Page 168-2012

Your "Survival" Guide to Using Time-Dependent Covariates

Teresa M. Powell, MS and Melissa E. Bagnell, MPH
Deployment Health Research Department, San Diego, CA

ABSTRACT

Survival analysis is a powerful tool with many strengths, like the ability to handle variables that change over time. Including time-dependent variables in survival analyses models, such as income, marital status, location, or treatment, can more accurately assess the data. This paper will give examples of the counting process syntax and programming statements which are the two methods to apply time-dependent variables in PROC PHREG. Coding techniques will be discussed as well as the pros and cons of both methods.

INTRODUCTION

Survival analysis is a robust method of analyzing time to event data. This type of analysis is useful for analyzing data when event times are known such as in medical, economic, and survey data. The Cox proportional hazards model is one method of analyzing time to event data. This model assumes that the hazards are proportional and uses partial likelihood, which is more generalized than the maximum likelihood, to estimate the model (Hosmer & Lemeshow, 1999).

When explanatory variables do not change over time or when data is only collected for the explanatory variables at one time point, it is appropriate to use static variables to explain the outcome. On the other hand, there are many situations where it is more appropriate to use time varying covariates. Using time varying explanatory variables, when appropriate, is more robust because it utilizes all available data (Allison, 2010).

This paper will first show how to use the Cox model to analyze data containing static explanatory variables. We will then show how to analyze survival data containing time varying explanatory variables using both programming statements and the counting process syntax. More specifically, we will show the association of hypertension, as both static and time varying, with coronary heart disease (CHD). Though not discussed here, prior to running these analyses, it is important to check necessary assumptions, such as proportionality, non-informative censoring, and independent observations (Allison, 2010). This paper will focus on SAS® PROC PHREG.

DATA

The Millennium Cohort Study, a prospective longitudinal study, consists of more than 150,000 military personnel. Participants are asked to fill out a questionnaire approximately every 3 years, and survey topics range from self-reported behavioral characteristics to mental health to sleep (Smith et al., 2011). Due to confidentiality, variables and data have been scrambled. The data set contains the following variables:

SID	Study ID
TIME	Follow-up time in years
CENSOR	Event indicator with value 1 for CHD development time and value 0 for censored time
AGE	Age in years from birth to study start
SEX	Male or female

WEIGHT	Weight at beginning of study
MARITAL	Married or not married
RACE	Race or ethnic group (non-Hispanic White, non-Hispanic black, Hispanic, Asian/Pacific Islander, Other)
HTN_1-HTN_8	Time varying hypertension status

SID	sex	marital	race	age	weight	htn_1	htn_2	htn_3	htn_4	htn_5	htn_6	htn_7	htn_8	censor	time
1	1	0	4	60	135	0	0	0	0	0	0	0	0	1	7
2	0	1	3	51	200	1	1	1	1	1	1	1	0	0	8
3	0	0	4	39	200	0	0	0	0	0	0	0	0	0	8
4	1	1	2	50	165	1	1	1	1	1	1	1	1	0	8
5	0	0	4	48	165	1	0	0	0	0	0	0	0	0	8
6	0	1	1	45	173	1	1	1	1	1	1	0	0	0	8
7	0	1	4	38	165	1	1	1	1	1	1	1	1	0	8
8	0	1	4	37	225	1	1	1	0	0	0	0	0	1	3
9	1	0	4	38	185	0	0	0	0	0	0	0	0	0	8
10	0	1	4	32	165	1	1	1	0	0	1	1	1	0	8

Table 1: Snapshot of data

After taking a random sample from the Millennium Cohort Study data set, our final data set contained 30,000 individuals. A snapshot of this data can be seen in Table 1. Table 1 displays all variables, including the time varying hypertension covariates. When analyzing hypertension as a static covariate, we used the data at HTN_1.

STATIC VARIABLES EXAMPLE

We will first look at the example using all static explanatory variables. The model with static explanatory variables can be expressed in the following way:

$$h_i(t) = \lambda_0(t)e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{1k} x_{ik}}$$

where $\lambda_0(t) = e^{\alpha(t)}$ is the baseline hazard function at time t and $e^{\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{1k} x_{ik}}$ is the risk associated with the covariate values. When you take the ratio of the hazards, the baseline hazard cancels out and the hazards are proportional at any time, t, giving the proportional hazards model.

The following SAS code illustrates how to use PROC PHREG with CLASS and MODEL statements. We have also included the TIES=EFRON option, which accounts for ties present in the data that occurs when two events happen at the same time, and the RL option which prints the 95% confidence intervals in the output. Though the CLASS statement is red, do not be alarmed! For some reason, this is normal in PROC PHREG.

```
PROC PHREG DATA = STATIC_SURVIVAL;
CLASS RACE;
```

```
MODEL time*censor(0) = sex marital race age weight htn_1/ TIES = EFRON RL;
RUN;
```

PHREG Output

Data Set	WORK.STATIC_SURVIVAL
Dependent Variable	time
Censoring Variable	censor
Censoring Value(s)	0
Ties Handling	EFRON

Table 2.1: Model Information

Class	Value	Design Variables				
race	American Indian	1	0	0	0	0
	Asian/Pacific Islander	0	1	0	0	0
	Black, non-Hispanic	0	0	1	0	0
	White, non-Hispanic	0	0	0	1	0
	Hispanic	0	0	0	0	1
	Other (reference)	0	0	0	0	0

Table 2.2: Class Level Information

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Male	1	0.05054	0.05249	0.9271	0.3356	1.052	0.949	1.166
Married	1	0.14610	0.06265	5.4373	0.0197	1.157	1.024	1.309
American Indian	1	-0.30073	0.25292	1.4138	0.2344	0.740	0.451	1.215
Asian/Pacific Islander	1	-0.08850	0.16127	0.3012	0.5832	0.915	0.667	1.256
Black, non-Hispanic	1	0.27202	0.15456	3.0976	0.0784	1.313	0.970	1.777
White, non-Hispanic	1	-0.23833	0.15092	2.4938	0.1143	0.788	0.586	1.059
Hispanic	1	-0.13217	0.16631	0.6316	0.4268	0.876	0.632	1.214
Age at start	1	0.04229	0.00194	475.9528	<.0001	1.043	1.039	1.047
Weight at start	1	0.01080	0.000625	298.6053	<.0001	1.011	1.010	1.012
Hypertension	1	-0.04684	0.05821	0.6474	0.4210	0.954	0.851	1.070

Table 2.3: Analysis of Maximum Likelihood Estimates

Table 2.1 provides basic information about the variables and data used in the analysis. It includes the dependent variable, censoring variable, censored value, and the method for handling ties. We used the EFRON option to handle ties on our data. For more information, see the SAS/STAT documentation.

Table 2.2 shows information about the CLASS variable RACE. This table displays the design matrix, which provides information regarding which variable has been designated the reference.

Table 2.3 is the Analysis of Maximum Likelihood Estimates. This table provides information regarding the model degrees of freedom, parameter estimates, standard error, chi-square, p-value for the chi-square, hazard ratio, and the 95% confidence interval for the hazard ratio. We received the confidence interval by specifying the /RL option in the MODEL statement. Note that there is no statistically significant association between CHD and sex, race/ethnicity, or hypertension. Stay tuned - we will see if this changes when we make hypertension into a time varying covariate.

INCORPORATING TIME-DEPENDENT VARIABLES IN COX PROPORTIONAL HAZARD MODELING

Time-dependent variables are those that can change value over the course of the observation period. Variables such as body weight, income, marital status, marketing promotions, hypertension status, are a few examples that could vary over time. While researchers can hold the values of such variables fixed at a certain point in time, say baseline, the changing values may yield a different, dare we say, a more accurate analysis of the data simply because we use as much data as possible.

To extend the logged hazard function to include variables that change over time, all we need to do is put a (t) after all the x's that are time-dependent variables. To write the equation that has one static and one time-dependent variable, we have

$$\log h_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2}(t)$$

This function now means that the hazard at time t depends on the value of x_1 and on the value of x_2 at time t, where x_1 is static and x_2 is time varying.

PROC PHREG has two different methods to handle time-dependent variables: the counting process and programming statements. Both methods will yield the same results if correctly coded. We will show examples of how to use these methods by allowing our hypertension variable to vary over time.

