Sample size impact on the categorisation of continuous variables in clinical prediction

Irantzu Barrio, Inmaculada Arostegui, and María-Xosé Rodríguez-Álvarez

Departamento de M.A. y Estadística e I.O. Universidad del País Vasco UPV/EHU
irantzu.barrio@ehu.eus

Departamento de M.A. y Estadística e I.O. Universidad del País Vasco UPV/EHU
BCAM- Basque Center for Applied Mathematics
inmaculada.arostegui@ehu.eus

Departamento de Estadística e I.O and Biomedical Research Centre (CINBIO). Universidade de Vigo
mxrodriguez@uvigo.es

Abstract: Recent advances in information technologies are generating a growth in the amount of available biomedical data. In this paper, we studied the impact sample size may have on the categorisation of a continuous predictor variable in a logistic regression setting. Two different approaches to categorise predictor variables were compared.

1 Motivation

Recent advances in information technologies are generating a growth in the amount of available biomedical information and data, what is known as *Big Data*. This fact makes that the data available in some biomedical research studies is getting larger in the last years. The collection and analysis of this data allows improving the quality and efficiency of health care services and enhance the quality and longevity of life [**1**]. Furthermore, the development of prediction models to estimate the risk of developing a particular disease are nowadays relevant in the decision making process [**2**], with a significant growth in the number of predictive models developed in the last years. When developing prediction models to be used in clinical practice, categorised versions of continuous predictor variables are commonly used by clinical researchers [**3**]. It is possible that research studies with a big amount of data may require or use a categorisation of the continuous predictor variables and hence, we think it is necessary to evaluate the impact the sample size may have on the selection of the cut points to categorise the predictor variable.

Recently two different methodologies have been proposed to categorise a continuous predictor variable in a logistic regression setting [**4, 5**]. The first approach is based on a graphical display using generalised additive models (GAM [**6**]) with P-spline smoothers to determine the relationship between the continuous predictor and the outcome. The second approach proposes to select the optimal cut points based on the maximisation of the area under the ROC curve (AUC) of the logistic regression model for the categorised variable. When developing a prediction model to predict the risk of poor evolution of patients with chronic obstructive pulmonary disease (COPD) in the IRYSS-COPD study [**7**], we categorised the predictor variable respiratory rate into three categories using both methods. The same categorisation proposal was obtained with both methods, being 20 and 24 the cut points. In this case, the sample size we had was of 1350 patients. However, we wondered whether same results would be obtained with higher sample sizes. Therefore, this question motivated the work presented in this paper, where the aim is to study how sample is related to the location of the cut points to categorise a predictor variable in a logistic regression setting.

2 Methods

In this section, we briefly describe the two methods considered for the categorisation of a continuous predictor variable. Let us assume that there is a dichotomous response variable Y and a continuous predictor variable X which we wish to categorise in a logistic regression setting.

Categorisation proposal based on GAM with P-spline smoothers:

Barrio et al. (2013) [**4**] proposed a methodology to categorise a continuous predictor variable which consists of creating at least one average-risk category along with high- and low-risk categories based on a GAM with P-spline smoothers. Let $logit(p) = \beta_0 + f(X)$ be the logistic GAM for X where $p = P(Y = 1|X)$ and $f()$ is the smooth function of the GAM regression model. The average-risk category is created by building an interval around the point $x_0 \in X$ for which $f(x_0) = 0$. Lets denote $\hat{\pi}_0 = logit^{-1}(\beta_0 + f(x_0)) = logit^{-1}(\beta_0)$ and $\left(\hat{\pi}_{0_{inf}}, \hat{\pi}_{0_{sup}}\right)$ its 95% confidence interval. The average-risk category $\left(x_{0_{inf}}, x_{0_{sup}}\right)$ is obtained as $f^{-1}(logit(\hat{\pi}_{0_{inf}}) - \beta_0) = x_{0_{inf}}$ and $f^{-1}(logit(\hat{\pi}_{0_{sup}}) - \beta_0) = x_{0_{sup}}$. Thus, this categorisation proposal considers at least three categories. In a context in which only two categories are considered, Hin et al. (1999) [**8**] proposed to dichotomise a continuous variable with x_0 as the optimal cut point.

