A SAS MACRO FOR THE ADDITIVE HAZARDS REGRESSION MODEL

Alicia M. Howell and John P. Klein, Aurora Health Care, Medical College of Wisconsin
Alicia M. Howell, Aurora Health Care, 3031 West Montana Street, PO Box 343910, Milwaukee, WI 53234-3910

KEY WORDS: Survival analysis

Regression models for survival data have traditionally been based on a proportional hazards model. The effect of the covariates on survival is to act multiplicatively on some unknown baseline hazard rate which makes it difficult to model covariate effects that change over time. An alternate model is Aalen's (1980) additive model in which the covariates act in an additive manner on an unknown baseline hazard rate. This model allows for covariate effects to vary over time. Aalen's additive model is not yet widely used. One reason for this is the model is not available in any commonly used computer packages, such as SAS, SPSS, or BMDP. Presented here is a SAS macro that performs the additive hazards regression. Estimates of the cumulative covariate function and the respective standard deviations are computed. The macro provides graphical summaries of the covariate effects and tests the hypothesis of no covariate effect, as well as of contrasts of parameter vectors. An example data set is used to illustrate the macro.

1. Introduction

Survival data is time-to-event data, such as time to death, appearance of a tumor, or recurrence of a disease. Regression models for survival data have traditionally been based on a proportional hazards model, the most common being the Cox model (Cox 1972). The survival times of each individual is assumed to follow its own hazard function, $h_i(t)$. The Cox proportional hazards model is given by:

$$h(t \mid \mathbf{Z}_i(t)) = h_o(t)\exp(\mathbf{Z}_i(t)'\beta)$$

where h_o is the baseline hazard function, $\mathbf{Z}_i(t)$ is a vector of measured explanatory variables for the ith individual at time t, and β is a vector of unknown regression parameters which are assumed to be the same for all individuals. The data available in regression problems for right-censored time data are independent observations on the triple (X, δ, **Z**), where X is the minimum of the death and censoring time pair (T, U), $\delta = I_{\{T \le U\}}$ is the indicator of whether or not a death has been observed (censoring indicator), and $\mathbf{Z} = (Z_1, \ldots, Z_p)'$ is a p-dimensional column vector of covariates. The vector **Z** may be a function of t, but the only requirement is that **Z**(t) can be determined from the observations up to time t.

One alternate model is Aalen's (1980) additive hazards model. This model assumes that the covariates act in an additive manner on an unknown baseline hazard rate. The unknown risk coefficients are allowed to be functions of time so that the effect of a covariate may vary over time. For example, in studies of excess risk where the background risk and excess risk typically can have very different temporal forms, additive hazard models seem to be biologically more plausible than proportional hazards models (Buckley 1984).

Even though there are many advantages using the additive hazards model, it is not widely used. One reason for this is the model is not available in any commonly used statistical packages, such as SAS, SPSS, or BMDP. This is the inspiration for writing a SAS macro that will perform the additive hazards regression.

Section 2 outlines the additive hazards model, including estimation of the cumulative regression functions, confidence intervals, and testing. Section 3 describes the SAS macro and how it can be implemented by other users. Section 4 gives several examples to illustrate the macro, using data from Kardaun (1983) who reported data on 90 males diagnosed with cancer of the larynx. The conclusion summarizes the additive hazards model and the SAS macro written.

2. Aalen's Additive Hazards Regression Model

Applications of Aalen's additive model have been given by Mau (1986, 1988) and by Anderson and Vaeth (1989). Further theoretical analysis was made by McKeague (1986), McKeague and Utikal (1988) and Huffer and McKeague (1987, 1991).

The data required consists of a sample (t_i, δ_i, $\mathbf{Z}_i$), i=1,.., n where t_i is the time on study for the ith individual, δ_i is the censoring indicator (1 if event, 0 otherwise), and $\mathbf{Z}_i(t)=\{Z_{i1}(t), \ldots, Z_{ip}(t)\}$ is a p-vector of covariates. Although the covariates may be time-dependent, the following macro will not work for time-dependent covariates. Therefore, we will only look at the fixed covariate case, $\mathbf{Z}_i=\{Z_{i1}, \ldots, Z_{ip}\}$ For the ith individual the conditional hazard rate at time t, given $\mathbf{Z}_i$, can be modeled by the following linear model:

$$h(t \mid \mathbf{Z}_i) = \beta_o(t) + \sum_{j=1}^{p} \beta_j(t)\mathbf{Z}_{ij}$$

where the $\beta_j(t)$'s are unknown parameter functions to be estimated. These functions measure the influence of the respective covariates. Because regression functions may vary with time, analysis of them may reveal changes in the influence of the covariates over time, which is one of the main advantages of the additive model.

