

A SAS MACRO FOR MODIFIED KAPLAN-MEIER ESTIMATES WITH TIME DEPENDENT COVARIATES

Lonni R. Schultz, Gordon R. Jacobsen, Edward L. Peterson
Henry Ford Health Sciences Center

1. INTRODUCTION

In survival analysis, it is desirable to graphically represent the survival curves for the covariates of interest. Kaplan-Meier (1) estimates have commonly been used for plotting survival curves for covariates which are categorical or can be categorized. These estimates assume that the value of the covariate is set at the beginning of follow-up and does not change over time. This assumption is reasonable for fixed covariates, but may not be appropriate for time dependent covariates. We present in Section 2 a modification of the Kaplan-Meier estimates that takes into account the changing values of these covariates. In Section 3, two examples are given in which the usual Kaplan-Meier and the modified Kaplan-Meier methods differ. A brief discussion is given in Section 4.

2. EQUATIONS

For the n subjects, let $t_1 < t_2 < \dots < t_k$ denote the k ($k \leq n$) distinct event times. Let d_j represent the number of events at t_j and n_j represent the number of subjects at risk at time t_j . For those subjects in which the event was not observed, let L_i denote the censoring time for the i^{th} subject. The Kaplan-Meier estimate for the survivor function $S(t)$ is defined as

$$\hat{S}(t) = \prod_{j: t_j \leq t} \frac{n_j - d_j}{n_j} \quad (2.1)$$

In this calculation, if a censoring time L_i and an event time t_j have the same value, then the censoring time is assumed to have occurred just after the event time.

In the above equation, the largest n_j is n_1 and the smallest is n_k . The number at risk (n_j) is set at time 0 and decreases with time. For the time dependent covariate, this feature of the number at risk is not true. The number at risk can increase or decrease depending on the given time.

We modified the K-M estimates by determining the number at risk in each level of the covariate at each event time. The new equation given below is the same as equation 2.1, except for n'_{ij} and d'_{ij} . n'_{ij} is the number at risk and d'_{ij} is the number of deaths at time t_j for the i^{th} level of the covariate. The modified K-M estimates for the survivor function $S(t)$ is

$$\hat{S}_{Mi}(t) = \prod_{j: t_j \leq t} \frac{n'_{ij} - d'_{ij}}{n'_{ij}} \quad (2.2)$$

where $i=1, \dots, \#$ of covariate levels.

Greenwood's formula (2) for variance estimation of the survivor function estimate can be used for this new method. The formula for this estimation is

$$\widehat{Var} [\hat{S}_{Mi}(t)] = \hat{S}_M(t)^2 \sum_{j:t_j \leq t} \frac{d'_{ij}}{n'_{ij}(n'_{ij} - d'_{ij})} \quad (2.3)$$

A limitation of this new method is when all of the subjects at risk for a given level of the covariate experience an event at the same time during follow-up before the last time point. In equation 2.2, the survival estimate is zero from that time point on, regardless of the fact that other subjects may be at risk at later times.

3. EXAMPLES

In some cases, the survival estimates from the two methods (usual K-M and modified K-M) are similar. The following two examples are cases for which the two methods differ.

Example 1

In a sample of young adult smokers ($n=424$), we were interested in assessing the effect of alcohol problems on smoking cessation (3). The subjects could be divided into three alcohol groups. The three groups are 1) never had an alcohol problem, 2) have an alcohol problem, and 3) had an alcohol problem that has been resolved. Each subject's status could change over time. For example, a subject could start out with no alcohol problem, develop an alcohol problem at age 20 and resolve the alcohol problem at age 25. This subject has been in all three alcohol groups, depending on what age is being considered. Hence, the alcohol status was a time dependent covariate.

The starting point for all subjects was their age at daily smoking. The endpoint of interest was smoking cessation. Subjects were censored if they had not stopped smoking for at least one year prior to follow-up. Years from daily smoking until smoking cessation or last follow-up was the time variable.

Cox regression with time dependent covariates was done to test the differences among the three alcohol groups. An overall difference among the three groups was detected ($p=0.0012$, Wald chi-square test). Table 1 contains the results of the pairwise comparisons among the three groups. The subjects who resolved an alcohol problem were most likely to stop smoking. While the subjects who currently had an alcohol problem were least likely to stop smoking.

We also wanted to graphically present the survival curves for the three groups. Figure 1 shows the usual Kaplan-Meier curves for the three groups. Each subject's alcohol status was defined as never having an alcohol problem, having an alcohol problem, and resolving an alcohol problem. This status was set at the time of daily smoking and did not change over time. The curves in figure 1 do not represent the findings from the Cox regression with time dependent covariates. In this figure, there appears to be no difference among the three groups.

