Understanding data in clinical research: a simple graphical display for plotting data (up to four independent variables) after binary logistic regression analysis

José Luis Mesa

Department of Medical Physiology, School of Medicine, University of Granada, E-18071 Granada, Spain

Received 4 March 2003; accepted 23 October 2003

Summary In clinical research, suitable visualization techniques of data after statistical analysis are crucial for the researches’ and physicians’ understanding. Common statistical techniques to analyze data in clinical research are logistic regression models. Among these, the application of binary logistic regression analysis (LRA) has greatly increased during past years, due to its diagnostic accuracy and because scientists often want to analyze in a dichotomous way whether some event will occur or not. Such an analysis lacks a suitable, understandable, and widely used graphical display, instead providing an understandable logit function based on a linear model for the natural logarithm of the odds in favor of the occurrence of the dependent variable, Y. By simple exponential transformation, such a logit equation can be transformed into a logistic function, resulting in predicted probabilities for the presence of the dependent variable, P(Y = 1). This model can be used to generate a simple graphical display for binary LRA. For the case of a single predictor or explanatory (independent) variable, X, a plot can be generated with X represented by the abscissa (i.e., horizontal axis) and P(Y = 1|X) represented by the ordinate (i.e., vertical axis). For the case of multiple predictor models, I propose here a relief 3D surface graphic in order to plot up to four independent variables (two continuous and two discrete). By using this technique, any researcher or physician would be able to transform a lesser understandable logit function into a figure easier to grasp, thus leading to a better knowledge and interpretation of data in clinical research. For this, a sophisticated statistical package is not necessary, because the graphical display may be generated by using any 2D or 3D surface plotter.

Introduction

In clinical research, suitable visualization techniques of data after statistical analysis are crucial for the researches’ and physicians’ understanding. Common statistical techniques to analyze data in clinical research are logistic regression models. Concretely, the application of binary logistic regression analysis (LRA) in medical research has greatly increased during past years [1]. This is because: (1) researchers and physicians often want to analyze in a dichotomous way whether some event will occur or not by using other variables, and (2) binary LRA has been showed to provide the same or better diagnostic accuracy than other methods such as multistep algorithms [2], ordered categorical models [3], or even neural networks [4].

Unfortunately, the binary LRA lacks a simple, understandable, and widely used graphical display. Consequently, LRA is often misinterpreted in medical research [5], thus leading to: (1) misleading
conclusions by researchers; and (2) misinterpretations of published data by physicians and general readers. Therefore, suitable visualization techniques of data for binary LRA are needed. These techniques (if simple) would improve the researchers’, physicians’ and general readers’ understanding of data and results in clinical research. This paper provides a simple graphical display for binary LRA.

**Understanding the binary logistic regression model**

Binary LRA is used to predict (and hence to explain) the presence or not of any event (i.e., dependent variable, "Y") by other variables (i.e., independent or explanatory variables, "X_1, . . . , X_p"). The dependent variable is a discrete variable (i.e., dichotomous, with dummy variables coded 0, 1, where 0 is either the presence or absence of the event, and 1 is the opposite). The independent or explanatory variables may be continuous or discrete (with dummy variables). Since the model for binary LRA assumes that the dependent variable, Y, is dichotomous, the binary LRA does not model this dependent variable directly as other non-logistic regression models (e.g., linear models). Rather, binary LRA is based on probabilities associated with the two possible values of Y (either 0 or 1). In theory, the hypothetical population proportion of cases for which Y = 1 is defined as P = P(Y = 1).

Then, the theoretical proportion of cases for which Y = 0 is 1 - P = P(Y = 0). It is assumed that there is a set of explanatory variables, X_1, . . . , X_p, that are related to Y and, therefore, provide information for predicting the probability of the presence/absence of the event. Since probabilities and odds obey multiplicative rules, taking the logarithm of the odds allows for the simpler, additive model (because logarithms convert multiplication into addition). This is the model for logit analysis.