Estimation of the risk coefficients is based on a least squares technique. This differs from estimation in the proportional hazards model which is based on a partial or conditional likelihood. Derivation of these estimators can be found in Aalen (1989). It is much easier to estimate the cumulative regression functions than the regression functions themselves. The column vector B(t), with elements $B_j(t)$, j=0, 1, ..., p , will be estimated, where

$$\mathbf{B}_j(t) = \int_0^t \beta_j(s)\,ds.$$

To obtain the estimates we first compute the n×(p+1) matrix **Y**(t) which is defined as follows: if the ith individual is a member of the risk set at time t (event has not happened and the individual is not censored) then the ith row of **Y**(t) is the vector $\mathbf{Z}_i=(1,Z_{i1},Z_{i2},...,Z_{ip})'$. If the ith individual is not in the risk set at time t, then the corresponding row of **Y**(t) contains only zeros. It should be noted that one should use the value of **Y** just *before* a relevant event time.

Let $T_1<T_2<...$ be the ordered observed times when at least one event occurs. A reasonable estimator of B(t) is given by

$$\mathbf{B}(t) = \sum_{T_k \le t} \mathbf{X}(T_k)\mathbf{A}_k$$

where **X**(t) is a (p+1)×n generalized inverse of **Y**(t) and $\mathbf{A}_k$ is a vector of zeros except for ones in the rows corresponding to the subjects who experience an event at time T_k. The generalized inverse of **Y**(t) used here was suggested by Aalen (1989):

$$\mathbf{X}(t) = [\mathbf{Y}(t)'\mathbf{Y}(t)]^{-1}\mathbf{Y}(t)'.$$

It should be noted that the estimator **B**(t) is only definable as long as **Y**(t) has full rank and therefore $\mathbf{Y}'\mathbf{Y}$ is invertible. Therefore, estimates are restricted to the time interval where **Y** is not singular. The upper boundary on this restricted time interval will be represented by τ. Also, the estimates of the baseline hazard rate are not constrained to be non-negative.

This macro allows more than one event at a given time T_k, or tied event times. The **A**(t) vector has a one in the rows corresponding to the subjects who experience an event at time T_k.

The following estimator for the covariance matrix of **B**(t) (Aalen 1989) is used:

$$Cov = \sum_{T_k \le t} \mathbf{X}(T_k)\mathbf{A}_k^D\mathbf{X}(T_k)', \qquad i=1, ..., k,$$

where $\mathbf{A}_k^D$ is a diagonal matrix with $\mathbf{A}_k$ as the diagonal.

Confidence intervals for **B**(t) can be constructed in the usual fashion:

$$B_j(t) \pm Z_{1-\alpha/2}[\mathrm{var}(B_j(t))]^{1/2}.$$

Testing can also be done in the additive hazards model. Aalen (1989) discusses testing the hypothesis of no regression effect for one or more of the covariates. This corresponds to testing the following null hypothesis for some j = 1, ..., p:

$$H_{O_j}: \beta_j(t) = 0, \quad \text{for all } t \le \tau.$$

Testing this hypothesis can only be done in the range where **Y**(t) has full rank. A test statistic for H_{O_j} is given by the jth element, U_j, of the vector

$$\mathbf{U} = \sum_k \mathbf{K}(T_k)\mathbf{X}(T_k)\mathbf{A}_k$$

where summation takes place over all event times. **K**(t) is a (p+1)×(p+1) diagonal matrix of weight functions. Using the suggestion of Aalen (1989), the weight matrix used here is:

$$\mathbf{K}(t) = \{\mathrm{diag}[(\mathbf{Y}(t)'\mathbf{Y}(t))^{-1}]\}^{-1}.$$

An estimator of the covariance matrix of **U** is given by

$$\mathbf{V} = \sum_k \mathbf{K}(T_k)\mathbf{X}(T_k)\mathbf{A}_k^D\mathbf{X}(T_k)'\mathbf{K}(T_k).$$

