STUDENTS' UNDERSTANDING OF LINEAR REGRESSION

Carmen Capilla[1]

[1]*Polytechnic University of Valencia (SPAIN)*

Abstract

This paper analyzes students' beliefs and difficulties in understanding linear regression methods. Two groups of students are considered. The first one is involved in an introductory statistics course in the computer science engineering degree. The other one is formed by students in the fourth year of the environmental science degree. In this second group the statistic course contains more advanced topics. Learning and assessment activities of both courses are presented. Results show that reasoning about association when using graphical and numerical methods, implies misunderstandings such as evaluating the degree of correlation, and interpreting that it implies causation. Teaching of linear regression models is designed to understand that these models are useful approximations to reality, but they should never be considered the final word.

Keywords: Statistics education, linear regression models, correlation, computer science degree, environmental science degree, undergraduate students.

1 INTRODUCTION

Covariation is a term used in statistics to measure the degree of statistical association between the two components of a bidimensional quantitative variable. Correlation is the sample parameter that adimensionally evaluates this association. These statistical concepts are included as important topics in introductory statistics courses at all education levels ([1]). The Accreditation Board of Engineering and Technology (ABET, [2]) has recommended teaching statistical methods to analyze association and regression models in undergraduate engineering education. Reasoning about association is an important cognitive activity and plays an important role for decision-making in real world problems ([3]). Knowledge of students' conceptions and misunderstandings of these topics, is important to develop effective teaching methodologies ([4]).

In this paper we analyze students' beliefs and difficulties in understanding association, correlation and linear regression methods, in two different degrees at the Polytechnic University of Valencia (Spain). The first one is the computer science engineering degree. Statistics is taught in the first semester as an introductory course. Statistics education for computer science engineers tries to give them a solid foundation to ensure correct use of methods ([5]). Graphical and numerical methods to assess correlation and association, and simple linear regression are taught as part of the curriculum. A brief introduction to inference in the regression model is also included.

The second degree is the environmental science program. In this case the course is in the fourth year of the degree. Its contents are more advanced statistical methods: analysis of variance, multiple regression, descriptive analysis of time series and some multivariate methods. These students have previously been taught introductory courses in statistics with descriptive and probability methods. The multiple regression methods are taught after and introduction with a review of correlation and association, and simple regression. The topic covers questions such as polynomial regression, use of dummy variables to model the effects of qualitative factor, analysis of interactions between factors, and inference on the model. In the descriptive analysis of time series, regression is introduced to estimate temporal trends and seasonal components. Piegorsch & Edwards [6] discuss the design of undergraduate environmental statistics courses, and in their opinion these courses should cover topics such as correlation and regression. In next sections learning and assessment activities for teaching these concepts in both degrees are presented. The evaluation results are commented.

2 TEACHING METHODOLOGY

2.1 Course contents

In both courses the teaching approaches focus on the application of the methods to real data. Emphasis is placed in the use and interpretation of statistical software outputs for estimation of statistical parameters and regression models. Table 1 shows the statistic course contents at the environmental science degree related with this topic and the number of delivery hours. Table 2 gives the same information in the computer science degree. In this case the course only includes an introduction by means of the simple linear regression model, with a brief explanation of model inference.

Table 1. Course contents related with linear regression, and number of hours shown by classroom lectures and in-class activities, and computer laboratory sessions, in the environmental science degree.

Content	Number of hours		
	Classroom lectures and In-class activities	Computer room sessions	**TOTAL**
Bivariate descriptive analysis Contingency tables Scatterplots Covariance Correlation Introduction to simple linear regression	3	2	5
Linear regression Simple linear regression Multiple linear regression (polynomial regression, dummy variables, interactions, inference)	7	4	11
Introduction to time series analysis Estimation of linear trends Joint estimation of trend and seasonal components	2	1	3

In both degrees, the idea of association is introduced using contingency tables. The first examples are 2x2 tables for qualitative bidimensional variables. More complex examples with a greater number of possible values are then studied. The application of this method is extended to quantitative bidimensional variables, explaining that in this case it is necessary to group the components values in a small number of intervals before building the table, especially for continuous measures. The examples studied are used to introduce the concepts of marginal and conditional distributions for bidimensional variables. The notion of statistical dependence and independence between the two components is illustrated with several examples.

