Paper 261-28

Cutpoint Determination Methods in Survival Analysis using SAS®

Mandrekar J. N., Section of Biostatistics, Mayo Clinic, Rochester, MN
Mandrekar S. J., The Ohio State University, Columbus, OH
Cha S. S., Section of Biostatistics, Mayo Clinic, Rochester, MN

ABSTRACT

In the analysis involving data from clinical or epidemiological studies, significant attention is given to continuous variables such as patient's age, blood pressure etc., but the predictive importance of such variables cannot be established easily. Transforming a continuous variable into a categorical variable, usually binary, makes the model more interpretable. The choice of a cutpoint to convert a continuous covariate to a binary covariate needs attention and is often based on biological knowledge about the particular risk factor or on the results already published in other studies. Several data-oriented techniques such as median and upper quartile, and outcome-oriented techniques based on score, Wald and likelihood ratio tests are commonly used in the literature. Contal and O'Quigley (1999) presented a technique that uses log rank test statistic in order to estimate the cutpoint. Their method is computationally intensive and hence is overlooked due to the unavailability of built in options in standard statistical software. We provide a SAS® macro that is easy to implement and does all the necessary computations in a relatively short amount of time. In addition, we critically compare this method with some of the existing methods and discuss the use and misuse of categorizing a continuous covariate.

INTRODUCTION

The variables included in medical or epidemiological data set can either be continuous or categorical. However, it is common strategy in medical and epidemiological research to categorize a continuous variable before evaluating its prognostic impact on the clinical outcome of interest. The variables of interest such as patient's age, blood pressure, cholesterol etc. are often times dichotomized by grouping patients based on the values of the variables, for example, by selecting a cutpoint to classify into two groups, namely high risk and low risk. The interpretations of the terms such as relative risk and/or odds ratio that are based on a dichotomized variable are easier to understand than that involving the original continuous variable.

Often times the risk factors that are measured on a continuous scale are categorized into two or more groups for the purposes of efficient data summarization. It is an investigator's decision to decide how many categories to make and where the category boundaries should be, based on the sample size. Cochran (1968) stated that in many cases, about five well-chosen categories would be sufficient to adequately capture the information in the continuous covariate and control for any potential confounding that arises from categorization. However, in the case of sparse data, it would be impractical to use many categories. There is also no universally accepted method on where to draw the boundaries between categories. Some natural categories may arise from the mode(s) of the underlying distribution of the data. Considerable attention needs to be given to the problems of residual confounding when using open–ended categories (refer Rothman and Greenland, 1998 for a detailed discussion on these issues).

Although the categorization of a continuous covariate can be used in case of different types of regression, in our study we mainly focus our attention to the survival analysis with censored data. The main aim of almost all survival analysis studies is to investigate the importance of potential prognostic factors on a failure time outcome variable like overall survival or disease free survival. In this investigation, we will illustrate some approaches to determine a single cutpoint that helps transform a continuous covariate into a binary variable.

IDENTIFYING A CUTPOINT

The choice of a cutpoint to convert a continuous covariate to a binary covariate needs special attention. There is no single method or criterion to specify which criterion is best and thus the results of analyses from different categorization methods may be different. The choice of a cutpoint is often based on biological knowledge about the particular risk factor or physician's experience or the results already published in other studies. However the cutpoints suggested from other studies in literature may differ from the current study depending on several factors such as the purpose of the study, population under study, study design etc. Standard cutpoints are not readily available for some newly identified or previously unexplored risk factors. In such a situation when the cutpoint is not readily available, statistical methods that determine the cutpoint need to be used.

The statistical methods for cutpoint determination fall into two broad categories: data-oriented and outcome-oriented. The data-oriented methods include dichotomizing a continuous covariate based on certain quantile such as median or upper quartile. The outcome-oriented methods provide a value of a cutpoint that correspond to the most significant relation with outcome. Some of the outcome-oriented methods are based on log rank, score, likelihood ratio and Wald statistics. Generally, the outcome-oriented methods are expected to have better statistical indicators than data-oriented methods (Kuo, 1997).

AN OUTCOME-ORIENTED APPROACH

In this paper, we focus on the method proposed by Contal and O'Quigley (1999), which is based on the log rank test statistic. Let R be the risk factor of interest measured as a continuous variable and T be the outcome variable. In case of survival analysis, the outcome of interest T, is oftentimes time to death but it can also be time to some other event of interest. The population is divided into two groups based on the cutpoint: subjects with the value of the risk factor less than or equal to the value of the cutpoint and subjects with the value of the risk factor greater than the cutpoint. Let $t_{(1)} < t_{(2)} <.....< t_{(k)}$ be the ordered observed event times of the outcome variable T. Let C be the set of K distinct values of the continuous covariate R. Then, based on one hypothetical cutpoint from C, let d_i be the number of events at time $t_{(i)}$, r_i be the number of subjects at risk prior to time $t_{(i)}$ and d_i^+ and r_i^+ be the number of events at time $t_{(i)}$ in group R > C and number of subjects at risk just prior to $t_{(i)}$ in the group R > C. Similarly, d_i^- and r_i^- be the number of events at time $t_{(i)}$ in group R ≤ C and number of subjects at risk just prior to $t_{(i)}$ in the group R ≤ C. Thus, the log rank statistic for some fixed C is given by:

$$\text{Log Rank Statistic} = L_K(t) = \sum_{i=1}^{k} \left(d_i^+ - d_i \frac{r_i^+}{r_i} \right)$$

The optimal cutpoint is that value of C, C_k, that maximizes the absolute value of $L_k(t)$. C_k therefore gives the value of the

continuous covariate that gives the maximum difference between the subjects in the two groups defined by the cutpoint. In order to test the significance of the obtained cutpoint, Contal and O'Quigley proposed the following:

$$\text{Test Statistic} = q = \frac{1}{s\sqrt{k-1}} \max \left| L_k(t) \right|$$

where, $s^2 = \frac{1}{(k-1)} \sum_{j=1}^{k} a_j^2$, and a_j's are the scores associated with the j^{th} death, given by $1 - \sum_{i=1}^{j} \frac{1}{k - i + 1}$. Such a maximization of the statistic enables the estimation and evaluation of the significance of the cutpoint (refer Contal and O'Quigley (1999) for further details) and is adjusted for the bias created by the fact that the optimal cutpoint C_k is chosen such that it gives the maximum separation between the two groups. For $q > 1$, the p-value is approximately given by $2e^{-2q^2}$ and for $q \leq 1$, the p-value is at least 0.33. Thus, this procedure considers all possible values of the continuous covariate as potential cutpoints and our SAS® macro does all of the above computations in a very short amount of time.

ILLUSTRATION

We use data from a multicenter trial of bone marrow transplant patients with a radiation-free conditioning regimen (see Copelan et al., 1991 for details on this study). A total of 137 patients were classified into three disease groups: acute lymphoblastic leukemia (ALL, n = 38), acute myelocytic leukemia (AML) with low risk of first remission (n = 54), and AML with a high risk of second remission or untreated first relapse or second or greater relapse or never in remission (n = 45). Several potential risk factors were measured at the time of transplantation like recipient (patient) and donor sex, recipient and donor immune status, recipient and donor age (in years), waiting time (in months) from diagnosis to transplantation etc. However, for the purposes of this study, we only consider the following variables: patient's age, disease group, the outcome variable of interest, which is time to relapse or death (in months) along with a censoring indicator for relapse or death (Klein and Moeschberger, 1997).

Table 1 gives the summary statistics of age for all the three disease groups (N = sample size, Std Dev = standard deviation of age, Min = minimum age, Max = maximum age).

Group	N	Mean	Std Dev	Median	Min	Max
ALL	38	24.42	7.295	22.50	15	42
AML-Low	54	29.41	8.764	29.50	13	50
AML-High	45	30.44	11.220	32.00	7	52

Table 1: Summary statistics of age for the three disease groups.

We can use the median age as a simple data-oriented approach of cutpoint determination for the patient's age in each of the three groups, which is 22.5, 29.5 and 32 years for ALL, AML-Low and AML-High respectively. In addition, if we use the information about relapse or death, then the classification and regression tree (CART) approach can be used to get the cutpoints, which gives 27.5, 27.5, and 17.5 years as the cutpoints for the patient's age in the three disease groups respectively.

Next, we consider the lowess smoothed plot of the martingale residuals as the first outcome-oriented approach to determine a cutpoint for the patient's age from the three disease groups. A stochastic process with a property that it's expected value at time t, given it's history at time s < t, is equal to it's value at time s, is called a martingale. Martingale residuals are used to determine the functional form of a covariate (see Therneau et al. (1990), and Klein and Moeschberger (1997) for derivation and discussion of the properties of martingale residuals). Following is a sample SAS® code for generating the lowess smoothed plot of martingale residuals for the disease group ALL:

```
PROC PHREG DATA=all;
  MODEL time*censor(0) = ;
  OUTPUT OUT=residuals RESMART=resmart;
PROC SORT DATA=residuals;
  BY time;
DATA age;
SET all;
KEEP time age;
PROC SORT DATA=age;
  BY time;

DATA residage;
MERGE residuals age;
  BY time;
ODS LISTING CLOSE;
PROC LOESS DATA=residage;
  MODEL resmart=age / SMOOTH=0.65;
ODS OUTPUT OUTPUTSTATISTICS=myout;
run;
ODS LISTING;
QUIT;

PROC SORT DATA=myout;
  BY age;
run;
GOPTIONS RESET=all;
SYMBOL1 c = black i=none v=star h=0.8;
SYMBOL2 c=blue i=join v=none;
axis LABEL=(a=90 'Martingale Residuals');
PROC GPLOT DATA=myout;
  FORMAT DepVar f4.0 age f4.0;
  PLOT DepVar*age=1 Pred*age=2 / overlay
       vaxis=axis1;
run;
```

PROC LOESS option in SAS® performs lowess smoothing with default smoothing parameter as 0.5 (Hosmer and Lemeshow, 1999; and Cohen, 1999) . There are several strategies that can be used to select the smoothing parameter (Cohen, 1999). In our illustration we examine plots of the fitted residuals versus the predictor variable and choose the largest smoothing parameter that yields no clearly discernible trends in the fit residuals. Figures 1, 2, and 3 give the lowess smoothed residuals for the three disease groups: ALL, AML-Low, and AML-High respectively.

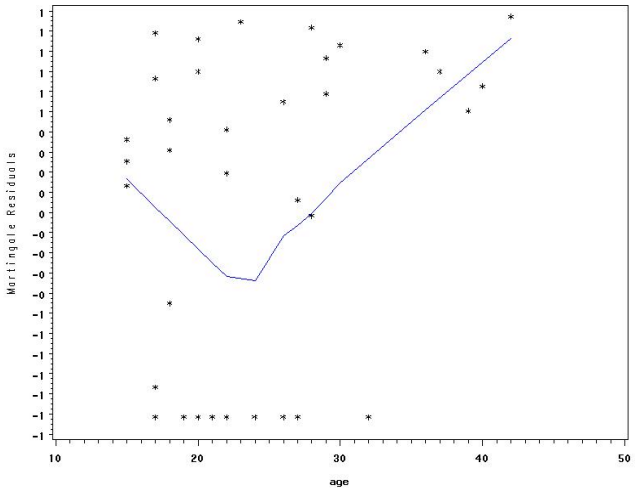

Figure 1: Plot of martingale residuals versus age and lowess smooth for ALL disease group.

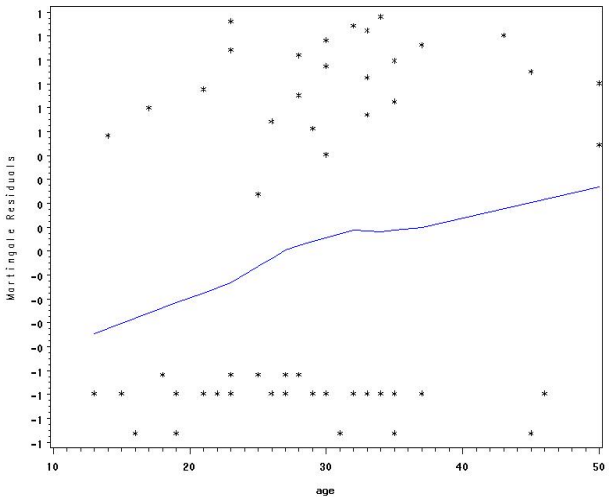

Figure 2: Plot of martingale residuals versus age and lowess smooth for AML-Low disease group.

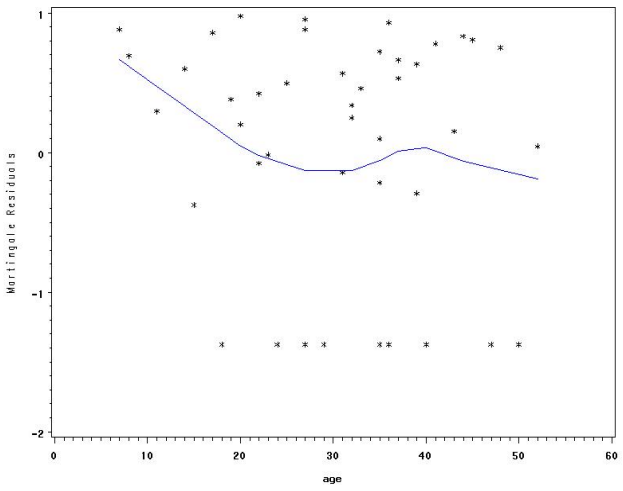

Figure 3: Plot of martingale residuals versus age and lowess smooth for AML-High disease group.

The display of both the smooth fit and the individual residuals provides insight into the influence of specific individuals on the estimate of the functional form. Figure 1 suggests that treating age as linear is inappropriate for the ALL disease group. The smoothed curve is roughly zero up to about 24 years and increases linearly up to about 42 years. This suggests that patient's age can be coded as an indicator variable in the Cox proportional hazards model. For distinct values of age, we create an indicator variable and then fit the Cox model with this new covariate to get the log-likelihood. The value of age that maximizes the log-likelihood gives the optimal cutpoint. For the ALL group, this occurs at 28 years as can be seen from Figure 4. However, in case of the AML-Low and AML-High groups, the lowess smooth values are nearly a straight line and support treating age as linear in the model (see Figures 2 and 3). Therefore, based on this approach, it is not appropriate to convert age into a categorical variable for the AML-Low and AML-High disease groups.

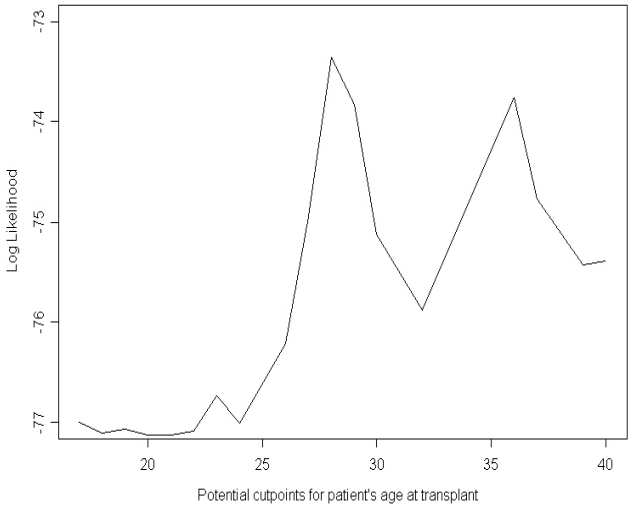

Figure 4: Plot of log-likelihood versus distinct patient ages at transplant for ALL disease group.

We now use the Contal and O'Quigley's method of categorizing patients into high or low risk groups for disease free survival based on the patient's age at transplantation for the three groups and also assess the significance of the cutpoint. In the ALL group, there are 20 distinct ages, any of which can be a potential cut point. There are 23 distinct times when death or relapse occurs, which gives $s^2 = 0.8757$. The maximum value of $|L_k|$ occurs at age 28 with q = 1.2946 and p-value of 0.07 (see Table 2). This suggests that the cutpoint obtained is significant and that age is related to time to disease free survival for ALL group (note: 10% level of significance is used due to the small sample size).

Obs	Distinct Ages	L_k	$\|L_k\|$	q
1	15	0.0000	0.0000	0.0000
2	17	-0.7503	0.7503	0.1709
3	18	-0.4232	0.4232	0.0964
4	19	-0.7427	0.7427	0.1692
5	20	0.2725	0.2725	0.0621
6	21	-0.2736	0.2736	0.0623
7	22	0.7416	0.7416	0.1690
8	23	2.1637	2.1637	0.4930
9	24	1.2171	1.2171	0.2773
10	26	3.2475	3.2475	0.7399
11	27	4.7287	4.7287	1.0773
12	**28**	**5.6822**	**5.6822**	**1.2946**
13	29	4.7785	4.7785	1.0887
14	30	3.4222	3.4222	0.7797
15	32	2.5916	2.5916	0.5904
16	36	3.6068	3.6068	0.8217
17	37	2.8075	2.8075	0.6396
18	39	2.1071	2.1071	0.4801
19	40	1.6014	1.6014	0.3648
20	42	0.9737	0.9737	0.2218

Table 2: Results from the Contal and O'Quigley's method for the ALL disease group.

In the case of AML-Low group, there are 26 distinct ages, any of which can be a potential cutpoint. There are 25 distinct times when death or relapse occurred which gives $s^2 = 0.8827$. The maximum value of $|L_k|$ occurs at age 28 with q= 0.983. However, the high p-value (> 0.33) suggests that the cutpoint obtained is not significant and age is not related to time to disease free survival for AML-Low risk group. This is also the case for AML-High risk group (31 distinct ages, 33 distinct times when death or relapse occurred, $s^2 = 0.9035$, q = 0.1464, p-value at least 0.33) and again the cutpoint of 23 years is not significant for AML-High risk group.

In our example, the two outcome-oriented approaches give the same cupoint, however, this need not be true in general. In situations when the estimated cutpoint is close to boundaries, we should carefully examine the reasons behind it as the cutpoint obtained may be real or may be due to the presence of outliers.

SOME LIMITATIONS

We have only focused on the dichotomization of a continuous covariate with the assumption that such a dichotomization is possible from biological point of view, however, in reality, more than one cutpoint may exist. In addition, ideally, this cutpoint search has to be done within the framework of a multiple regression model to eliminate the potential influence of other prognostic factors on the cutpoint. As stated before, one has to be also aware of potential confounding that might arise from categorization and using open-ended categories (Rothman and Greenland, 1998). The obtained cutpoint(s) may differ across studies depending on several factors including which data or outcome-oriented approach is used and therefore the results may not be comparable. Lastly, there is always the possibility of loss in information from categorizing a continuous covariate, possible loss of power to detect actual significance and can sometimes lead to biased estimates in regression settings, all of which need to be sufficiently addressed (Selvin S., 1987, and Altman, 1998).

CONCLUSION

Given the widespread use of categorizing a continuous covariate, there is very little attention given to this topic in statistical and epidemiological textbooks and in the literature. Our current work provides an insight into some of the data and outcome-oriented cutpoint determination methods along with accessibility to some of the computer programs using SAS®. Keeping in mind the frequent use of dichotomizing a continuous covariate in a survival analysis setting encountered in the biostatistical and epidemiological research, our SAS® macro makes this a practical proposition for the very first time.

REFERENCES

Altman, D. G. (1998), "Categorizing continuous variables," in Armitage, P. and Colton, T. (eds), *Encyclopedia of Biostatistics*, Chichester: John Wiley, 563 - 567.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatter Plots," *Journal of the American Statistical Association,* 74, 829 - 836.

Cochran, W. G. (1968), "The effectiveness of adjustment by subclassification in removing bias in observational studies," *Biometrics*. 24, 295 - 313.

Cohen, R. A. (1999), "An introduction to PROC LOESS for Local Regression," *Proceedings of the 24th SAS Users Group International Conference (SUGI)*, Paper 273, 1584 - 1592.

Contal, C., and O'Quigley, J. (1999), "An application of changepoint methods in studying the effect of age on survival in breast cancer," *Computational Statistics and Data Analysis,* 30, 253 - 270.

Copelan, E. A., Thompson, J. M., Crilley, P., Szer, J., Klein, J. P., Kapoor, N., Avalos, B. R., Cunningham, I., Atkinson, K., Downs, K., Harmon, G. S., Daly, M. B., Brodsky, I., Bulova, S. I., and Tutschka, P. J. (1991), "Treatment for Acute Myelocytic Leukemia with Allogenic Bone Marrow Transplantation Following Preparation with Bu/Cy," *Blood*, 78, 838 - 843.

Grambsch, P.M., Therneau, T. M., and Fleming, T. R. (1995), "Diagnostic plots to reveal functional for covariates in multiplicative intensity models," *Biometrics,* 51, 1469 - 1482.

Hosmer, D. W. Jr., and Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, New York: Wiley.

Klein, J. P., and Moeschberger, M. L. (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer-Verlag.

Kuo, Y. (1997), "Statistical methods for determining single or multiple cutpoints of risk factors in survival data analysis," *Dissertation*, Division of Biometrics and Epidemiology, School of Public Health, The Ohio State University.

Lausen, B., and Schumacher, M. (1996), "Evaluating the effect of optimized cutoff values in the assessment of prognostic factors," *Computational Statistics and Data Analysis*, 21, 307 - 326.

Rothman, K. J., and Greenland, S. (1998), *Modern Epidemiology*, 2nd Edition, Philadelphia: Lippincott-Raven.

Schulgen, G., Lausen, B., Olsen, J. H., and Schumacher, M. (1994), "Outcome-oriented cutpoints in analysis of quantitative exposures," *American Journal of Epidemiology*, 140, 172 - 184.

Selvin, S. (1987), "Two issues concerning the analysis of grouped data," *European Journal of Epidemiology,* 3, 284 - 287.

Therneau, T. M., Grambsch, P.M., and Fleming, T. R. (1990), "Martingale-based residuals for survival models," *Biometrika,* 77, 147 - 160.

Therneau, T. M., and Grambsch, P.M. (2000), *Modeling survival data: Extending the Cox model*, New York: Springer-Verlag.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author for information about the SAS® macro at:

Jayawant N. Mandrekar, Ph.D.
Mayo Clinic
200 First Street SW
Harwick 7
Rochester MN 55905
Work Phone: (507) 266 0573
Fax: (507) 284 9542
Email: mandrekar.jay@mayo.edu