TEACHING REGRESSION MODELS: AN APPLICATION WITH SIMULATIONS

Irene Schiattino[1], Rosa Montaño[2], Claudio Silva[1], Carmen Acuña[3] and Isabel Ormeño[2]
[1]Escuela de Salud Pública, Universidad de Chile, Chile
[2]Universidad de Santiago de Chile, Chile
[3]Bucknell University, United States of America
ischiattino@med.uchile.cl

The availability of efficient statistical software makes it possible to enhance the strategies for teaching statistics with activities based in methods of stochastic simulation. The objective of this work is to offer, from an educational perspective, a review and presentation of several articles that address the subject of simulations. Simulated data sets with different features, but that result in the same estimates of the regression parameters are presented. Extensions to multivariate methods, such as multiple regression, are reviewed. Some ad-hoc programs written in R, the PROC IML of SAS, and PROC MATA of STATA are produced and used in the illustrations presented.

INTRODUCTION

The availability of efficient statistical software makes it possible to enhance the strategies for teaching statistics with activities based in methods of stochastic simulation (Hair, 2006; Laviolette 1994.) The use of simulation facilitates the exploration of the underlying structures in a dataset and the understanding of abstract concepts (Chance & Rossman, 2006). In this line of thought a sequence of works that stands out begins with Edwards (1959).

Regression analysis is one of the most frequently used methods in biomedical research. Unfortunately, sometimes such analyses are careless, without paying attention to the real underlying structure. The aim of this work is to present a selection of statistical papers that we deem adequate to warn our students against such analyses. Our selected papers refer to the simulation of datasets with different features, but with the same estimates for the regression parameters. Programs in R, SAS and STATA were used to generate the datasets in our examples.

METHODS

A qualitative bibliographical search was done using the databases EBSCOhost and Ingentaconnect. The keywords used were: “simulated datasets,” “equal estimates of regression parameters,” “teaching regression,” “regression analysis,” “statistics,” and “graphical methods.” The articles reviewed in this work were selected because they present algorithms for generating univariate or multivariate data with some specified features, but that resulted in the same estimates of the regression parameters. Computational programs in SAS, R, and STATA were written and used for preparing the applications presented. Because of space limitations these programs cannot be included in this report, but they can be obtained by contacting the authors.

RESULTS

The search in EBSCOhost and Ingentaconnect using the keywords listed above yielded sixteen articles. They illustrate, by means of simulations, the effect that the modifications on the data have on the regression parameters, but only six of them present algorithms that produce different databases with the same estimates of the regression parameters.

Article 1: Constructing simple correlation problems with predetermined answers (Edwards, 1959)

Objectives: To determine two datasets that have, standard deviation, coefficient of correlation and regression coefficients with rational values.

Method: Manual computations organized in twelve steps for generating data according to the stated objective.

Article 2: Computer generation of data sets for homework exercises in simple regression (Searle & Firey, 1980)

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

Objective: To generate different datasets that in simple linear regression lead to the same estimates for the slope, the intercept, and the correlation coefficient, with the same ANOVA table, and a mean square error that is a perfect square.
Method: An extension of the methodology proposed by Edwards (1959) where n (x_i,y_i) pairs are generated, with the first (n-1) pairs generated using Edwards' method. The nine step algorithm is based on number theory and on the special form of the ANOVA table.
Computer program: REGDATA (written by the authors in Fortran and LP/1).

Article 3: Constructing Regressions with controlled features: A method of probing regression performance (Velleman & Ypelaar,1980)

Objective: To identify aspects of the data structure that can influence the performance of robust regression procedures.
Method: Algorithm for constructing regression problems in which each of the factors affecting the performance of robust regression procedures (regression dimensions, coefficient of determination, least squares coefficients, collinearities, and the residual and leverage of a certain number of key points) can be controlled separately.
Computer program: ROSEPACK, a publicly available package for robust regression is used to illustrate the method presented in M-estimate regression.

Article 4: Creating Realistic Data Sets with Specified Properties via Simulation (Goldman & McKenzie, 2009)

Objective: To generate datasets in the following situations:

- One variable with specified average and standard deviation.
- Two variables with specified correlation coefficient, averages and standard deviations.
- Bivariate data with specified regression coefficients.

Method: Algorithms based on the properties of correlation and least square regression are presented.
Computer program: MINITAB. The algorithms can also be implemented in EXCEL.
To illustrate the usefulness of these algorithms, we created a dataset using the methodology proposed by Farnsworth, 2009 (original dataset). The regression equation is $y = 3.02 + 5.99x$ with r^2=0.99, and n=125. The average and standard deviation of the *y* variable are 20.26, and 9.08, respectively. The average and standard deviation of *x* are 2.88 and 1.51, respectively.

Figure 1 shows the scatterplot, the residual plot and the normal probability plot of these data.

