mathematics I. We focus our research on Mathematics I (mathematics for the Science and Technology Baccalaurate). The official curriculum (MEC, 2007) is composed of content and assessment criteria. The curriculum includes regression and correlation, and it also includes the following content: two-dimensional distributions, relationships between two variables, and linear regression. The assessment criteria explain that it is important to check the student´s ability to estimate and associate parameters related to correlation and regression and to apply the parameters in related situations. These statements provide a high degree of freedom when they are developed in textbooks, because those are too vague. Hence, this underlines the importance of Chevellard´s didactic transposition study (Chevellard, 1985), that is, to transform content, outcomes and assessment criteria into skills for students.

On the other hand, the official curriculum (MEC, 2007) establishes other items that are important in the didactic transposition such as the necessity of solving problems, which helps students to explain, predict and model real situations using scientific knowledge. In addition, it remarks that problems have to be related to other subjects or sciences in order for students to use their knowledge and skills in real contexts in their lives. So, we will consider that those textbooks which encourage statistical reasoning, explain relationships between concepts, and stimulate critical thinking have introduced innovate key points, rather than those texts in which prevailing approaches are based on calculus.

## RESEARCH METHODOLOGY

We examined five textbooks of Mathematics I (but regarding regression, curriculum is similar also in Applied Mathematics to Social Sciences I). We carried out an informal survey in several Secondary schools in our region, and we chose five textbooks among the most frequently used. The selected publishers are: Anaya (Colera et al., 2008), Edelvives (Monteagudo et al., 2008), Editex (Ruiz et al., 2008), Pearson (Ortega et al., 2010) and SM (Vizmanos et al., 2008).

Our study will use the model in Sierra (2011) for the analysis of textbooks and the classification of problems by Butts (1980). Sierra (2011) suggests carrying out a triple analysis for each conceptual unit: conceptual, educational-cognitive and phenomenological analyses. In this work, we constrained ourselves to linear regression, which only includes working with the regression line and making estimations. We assume correlation would constitute a different conceptual unit; thus, we are not focusing on correlation here, and we plan to analyze it in a forthcoming paper.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org 

ANALYSIS OF RESULTS

We are organizing our results into four different subsections: conceptual, educational-cognitive, phenomenological analyses, and exercises and problems. The conceptual analysis block is related to the order of the contents, to the definitions and to the symbolic or graphic representations used in the textbook. The educational-cognitive analysis refers to the author´s didactic approach, and in the phenomenological analysis we study which phenomena are used in the explanations, not only related to statistics but also to other sciences. Although Sierra (2011) includes exercises in the conceptual analysis, we study them in a single section for clarity.

*Conceptual Analysis*

We collected the information from different publishers in a table (not presented here due to space constraints), and four concepts associated with regression analysis emerged from our analysis: definition, construction and graphical methods, interpretation and estimation, and examples. Interpretation connects a regression line with the correlation coefficient because they are always studied together. However, we considered them different conceptual units, each one with its own identity, to be analyzed separately.

*Regression's definition*: the concept is introduced in a similar way, as the best fitting line for the set of dots. Some books specify that the aim is finding the mathematical function that allows us to relate the variables *X* and *Y* and to predict results. It is important to note the order of the contents in a textbook: the regression line is typically the last item in the unit and understanding the regression line relies on many previous taught concepts, such as two-dimensional tables, parameters, scatter plots and covariance.

*Least squares method (LSM)*: three textbooks (Colera et al., 2008; Ortega, 2010; Vizmanos et al., 2008) explain the method in depth, using graphics to support the explanation. Colera et al. (2008) consider all linear functions $y = A + Bx$ and, then chooses the one minimizing the sum of squares distances $\sum d_i^2$. It does not develop the method because it exceeds the level of the course (students have not yet studied partial derivatives). Graphics shows distances from real values to estimates. In addition, the text presents the line *Y* over *X*, showing different positions between this regression line and the another one, X over Y. Vizmanos et al. (2008) uses a similar approach, but it uses different notation: $\hat{y} = mx_i + n$ for the best line, and $d_i = y_i - \hat{y}_i$ for errors. It also includes supporting graphics, but it does not analyze the regression line, despite the presentation of a graphic where we can see both lines, as in the previous textbook. Ortega et al. (2010) details the proof of LSM. It uses one graphic to point out distances to the line and another one to show the estimation error, together with the estimate and the real value. It does not develop the method for the other line (X over Y) but it explains it would be analogue and shows the equation line. Two other textbooks (Monteagudo et al., 2008; Ruiz et al., 2008) present the regression line without explaining the method to determine it, nor mentioning regression errors. Both show the equations of the two regression lines (Y over X, and X over Y), and Ruiz et al. (2008) incorporates a graphic with the two lines. So, in these cases, students could think the regression line is always the same, independent of the data they have.