**The logit analysis**

The logit analysis of LRA is based on a linear model for the natural logarithm of the odds (i.e., the log-odds) in favor of Y = 1:

\[
\log \left( \frac{P(Y = 1|X_1, . . . , X_p)}{1 - P(Y = 1|X_1, . . . , X_p)} \right) = \log \left( \frac{P}{1 - P} \right) = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p
\]

\[= \alpha + \sum \beta_j X_j, \quad \forall j = 1, . . . , p.\]

There is a (relatively) simple exponential transformation for converting log-odds back to probability. This inverse transformation is the logistic function for LRA, easier to grasp.

**The logistic function**

Since \( \log_a X = Y \) implies that \( a^Y = X \), the logit analysis can be transformed into:

\[
[P/(1 - P)] = e^{\sum \beta_j X_j}. \quad \forall j = 1, . . . , p.
\]

Then, the value of \( P' = P(Y = 1) \) is:

\[
P' = e^{\sum \beta_j X_j}/(1 + e^{\sum \beta_j X_j}).
\]

On the other hand, if the interest is in knowing the \( P(Y = 0) \), due to the mathematical relation, \( 1 - e^a/(1 + e^a) = 1/(1 + e^a) \), the probability for a 0 response is:

\[
1 - P' = P(Y = 0) = 1/(1 + e^{\sum \beta_j X_j}).
\]

This form of the model (the logistic function) is the most practical because it results in predicted probabilities for the presence/absence of the event of interest. Additionally, the logistic function permits to design a graphical display for LRA.

**Graphical display for single predictor model**

For the case of a single predictor or explanatory variable, X, a plot can be generated with X represented by the abscissa (i.e., horizontal axis) and \( P(Y = 1|X) \) represented by the ordinate (i.e., vertical axis). The resulting curve is S-shaped. To illustrate an example, suppose a relation between the probability of survival/death within 7 yr and the body mass index (BMI) in humans related by the logit analysis:

\[
\log \left( \frac{P(Y = \text{survival|BMI})}{1 - P(Y = \text{survival|BMI})} \right) = 13.7 - 0.477 \text{BMI}.
\]

Note that the dependent variable is dichotomous with "1" representing survival, and "0" death. In order to plot the graphical function relating the probability of survival within 7 yr and BMI, the logit equation must be transformed into a logistic function (see the logistic function), of the form:

\[
P(Y = \text{survival|BMI}) = e^{13.7 - 0.477 \text{BMI}}/(1 + e^{13.7 - 0.477 \text{BMI}}).
\]

A scatter diagram of this relation is easily generated by using any dispersion grapher, with BMI represented by the abscissa and \( P(Y = \text{survival|BMI}) \) represented by the ordinate (Fig. 1).
Graphical display for multiple predictor models

A simple and understandable graphical display for binary LRA presenting two, three, or even four explanatory variables is possible.

Two explanatory variables

Suppose an additional explanatory variable (gender) in the previous equation, yielding the logistic function:

\[
P(Y = \text{survival} | \text{BMI}, \text{gender}) = \frac{e^{13.5 - 0.475 \times \text{BMI} + 0.4 \text{gender}}}{(1 + e^{13.5 - 0.475 \times \text{BMI} + 0.4 \text{gender})}.
\]

Due to the dichotomous nature of gender (0 = male, 1 = female), the equation can be separated into two logistic functions, one for males and another one for females:

\[
P(Y = \text{male survival} | \text{BMI}) = e^{13.5 - 0.475 \times \text{BMI}}/(1 + e^{13.5 - 0.475 \times \text{BMI}})
\]

\[
P(Y = \text{female survival} | \text{BMI}) = e^{13.5 - 0.475 \times \text{BMI} + 0.4}/(1 + e^{13.5 - 0.475 \times \text{BMI} + 0.4}).
\]

Therefore, there are two equations easily generated in the same scatter diagram (Fig. 2(a)). Note that there are two equations (and hence two graphical functions) due to the dichotomous nature of gender; if the categorical variable had 10 possible values, the graphical display would generate 10 different graphical functions in the same scatter diagram. This graphic is useful when there are one continuous and another one categorical explanatory variables in the equation. Suppose now a logistic function defined by two continuous explanatory variables (BMI and age):

\[
P(Y = \text{survival} | \text{BMI}, \text{age}) = e^{17.5 - 0.475 \times \text{BMI} - 0.1 \times \text{age}}/(1 + e^{17.5 - 0.475 \times \text{BMI} - 0.1 \times \text{age}}).
\]

The graphical display may be generated by using any 3D surface grapher. For this, probability of survival (i.e., the dependent variable) is represented by the ordinate (vertical axis) and both BMI and age (i.e., explanatory variables) are represented by the two horizontal axes (Fig. 2(b)). As can be seen in this figure, probability of survival depends on BMI and age.