To test an individual H_{O_j}, the test statistic $U_j V_{jj}^{-1/2}$ can be used. It has an asymptotic standard normal distribution under the null hypothesis. The global test statistic for testing simultaneously H_{O_j} for all j = 1, ..., q, with $q \le p$, can be done by constructing the q-vector $\mathbf{U}_q = (U_1,...,U_q)'$ and the $q \times q$ matrix $\mathbf{V}_q = ((V_{ge}),$ g=1, ..., q, e=1, ..., q). The test statistic is the quadratic form

$$\mathbf{U}_q'\mathbf{V}_q^{-1}\mathbf{U}_q$$

which has an asymptotic chi-square distribution with q degrees of freedom if the null hypothesis is true.

It is also possible to generalize testing to that of contrasts, or linear combinations of the β's. Let **C** be a r×(p+1) matrix of r contrasts. The hypothesis tested will be

$$H_O: \mathbf{C}\beta(t) = \mathbf{0}, \quad \text{for all } t \le \tau.$$

The formulas for **U, K,** and **V** change slightly:

$$\mathbf{U}_c = \sum_k \mathbf{K}_c(T_k)\mathbf{C}\mathbf{X}(T_k)\mathbf{A}_k$$

$$\mathbf{K}_c(t) = \{\mathrm{diag}[\mathbf{C}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{C}']\}^{-1}$$

$$\mathbf{V}_c = \sum_k \mathbf{K}_c(T_k)\mathbf{C}\mathbf{X}(T_k)\mathbf{A}_k^D\mathbf{X}(T_k)'\mathbf{C}'\mathbf{K}_c(T_k).$$

The test statistic for H_O is $\mathbf{U}_c'\mathbf{V}_c^{-1}\mathbf{U}_c$ which has a limiting chi-square distribution with r degrees of freedom if the null hypothesis is true.

3. The SAS Macro for the Additive Hazards Regression Model

A SAS macro is stored text that performs one or more functions and is identified by a specified name. When a macro is invoked, or executed, values are passed from the user's program to the macro through variables called macro parameters.

The macro presented here is called "additive." The additive macro fits the additive hazards model for continuous or binary covariates. The test statistic for the global test of no effects is calculated, and an analysis of variance table is printed. Options such as testing linear contrasts, plotting parameter estimates over time, and creating output data sets are available. This macro is written in IML (SAS/IML®1989) which is a matrix language built into SAS.

Even though the additive hazards model can have time-dependent covariates, this macro is not equipped to handle them. Only fixed covariates should be used. Before any estimation begins, the macro will check for missing data. If an observation has a missing value for the event time, censoring indicator, or any of the covariates then that observation will be deleted and will not be used in any calculations. The output will contain a message stating which cases, if any, were deleted due to missing values.

This macro uses a data set specified by the user. In order for the program to run correctly, the data set must meet certain criteria. First, the data set must have the following ordering of its variables: Time, Censoring Indicator, Covariates. Second, the data set must be sorted by ascending time. Third, the censoring indicator needs to specified as follows:

0 - censored; 1 - event.

Finally, all categorical variables should be defined as binary variables. The user now needs to be in IML, which is done by issuing the statement "proc iml" in the program. The additive macro is included in the program via the "%include 'filename'" statement where 'filename' is the name of the additive macro file. Next the eight parameters are defined:

1. Data set parameter: name of user's data set.

2. Confidence level parameter: defines α for $(1-\alpha)\times 100\%$ confidence intervals.

3. Time unit parameter: defines the units of the time variable.

4. Variable list parameter: lists the names of the covariates, in the order they are read in the data set.

5. Option parameter: defines which options of the five possible options the user wants by placing a "y" for "yes" or a "n" for "no" in the position of the desired option.

Options:

(1) Printing of the cumulative parameter estimates and respective standard deviations for each event time.

(2) Testing of general linear contrasts.

(3) Output data set containing parameter estimates, standard deviations, and confidence limits.

(4) Line plots of each parameter estimate versus time.

(5) Output data set containing **U** and **V** for test statistics.

