Paper 3249-2015

Cutpoint Determination Methods in Survival Analysis using SAS®: Updated %FINDCUT macro

Meyers J. P. and Mandrekar J. N., Mayo Clinic, Rochester, MN

ABSTRACT

Statistical analysis that uses data from clinical or epidemiological studies, include continuous variables such as patient's age, blood pressure, and various biomarkers. Over the years there has been increase in studies that focus on assessing associations between biomarkers and disease of interest. Many of the biomarkers are measured as continuous variables. Investigators seek to identify the possible cutpoint to classify patients as high risk versus low risk based on the value of the biomarker. 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 used log rank test statistic in order to estimate the cutpoint. Their method was computationally intensive and hence was overlooked due to the unavailability of built in options in standard statistical software. In 2003, we had provided the %FINDCUT macro that used Contal and O'Quigley's approach to identify a cutpoint when the outcome of interest was measured as time to event. Over the past decade demand for this macro has continued to grow that has led us to consider updating the %FINDCUT macro to incorporate new tools and procedures from SAS such as array processing, Graph Template Language, and the REPORT procedure. New and updated features will include: results presented in a much cleaner report format, user specified cut points, macro parameter error checking, temporary data set clean-up, preserving current option settings, and increased processing speed. We intend to present the utility and added options of the revised %FINDCUT macro using a real life dataset. In addition, we will critically compare this method with some of the existing methods and discuss the use and misuse of categorizing a continuous covariate.

INTRODUCTION

Data collected as a part of clinical and epidemiological studies consists of both continuous and categorical variables. It is routine practice to categorize a continuous variable in order to assess the impact on the clinical outcome of interest. Dichotomization of a continuous variable into finite number of risk groups is desirable for ease of interpretation and for treatment decision making, in addition to communicating findings to patients. These continuous variables may include patient's age, cholesterol, blood pressure, tumor volume, or many of the newly identified biomarkers. Dichotomized variables used in the analysis such as logistic regression or Cox proportional hazards regression models allow for estimation of odds ratio or hazards ratio which may improve interpretability instead of using original continuous variable.

Categorization of a continuous covariate is also desirable in many types of regression settings, however, in the %FINDCUT macro, we focus our attention only to time to event data i.e. survival analysis with censored data. We are also limiting ourselves to an assessment of a single cutpoint that will help us in dichotomizing a continuous variable of interest. The focus of this paper is to outline new enhanced features of the macro. In order to avoid redundancy with prior publication, we refer readers to Mandrekar et al (2003) from SAS User's Group International conference for any technical details on statistical approach of Contal and O'Quigley (1999) as well as discussion on the limitations of categorizing a continuous covariate.

IDENTIFICATION OF A CUTPOINT

Choice of a cutpoint can be based on prior published studies, biological background about particular risk factor or clinician or expert's opinion. However one needs to note that the cutpoints for similar disease may vary based on various factors. For example, cutpoint for a variable for pediatric patients may be different from the cutpoint for the same variable among adult patients. Also, cutpoints may vary based on study design, purpose of the study, population under the study etc. There is no single approach that may be best and thus one may have to consider various methods prior to arriving at a cutpoint especially if the

continuous variable being studied is a new biomarker or a known biomarker being considered for a different disease.

Cutpoints can be identified based on data oriented methods such as mean, median, quartile or certain percentiles. These are thus independent of outcome. However, one can use outcome oriented methods which may suggest a value of a cutpoint that correspond to the most significant relation with outcome. Kuo (1997) suggest that generally the outcome-oriented methods are expected to have better statistical indicators than data-oriented methods. Outcome-oriented methods that can be explored in the context of time to event data are based on log rank, score, likelihood ratio and Wald statistics or can also use minimum p-value or maximum hazards ratio as a statistic of interest. We had written a %FINDCUT macro to implement outcome oriented methods proposed by Contal and O'Quigley (1999), which is based on the log rank test statistic. Technical and practical details of this approach can be found in Contal and O'Quigley (1999) and Mandrekar et al (2003) respectively.

NEW FEATURES OF THE MACRO

In this section we provide a short summary of the changes and additional built in feature of macro. The macro %FINDCUT has been improved in several ways. Firstly the efficiency of the macro has been vastly improved by condensing multiple data steps and summary procedures into far less data steps that utilize ARRAY processing combined with data step summary functions. This results in far less procedure calls allowing the macro to finish quicker. Secondly the macro will now save user options and titles and restore them at the end of the macro call where previously several titles, options and graphics options were modified. Error checking has been added in addition to this to prevent the macro from crashing partway through processing. The error checking makes sure variables exist, are the right type of variables, and that the entered macro parameters match what is allowed by the macro. Thirdly the macro now uses the SAS SG graphics engines with the SGPLOT procedure and the SGRENDER procedure (along with the Graph Template Language) to make high quality graphics. The macro will often combine plots into one image where possible to reduce the total number of images generated. Lastly the macro has made the cut-points easier to generate in two ways.