It is shown that contingency tables are not informative enough to analyze associations for quantitative measures. Scatterplots are presented as alternative and more effective means to study associations with this type of data. Concepts and interpretation of several plots is exposed with negative and positive linear relationships, or other non linear associations. The intuitive concept of the equation to compute the covariance sample parameter, is explained using a scatterplot. The covariance matrix and statistical software to obtain it, are introduced. Covariance value depends on the measurements units. Therefore it is useless to interpretation and compare the association between variables measured in different scales. As an alternative the correlation sample coefficient is studied as well as its properties, with several examples.

All these concepts are given in the module of bivariate descriptive analysis in the two degrees (see Tables 1 and 2). In the environmental science degree they are studied at the beginning of the semester, and at the end of the subject the simple regression model is treated from a descriptive point of view. Formulas are given to estimate the regression line parameters using the sample parameters.

In the computer science engineering degree, these concepts are studied at the end of the semester. After the explanation of the bivariate descriptive analysis, the simple regression model is introduced.

Students are taught the different steps involved in building a model: clear statement of the problem, data collection and initial exploration (application of scatterplots and correlation), model formulation (practical interpretation of parameters), tentative model fitting and inference, validation of assumptions, and the use of the model to make predictions and decisions regarding the problem. The estimation step involves the explanation of the least squares method, and the formulas using the sample descriptive parameters. The inference methods are the determination coefficient R^2, the analysis of variance of the model and the t-student hypothesis test on the intercept and slope. Residual analysis is given to validate model assumptions. The last section of the subject introduces the model extension to include quadratic effects.

Table 2. Course contents related with linear regression, and number of hours shown by classroom lectures and in-class activities, and computer laboratory sessions, in the computer science engineering degree.

Content	**Number of hours**		
	Classroom lectures and In-class activities	Computer room sessions	**TOTAL**
Bivariate descriptive analysis Contingency tables Scatterplots Covariance Correlation	2,5	1,5	4
Introduction to linear regression models Model formulation and estimation Goodness of fit of the model: R^2 coefficient and analysis of variance Hypothesis test on the model parameters Validation of the model: residuals analysis Modelling quadratic effects	6	1,5	7,5

In the environmental science degree the linear regression approach is studied in a module that includes simple and multiple models, at the end of the semester before descriptive analysis of time series and more advanced methods (principal component analysis). The multiple regression models are introduced as an extension of the simple one. The first extension is to include quadratic, cubic or other orders polynomial effects. Qualitative factors effects are modelled using dummy variables. Next step is the inclusion of interactions. Estimation is done with the least squares method. Besides the inference methods already mentioned for the computer science engineering degree, the adjusted R^2 coefficient and the residuals sum of squares increment hypothesis test are also given. Validation step involves several graphical methods for the residuals. Linear regression models are also applied in the time series module to estimate linear trends and the joint estimation of trend and season components, from a descriptive point of view.

2.2 In-class activities

In both courses the teaching approaches focus on the application of the methods to real data. At the beginning of the course, an activity similar to one exposed in [7] is done in the lecture class. Students are asked to answer an anonymous questionnaire on their age, gender, weight, height, their opinion regarding the most serious problems affecting the country in general, etc. These data are used during the course to illustrate the methods application. Environmental science degree students apply the methods to atmospheric pollution data gathered by the automatic network of Valencia, meteorological observations or water quality measurements. These data allow making connections with other compulsory subjects on the syllabus such as Atmospheric Pollution, Meteorology, or Hydrology. In this degree there are two computer laboratory sessions. The first one is made at the beginning of the semester and is related with scatterplots, covariance, correlation and simple linear regression. The other one is made at the end of the semester and students apply the multiple linear regression model. Emphasis is placed in the use and interpretation of statistical software outputs for plotting, estimation of statistical parameters and regression models. At the Polytechnic University of Valencia, there are two programs available in the campus net to perform statistical analysis: Statgraphics and SPSS. The software has the features to be adequate for supporting learning in statistics courses [8]. Two examples of the activities that they have to do to evaluate their understanding of the concepts, are the following.