Optimal categorisation based on the maximisation of the AUC:

Given $k = 2$ the number of cut points set for categorising X in 3 intervals, lets denote as X_{cat} the categorised variable taking values from 0 to 2. Barrio et al. (2015) [**5**] proposed that the vector of 2 cut points $\boldsymbol{v} = (\mathsf{x}_1, \mathsf{x}_2)$ which maximises the AUC of the logistic regression model $P(Y = 1|X_{cat}) = logit^{-1}(\beta_0 + \beta_1 1_{\{X_{cat}=1\}} + \beta_2 1_{\{X_{cat}=2\}})$, is thus the vector of the optimal cut points. In general, this method allows to search for any possible number of cut points, nevertheless in order to compare both method we will focus on $k = 2$ number of cut points.

For ease of notation and interpretation we will refer to these two approaches as the “GAM approach” and the “AUC approach”, respectively.

3 Simulation study

A simulation study was performed under known theoretical conditions that verify linear effects in the logistic regression model. We used this setting to evaluate the performance of the GAM approach and the AUC approach when different sample sizes were used.

Scenarios and set up

The predictor variable X was simulated from a normal distribution separately in each of the populations defined by the outcome ($Y = 0$ and $Y = 1$). Specifically, we considered $X|(Y = 0) \simeq N(0, 1)$ and $X|(Y = 1) \simeq N(1.5, 1)$. When the aim is to maximise the AUC, the theoretical location of cut points to categorise the predictor variable is known [**9**], as well as the AUC associated with the corresponding categorical covariate.

The simulations were performed for different sample sizes assuming the same number of individuals in $Y = 0$ and $Y = 1$. As far as the number of cut points is concerned, $k = 2$ were considered. When using the AUC approach, we considered the *Genetic* algorithm to estimate the optimal number of cut points.

Results

Figure 1 depicts the boxplot of the estimated optimal cut points over 200 simulated data sets for the different sample sizes and each of the categorisation approaches considered. Different results were obtained when the sample size was increased with the two categorisation proposals considered. The AUC approach identified the optimal cut points for any sample size considered (see Figure 1(a)). Under this scenario the theoretical cut points were 0.227 and 1.274.

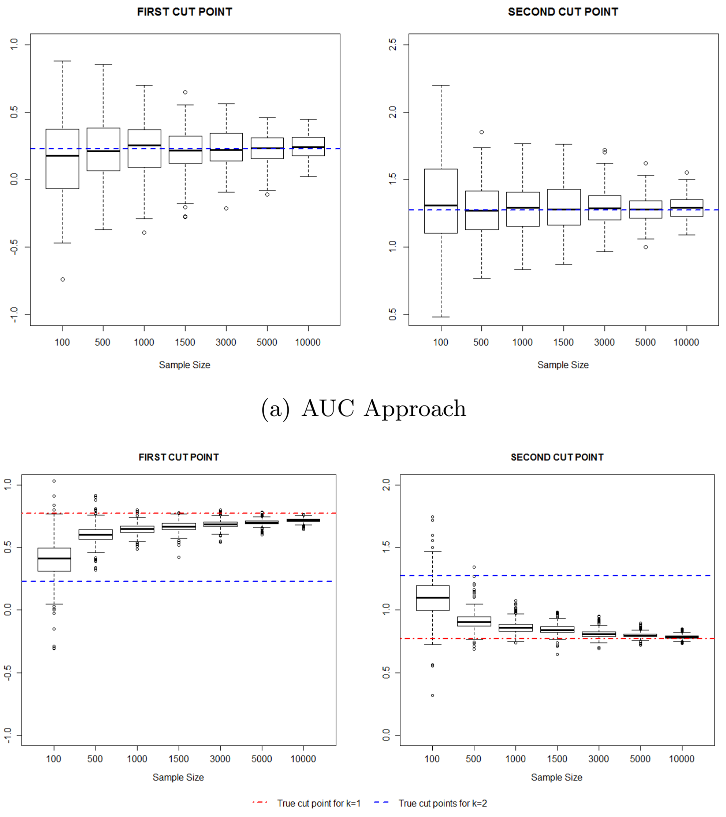
Table 1. Numerical results obtained over 200 simulated data sets when the GAM approach was used to estimate the cut points. Mean, standard deviation and median for the estimated x_0 point together with the mean and standard deviation for the average-risk category values are reported.