1. Character type variables are now allowed to be used as a cut-point variable. One caveat with this is that by default the macro will assume the variable is alphabetically ordinal. To counter this, a new parameter CUTORDER, was added to allow the user to specify the order of the variable values.
2. A new parameter, CUTPOINTS, was added to allow a user to specify numeric cut-points, in multiple formats, to be used instead of the cut-point variable values. For example, suppose the interest is in identifying a cutpoint for age as a continuous variable. And the user might specifically want to test the decade cutpoints (30, 40, 50, 60, etc.), but these are not the kind of values that are in the data set for the age variable. With the new update, cutpoints by decade can be tested specifying these cut-points with the CUTPOINTS parameter.

In the next section, we illustrate these features with the same real life dataset that was explored in our prior publication (Mandrekar et al 2003), to enable readers to see the advantages of this revised macro.

ILLUSTRATION

A 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) will be used for the illustration purposes. 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, 2005). To avoid redundancies with our prior publication and to limit the focus to new and

enhanced features of the macro we will only limit ourselves to only ALL group. Table 1 gives the summary statistics of age for ALL disease group (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	7.30	22.50	15	42

Table 1: Summary statistics of age for the ALL group.

One can use the median age (22.5 years for ALL group) as a simple data-oriented approach for a cutpoint determination for the patient's age. In addition, if we use the information about relapse or death, then recursive partitioning approach can be used to get the cutpoints, which gives 27.5 years as the cutpoints for the patient's age in this disease group. The main feature of our macro is to identify the cutpoints based on Contal and O'Quigley's (1999) method. The Contal and O'Quigley's (1999) method of categorizing patients into high or low risk groups for disease free survival based on the patient's age at transplantation for the ALL patients using %FINDCUT macro gives following results.

In the ALL group, there are 20 distinct ages, any of which can be a potential cutpoint. There are 23 distinct times when death or relapse occurs, which gives $s^2 = 0.8757$. The maximum value of |Sk| 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). This output table is similar to what was in the prior version of %FINDCUT and is presented in Table 2. The macro call requires only 4 parameters; name of the dataset used, time variable, event variable and continuous variable for which the cutpoint is being investigated.

```
%findcut(data=all, time=time, event=event, cutvar=ageyrs);
```

Cut-Points		Contal and O'Quigley Method				
Cut Level	AGEYRS	SK	Absolute SK	Q Statistic	P-value	Selected Cut-Point
1	15	0	0	0	0.3000	
2	17	-0.750263	0.7502634	0.1709312	0.3000	
3	18	-0.42321	0.42321	0.0964192	0.3000	
4	19	-0.742711	0.7427106	0.1692104	0.3000	
5	20	0.272498	0.272498	0.0620827	0.3000	
6	21	-0.273608	0.2736081	0.0623356	0.3000	
7	22	0.7416005	0.7416005	0.1689575	0.3000	
8	23	2.1637208	2.1637208	0.4929567	0.3000	
9	24	1.2170637	1.2170637	0.2772814	0.3000	
10	26	3.2474808	3.2474808	0.7398678	0.3000	
11	27	4.7286924	4.7286924	1.0773295	0.1963	
12	28	5.6821864	5.6821864	1.2945624	0.0700	<====
13	29	4.7785156	4.7785156	1.0886807	0.1869	
14	30	3.4222249	3.4222249	0.7796794	0.3000	
15	32	2.5916317	2.5916317	0.5904468	0.3000	
16	36	3.6068403	3.6068403	0.82174	0.3000	
17	37	2.8074971	2.8074971	0.6396271	0.3000	
18	39	2.1070787	2.1070787	0.480052	0.3000	
19	40	1.6013513	1.6013513	0.364833	0.3000	
20	42	0.9736842	0.9736842	0.2218327	0.3000	

Table 2: Results from Contal and O'Quigley's method

A new feature added to the macro call is that the user can specify the cutpoints of their interest to be assessed. There are 2 ways in which user can specify the values. For example, fixed intervals as in the following case where interest is in exploring values of age at 4 unit increments ranging between 20 years to 36 years as potential cutpoints.

```
%findcut(data=all, time=time, event=event, cutvar=ageyrs, cutpoints=20 to 36
by 4);
```

This can also be accomplished by simply specifying the values of interest,

```
%findcut(data=all, time=time, event=event, cutvar=ageyrs, cutpoints= 20 24 28
32 36);
```

The relevant output is presented in Table 3 below.

Cut-Points		Contal and O'Quigley Method				
Cut Level	AGEYRS	SK	Absolute SK	Q Statistic	P-value	Selected Cut-Point
1	20	0.272498	0.272498	0.0620827	0.3000	
2	24	1.2170637	1.2170637	0.2772814	0.3000	
3	28	5.6821864	5.6821864	1.2945624	0.0700	<====
4	32	2.5916317	2.5916317	0.5904468	0.3000	
5	36	3.6068403	3.6068403	0.82174	0.3000	

Table 3: Results from Contal and O'Quigley's Method assessing user specified cutpoints

Another new feature is that it allows user specify as a cutpoints a value of the variable that may not necessarily be a value in the dataset. For example, our dataset does not have a patient with age 31 years.

OTHER METHODS TO IDENTIFY CUTPOINTS

In this section we will display results from a few other graphical approaches that we have incorporated as a part of %FINDCUT macro so researcher has access to cutpoints obtained using a few other outcome oriented methods. This includes plots of martingale and deviance residuals versus the predictor variable, plots of p-values or hazards ratio from Cox proportional hazards regression model treating each value of a continuous variable in the dataset as a possible cutpoint. We have also included false discovery rate corrected p-value plot to adjust for multiple comparisons. Martingale residuals are used to determine the functional form of a covariate (see Therneau et al. (1990, 2000), and Klein and Moeschberger (2005) for derivation and discussion of the properties of martingale residuals).

The %FINDCUT macro requires PLOTOP and RESIDOP parameters in the macro call to generate these plots:

```
%findcut(data=all, time=time, event=event, cutvar=ageyrs, plotop=1,
residop=1);
```

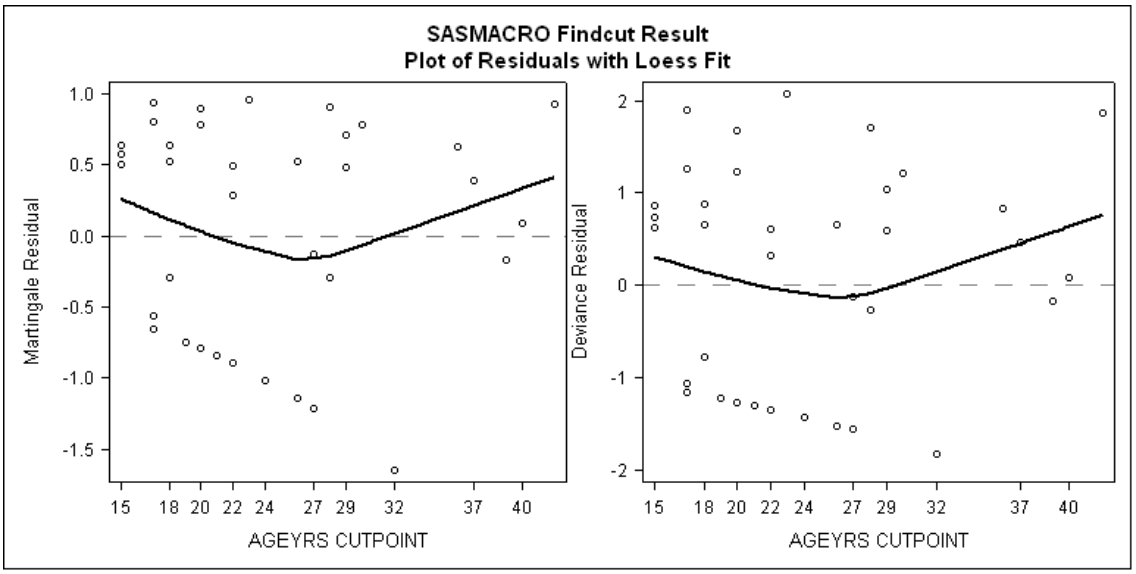

Figure 1: Plot of martingale and deviance residuals versus age with lowess smooth for ALL disease group

Figure 1 gives the lowess smoothed residuals for the ALL 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 decreases 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 can 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 based on a plot of log-likelihood versus distinct patient ages (Note: plot is not shown here, refer Mandrekar et al 2003).

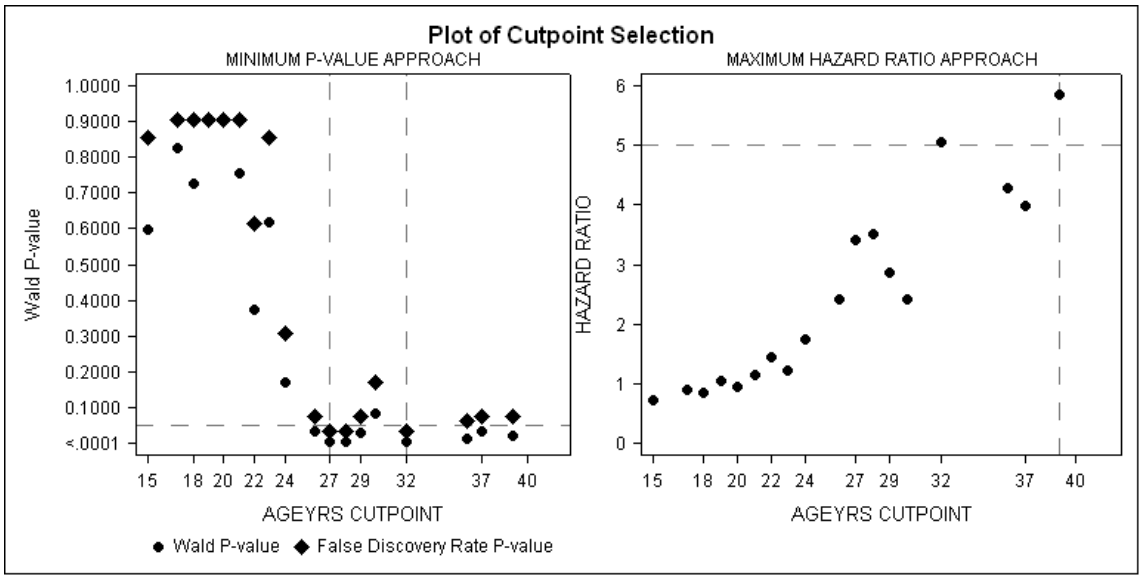

Figure 2: Plot of p-values and hazard ratios versus age for ALL disease group

A graphical display of p-values and hazard ratios from individual Cox proportional hazards regression model is presented in Figure 2. Using raw p-values or false discovery adjusted p-values one may consider 27 years or 32 years as possible candidates for cutpoint. Plot of hazards ratio needs to be interpreted with caution as the hazard ratio will have extreme values at the lower or higher end of age range due to limited number of patients and events. This plot suggests 32 years of age as a possible cutpoint.

In our example with only 38 patients, various outcome oriented methods have yielded slightly different cutpoint in the range of 27 years to 32 years of age. 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.

LIMITATIONS OF DICHOTMIZING A CONTINUOUS VARIABLE

Dichotomization of a continuous variable can result in information loss, 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). Also, not all continuous variables have a single cutpoint. There is the possibility of 2 or more cutpoints, for example to classify patients into no risk, low risk, moderate risk and high risk. We have not focused our attention to multiple cutpoint scenarios. Also, if a single cutpoint exists and it is statistically significant, one also needs to consider the clinical relevance of such a cutpoint.

As stated in our prior publication (Mandrekar et al 2003), 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. Researcher has to be also aware of potential confounding that might arise from categorization and using open-ended categories (Rothman and Greenland, 1998). Also, obtained cutpoint(s) may differ across studies depending on several factors which also include type of data or outcome-oriented approach used for assessing a cutpoint and therefore the results may not be comparable.

CONCLUSION

Demand for investigation of a possible cutpoint continues to grow with research focused on biomarkers for various diseases. Patients, clinicians and researchers want to distinguish between high versus low risk. Our revised %FINDCUT macro has added many features that will allow statisticians to explore various new options to make better decisions on whether the cutpoint exists and if it is statistically significant or not. Some of the results are displayed in graphical format for visual representation and easier interpretation.

We have responded to the increase in demand for our original %FINDCUT macro over the past decade and thus have revised the macro using new tools and procedures from SAS such as array processing, Graph Template Language, and the REPORT procedure. New and updated features provide results in a much cleaner report format, allowing for checking of user specified cut points with multiple error checking features for macro parameters yet offering increased processing speed. Application of this macro is directly translatable to non-clinical areas as well, where the goal is to assess possibility of a cutpoint for a continuous predictor when outcome of interest is measured as a time to event.

REFERENCES

Mandrekar JN, Mandrekar SJ, Cha SS. 2003. Cutpoint determination methods in survival analysis using SAS®. (Paper 261-28). Proceedings of the 28th SAS Users Group International Conference (SUGI 28).

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

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.

Klein, J. P., and Moeschberger, M. L. 2005. Survival Analysis: Techniques for Censored and Truncated Data 2nd edition, New York: Springer.

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.

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

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 at:

Jayawant N. Mandrekar, Ph.D.
Professor of Biostatistics
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