*Interpretation and estimations*: all textbooks remark on the importance of a high correlation to obtain good prediction estimates and the necessity to be careful with extrapolation. Nevertheless, no ideas are shown about deciding what line to use (Y over X or X over Y), and they just refer to what value to estimate (*x* or *y*). This is an important source of mistakes among students. Since they are used to calculus methods, they tend to estimate *x* from *y* by solving for *x* the regression line *Y* over *X*, and they do not realize that this procedure is not valid. None of the textbooks works with the idea of positive or negative covariance connected to the sign of the slope of regression lines. All the books insist on calculating estimates only when correlation is close to 1 or –1, but students must be able to decide by themselves when it is closer enough. Vizmanos et al. (2008) is the only one giving an alternative (Tukey's line) when covariance is near 0.

*Examples*: we define as examples those small clarifications about an explanation, which help students to understand a definition or a method. We distinguish them from solved exercises where students can observe the solution of an exercise involving many of the concepts and methods studied in the unit. We found that books have solved exercises but few examples. The solved exercises presented in all the textbooks have similar structure: plotting the scatter plot, calculating

covariance, calculating both regression lines and estimating and interpreting different values in order to use both regression lines. The contexts are also very similar: grades in mathematics and physics courses, children's age and weight, etc. Only one (Ruiz et al., 2008) includes different situations where there would be either a functional or statistical relationship.

*Extra information*: some textbooks go much further than the official curriculum (MEC, 2007). Ortega et al. (2010) define the determination coefficient deducting their formula through the regression coefficients; in addition, they also define the standard error of the estimation. Ruiz et al. (2008) add a section about the use of calculators. Vizmanos et al. (2008) present three extra sections: determination coefficient, linearization by logarithms and Tukey's line.

*Educational-cognitive Analysis*

Analyzed textbooks could be classified depending on how they explain the concepts. Monteagudo et al. (2008) and Ruiz et al. (2010) present the concepts by a using a deductive process: going from definitions to examples and finally solving related exercise. Both textbooks introduce lessons by a summary with some historical notes.

Colera et al. (2008), Ortega et al. (2010), and Vizmanos et al. (2008) use an inductive approach. Colera et al. (2008) use many examples to link different content; they also introduce lessons with a summary of historical notes about different contents. Ortega et al. (2010) develop a historical introduction to the unit, but exercises are less structured, more open and do not guide students' learning as much as in other textbooks. They encourage mathematical thinking and reasoning instead of repeating the same structures in all exercises. Vizmanos et al. (2008) also use an inductive approach but, in our opinion, they use too many extra content areas, which can be confusing for medium students.

*Phenomenological Analysis*

Statistics allows students to easily work in real situations with real data. This is easier in statistics than in other parts of mathematics, and authors should take advantage of this fact. In the considered textbooks, problems are usually about situations close to students: education (grades, performance), biology (anthropometric measures), economics (inflation, incomes), sociology (birth rates, phone calls), physics, technology, etc. This is a must when trying to equip students with statistical literacy. To have a look at data is the first step of the statistical method.

*Exercises and Problems*

We classified exercises and problems by using Butts' criterion (1980). The results are displayed in Table 1.