Figure 2 shows the modified Kaplan-Meier curves for the three groups. In this figure, a subject's alcohol status was determined each time someone stopped smoking. Figure 2 agrees with the findings from the Cox regression with time dependent covariates. It shows that the

subjects who resolve alcohol problems are more likely to stop smoking followed by the subjects never having an alcohol problem and ending with the subjects who still have an alcohol problem.

Example 2

In the same group of young adults, we were also interested in assessing the effect of major depression (MDD) on smoking cessation within subjects who were nicotine dependent ($n=282$). In this example, subjects were divided into two categories, with and without MDD. As with the alcohol problem status, each subject's MDD status could change over time. Thus, MDD status was considered a time dependent covariate.

In this analysis, the starting point was age at nicotine dependence. The endpoint was again smoking cessation. The same censoring scheme from example 1 was used. Years from nicotine dependence until smoking cessation or last follow-up was the time variable.

Cox regression with a time dependent covariate for MDD was done to test the differences between subjects with and without MDD. The difference was not statistically significant ($\beta=-0.181$, $se=0.328$, $RR=0.83$, 95% $CI=(0.44,1.59)$, $p=0.581$).

As in example 1, we wanted to graphically present the survival curves for the two groups. Figure 3 shows the usual Kaplan-Meier curves, where subjects who ever had a MDD were considered yes and all other subjects considered no. From this figure, it would appear that the survival experience for the two groups may be different. The p -value for the log-rank test was 0.08. As in example 1, the usual Kaplan-Meier curves do not agree with the results from the Cox regression.

Figure 4 shows the modified Kaplan-Meier curves for the two groups. This figure does agree better with the Cox regression, which indicated that the difference between the two groups was not significant.

4. DISCUSSION

The modified Kaplan-Meier method presented here is similar to the method given in Mantel and Byar (4). In this paper, a tally sheet is presented to show how subjects move from one state to another over time. Mantel and Byar use this tally sheet to carry out a test that compares the survival experiences for subjects with and without heart transplant while taking into account time of heart transplantation. This test is similar to the usual log-rank test. We used this same idea of a tally sheet and calculated the survivor function for the different levels (or states) of the covariate.

This modified Kaplan-Meier estimation appears to be very useful for graphically representing survival curves for time dependent covariates. It gives us another graphical tool for displaying data.

REFERENCES

- (1) Kaplan, E.L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457-481.
- (2) Greenwood, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*, **33**, London: Her Majesty's Stationery Office, 1-26.
- (3) Breslau, N., E. Peterson, L. Schultz, P. Andreski, H. Chilcoat. (In Press) Are smokers with alcohol disorder less likely to quit? *American Journal of Public Health*.
- (4) Mantel, N. and D.P. Byar (1974). Evaluation of response-time data involving transient states: an illustration using heart-transplant data. *J. Am. Stat. Assoc.* **69**, 81-86.

Table 1
Pairwise Comparisons from Cox Regression for Example 1

<u>Comparison</u>	<u>Beta</u>	<u>S.E.</u>	<u>R.R.</u>	<u>95% CI</u>	<u>p-value*</u>
Continue vs Never	-0.814	0.276	0.44	(0.26,0.76)	0.0032
Resolved vs Never	0.433	0.251	1.54	(0.94,2.52)	0.0848
Resolved vs Continue	1.238	0.399	3.45	(1.58,7.54)	0.0019

Figure 1

Usual Kaplan—Meier Estimates

Effect of Alcohol Problems on Smoking Cessation
Probability of No Smoking Cessation