COUNTING PROCESS EXAMPLE

The counting process method has multiple records for each individual, with each record corresponding to an interval of time during which all covariates remain constant. If the interval does not end in an event, code it as censored. The first step is to create some indicator variables that tell us when an individual's hypertension status changes. There are a total of seven possible times that the hypertension status can change. We will call these variables CHNG1-CHNG7.

```
DATA change;
SET static_survival;
ARRAY htn_(*) htn_1-htn_8; *call in the time-varying hypertension variables;
ARRAY chng(7); *the new indicator variables;
t=1; initialize the position variable for the indicator variables;
DO i = 2 TO 8;
   IF htn_(i) NE htn_(i-1) THEN DO; *detects whether there is a change in
                                     hypertension status;
      chng(t) = i-1; *assigns the last year the status remained constant;
      t=t+1;
   END;
END;
RUN;
```

SID	chng1	chng2	chng3	chng4	chng5	chng6	chng7	htn_1	htn_2	htn_3	htn_4	htn_5	htn_6	htn_7	htn_8
1	.	.	.	.	.	.	.	0	0	0	0	0	0	0	0
2	7	.	.	.	.	.	.	1	1	1	1	1	1	1	0
3	.	.	.	.	.	.	.	0	0	0	0	0	0	0	0
4	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1
5	1	.	.	.	.	.	.	1	0	0	0	0	0	0	0
6	6	.	.	.	.	.	.	1	1	1	1	1	1	0	0
7	.	.	.	.	.	.	.	1	1	1	1	1	1	1	1
8	3	.	.	.	.	.	.	1	1	1	0	0	0	0	0
9	3	5	.	.	.	.	.	1	1	1	0	0	1	1	1
10	1	4	7	.	.	.	.	0	1	1	1	0	0	0	1

Table 3: Indicator and Time-varying Variables

Note that if an individual does not change hypertension status at any point in the study period, all indicator variables (CHNG1-CHNG7) are missing (Table 3). We see that individuals 1, 3, 4, and 7 did not have any changes in hypertension status. Individuals 2, 5, 6, and 8 had one change in hypertension, and individuals 9 and 10 had more than one change. We can interpret the indicator variable for individual 8 as the individual had hypertension from the beginning of the study through year 3, and then did not have hypertension for the rest of the study period.

This step is necessary if your time-varying covariate can switch values more than once. In some studies, individuals may only be able to change status once, such as heart surgery where a person previously never had heart surgery and then changed status to having heart surgery, or you may want to take into account a more cumulative history of changes. If you have these types of time-varying covariates, you can still use either the counting process or the programming statements to perform proportional hazard modeling on your data.

The next step is to output a record for each time period in which the hypertension variable stays constant.

```
DATA count;
SET change;
ARRAY htn_(*) htn_1-htn_8; *call in the time-varying hypertension variables;
ARRAY chng(*) chng1-chng7; *call in the indicator variables;
start = 0; *initialize the beginning time for the study;
censor2 = 0; *initialize the new censor variable;
t = 1; *initialize the position variable for the indicator variables (chng1-
chng7);

DO i=1 TO time;
   *makes sure we only output the records that hypertension status remains
   constant;
   IF (chng(t) >. and chng(t) < time) or i=time THEN;
      *assign the value of hypertension status;
      IF chng(t)>. THEN hyper = htn_(chng(t));
      ELSE hyper = htn_(time);

      *assign the end time – this will either be the end of the interval that
      the hypertension status remained constant, the year that CHD occurred,
      or the end of the study period;
      stop = min(chng(t), time);

      *assign the value of the censor variable;
      IF i = time THEN censor2 = censor;

      *assign the new start time if a change in hypertension status has
      occurred;
      IF t>1 THEN start = chng(t-1);

      *move the position variable to go to the next chng1-chng7 variable;
      t = t+1;
      OUTPUT; *output the record to the new dataset;
   END;
END;
RUN;
```

Here is a brief explanation of the above code. The two ARRAY statements call in our time-varying hypertension status (HTN_1-HTN_8) and change indicator variables (CHNG1-CHNG7). The first two assignment statements initialize the starting time (START) and the censor variable (CENSOR2) at zero. The next assignment statement initializes a position variable (T) that will tell us which change indicator variable (CHNG1-CHNG7) we should use.

The DO statement loops from 1 to either the end of the study period or the year that CHD occurred. The IF-THEN statement makes sure that we only output the records that hypertension status remains constant. The next IF-THEN statement assigns the new hypertension status, HYPER, the value for the interval it remains constant. The end time (STOP) is then assigned to either the end of the interval that the hypertension status remained constant, the year that CHD occurred, or the end of the study period. We then check if we are at the end of the follow-up time. If we are, then we specify CENSOR2 to have the same value as the old CHD CENSOR variable. The last assignment statement increments T to move on to the next CHNG1-CHNG7 variable. Finally, we output the record to the new dataset. By default, all other variables in the original dataset are included in the record. The new dataset has 46,713 records. Below is a snapshot of the dataset created by the counting process code.

