Expert Systems with Applications 39 (2012) 8573–8583

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Decision support for selecting optimal logistic regression models

Hans van der Heijden *

Surrey Business School, University of Surrey, Guildford GU2 7XH, United Kingdom

ARTICLE INFO

Keywords:
Decision support systems
Logistic regression
Model selection

ABSTRACT

This study concerns itself with providing user support for a decision problem in logistic regression analysis: given a set of metric variables and one binary dependent variable, select the optimal subset of variables that can best predict this dependent variable. The problem requires an evaluation of competing models based on heuristic selection criteria such as goodness-of-fit and prediction accuracy. This paper documents the heuristics, formalizes the algorithms, and eventually presents an interactive decision support system to facilitate the selection of such an optimal model.

This study adds to the sparsely studied domain of expert systems for social science researchers, and makes three contributions to the literature. First, the study formalizes a number of heuristics to arrive at optimal logistic regression models. Second, the study presents two computational algorithms that incorporate these formalized heuristics. Third, the paper documents an implementation of these algorithms through an interactive decision support system. The study concludes with a discussion on the risks of relying too heavily on the system and with future opportunities for research.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

This study concerns itself with the problem of model specification in logistic regression analysis. Logistic regression is a special form of regression, where the dependent variable (DV) represents the success or failure of a certain event. Unlike traditional regression, where the DV has an interval or ratio scale, the logistic DV has only two outcomes: true (the event occurred) or false (the event did not occur). Logistic regression analysis is a widely used research method in many areas of the social sciences (see e.g., Peng, Lee, & Ingersoll, 2002). For example, in business and management, logistic regression can predict whether potential customers will convert (Akinci, Kaynak, Atilgan, & Aksoy, 2007), or whether companies will go bankrupt (Chen, 2011; Ge & Whitmore, 2009).

To establish the likelihood of success or failure of a certain event, the researcher typically collects data on a number of variables. After the data is collected, the researcher must formulate a logistic regression model, consisting of a subset of those variables, and use this model to predict whether the event either happens or does not happen. The specification of this model is a crucial activity in logistic regression analysis, because frequently a number of different models will compete for suitability, and yet the outcomes of each of these models may vary significantly. This specific activity, the selection of an optimal logistical model, is the focus of the present study.

The study documents the development of a decision support system (DSS) that helps the researcher to specify logistical regression models and analyze these models on statistical performance measures. It does so by scrutinizing all possible variables, calculating performance measures for all possible models, and highlighting the aspects on which each of these models compete with each other. These aspects include the suitability of the independent variables selected, the goodness-of-fit of each model with the underlying data values, and the classification accuracy of the model in correctly predicting the success or failure of a certain event.

The study makes three contributions to the literature. First, the study reviews and formalizes a number of heuristics to arrive at optimal logistic regression models. Second, the study presents two computational algorithms that incorporate these formalized heuristics. Third, the paper documents an implementation of these algorithms in an interactive decision support system.

Viewed from a wider angle, the study also adds to the application domain of expert systems for social science research. This application domain is notably understudied compared to other

* Tel.: +44 1483 686302; fax: +44 1483 686306.
E-mail address: h.vanderheijden@surrey.ac.uk

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.168

domains such as, say, healthcare and finance. Applications in the social science domain are rare. A recent exception is Chu, Tseng, Tsai, and Luo (2009), who develop a system to explore the statistical significance of group mean differences in a questionnaire data set.

The rest of this paper is organized as follows. First, the paper analyzes and reviews the heuristics that lead to optimal model selection. This analysis results in the formulation of two algorithms. The paper then presents a demonstration of the decision support system, in which each of these algorithms is illustrated with a real data set. A discussion of risks of using the system and opportunities for further research conclude the paper.

2. Theory and algorithms

Logistic regression analysis involves a set of n independent variables and one binary dependent variable. The classic references are Cox and Snell (1989), Hosmer and Lemeshow (1989). Hair, Anderson, Tatham, and Black (1998), Kleinbaum, Kupper, and Muller (2008), Menard (1995) provide detailed, step-by-step introductions to the analysis. To avoid redundant coverage this material will not be repeated here. The appendix of this paper provides a review of the key concepts of logistic regression analysis, insofar as necessary to understand the rest of the paper.