Activity 1: analysis of the relationship between daily gas consumption and mean temperature with Statgraphics

This exercise introduces Statgraphics options for the descriptive analysis of tye relationship between the two components of a numerical bidimensional variable. A sample of daily observations of energy consumption and mean temperature is used [9]. Data are in the file GAS.ASF.

First step of the analysis: Represent the scatterplot of the two characteristics (see Fig 1)

Question 1: Interpret the plot. Which type of relationship can be assumed between the two components? Is there a strong or weak correlation?

Question 2: Compute the sample covariance and correlation coefficient (see Table 3). Do these values reflect the information shown in the scatterplot?

Question 3: Perform the simple linear regression fitting:

CONSUMPTION= a + b TEMPERATURE

And interpret the practical meaning of the two parameters a and b (see Table 4)

Question 4: Use the estimated model to forecast the average gas consumption for days with a mean temperature of 7ºC.

Students have to repeat the analysis using the other program SPSS. This example is used to illustrate that correlation and association may sometimes indicate causation. Other example are presented to show that this is not necessarily true in all cases when correlation is detected.

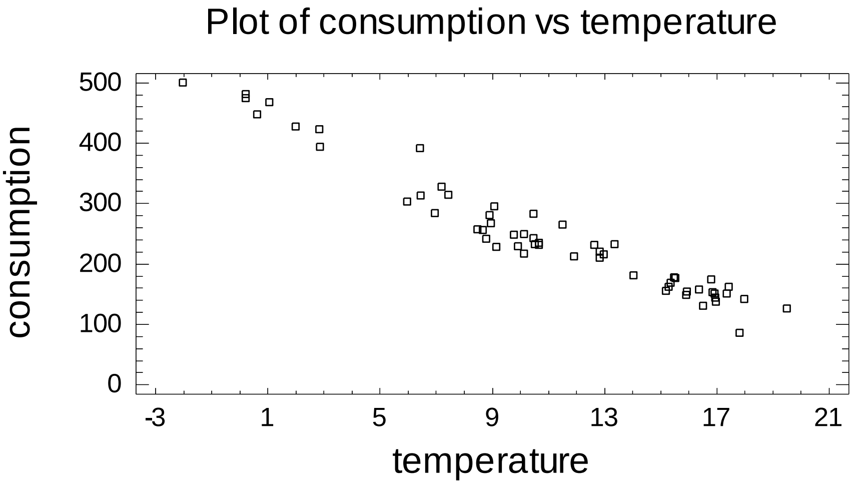

Fig. 1 Example of scatterplot that environmental science students obtain to evaluate this technique

Table 3. Covariance and correlation matrix for the consumption and temperature data

```
Covariances
               CONSUMO           TEMPER
--------------------------------------------------------
CONSUMO   10490,3             -535,59
            (   57)              (   57)

TEMPER      -535,59              29,0832
            (   57)              (   57)
--------------------------------------------------------
Covariance
(Sample Size)

Correlations
               CONSUMO           TEMPER
--------------------------------------------------------
CONSUMO                          -0,9697
                                  (   57)
TEMPER     -0,9697
          (   57)
--------------------------------------------------------
Correlation
(Sample Size)
```

Table 4. Estimation of the intercept a and the slope b of the simple linear regression fitting

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------
Dependent variable: CONSUMPTION
Independent variable: TEMPERATURE
---------------------------------
Parameter       Estimate
---------------------------------
Intercept       448,426
Slope          -18,4158
---------------------------------
```

Activity 2: Effect of rain water ph on the productivity of soybeans

An experience has been done [10] to analyze the effect of acid rain on the productivity of field-grown soybeans. Several plants were exposed to water with different ph, and half of them were protected. The response variable is the soybeans weight (gr) per plant.

Question 1: Formulate the multiple linear regression model that could be used to study protection, linear effect of ph, and their interaction. Explain the practical meaning of the model parameters.

Data are:

Table 5. Activity 2 data

	Rain water pH in the experiment			
Protection	2,7	3,3	4,1	5,6
Yes	10,1	10,9	11,7	13,1
No	10,55	10,62	11,11	11,42

Question 2: Estimate the model parameters using the software and apply the inference methods to study whether they are significant or not (see Table 6 and Fig. 2).

Table 6. Multiple regression fitting results for activity 2