SID	start	stop	hyper	censor2	chng1	chng2	chng3	htn_1	htn_2	htn_3	htn_4	htn_5	htn_6	htn_7	htn_8	time
1	0	7	0	1	.	.	.	0	0	0	0	0	0	0	0	7
2	0	7	1	0	7	.	.	1	1	1	1	1	1	1	0	8
2	7	8	0	0	7	.	.	1	1	1	1	1	1	1	0	8
3	0	8	0	0	.	.	.	0	0	0	0	0	0	0	0	8
4	0	8	1	0	.	.	.	1	1	1	1	1	1	1	1	8
5	0	1	1	0	1	.	.	1	0	0	0	0	0	0	0	8
5	1	8	0	0	1	.	.	1	0	0	0	0	0	0	0	8
6	0	6	1	0	6	.	.	1	1	1	1	1	1	0	0	8
6	6	8	0	0	6	.	.	1	1	1	1	1	1	0	0	8
7	0	8	1	0	.	.	.	1	1	1	1	1	1	1	1	8
8	0	3	1	1	3	.	.	1	1	1	0	0	0	0	0	3
9	0	3	1	0	3	5	.	1	1	1	0	0	1	1	1	8
9	3	5	0	0	3	5	.	1	1	1	0	0	1	1	1	8
9	5	8	1	0	3	5	.	1	1	1	0	0	1	1	1	8
10	0	1	0	0	1	4	7	0	1	1	1	0	0	0	1	8
10	1	4	1	0	1	4	7	0	1	1	1	0	0	0	1	8
10	4	7	0	0	1	4	7	0	1	1	1	0	0	0	1	8
10	7	8	1	1	1	4	7	0	1	1	1	0	0	0	1	8

Table 4: Snapshot of Counting Process Data

In Table 4, each record with the same SID number corresponds to one individual. This means that individual 1 only has one record since their hypertension status did not change. However, individual number 10 has 4 records since their hypertension status changed 4 times. Notice how number 10's START and STOP values correspond to CHNG1-CHNG3 and that only at the last record is the individual's CENSOR2 value =1. This happens since they did not have CHD until year 8.

It is important to note that in SAS the (START, STOP] intervals are open on the left and closed on the right. This implies that the STOP time is included in the interval, but the START time is not.

To estimate the model, we use the counting process syntax. We use the same model options as we did in the static variable model. The only difference is that this syntax requires us to specify a starting and stopping time for each record:

```
PROC PHREG DATA = count;
CLASS race;
MODEL (start, stop)*censor2(0) = sex marital race age weight hyper/ TIES =
EFRON RL;
RUN;
```

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Male	1	0.07582	0.05265	2.0744	0.1498	1.079	0.973	1.196
Married	1	-0.01603	0.04749	0.1140	0.7356	0.984	0.897	1.080
American Indian	1	-0.31035	0.25290	1.5059	0.2198	0.733	0.447	1.204
Asian/Pacific Islander	1	-0.09215	0.16125	0.3266	0.5677	0.912	0.665	1.251
Black, non-Hispanic	1	0.27832	0.15455	3.2432	0.0717	1.321	0.976	1.788
White, non-Hispanic	1	-0.23661	0.15091	2.4583	0.1169	0.789	0.587	1.061
Hispanic	1	-0.13460	0.16631	0.6550	0.4183	0.874	0.631	1.211
Age at start	1	0.04335	0.00196	490.4556	<.0001	1.044	1.040	1.048
Weight at start	1	0.01092	0.000627	302.9248	<.0001	1.011	1.010	1.012
hyper	1	0.22752	0.04213	29.1584	<.0001	1.255	1.156	1.364

Table 5: Maximum Likelihood Estimates for Time-varying Covariates – Counting Process

The results in Table 5 indicate that hypertension has a significant effect on CHD! We can interpret this as: in any given year, if an individual has hypertension, their hazard of CHD is 25.5% greater than if they did not have hypertension. Let's remember our example where hypertension was a static explanatory variable, hypertension was not significantly associated with the development of CHD (p=0.42). We would not have found the conclusion that hypertension is indeed associated with CHD if we had not allowed hypertension status to vary over time.