Table 1. Exercises and problems classification

| | Recognition | Algorithmic | Application | Open search | Real Problem | Total |
|---|---|---|---|---|---|---|
| Colera et al. (2008) | 4 | 11 | 4 | 0 | 0 | 19 |
| Monteagudo et al. (2008) | 2 | 16 | 1 | 0 | 0 | 19 |
| Ruiz et al. (2008) | 0 | 5 | 1 | 0 | 0 | 6 |
| Ortega et al. (2010) | 9 | 18 | 9 | 0 | 0 | 36 |
| Vizmanos et al. (2008) | 3 | 22 | 4 | 0 | 1 | 30 |
| Total | 18 | 72 | 19 | 0 | 1 | 110 |

We can see that the number of algorithmic exercises is very high. Moreover, most of them repeat the same structure. Therefore, this may lead students to a very mechanical learning, often not sustained, and they possibly do not acquire the necessary statistical skills to solve a real problem that is different from the schematic exercises in the book.

These kind of exercises clearly contradict the official curriculum (MEC, 2007), where students are encouraged to cope with new problems, reflect and argue about their opinions. This is hard to achieve with so many algorithmic problems and exercises and no space for application exercises or real problems.

In addition, even though the official curriculum (MEC, 2007) suggests the necessity of working with new technologies and computer programs, only Ruiz et al. (2008) includes a section about calculators, but they do not insist on calculator use in the rest of the exercises. Monteagudo et al. (2008) includes a section about using Microsoft Excel, but they only propose one exercise to be solved with this technology. The rest of textbooks do not include any reference to computing technologies.

## CONCLUSIONS

A general overview shows that prevailing approaches are more based on calculus rather than on statistics. Thus, mechanisms for calculating regression are over-represented in contrast with the statistical meaning of regression. This fact endorses the claim in Agnelli et al. (2009) about the difficulties of linking a deterministic method such as LSM with the underlying randomness in every statistical process. As we have demonstrated, textbooks tend to pay more attention to the calculus of the regression line, often leaving aside the statistical concept: different data will produce different regression lines. That is, we are not obtaining a universal equation valid for any case, but an estimated regression line. In our opinion, this must be stressed when working with regression at this educational level. Moreover, it should be reinforced with a deeper analysis of the regression lines in connection with covariance (sign of the slope, type of relationship between both variables). Regarding educational approaches, we are convinced that inductive approaches based on experiences with real data reinforce a holistic approach to the statistical process. Thus, this approach should be supported by problems and exercises related to this particular approach. Some algorithmic exercises can be used to get students familiar with the procedure, but much more application and, especially, real problems must be posed in the textbooks. Hence, students should face real situations in which data are redundant or lost, and they should make decision under uncertainty, supporting their decisions using information provided by regression.

Finally, we plan to develop a similar analysis about correlation, which constitutes another facet when talking about regression, and, therefore, it also would need a careful analysis.

## REFERENCES

Agnelli, H., Konic, P., Peparelli, S., Zón, N., & Flores, P. (2009). La función lineal obstáculo didáctico para la enseñanza de las regresión lineal. *Revista Iberoamericana de Educación Matemática, 17*, 52-61.

Butts, T. (1980). *Posing problems properly.* In S. Krulik and R. E. Reys (Eds.), *Problem solving in school mathematics* (pp. 22-33). Reston, VA: National Council of Teachers of Mathematics.

Chevallard, Y. (1985). *La transposition didactique: du savoir savant au savoir enseigné*. Grenoble: Le Pensée Sauvage.

Colera, J., García, R., Oliveira, M. J., & Santaella, E. (2008). *Matemáticas I*. Madrid: Anaya.

Monteagudo, M.F., & Paz, J. (2008). *Matemáticas.* Zaragoza: Edelvives.

MEC: Ministerio de Educación y Ciencia (2007). Real Decreto 1467/2007, de 2 de noviembre, por el que se establece la estructura del Bachillerato y se fijan sus enseñanzas mínimas. *Boletín Oficial del Estado*, *266*, 45381-45477.

Ortega, P., Serra, J. F., Prieto, J. J., & Bautista, A. (2010). *Matemáticas I. Integra*. Madrid: Pearson.

Ruiz, M. J., Llorente, J., & González, C. (2008). *Matemáticas*. Madrid: Editex.

Sierra, M. (2011). Investigación en Educación Matemática: objetivos, cambios, criterios, métodos y difusión. *Educatio Siglo XXI, 29*(2), 173-198.

Vizmanos, J. R., Hernández, J., & Alcaide, F. (2008). *Matemáticas 1*. Madrid: SM.