Sample Size	x_0		**Average-risk category**	
			Low limit	**Upper limit**
	mean (sd)	**median**	**mean (sd)**	**mean (sd)**
100	0.751 (0.198)	0.759	0.395 (0.183)	1.099 (0.181)
500	0.760 (0.087)	0.752	0.605 (0.085)	0.915 (0.087)
1000	0.754 (0.050)	0.754	0.645 (0.048)	0.863 (0.054)
1500	0.755 (0.046)	0.755	0.665 (0.046)	0.845 (0.048)
3000	0.749 (0.036)	0.749	0.685 (0.036)	0.813 (0.037)
5000	0.751 (0.026)	0.748	0.701 (0.026)	0.800 (0.027)
10000	0.751 (0.019)	0.751	0.716 (0.019)	0.786 (0.019)

On the other hand, the cut points obtained with the GAM approach differed more from those theoretical cut points as the sample size was increased (see Figure 1(b)). In fact, for larger sample sizes, the average-risk category obtained converged to the point x_0 for which $f(x_0) = 0$ which turned out to be close to the theoretical cut point for $k = 1$, i.e, 0.773. This results can be seen in Table 1, where the numerical results obtained with the GAM approach are shown.

4 Conclusions

To summarise, we have seen that the sample size has an impact on the categorisation of a continuous predictor variable. For a large sample size, the GAM approach leads to a very narrow average-risk category which can be interpreted as a unique cut point, being thus equivalent to the proposal of Hin et al. (1999), this is, a dichotomisation of the continuous variable. On the other hand, the AUC approach performs satisfactorily in large sample sizes when looking for two cut points, i.e. categorising the predictor variable into three categories. In general, as long as it is feasible, we recommend the use of the AUC approach. Otherwise, we should take into account that for large sample sizes the GAM approach does not provide an optimal categorisation when the goal is to



(a) AUC Approach

(b) GAM Approach

FIGURE 1. Boxplot of the estimated cut points for $k = 2$, based on 200 simulated data sets for different sample sizes ($N = 100$, $N = 500$, $N = 1000$, $N = 1500$, $N = 3000$, $N = 5000$ and $N = 10000$). From top to bottom: results obtained with the AUC approach and the GAM approach, respectively. True cut points are represented with a dashed line which are $\boldsymbol{v_2} = (0.227, 1.274)$ for $k = 2$ and $\boldsymbol{v_1} = (0.773)$ for $k = 1$.

categorise the predictor variable into three categories.

Funding: this study was supported by grants IT620-13, MTM2011-28285-C02-01, UFI11/52, MTM2013-40941-P, MTM2014-55966-P and Agrupamento INBIOMED from DXPCTSUG-FEDER unha maneira de facer Europa (2012/273).

References

[1] Costa FF. Big data in biomedicine. *Drug Discov Today* 2014; 19: 433-440.

[2] Steyerberg EW. *Clinical prediction models. A practical approach to development, validation, and updating.* New York: Springer, 2009.

[3] Turner E, Dobson J and Pocock J. Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiol Perspect Innov* 2010; 7: 9.

[4] Barrio I, Arostegui I, Quintana JM, et al. Use of generalised additive models to categorise continuous variables in clinical prediction. *BMC Med Res Methodol* 2013; 13: 83.

[5] Barrio I, Arostegui I, Rodríguez-Álvarez MX and Quintana JM. A new approach to categorising continuous variables in prediction models: Proposal and validation. *Stat Methods Med Res* (in press).

[6] Hastie T and Tibshirani R. *Generalized additive models.* London: Chapman & Hall, 1990.

[7] Quintana JM, Esteban C, Barrio I, et al. The IRYSS-COPD appropriateness study: objectives, methodology, and description of the prospective cohort *BMC Health Serv Res* 2011; 11:322.

[8] Hin LY, Lau TK, Rogers MS, et al. Dichotomization of continuous measurements using generalized additive modelling - application in predicting intrapartum caesarean delivery. *Stat Med* 1999; 18: 1101-1110.

[9] Tsuruta H and Bax L. Polychotomization of continuous variables in regression models based on the overall C index. *BMC Med Inform Decis Mak* 2006; 6: 41.