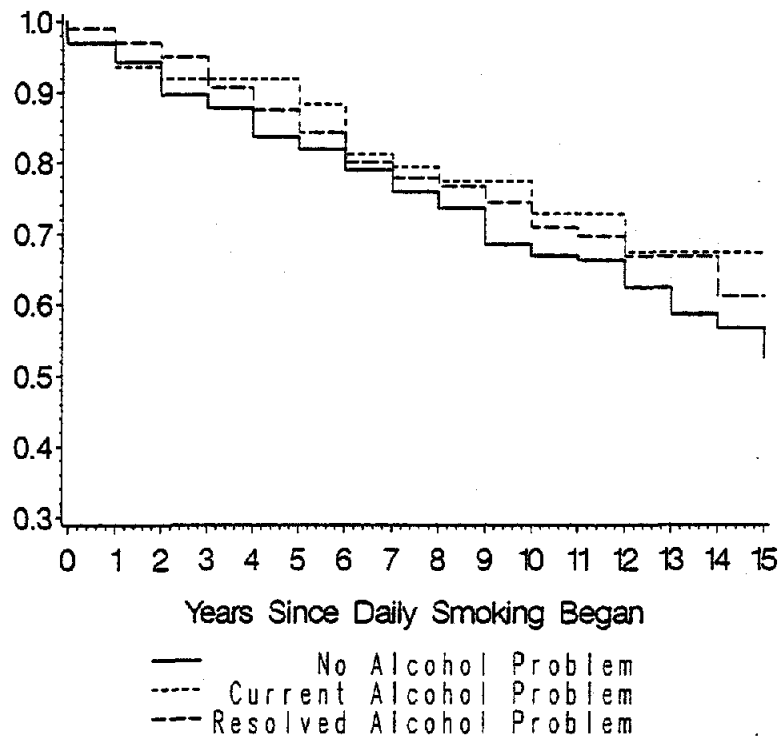


Figure 2

Modified Kaplan—Meier Estimates

Effect of Alcohol Problems on Smoking Cessation
Probability of No Smoking Cessation

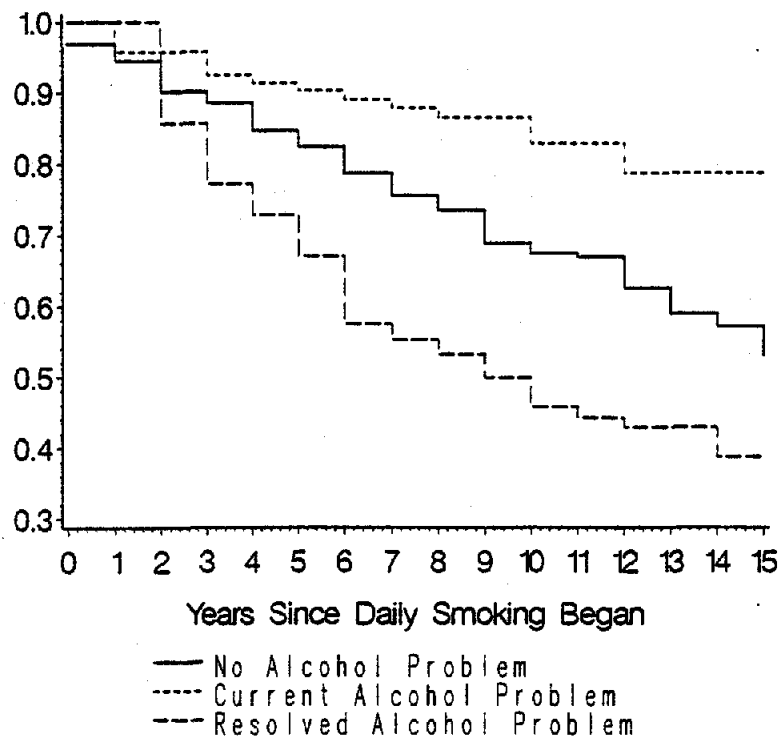


Figure 3
Usual Kaplan—Meier Estimates
Effect of MDD on Smoking Cessation
Probability of No Smoking Cessation

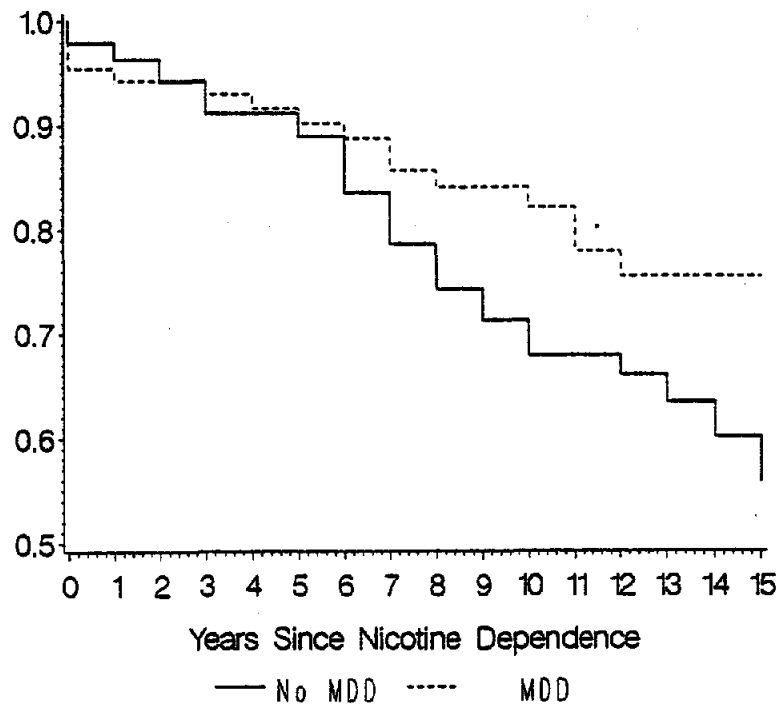


Figure 4
Modified Kaplan—Meier Estimates
Effect of MDD on Smoking Cessation
Probability of No Smoking Cessation