PROGRAMMING STATEMENT EXAMPLE

The programming statement method has only one record for each individual. The time-varying covariates are defined in programming statements (hence, the method name) that are part of the PHREG step. Here's how we estimate the model using this method:

```
PROC PHREG DATA = static_survival;
CLASS race;
MODEL time*censor(0) =  sex marital race age weight hyper2/ TIES = EFRON RL;
   ARRAY htn_(*) htn_1-htn_8;
   hyper2 = htn_[time];
RUN;
```

We use the same model options as done in the previous two examples. But now we define the time-varying hypertension covariate (HYPER2) using only two extra lines of code. First, we have an ARRAY statement that calls in the different hypertension statuses (HTN_1-HTN_8). Then we have an

assignment statement that defines the time-varying covariate (HYPER2). This syntax requires that brackets are used in the assignment statement. However, unlike an assignment statement in the data step that is used only once, this statement reassigns the value of HYPER2 at each TIME that there is an event. For example, if an individual has CHD at TIME=3 this statement assigns HYPER2 the status of hypertension at TIME=3 (HYPER2=HTN_3) for all individuals in the risk set. Then, if a different individual has CHD at TIME=6, it assigns HYPER2=HTN_6 for all those left in the risk set, and so on.

The results are identical to those using the counting process (Table 5 and 6).

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
Male	1	0.07582	0.05265	2.0744	0.1498	1.079	0.973	1.196
Married	1	-0.01603	0.04749	0.1140	0.7356	0.984	0.897	1.080
American Indian	1	-0.31035	0.25290	1.5059	0.2198	0.733	0.447	1.204
Asian/Pacific Islander	1	-0.09215	0.16125	0.3266	0.5677	0.912	0.665	1.251
Black, non-Hispanic	1	0.27832	0.15455	3.2432	0.0717	1.321	0.976	1.788
White, non-Hispanic	1	-0.23661	0.15091	2.4583	0.1169	0.789	0.587	1.061
Hispanic	1	-0.13460	0.16631	0.6550	0.4183	0.874	0.631	1.211
Age at start	1	0.04335	0.00196	490.4556	<.0001	1.044	1.040	1.048
Weight at start	1	0.01092	0.000627	302.9248	<.0001	1.011	1.010	1.012
hyper2	1	0.22752	0.04213	29.1584	**<.0001**	**1.255**	**1.156**	**1.364**

Table 6: Maximum Likelihood Estimates for Time-varying Covariates – Programming Statements

DISCUSSION

In this paper we have shown an example of the two methods SAS uses to handle time-varying covariates in Cox Proportional Hazard modeling and how the results compared to modeling only time-invariant covariates. Both the counting process and programming statements yield the same results, but each of them has pros and cons. As shown, the counting process requires substantially more coding. However, this upfront effort makes it easier to detect and correct errors because a data set is created and can be debugged. The programming statements are faster to code, but the coding is tricky and there is no way to detect if the time-varying covariates have been coded correctly. Furthermore, when using the programming statement method, a temporary dataset containing the time-varying covariates has to be created each time PROC PHREG is run. Depending on how large the dataset is, this could drastically increase computing time. For these reasons, we prefer the counting process.

Our results from the examples illustrated how impactful time-dependent variables can be in Cox Proportional Hazard modeling. When we only used static variables in the model, hypertension had no effect on CHD. But when we used more time points of hypertension status, we saw a very significant effect of hypertension on CHD.

In conclusion, SAS has very efficient methods to incorporate time-dependent variables in proportional hazard modeling. These methods should be utilized to ensure the most accurate model has been created.

REFERENCES

Allison, P. D. (2010). *Survival Analysis Using SAS: A Practical Guide*: Sas Inst.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: regression modeling of time to event data*: Wiley.

Smith, T. C., Jacobson, I. G., Hooper, T. I., Leardmann, C. A., Boyko, E. J., Smith, B., et al. (2011). Health impact of US military service in a large population-based military cohort: findings of the Millennium Cohort Study, 2001-2008. *BMC Public Health, 11*, 69.

ACKNOWLEDGEMENTS

We'd like to thank Dr. Nancy Crum-Cianflone, Cynthia LeardMann, and Carter Sevick for their thoughtful review of this paper. We would also like to thank the Millennium Cohort Team for their continued support.

CONTACT INFORMATION

Teresa M. Powell, MS
Naval Health Research Center
Deployment Health Research Department – Dept 164
140 Sylvester Road
San Diego, CA 92106
619-553-0684
Teresa.Powell@Med.Navy.Mil

Melissa E. Bagnell, MPH
Naval Health Research Center
Deployment Health Research Department – Dept 164
140 Sylvester Road
San Diego, CA 92106
619-553-7980
Melissa.Bagnell@Med.Navy.Mil

DISCLOSURE

The views expressed in this research are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government. Human subjects participated in this study after giving their free and informed consent. This research has been conducted in compliance with all applicable Federal Regulations governing the Protection of Human Subjects in Research.