Three explanatory variables

Suppose the same two previous explanatory variables, but adding the dichotomous variable gender ("0" representing male, "1" female), yielding the logistic function:

\[
P(Y = \text{survival} | \text{BMI}, \text{age}, \text{gender}) = e^{17.5 - 0.475 \times \text{BMI} - 0.1 \times \text{age} + 0.23 \text{gender}}/(1 + e^{17.5 - 0.475 \times \text{BMI} - 0.1 \times \text{age} + 0.23 \text{gender}}).
\]
Since the variable introduced (gender) is dichotomous, the equation could be separated into two logistic functions:

\[
P(Y = \text{male survival}|\text{BMI}, \text{age}) = \frac{e^{17.5 - 0.475 \text{BMI} - 0.1 \text{age}}}{(1 + e^{17.5 - 0.475 \text{BMI} - 0.1 \text{age}})},
\]

\[
P(Y = \text{female survival}|\text{BMI}, \text{age}) = \frac{e^{17.5 - 0.475 \text{BMI} - 0.1 \text{age}}}{1 + e^{17.5 - 0.475 \text{BMI} - 0.1 \text{age}} + e^{0.23}}.
\]

Although both functions can be generated in the same graphic (see two explanatory variables), it is not advisable due to the overlay of both surface graphics. We have solved this problem designing only one, but relief, surface graphic (Fig. 3). In this graphic, the low points in the relief surface represent the function for males (gender = 0), and the high points represent the function for females (gender = 1). Therefore, at higher differences in the relief surface, higher differences with gender on probability of survival for a determined BMI and age.

Four explanatory variables

Suppose that the dichotomous variable lipid profile ("0" representing healthy, "1" unhealthy lipid profile) is also added, yielding the logistic function:

\[
P(Y = \text{survival}|\text{BMI, age, gender, lipid profile}) = \frac{e^{17.5 - 0.475 \text{BMI} - 0.1 \text{age} - 0.23 \text{gender} - 1 \text{ lipid profile}}}{(1 + e^{17.5 - 0.475 \text{BMI} - 0.1 \text{age} - 0.23 \text{gender} - 1 \text{ lipid profile}}).
\]

Due to the dichotomous nature of lipid profile (healthy = 0, unhealthy = 1), two different functions can be generated:

\[
P(Y = \text{survival presenting healthy lipid profile}|\text{BMI, age, gender}) = \frac{e^{17.5 - 0.475 \text{BMI} - 0.1 \text{age} + 0.23 \text{gender}}}{(1 + e^{17.5 - 0.475 \text{BMI} - 0.1 \text{age} + 0.23 \text{gender}}),}
\]

\[
P(Y = \text{survival presenting unhealthy lipid profile}|\text{BMI, age, gender}) = \frac{e^{16.5 - 0.475 \text{BMI} - 0.1 \text{age} + 0.23 \text{gender}}}{(1 + e^{16.5 - 0.475 \text{BMI} - 0.1 \text{age} + 0.23 \text{gender}}).
\]

These two functions are two relief surface graphics, which can be generated as two adjacent graphics in the same figure (Fig. 4). If the variable added had more than two possible values, the number of graphical functions generated would be equal to the categories of the variable.
Conclusion

This paper provides a simple way of expressing results graphically after binary logistic regression analysis (up to four explanatory variables: two continuous and two discrete) in clinical research. Therefore, any researcher or physician will be able to transform a lesser understandable logit function into a figure easier to grasp, thus leading to a better knowledge and interpretation of data. For this, a sophisticated statistical package is not necessary, because the graphical display may be generated by using any 2D or 3D surface plotter.

Acknowledgements

No present or past conflict of interest exists for the author or his institution.

References