Logistic regression analysis is available in a number of commercial statistical packages such as *SPSS* and *SAS*. These packages allow the researcher to manually specify the model, following which the results are analyzed and reported. The crucial difference between these packages and the system reported here is that the packages take a certain regression model as input, whereas the DSS presented here produces a regression model as output. The relationship is therefore complementary, as both types of systems can be used in sequence: first, the DSS can be used to select the most suitable model, and then this selected model can be further analyzed using the statistical package of choice.

In terms of selecting which model is the most appropriate, it is useful to decompose this selection process into three stages. The first stage is to go through the set of available variables and select potential independent variables. The study will refer to these independent variables as the candidate independent variables, or candidate IVs. The second stage is to construct different models that can be assembled using these candidate IVs. For example, candidate variables $v1$ and $v2$ may lead to three models: $(v1)$, $(v2)$ and $(v1, v2)$. The third stage is to go through these various models, and to select the model that performs best on a chosen performance measure. The study refers to the best model on a performance measure as the optimal model. Because there are several measures available, it is possible to have several optimal models.

The study develops two algorithms, one for the first stage and one for the last stage. The algorithm to recommend candidate independent variables is presented as Algorithm 1 on page 4. The algorithm to recommend an optimal model is presented as Algorithm 2 on page 6. The intermediate stage has no algorithm because the study adopts a brute force approach to this stage: it simply calculates every single possible model that can be formed given a set of candidate variables.

The algorithms incorporate a range of heuristic statistical diagnostics. The sources of these heuristics are Hair et al. (1998) and especially Hosmer and Lemeshow (1989). A description of each algorithm now follows.

Algorithm 1: Algorithm to recommend candidate independent variables (IVs)

```
    Input: $Variables$, $DV$
    Output: $CandidateIVs$
1   CRITICAL_SIG ← .25;
2   $CandidateIVs \leftarrow \{\emptyset\}$;
3   foreach $V \in Variables$ do
4       $VwithDVfalse \leftarrow$ subset($V$) where $DV$ is false;
5       $VwithDVtrue \leftarrow$ subset($V$) where $DV$ is true;
6       $Ttest \leftarrow$ calculateTtest($VwithDVfalse$, $VwithDVtrue$);
7       $include \leftarrow (Ttest.sig <$ CRITICAL_SIG);
8       $logResults \leftarrow$ logisticRegression(DV = $DV$, IV = $V$) ;
9       $ok \leftarrow$ not($logResults$.isHessianSingular());
10      if $ok$ then
11          $G \leftarrow logResults$.getLogLikelihoodBase() ;
12          $2LL \leftarrow logResults$.getLogLikelihoodModel() ;
13          $Chi2Test \leftarrow$ CalculateReduction($G$, $2LL$) ;
14          if $not(include)$ then
15              $include \leftarrow (Chi2test.sig <$ CRITICAL_SIG);
16          end
17          $Wald \leftarrow logResults$.getWald($V$);
18          if $not(include)$ then
19              $include \leftarrow (Wald$.sig $<$ CRITICAL_SIG);
20          end
21      end
22      else
23          $include \leftarrow$ FALSE
24      end
25      if $include$ then
26          $CandidateIVs$ += $V$
27      end
28  end
```

Algorithm 1 assumes that a pool of n variables is available to act as candidate variables for inclusion in the logistical regression model. The algorithm loops through these variables and examines them one by one on their suitability for inclusion. Thus, the algorithm gradually builds a set of included variables (*CandidateIVs*) which it eventually outputs.

The first step is to split each independent variable into two groups, one group, where the DV is true and one, where the DV is false. These two groups will have different means and standard deviations. If the difference between the two is statistically significant, then the IV should be informative for the DV, and consequently, the IV becomes a candidate independent variable (step 4–7).

To examine whether these differences are statistically significant, the algorithm uses the independent sample *t*-test. A liberal 25% critical rejection level ensures that even variables with modest informativeness are included at this stage. This level can be adjusted (step 1) if the algorithm exhibits either too much or too little sensitivity with respect to including variables.

A second step is to conduct a logistic regression analysis with just the variable under study operating as the independent variable. The results allow us to check if the Hessian matrix was singular (step 9). As highlighted in the Appendix, the Newton–Raphson method of searching for maximum likelihood requires an inverted Hessian, and if it is not invertible it is singular. In this specific case, a singular Hessian matrix would tell us that the independent variable is able to predict the dependent variable to such a high degree that it is perfect or near perfect. An example of where this occurs is when the independent variable looks different to the dependent variable but is in reality just another manifestation of it (for example, sales expressed in dollars versus sales expressed in euros).

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات