Abnormally wide confidence intervals in logistic regression: interpretation of statistical program results[1]

Jokin de Irala,[2] Rafael Fernandez-Crehuet Navajas,[3] and Amparo Serrano del Castillo[4]

ABSTRACT *This study describes the behavior of eight statistical programs (BMDP, EGRET, JMP, SAS, SPSS, STATA, STATISTIX and SYSTAT) when performing a logistic regression with a simulated data set that contains a numerical problem created by the presence of a cell value equal to zero. The programs respond in different ways to this problem. Most of them give a warning, although many simultaneously present incorrect results, among which are confidence intervals that tend toward infinity. Such results can mislead the user. Various guidelines are offered for detecting these problems in actual analyses, and users are reminded of the importance of critical interpretation of the results of statistical programs.*

Logistic regression is a method of statistical analysis commonly used in epidemiology. It is an attractive technique, since it permits relatively simple prediction of odds ratios (*1*). An odds ratio (OR) is a measure of association between categorical responses, something that is important in epidemiology because it represents a relative estimate of risk when no direct risk estimate is possible (*2–5*).

For each study subject, logistic regression predicts the probability, Π, of obtaining the desired outcome, conditioned by the values of the independent variables for such subjects and following a specific model (*6*). The formula used for this purpose is as follows:

$$\prod = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_n x_n}}$$

where β_0 is the model's constant, β_n are the coefficients of the model's independent variables, and x_n are the independent variables. Currently, statistical software packages can predict the beta coefficients of this model and its corresponding standard errors through iterations that solve probability equations. Once such values have been obtained, it is possible to estimate different odds ratios (OR) and their confidence intervals (CI). For a qualitative variable, such programs can calculate the OR of the presence of a response in a category of this variable relative to the reference category, on the condition that the qualitative variable is codified using the marginal method (*6, 7*). With continuous variables, it is possible to estimate the OR for the presence of a response at each level of interest of this variable (for example, for every increment of 10 units of the continuous variable) (*6, 7*).

The complexity of the results available to the user is determined by the statistical program's level of sophistication; however, programs generally produce beta coefficients and their standard errors as a minimum. With these results, the user can estimate the OR by calculating the exponential function of each coefficient and can obtain the corresponding CI from the standard errors if the program does not produce such data automatically (*6, 7*).

[1] A Spanish version of this article has previously been published in this journal (Vol. 1, No. 3, 1997, pp. 230–234) under the title "Intervalos de confianza anormalmente amplios en regresión logística: interpretación de resultados de programas estadísticos."

[2] University of Córdoba, Faculty of Medicine, Córdoba, Spain. Mailing address: Universidad de Córdoba, Facultad de Medicina, Departamento de Medicina Preventiva y Salud Pública, Córdoba 14004, Spain. Fax: (349) 57-218278.

[3] Queen Sofía Hospital *(Hospital Reina Sofía)*, Department of Preventive Medicine, Córdoba, Spain.

[4] University of Córdoba, School of Medicine, Córdoba, Spain.

Despite the ease with which these programs produce results, it is essential to properly understand the underlying logistic regression methods. A relatively important matter which frequently confuses the logistic regression user is that of numerical problems. These problems include, *inter alia*, the instability of the model due to inadequate sample size; the problem of "complete separation" that occurs when all subjects whose outcome variable is equal to 1 can be perfectly separated from those whose outcome variable is equal to 0, based on their characteristics; the colinearity problem; and, finally, the problem of profiles with a frequency equal to 0.

This study focuses on the last case, more specifically on the problem of the presence of a cell with a value equal to 0 in the context of categorical independent variables. To better understand this problem, consider the simple case of calculating the OR from a four-cell table for a hypothetical case-control study. The values of the four cells would be: *a* (risk factor present, disease present); *b* (risk factor present, disease absent); *c* (risk factor absent, disease present); and *d* (risk factor absent, disease absent). If cell *b* or *c* has a value equal to 0, the OR cannot be predicted using the ratio of cross-products method (*ad/bc*), since the resulting value would be infinite.

A similar situation can occur in logistic regressions performed on actual data sets. While statistical programs differ in the way they approach this problem, the programs often yield incorrect results. Therefore, it is important to draw attention to this fact to prevent such results from being accepted and published.

The goals of the study reported here were to evaluate and describe the performance of eight statistical programs when cells with a frequency equal to 0 are present in a data set; to point out the warning signs that could alert the user to the presence of numerical problems; and, on this basis, to note the importance of critical interpretation of results obtained from statistical programs.

MATERIALS AND METHODS

The study was carried out using a simulated data set that is summarized in Table 1. This set comprised 60 subjects, had a dependent and dichotomous variable coded 0 and 1 to represent the absence and presence of response, respectively, and had a qualitative independent variable x with three categories. In the logistic regression model, this qualitative variable is represented by two indicator variables, x_2 and x_3, created using the marginal method and representing categories 2 and 3, respectively. The reference category is x_1. Group 3 has no subjects with a dichotomous variable response of 0, a circumstance creating a cell with a value of 0.