```
Comparison of Regression Lines

Dependent variable: Productivity
Independent variable: pH-2,7
Level codes: protec

Number of complete cases: 8
Number of regression lines: 2

Multiple Regression Analysis
-----------------------------------------------------------------------------------------
                                      Standard         T
Parameter                Estimate       Error       Statistic       P-Value
-----------------------------------------------------------------------------------------
CONSTANT                  10,5333     0,0890214   118,324         0,0000
pH-2,7                   0,319725     0,0543532    5,88236        0,0042
protec=1                -0,328419     0,125895    -2,60867         0,0595
pH-2,7*protec=1          0,696668     0,076867     9,0633          0,0008
-----------------------------------------------------------------------------------------

                    Analysis of Variance
-------------------------------------------------------------------------------------------------
Source          Sum of Squares    Df  Mean Square    F-Ratio      P-Value
-------------------------------------------------------------------------------------------------
Model                5,91828       3     1,97276       141,25       0,0002
Residual           0,0558652       4    0,0139663
-------------------------------------------------------------------------------------------------
Total (Corr.)        5,97415       7

R-Squared = 99,0649 percent
R-Squared (adjusted for d.f.) = 98,3635 percent
Standard Error of Est. = 0,118179
```

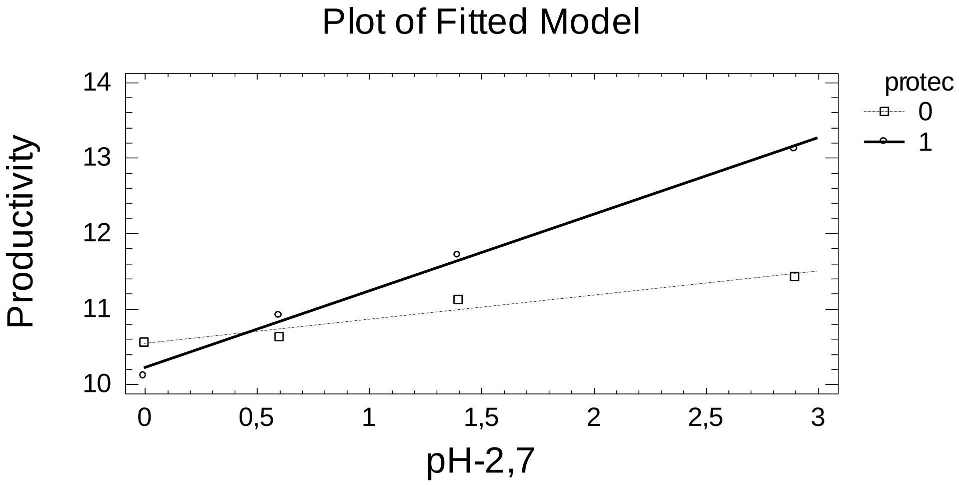

Fig. 2 Plot of the two multiple regression lines fitted.

Question 3: Using the final model, obtain the prediction of average productivity for plants protected when pH is 5.

In the next part of the course they apply again these models to analyze trends and seasonality in environmental time series. In the computer science engineering course there is one laboratory session to study correlation and simple regression. They only use the software Statgraphics. One of the activities they have to develop in the laboratory is:

Activity 3: Association between a system run time and the number of users

File pract8.sf3 contains a sample of 25 observations of run time of a benchmark and number of users of a system. After opening the file from Statgraphics data sheet, answer the following questions.

Question 1: In light of the scatterplot information (Fig. 3), is there any relationship between the two components? Is this association linear or exponential? Give a tentative value for the correlation coefficient.

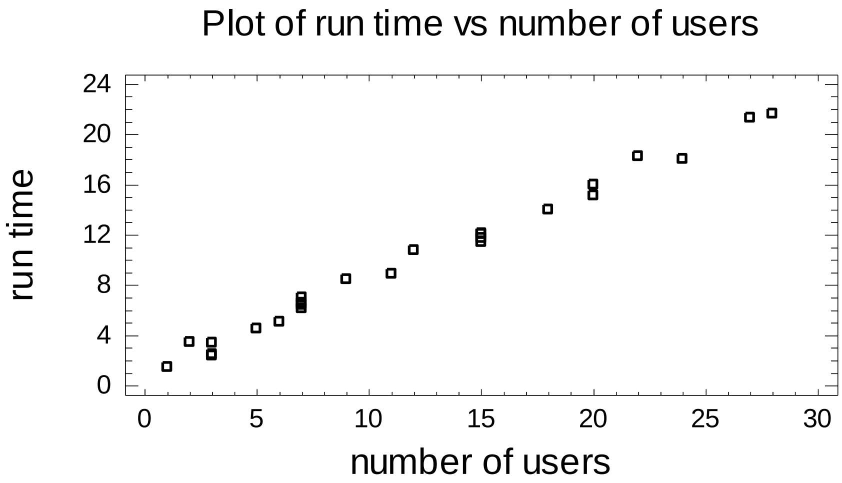

Fig. 3. Scatterplot of a benchmark run time versus number of users of a system

Question 2: Compute the correlation coefficient using the software (Table 7), and compare the value with the answer you gave in the first question.

Table 7. Correlation matrix of activity 3 data

