ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A new explanatory index for evaluating the binary logistic regression based on the sensitivity of the estimated model

Héctor M. Ramos *, Jorge Ollero, Alfonso Suárez-Llorens

Departamento de Estadística e Investigación Operativa, Facultad de CC. Económicas, av. Duque de Nájera 8, CP 11002, Cádiz, Spain

ARTICLE INFO

Article history:
Received 17 May 2016
Received in revised form 22 August 2016
Accepted 30 August 2016
Available online xxxx

Keywords:
Binary logistic regression
McFadden index
ROC curve
Sensitivity index

ABSTRACT

We propose a new explanatory index for evaluating the binary logistic regression model based on the sensitivity of the estimated model. We previously formalized the idea of sensitivity and established the principles a statistic should comply with to be considered a sensitivity index. We apply the results to a practical example and compare the results with those obtained utilizing other indices.

1. Introduction

Binary logistic regression is a frequently applied procedure used to predict the probability of occurrence for some binary outcome using one or more continuous or categorical variables as predictors. The logistic model relates the probability of occurrence P of the outcome counted by Y to the predictor variables X_i, with the occurrence of an event normally indicated by one and nonoccurrence by zero. The model takes the form

$$P(Y = 1) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)]}.$$

The regression parameters are typically obtained using maximum likelihood estimation. Hosmer and Lemeshow (2000) provide a detailed discussion of the goodness of fit of the logistic regression model, particularly the well-known and commonly used Hosmer–Lemeshow goodness-of-fit test. When the predicted probabilities resulting from logistic regression are for classification purposes, there is a need for additional indices of model fit. Known as pseudo-R^2 indices, these indices play a role similar to R^2 in ordinary least squares (OLS) regression. Some indices, such as those formulated by Cragg and Uhler (1970), McFadden (1974), Maddala (1983), Cox and Snell (1989), and Nagelkerke (1991) compare the likelihood functions for an intercept-only and full model. In particular, the McFadden pseudo-R^2 is defined as follows:

$$R^2_{MF} = 1 - \frac{\log L(full)}{\log L(null)}.$$

McKelvey and Zavoina (1975) propose a pseudo-R^2 based on a latent model structure, where the binary outcome results from discretizing a continuous latent variable relate to the predictors through a linear model. This pseudo-R^2 is then the proportion of the variance of the latent variable explained by the covariate. Cameron and Windmeijer (1997) define yet

* Corresponding author.
E-mail addresses: hector.ramos@uca.es (H.M. Ramos), jorge.ollero@uca.es (J. Ollero), alfonso.suarez@uca.es (A. Suárez-Llorens).

http://dx.doi.org/10.1016/j.spl.2016.08.022
0167-7152/

another pseudo-R^2 index as the proportionate reduction in uncertainty, as measured by the Kullback–Leibler divergence, given the inclusion of the regressors. Windmeijer (1995) and Smith and McKenna (2013) provide broad and detailed studies of the different pseudo-R^2 indices available for binary choice models. Theoretical results regarding the convergence and asymptotic normality of pseudo-R^2 indices are available in Hu et al. (2006).

It is worth mentioning that there does not exist an equivalent statistic to the classical R^2 coefficient in OLS regression when analyzing data with a logistic regression. It is well known that estimates arrived at through an iterative process and they are not computed to minimize variance, hence the OLS approach to goodness-of-fit does not apply. All previous indices are called pseudo-R^2 because they look like the classical R^2 in the sense that they are on a similar scale, ranging from 0 to 1, with higher values indicating better model fit.

Alternatively to pseudo-R^2 indices, there exist other exploratory methods for evaluating a logistic regression model. In accordance with the usual interpretation of R^2 for linear models they try to capture the model's ability to predict a single observation. Mainly, those methods take into account the differences between observed and predicted outcomes and a good model is defined when having a high explanatory power, i.e., a good prediction of an observation is only possible when the success probability is close to 1 or 0. Among others, we first highlight the coefficient of discrimination de Tjur (2009). It has a lot of intuitive appeal and the definition is very simple. For each of the two categories of the dependent variable is computed the average of the estimated probabilities and then the difference between them is computed. Its interpretation is based on the histograms of the empirical distributions of both the fitted values for the failures and the fitted values for the successes. Intuitively, the greater is the difference, the better is the model. Secondly, it is also worth to mention some indices based on the concept of concordance and discordance. Basically, concordance tells us the association between actual values and the values fitted by the model in percentage terms, namely, we compute the number of pairs where the one had a higher model score than the model score of zero and the opposite for discordance. Some examples are the classical well-known Kendall's tau, Goodman–Kruskal Gamma and Somers' D (Somers, 1962). Finally, another valuable contribution is given by the receiver operator characteristic (ROC) curve. The ROC curve represents "true positive" and "false positive" classification rates as a function of different classification cutoff values for the predicted probabilities resulting from the logistic regression. In literature, several indices of accuracy have been proposed to summarize ROC curves. In particular, the area under the curve (AUC) index is one of the most commonly used, see, for instance, Hosmer and Lemeshow (2000), Metz (1978) and Fawcett (2006) for a detailed explanation of the basic principles of ROC analysis. The AUC index is related to the Somers' D by the following relationship: $D_{YX} = 2AUC - 1$ (see Newson, 2002).