6. Contrast matrix parameter: $r\times(p+1)$-matrix that defines the r linear contrasts to be tested.

7. Parameter naming the first output data set: contains the output data set requested from Option (3) above.

8. Parameter naming the second output data set: contains the output data set requested from Option (5) above.

The contents of the first output data set are as follows. It contains parameter estimates over time as well as the respective standard deviations and $(1-\alpha) \times 100\%$ confidence intervals. The first column is time. The next $(p+1)$ columns are the parameter estimates (in order), the next $(p+1)$ columns are the respective standard deviations, the next $(p+1)$ columns are the respective lower confidence limits, and the final $(p+1)$ columns are the respective upper confidence limits.

The contents of the second output data set are as follows. It contains the vector **U** and the respective variance-covariance matrix **V**. The first column is the individual $\mathbf{U}_j$ and the next p columns are the variance-covariance matrix **V**.

The additive macro is invoked using the following command:

%additive(data set parameter,confidence level parameter,time unit parameter, variable list parameter, option parameter, contrast matrix parameter, parameter naming the first output data set,parameter naming the second output data set).

Now that the macro has been included in the program and the parameters have been defined, the macro can be invoked, or called, into the program. It is crucial that the user has the parameters listed in the following order. The parameters MUST be listed in this order because they are positional. This means that the additive macro reads in the first parameter as the data set parameter, the second parameter as the confidence level parameter, and so on. If the parameters are NOT listed in this order, the program will not work. It is also necessary that there are eight parameters in the invocation statement. Even if the user does NOT choose the contrast testing option or the output data set options, dummy

parameters need to be included in the invocation statement in these positions because the positions in name-style invocation are positional.

The macro calculates the parameter estimates for each unique time. The parameters are only estimable in the range where the **Y**(t) matrix is of full rank. The macro checks if this matrix is of full rank at each time and stops estimating once the matrix is singular. The output will display a message that defines the range of estimability. Using the given data set, the macro will automatically calculate the chi-square statistic for the global test of no effects. The hypothesis is

H_{O_j}: $\beta_j(t) = 0$ for all $j = 1, ..., p$ and all $t \leq \tau$.

The heading "Global Test" and the test statistic, degrees of freedom, and the corresponding p-value are printed. An analysis of variance table is printed that will list the individual effects and each effects' chi-square statistic, degrees of freedom, and corresponding p-value. The global test and analysis of variance table are the only items calculated and printed automatically. Whatever options specified by the user will then be processed.

4. Examples of The SAS Additive Hazards Regression Macro

Several examples will be given to illustrate the additive macro. The data used is from Kardaun (1983) who reported data on 90 males diagnosed with cancer of the larynx during the period 1970-1978 at a Dutch hospital. Times recorded are the intervals (in years) between first treatment and either death or the end of the study (January 1, 1983). Also recorded are the patients age in years at the time of diagnosis and the stage of the patient's cancer. The four stages of disease in the study were based on the T.N.M. classification used by the American Joint Committee for Cancer Staging (1972). The four groups are labeled Stage 1 through Stage 4, which is ordering the stages from least serious to most serious. Stage 1 is the baseline. The variables Stage2 through Stage4 are binary variables created to define stage. They are defined as follows:

Stage2= 1 if Stage 2 disease
0 if Stage 1, 3, or 4 disease

Stage3= 1 if Stage 3 disease
0 if Stage 1, 2, or 4 disease

Stage4= 1 if Stage 4 disease
0 if Stage 1, 2, or 3 disease.

The age variable, age, has been centered at its mean:

Age = age at diagnosis - 64.11.

The first example is the simplest possible program, one that defines "no" for all five options in the option parameter. Output follows the SAS program.

Example 1: Basic program

```
options nocenter pagesize=59 linesize=80;
 data cancer;
  infile 'larynx.dat';
  input stage time age year censor;
  stage2=0; if stage=2 then stage2=1;
  stage3=0; if stage=3 then stage3=1;
  stage4=0; if stage=4 then stage4=1;
  age=age-64.11;
 proc sort; by time;
* The following routine in proc iml creates the input
SAS data set for this problem;
proc iml;
  use cancer;
  read all var _num_ into temdat;
   dummy=j(90,6,0);
   dummy[,1]=temdat[,2];    * time;
   dummy[,2]=temdat[,5];    * censor;
   dummy[,3]=temdat[,6];    * stage 2;
   dummy[,4]=temdat[,7];    * stage 3;
   dummy[,5]=temdat[,8];    * stage 4;
   dummy[,6]=temdat[,9];    * age-64.11;
  create mydata from dummy;
  append from dummy;
   quit;
proc iml;
        %include 'addmacro';
         level=0.05;
         unit={"Years"};
         varlist={"Stage 2", "Stage 3", "Stage 4" ,"Age"};
          option={n,n,n,n,n};
%additive(mydata,level,unit,varlist,option,dummy1,
dummy2,dummy3);
         quit;
```