```
Correlations
               run time      N_users
---------------------------------------------
run time                      0,9955
                              (   25)

N_users         0,9955
               (   25)
---------------------------------------------
Correlation
(Sample Size)
```

Question 3: Would it be adequate to obtain the regression line? Compute the values of the intercept and the slope using the formulas given in the lectures. Repeat the estimation with the software (Table 8). Which is the interpretation of the two parameters?

Question 4: What average run time can be predicted when the number of users is 25?

Question 5: What percentage of the run time variance is associated with the linear effect of the number of users?

Question 6: Compute the residuals, and plot them in normal probability paper (Fig. 4). Can we assume normal distribution of data?

Table 8. Regression line estimation results

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: run time
Independent variable: N_users
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic       P-Value
-----------------------------------------------------------------------------
Intercept      1,04573        0,21043        4,96951        0,0001
Slope          0,736479       0,014546      50,6311         0,0000
-----------------------------------------------------------------------------


                        Analysis of Variance
-----------------------------------------------------------------------------------------------------
Source             Sum of Squares      Df  Mean Square      F-Ratio       P-Value
-----------------------------------------------------------------------------------------------------
Model                    859,077          1     859,077        2563,51        0,0000
Residual                   7,70771       23      0,335118
-----------------------------------------------------------------------------
Total (Corr.)            866,785         24

Correlation Coefficient = 0,995544
R-squared = 99,1108 percent
R-squared (adjusted for d.f.) = 99,0721 percent
Standard Error of Est. = 0,578894
```

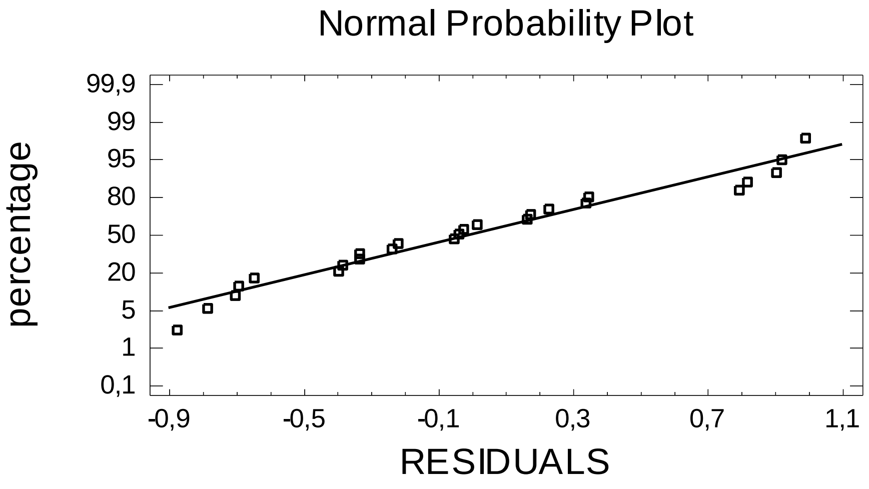

Fig. 4. Normal probability plot of residuals of the linear regression of activity 3.