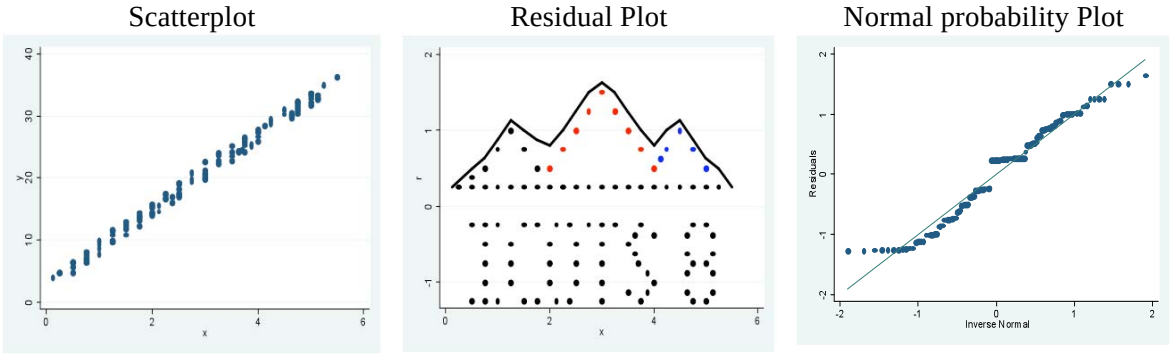

Figure 1. Characteristics of the original data set

Using computer programs for generating datasets with given correlation coefficients, averages and standard deviations written in STATA and R using the algorithm of Goldman and McKenzie (situation 2) we simulated new datasets from the original data. The descriptive statistics, correlation coefficient and regression equation of these simulated data are the same as those of the original data; the new variables differ from the original ones in their minimum and maximum

values, their structure and normality. Figure 2 shows the scatterplot, residual plot and normal probability plot of the simulated data. The regression assumptions are satisfied by the new dataset.

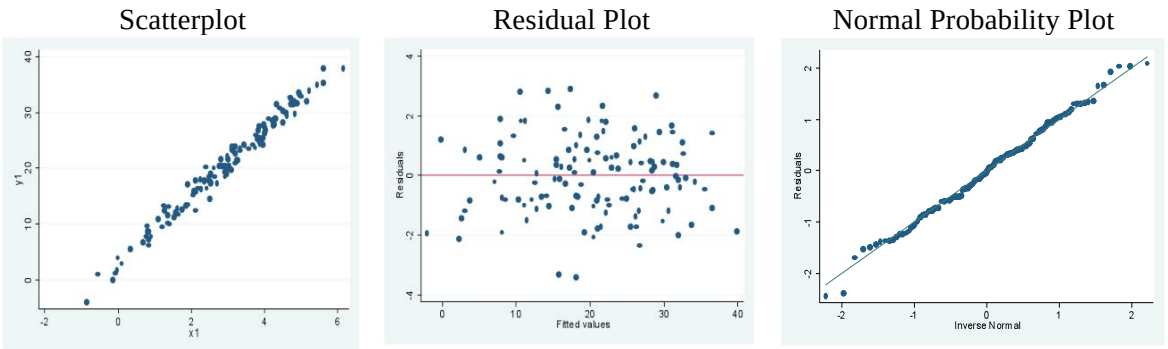

Figure 2. Characteristics of the first data set simulated from the original data

Article 5: Generating different data sets for linear regression models with the same estimates (Zocchi & Manly, 2006).

Objective: To generate different data sets that in a regression analysis lead to the same least squares estimates of the regression coefficients.

Method: A matrix algorithm based on the methodology of Huh and Jhun (2001) is developed.

Computer program: None is mentioned.

Computer programs were written using the procedure MATA in STATA for implementing the proposed matrix algorithm. These programs were applied to data from Kallow and Tang. (Daniel, 1999, p 478.) Table 1 shows part of the 19 original data points (y_1, x_1, x_2) and the corresponding generated values (y_2, x_1, x_2) (the differences appear in the response variables and the residuals); for both datsets the regression equation is $\hat{y}_1 = \hat{y}_2 = 4.52 - 0.0517x_1 + 0.170x_2$, r^2=0.79. The descriptive statistics for each of the variables are the same in both datasets, but the data differ in their structure. Figure 3 shows the matrix scatterplots of the original and simulated datasets.