This data set was used by eight statistical programs: BMDP, EGRET, JMP, SAS, SPSS, STATA, STATISTIX, and SYSTAT. A logistic regression of variable y over the qualitative variable x was performed by each program. A regression model was obtained from each program, and an attempt was made to predict the coefficients of the model's variables as well as the standard errors. In other words, an attempt was made to reach "convergence," through the process of solving probability equations by iterations.

RESULTS

The presence of a cell with a frequency equal to 0 means that an operation must be performed for which there is no definable outcome; where it is impossible to estimate the coefficients and standard errors; and where it makes no sense to estimate the OR for the category of the qualitative variable with a frequency equal to 0. Due to the numerical problem introduced, the programs fail; neither convergence nor correct results are obtained. In fact, because of the failure to achieve convergence, the calculation process is interrupted (at times the message "estimation terminated" is displayed).

Beyond this point, each program responded differently, producing the results indicated in Table 2. With the exception of STATISTIX, all the tested programs displayed some type of warning message, usually indicating that convergence could not be reached; and the SPSS program indicated that the process was interrupted because the probability logarithm showed an insignificant decline. However, none of the programs indicated that the difficulty was due to a cell with a frequency equal to 0. The SAS program revealed in greater detail and by the use of the pound sign (#) that the co-

TABLE 1. A simulated data set for 60 subjects; no subject in category 3 of variable *x* has variable *y*'s "0" attribute, so the frequency of subjects in cell (3, 0) is equal to 0

		Variable *y* = 0	Variable *y* = 1	
Variable *x*	1	13	7	20
	2	8	12	20
	3	0	20	20
		21	39	60

TABLE 2. Results obtained with the various statistical programs tested by applying logistic regression to the simulated data set shown in Table 1

Programs[a] and model variables	Coefficient	Standard error	Odds ratio	95% confidence interval	Warning message
BMDP LR (1990, PC)					
Constant	–0.62	0.469	0.54	(0.20, 1.38)	Yes
x_2	1.03	0.654	2.79	(0.77, 10.04)	"No convergence"
x_3	10.82	22.300	0.50E + 05	(0.2E – 14, 0.1E + 25)	
EGRET (1992, 0.03, PC)					
Constant	–0.62	?[c]	0.54	No[b]	Yes
x_2	1.03	?	2.79	No	"No convergence"
x_3	33.89	?	0.52E + 15	No	
JMP (3.01, MAC)					
Constant	–0.62	0.469	0.54	(.,.)	Yes
x_2	1.03	0.654	2.79	(0.79, 10.5)	Convergence by objective,
x_3	12.80	99.840	3.70E + 05	(.,.)	not by gradient
SAS (1987, 6.04, PC)					
Constant	–0.62	0.469	No	No	Yes
x_2	1.03	0.654	No	No	"No convergence"
x_3	38.21#[c]	.[c]	No	No	
SPSS (1990, 4.01, PC)					
Constant	–0.62	0.469	No	No	Yes
x_2	1.03	0.654	2.79	No	"Estimation interrupted"
x_3	10.82	36.730	0.50E + 05	No	
STATA (1990, 3.10, MAC)					
Constant	—[d]	—	No	No	Yes
x_2	1.03	0.654	2.79	(0.77, 10.04)	Variable x_3 not utilized
x_3	—	—	—	—	No results
STATISTIX (4.0, PC)					
Constant	–0.62	0.469	No	No	No
x_2	1.03	0.654	2.79	(0.77, 10.04)	
x_3	10.19	16.210	2.65E + 04	(0.00, 0.2E + 18)	
SYSTAT (1991, 2.00, MAC)					
Constant	–0.62	0.469	No	No	Yes
x_2	1.03	0.654	2.00	(0.77, 10.04)	"No convergence"
x_3	16.82	447.458	0.20E + 08	(0,.)	

[a] Program (year, version, PC/Macintosh).
[b] No = not regularly calculated by the program.
[c] The signs "#", "?", and "." are displayed by the corresponding statistical program; "#" = infinite estimate.
[d] "–" = no response.

efficient estimate of variable x_3 was actually an "infinite" or undeterminable estimate (38.21#). STATA reported that the indicator variable x_3 had been deleted.

The BMDP and STATISTIX programs yielded the results that are regularly calculated (beta coefficient, standard error, and OR confidence interval), whereas JMP, SPSS, and SYSTAT omitted the OR confidence interval. In contrast, EGRET, SAS, and STATA did not yield the standard error of the coefficient or the OR confidence interval; and STATA did not calculate the coefficient for the x_3 variable.

Three programs calculated the OR for the constant and one showed its corresponding CI, although these predictions are not useful.

DISCUSSION

It is important to emphasize that statistical programs which analyze data containing numerical problems, such as the problem involved in this study, do not actually reach convergence. Therefore, the results shown depend entirely on the particular moment at which the program interrupts its iterative mathematical process, i.e., the preestablished convergence criterion. This accounts for why the programs produced different values of the x_3 coefficient (see Table 2). The more attempts made by the program, the greater will be the value of the x_3 coefficient and subsequent estimates, with a tendency toward infinity.

