SciForum
Mol2Net

Categorization of Continuous Variables in a Logistic Regression Model Using the R Package CatPredi

Irantzu Barrio [1,]*, María-Xosé Rodríguez-Álvarez [2] and Inmaculada Arostegui [1,3]

[1] Departamento de Matemática Aplicada, Estadística e Investigación Operativa, Universidad del País Vasco UPV/EHU, Leioa, Spain; E-Mail: inmaculada.arostegui@ehu.eus

[2] Departamento de Estadística e Investigación Operativa, Universidade de Vigo, Vigo, Spain; E-Mail: mxrodriguez@uvigo.es

3 BCAM—Basque Center for Applied Mathematics, Bilbao, Spain

* Author to whom correspondence should be addressed; E-Mail: irantzu.barrio@ehu.eus;
Tel.: +34-94-601-2504.

Published: 4 December 2015

Abstract: Prediction models are gaining importance in many areas such as medicine, meteorology, finance, toxicology, etc. In this context, a common distribution for the response variable is the binomial distribution and hence the logistic regression model is a commonly used regression modeling approach. Although it is not recommended from a statistical points of view due to loss of information and power, the categorization of continuous variables is a common practice in the development of prediction models. However, there are no unified criteria for the selection of the cut points in the categorization process. In order to provide valid cut points whenever a categorization is going to be performed, we have developed a valid methodology to categorize continuous variables in a logistic regression model based on the maximization of the AUC. This methodology has been implemented in an `R` package called `CatPredi`. This is a package of R functions that allows the user to categorize a continuous predictor variable in a univariate or multiple logistic regression model. It provides the optimal location of cut points for a chosen number of cut points and returns the estimated and bias-corrected discriminative ability index for this model. Additionally, it allows a comparison of two categorization proposals for different number of cut points and the selection of the optimal number of cut points.

Keywords: categorization; R package; prediction model

Mol2Net YouTube channel: *http://bit.do/mol2net-tube*
YouTube link: *please, paste here the link to your personal YouTube video, if any.*

1. Introduction

Prediction models are gaining importance in many areas such as medicine, meteorology, finance, toxicology, etc. In this context, a common distribution for the response variable is the binomial distribution and hence the logistic regression model is a commonly used regression modeling approach. Although it is not recommended from a statistical points of view due to loss of information and power, the categorization of continuous variables is a common practice in the development of prediction models. However, there are no unified criteria for the selection of the cut points in the categorization process. In order to provide valid cut points whenever a categorization is going to be performed, we have developed a valid methodology to categorize continuous variables in a logistic regression model based on the maximization of the discriminative ability of the model measured by the area under the ROC curve – AUC.

2. Methods

We have developed a methodology to categorize continuous variables in a logistic regression model. The proposed methodology consists on the maximization of the AUC. Two alternative algorithms have been proposed to select the optimal cut points to categorize continuous variables named *AddFor* and *Genetic* respectively. This methodology has been presented elsewhere (Barrio et al. 2015). This methodology has been implemented in an `R` (R Core Team 2015) package which is explained below.

3. The `CatPredi` Package

`CatPredi` is a package of `R` functions that allows the user to categorize a continuous predictor variable either before or during the development of a prediction model. `The CatPredi` package can be used to categorize a predictor variable in a univariable or a multivariable setting. It provides the optimal location of cut points for a chosen number of cut points, fits the prediction model with the categorized predictor variable and returns the estimated and bias-corrected discriminative ability index for this model. Additionally, it allows a comparison of two categorization proposals for a different number of cut points and the selection of the optimal number of cut points.

The `CatPredi` package has been designed similarly to other packages in R. It has a main function called `catpredi()` which categorizes a continuous predictor variable in a logistic regression model.

Numerical and graphical summaries of the fitted objects can be obtained by using `print.catpredi`, `summary.catpredi` and `plot.catpredi` for `catpredi` type objects. Furthermore, one more main function has been developed `comp.cutpoints` to obtain the optimal number of cut points in a logistic regression model. Table 1 contains a description of all the functions available in the package.