Table 1. Kallow and Tang's data (n = 19), original and simulated data and descriptive statistics

Id	y_1	x_1	x_2	$\hat{y}_1$	r_1	y_2	x_1	x_2	$\hat{y}_2$	r_2
1	4.165	1	0	4.472	-0.307	3.572	1	0	4.472	-0.9
2	3.731	1	0	4.472	-0.74	4.449	1	0	4.472	-0.023
3	5.748	1	0	4.472	1.276	4.294	1	0	4.472	-0.178
4	4.437	1	0	4.472	-0.035	5.101	1	0	4.472	0.63
.	.	.	.	.	.	.	.	.	.	.
17	7.564	20	21.227	7.102	0.461	7.779	20	21.227	7.102	0.677
18	7.216	20	21.127	7.085	0.13	5.591	20	21.127	7.085	-1.494
19	13.5	24	63.213	14.042	-0.542	12.387	24	63.213	14.042	-1.654
Mean	6.28	9.32	13.16	6.28	0	6.28	9.32	13.16	6.28	0
Std.Dev	2.88	7.21	16.64	2.56	1.31	2.88	7.21	16.64	2.56	1.31
Min	3.37	1	0	3.75	-2.93	1.8	1	0	3.75	-2.78
Max	13.5	24	63.21	14.04	2.95	12.39	13.5	63.21	14.04	2.69

Article 6: *Generating Test Data with Independently Controllable Features for Multivariate General Linear Forms (Heiberger, Velleman & Ypelaar, 1983)*

Objective: To control some characteristics of a general a linear model observed in a given dataset and replicating them in a simulated dataset.

Method: To obtain the Singular Value Decomposition (SVD) of the Y, X, and E matrices of the observed linear model; then fix some of their factors, and reproduce the properties of interest in a simulated dataset.

Computer program: None is proposed

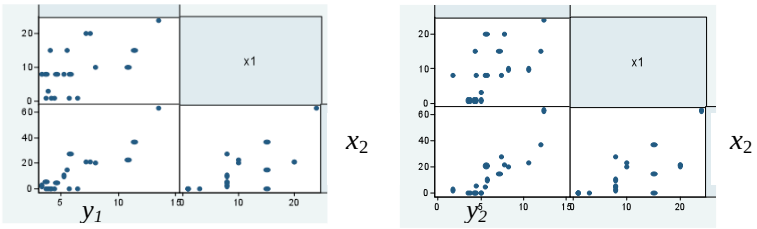

Figure 3. Scatterplots of the original and of the simulated datasets

CONCLUSION

The bibliographical search performed using the keywords described in the methods section yielded few articles that present procedures for generating datasets with different features that produce the same least squares estimates in a regression analysis. However, we found numerous articles that illustrate, by means of simulations, the effect that the modifications on the data have on the regression parameters. Both types of articles are useful for the teaching of statistical methods. To have available computer programs for constructing datasets for simple and multiple regression written for different statistical packages facilitates the educational activities (homework, tests, etc.), helps focus attention on the model assumptions and weigh the consequences of their violation, and allows students to experience situations in which the "algebraic" description of the data in a regression analysis is deceptive.

Our main purpose must be to get our students to recognize the importance of exploratory data analysis as the bases of any sound statistical analysis.

REFERENCES

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician, 27*(1), 17-21.

Chance, B., & Rossman, A. (2006). Using Simulation to Teach and Learn Statistics. *Proceedings of the 7th International Conference on Teaching Statistics*, Auckland, New Zealand: International Association for Statistical Education.

Daniel, W. (1999). *Biostatistics: A foundation for analysis in the Health Sciences* (7th Edition). New York: Wiley & Sons, Inc.

Edwards B. (1959). Constructing Simple Correlation Problems with Predetermined Answers. *The American Statistician, 13*(5), 25-27.

Farnsworth, D. L. (2009). Playing with Residuals. *Teaching Statistics, 31*(2), 81-84.

Goldman R. N., & McKenzie, J. D. Jr. (2009). Creating Realistic Data Sets with Specified Properties Via Simulation. *Teaching Statistics*, *31*(1), 7-11.

Hair, J. F. Jr. (2006). Successful Strategies for Teaching Multivariate Statistics. *Proceedings of the 7th International Conference on Teaching Statistics*, Auckland, New Zealand: International Association for Statistical Education.

Heiberger, R. M., Velleman P. F., & Ypelaar, M. A. (1983). Generating Test Data with Independently Controllable Features for Multivariate General Linear Forms. *Journal of the American Statistical Association, 78*(383), 585-595.

Laviolette, M. (1994). Linear Regression: The Computer as a Teaching Tool. *Journal of Statistics Education, 2*(2). Online: www.amstat.org/publications/jse/v2n2/laviolette.html.

Searle, S. R., & Firey, P. A. (1980). Computer Generation of Data Sets for Homework Exercises in Simple Regression. The American Statistician, *34*(1), 51-54.

Velleman, P. F., & Ypelaar, M. A. (1980). Constructing Regression with Controlled Features: A Method of Probing Regression Performance. *Journal of the American Statistical Association, 75*(372), 839-844.

Zocchi, S. S., & Manly, B. F. J. (2006). Generating Different Data Sets for Linear Regression Models with the Same Estimates. *Proceedings of the 7th International Conference on Teaching Statistics*, Auckland, New Zealand: International Association for Statistical Education.