Since the warning signs observed in these analyses are not specific, it is possible that their presence and significance could go undetected by an inexperienced program user. Programs that yield results, even though such results are incorrect (BMDP and STATISTIX), can be the most misleading for this type of user. Programs that do not produce OR confidence intervals for the problematic variable but do generate values for its coefficient and standard error (JMP, SPSS, and SYSTAT) are

less confusing. Nevertheless, the drawback remains that researchers can calculate the OR, estimating the exponential function of the given coefficient, and can also calculate the CI using the standard error, even though both procedures yield meaningless values within this context. Programs such as EGRET, SAS, and STATA would be the most appropriate for the inexperienced user, since it is impossible to calculate parameters for variables with numerical problems on these programs.

The differences that exist between these statistical programs would not pose obstacles to the everyday user of logistic regression, who would understand any warning sign in the results and would heed written warnings or messages in the outcome. The presentation of an abnormally high coefficient (in the tens), followed by a standard error of the same magnitude, could be a valuable clue to detecting a numerical problem. However, such clues are not restricted to problems with null value cells but are also found in situations involving such things as the presence of colinearity. The absence of a calculated parameter, such as the OR and its CI, in a program that regularly calculates these values, is also a cause for concern. Program estimates that tend toward infinity must be rejected. In general, it is necessary to insist that such results not be published, and that the underlying problem be resolved before applying the model.

It should also be recalled that numerical problems can arise when two variables have several categories with small frequencies and these variables are introduced together in a model, as in the case of an interaction term. By creating an interaction variable, the product of two variables, cells with no frequency can appear, along with subsequent numerical problems. To avoid this problem, it is recommended that there be sufficient numbers of study subjects possessing each attribute of the qualitative variables employed (*6*, *7*).

Some authors have stressed the importance of performing a univariate data analysis before conducting the multivariate analysis, so as to better understand the quality and nature of the variables used and to get a general idea of the associations existing at the univariate level (*8*). Moreover, performing a descriptive and univariate analysis with all the variables before doing the multivariate analysis permits detection of one or more variables or categories rendered invalid by insufficient frequencies. In such a case, the variable or a group of its categories should be eliminated, following a biologic guideline, in order that there be fewer groups and higher frequencies in each. In any case, one should not be satisfied with models automatically produced by statistical software packages. On the contrary, it is essential to always assess the goodness of fit before accepting the model's validity (*9*). It is also preferable to use programs that provide specific warnings, or that do not display values for coefficient predictions in situations such as those involving numerical problems like the ones cited in this study.

REFERENCES

1. Bautista LE. "Razón relativa" y "tasa relativa" como traducciones de odds ratio and hazard ratio [letter]. *Bol Oficina Sanit Panam* 1995;119: 278–280; and Tapia JA. Respuesta. *Bol Oficina Sanit Panam* 1995;119:280–282.
2. Rothman KJ. *Modern epidemiology*. 6th ed. Boston: Little Brown; 1986.
3. Hanley J. Utilizaciones adecuadas del análisis multivariante. *Rev Salud Publica (Barcelona)* 1989;1:45–74.
4. Hennekens CH, Buring JE. *Epidemiology in medicine*. 6th ed. Boston: Little Brown; 1987.
5. Kleinbaum DG, Kupper LL, Muller KE. *Applied regression analysis and other multivariate methods*. 2nd ed. Boston: PWS-KENT; 1988.
6. Hosmer DW. *Computer analysis of health sciences data: PH744 course*. Amherst: University of Massachusetts; 1992.
7. Hosmer DW, Lemeshow SA. *Applied logistic regression*. New York: Wiley; 1989.
8. Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med* 1993;118:201–210.
9. Hosmer DW, Lemeshow SA, Taber S. The importance of assessing the fit of logistic regression models. *Am J Public Health* 1991; 81:1630–1635.

Manuscript received on 18 December 1995. Revised version accepted for publication on 22 April 1996.

RESUMEN

Intervalos de confianza anormalmente amplios en regresión logística: interpretación de resultados de programas estadísticos

Este estudio describe el comportamiento de ocho programas estadísticos (BMDP, EGRET, JMP, SAS, SPSS, STATA, STATISTIX y SYSTAT), al realizar una regresión logística con una base de datos simulados en la cual existe un problema numérico creado por la presencia de una celda con frecuencia igual a 0. Los programas responden de manera heterogénea a este problema. La mayor parte de ellos ofrecen señales de alarma, aunque muchos presentan, simultáneamente, resultados incorrectos entre los cuales destacan los intervalos de confianza que tienden al infinito. Estos resultados pueden desorientar al usuario. Se describen diferentes criterios orientativos para detectar estos problemas en situaciones de análisis reales y se recuerda la importancia de la interpretación crítica de los resultados de programas estadísticos.