In this paper we propose a new exploratory index to measure the predictive power of a logistic regression model. From a theoretical point of view, it is not a proper pseudo-R^2 index and analogously to other exploratory methods mentioned before it is based on the differences between observed and predicted outcomes. This new index is based on the sensitivity of the estimated binary logistic regression model. The term sensitivity within this context implies the quality of the model to predict correctly the value of the dependent variable. Most statistical software packages provide, as a self-evaluation of the estimated model, the number of individuals in the sample that the model predicts correctly as a function of the critical values considered (the cutoff points). In other words, each cutoff point c_p provides the percentage of sampled individuals observed with values of one that the estimated model predicts correctly by assigning $P[Y = 1] > c_p$. These values, which decrease as c_p increases, are the components of a vector we refer to as S_1, whose dimension is determined by the number of cutoff points $c_{p_1}, c_{p_2}, \ldots, c_{p_n}$ considered. Associated with the vector we have a vector X_1 of the same dimension and of which the components are the number of sampled individuals observed with values of one that the model predicts accurately for each cutoff point c_{p_i}, but inaccurately for $c_{p_{i+1}}$ $(i : 1, \ldots, n-1)$. Likewise, self-evaluation of the model provides the percentage of individuals of the sample observed with values of zero that the estimated model predicts correctly by assigning $P[Y = 1] < c_p$. These values, which decrease as c_p decreases, are the components of a vector we refer to as S_0. In this case, we consider the components of S_0 in decreasing order. Associated with S_0, we have a vector X_0 of which the components are the number of individuals of the sample observed with values of zero that the model predicts accurately for each cutoff point c_{p_i}, but inaccurately for $c_{p_{i-1}}$.

To illustrate, consider a model estimated using a sample of 20 individuals observed with the value one and 10 individuals observed with the value zero, and that we select deciles $c_{0.1}, c_{0.2}, \ldots, c_{0.9}$ as cutoff points. Let us assume that S_1 and S_0 are:

$$S_1 = (1,\ 1,\ 0.95,\ 0.9,\ 0.8,\ 0.8,\ 0.7,\ 0.55,\ 0.35); \qquad S_0 = (1,\ 1,\ 1,\ 0.9,\ 0.9,\ 0.9,\ 0.7,\ 0.6,\ 0.3).$$

The corresponding X_1 and X_0 vectors will then be:

$$X_1 = (0,\ 0,\ 1,\ 1,\ 2,\ 0,\ 2,\ 3,\ 4,\ 7); \qquad X_0 = (0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 2,\ 1,\ 3,\ 3).$$

We may interpret the components of S_1 and S_0, within a ROC curve context, in terms of sensitivity and specificity, respectively. However, even though we have this common starting point derived from the ROC curve, we show later on Q3 that there are substantial methodological differences. We intend to introduce the idea of sensitivity as applied to any vector $X(x_i) \in R_n^+$. In the same way that there is a certain intuitive idea that the components of a vector X are "more nearly equal" than the components of another vector Y, we can talk about a certain intuitive idea that the components of a vector X present more sensitivity than the components of another vector Y.

Focus again on vector $X_1 = (0,\ 0,\ 1,\ 1,\ 2,\ 0,\ 2,\ 3,\ 4,\ 7)$. We know that we can measure the inequality of the components of X_1 via certain inequality measures, such as variance, where we consider the underlying dispersion, or the Gini index, where the vector components correspond to income distribution. These measurements and any others corresponding

Please cite this article in press as: Ramos, H.M., et al., A new explanatory index for evaluating the binary logistic regression based on the sensitivity of the estimated model. Statistics and Probability Letters (2016), http://dx.doi.org/10.1016/j.spl.2016.08.022