Below, we give a general overview of the package and its general use.

2.1 catpredi() function

The `catpredi()` function provides the optimal cut points to categorize a continuous predictor variable in a logistic regression model.

This function creates an object of class `catpredi`. The main arguments of this function are presented in Table 2. The call to the function is as follows:

```
catpredi(formula, cat.var, cat.points, data, method = c("addfor","genetic"),range=NULL, correct.AUC=TRUE, control = controlcatpredi())
```

In the formula argument users must specify the prediction model setting in which they want to categorize the predictor variable *X* specified in the `cat.var="X"` argument. If the model is a univariate logistic regression model, then the formula would be specified as `Y~1`, with *Y* being the response variable available in the data set specified in the argument `data`. However, if the model is a multiple logistic regression model, and the aim is to categorize the predictor variable *X* together with another predictor *Z*, then the formula would be specified as `Y~Z`. Additionally, in the argument `cat.points` the user must specify the number of cut points to look for. The range argument allows for modifying the range of the predictor variable *X* in which to look for the cut points. By default it would be NULL, which represents the entire range of *X*. Finally, if `correct.AUC` is set to TRUE, the bias-corrected AUC would be estimated.

A numerical summary of the results of the categorization method can be obtained by calling the functions `print.catpredi()` or `summary.catpredi()`. When the method selected is the *AddFor*, the summary returns the estimated AUC for each of the selected cut points. For example, if `cat.points = 2` is chosen, it returns the estimated AUC for one and two cut points. Additionally, if `correct.AUC=TRUE` is chosen it returns the bias-corrected AUC for two cut points. If the method selected is the *Genetic*, estimated cut points, AUC and bias-corrected AUC will be given only for the selected number of cut points.

Table 1. Summary of the functions in the `CatPredi` package.

Function	Description
`catpredi()`	Returns an object with the optimal cut points to categorize a continuous predictor variable in a logistic regression model.
`controlcatpredi()`	Function used to set several parameters to control the selection of the optimal cut points in a logistic regression model
`print.catpredi()`	Print method for objects of type `catpredi`.
`summary.catpredi()`	Produces a summary of the `catpredi` object.
`plot.catpredi()`	Plots the relationship between the continuous predictor and the response variable obtained by fitting a Generalized Additive Model (GAM), together with the location of the optimal cut points.
`comp.cutpoints()`	Compares two objects of type `catpredi`.
`print.comp.cutpoints()`	Print method for objects of type `comp.cutpoints`

Table 2. Summary of the arguments in the `catpredi()` function.

Argument	**Description**
`formula`	A formula giving the model to be fitted.
`cat.var`	Name of the continuous variable to categorize.
`cat.points`	Number of cut points to look for.
`data`	Data frame containing all needed variables.
`method`	The algorithm selected to search for the optimal cut points- `"addfor"` if the *AddFor* algorithm is chosen; otherwise, `"genetic"`.
`range`	The range of the continuous variable in which to look for the cut points. By default `NULL`, i.e., the entire range.
`correct.AUC`	A logical value. If `TRUE` the bias-corrected AUC is estimated.
`control`	Output of the `controlcatpredi()` function.

4. Conclusions

We have developed a user-friendly R package, named `CatPredi`, to obtain optimal cut points to categorize continuous predictor variables in a logistic regression model in practice, either in a univariable or multivariable setting. The `CatPredi` package can be freely download from https://sites.google.com/site/biostit/lineas-de-investigacion/software/catpredi.

Acknowledgments

This work was supported by grants MTM2010-14913, 2012111008, IT620-13, UFI11/52, MTM2011-28285-C02-01, MTM2013-40941-P MTM2014-55966-P. MXRA thanks the Agrupamento INBIOMED from DXPCTSUG-FEDER unha maneira de facer Europa (2012/273).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Barrio, I.; Arostegui, I; Rodríguez-Álvarez, MX; Quintana, JM. A new approach to categorising continuous variables in prediction models: Proposal and validation. *Statistical Methods in Medical Research* **2015 (in press)**.

2. R Core Team. R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. 2015.