2.3 Assessment results

Common misunderstandings when using these methods are considering that correlation does necessarily imply causation, visual interpretation of scattterplots, and visually quantifying the correlation. Other mistakes are done when interpreting the model parameters, in the step of definition of dummy variables, and when applying inferential methods to evaluate the models. Students have to understand that a model is a useful approximation to reality but that it should never be considered the final word. Table 9 shows the mean and standard deviation of the marks obtained by the environmental science course students in the two sessions related with linear regression. The first activity is bivariate descriptive analysis and the second multiple linear regression models, which posed more difficulties and had a lower average mark. Table 10 contains the analysis of variance of the marks. It can be seen that there significant differences between the two sessions but not between students. There were 17 students evaluated.

Table 9. Average and standard deviation of the marks obtained by environmental science students.

	Average	Standard Deviation
Biv.desc analysis	9,08824	0,712287
Multiple linear regression	8,11765	1,04999

Table 10. Analysis of variance of marks by student and activity

Analysis of Variance for mark

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
activity	8,00735	1	8,00735	12,74	0,0026
student	15,7022	16	0,981388	1,56	0,1910
RESIDUAL	10,0551	16	0,628447		
TOTAL	33,7647	33			

In the computer science engineering course the results of the laboratory class devoted to correlation and regression, are given in Tables 11 and 12. Table 11 shows the averages and standard deviations of the students' marks in the three groups that are considered. The total average is lower than in environmental science students. Table 11 contains the analysis of variance to compare the marks between groups. The comparison indicates that there is no significant difference between groups.

Table 10. Average and standard deviation of the marks obtained by computer science students.

Group	N.students	Average	Standard Deviation
A	26	7,46154	1,39229
B	47	8,19681	1,96016
C	32	7,42188	2,45232
Total	105	7,77857	2,02518

Table 11. Analysis of variance of marks by group.

Analysis of Variance for mark

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
group	14,906	2	7,45302	1,85	0,1630
RESIDUAL	411,633	102	4,03562		
TOTAL	426,539	104			

3 CONCLUDING REMARKS

The application of statistical methods to analyze associations, pose several difficulties. The interpretation of sample parameters and plots is a first step in the analysis that depends on the personal perception. A common misunderstanding is that correlation and association imply causation. Model formulation and subsequent explanation of the practical meaning of parameters is necessary to elaborate models that really represent the associations that descriptive analysis reveals. The assessment of the use of inferential tools to study effects significance, has had worse results in the introductory statistics course of the computer science degree. The introduction of the methods using real data examples improves the results.

REFERENCES

[1] Garfield, J. & Ben-Zvi, D. (Eds) (2004). The Challenge of Developing Statistical Literacy, Reasoning and Thinking. Dordrecht, The Netherlands: Kluwer Academic Publishers.

[2] ABET (2013). About ABET. Available on-line http://www.abet.org/about-abet/.

[3] McKenzie, C.R.M. & Mikkelsen, L.A. (2007). A Bayesian view of covariation assessment. Cognitive Psychology 54(1), pp. 33-61.

[4] Casey, S.A. (2014). Teachers’ knowledge of students’ conceptions and their development when learning linear regression. In K. Makar (Ed.), Proceedings of the 9th International Conference on Teaching Statistics, Flagstaff, USA.

[5] Al-yahya, S.A. & Abdel-halim, M.A. (2013). A Successful Experience of ABET Accreditation of an Electric Engineering Program. IEEE Transactions on Education, 56(2), pp.165-173.

[6] Piegorsch, W.W. & Edwards, D. (2002). What Shall We Teach in Environmental Statistics?. Environmental and Ecological Statistics, 9 (2), pp. 125-150.

[7] Scheaffer, R.L., Gnanadesikan, M., Watkins, A. & Witmer, J. (1996). Activity-Based Statistics. New York: Springer-Verlag.

[8] Biehler, R. (1997). Software for Learning and for Doing Statistics. International Statistical Review, 65(2), pp. 167-189.

[9] Romero, R. & Zúnica, L. (2005). Métodos Estadísticos en Ingeniería. Valencia, Spain: Polytechnic University Editorial.

[10] Evans, L.S., Lewin, K.F., Pattin, M.J. & Cunningham, E.A. (1983) Productivity of field-grown soybeans exposed to simulated acidic rain. New Phytologist, 93(3), pp.377-388.