The following output is the result from the first example:

```
The SAS System
          Additive Hazards Model

   No missing data: all observations were used in
analysis.
          90 observations used.

Estimates are restricted to the time interval 0  to   4.30

          Global Test
Chi-Square   d.f   p-value
10.9613        4    0.0270
```

Analysis of Variance

Effect	Chi-Square	d.f	p-value
Stage 2	0.1456	1	0.7027
Stage 3	3.0062	1	0.0829
Stage 4	8.4655	1	0.0036
Age	0.2333	1	0.6291

Example 2: Illustrating Some Options

This example illustrates the option of testing contrasts ("y" in position 2 in the option parameter) and creating the two output data sets ("y" in position 3 and position 5 in the option parameter). Output follows the SAS program.

Use the same code as in Example 1 to create the data set "mydata."

```
proc iml;
%include 'addmacro';
 sig=0.05;
unittime={"Years"};
covlist={"Stage 2", "Stage 3", "Stage 4","Age"};
 options={n,y,y,n,y};
contrast={0 1 -1 0 0, 0 0 1 -1 0};

%additive(mydata,sig,unittime,covlist,options,contrast,
newdat1,newdat2);
quit;
```

Output from Example 2:

The SAS System

Additive Hazards Model

No missing data: all observations were used in analysis.

90 observations used.

Estimates are restricted to the time interval 0 to 4.30

Global Test

Chi-Square	d.f	p-value
10.9613	4	0.0270

Analysis of Variance

Effect	Chi-Square	d.f	p-value
Stage 2	0.1456	1	0.7027
Stage 3	3.0062	1	0.0829
Stage 4	8.4655	1	0.0036
Age	0.2333	1	0.6291

Test of Linear Combinations

Contrast Matrix					
	0	1	-1	0	0
	0	0	1	-1	0

Chi-Square	d.f	p-value
6.8131	2	0.0332

Comments:

This particular contrast matrix,

0	1	-1	0	0
0	0	1	-1	0

is testing the hypothesis that

H_o: $\beta_1 = \beta_2 = \beta_3$

which is the test of no difference in survival between Stage 2, Stage 3, and Stage 4 patients.

It is noted that there is no output printed from using options 3 and 5 (creating the output data sets) but these data sets are now ready to be utilized in the user's program. An example of this is to use the estimates and confidence intervals given in the first output data set and create a detailed graph using SAS GRAPH ®(1990).

5. Discussion

The additive hazards regression model is very useful in modeling survival data. This SAS macro has made the model more accessible and hopefully many people will take advantage of it. This macro calculates the parameter estimates and respective standard deviations and confidence intervals. Testing the hypothesis of no regression effect for one or more of the covariates can be done as well as tests of contrasts. Line plots can be printed of each parameter estimate versus time to view how the covariate effects may change over time. Output data sets that include the parameter estimates, standard deviations, and confidence intervals can be created and used for further analysis.

This macro is available from the World Wide Web site http://biostat.mcw.edu/Software.html .

ACKNOWLEDGMENTS

This research was supported by contract 5 R01 CA54706-05 from the National Cancer Institute.

BIBLIOGRAPHY

1. Aalen, O.O. (1980) 'A model for nonparametric regression analysis of counting processes', Lecture Notes In Statistics 2:1-25.

2. Aalen, O.O. (1989) 'A linear regression model for the analysis of life time', Statistics in Medicine 8:907-925.

3. Aalen, O.O. (1993) 'Further results on the non-parametric linear regression model in survival analysis', Statistics in Medicine 12: 1569-1588.

4. American Joint Committee for Cancer Staging and End-Result Reporting (1972). Manuel for staging of cancer.

5. Anderson, P.K. and Vaeth, M. (1989) 'Simple parametric and non-parametric models for excess and relative mortality', Biometrics 45: 523-535.

6. Buckley, J.D. (1984) 'Additive and multiplicative models for survival rates', Biometrics 40: 51-62.

7. Cox, D.R. (1972) 'Regression models and life tables (with discussion)', Journal of the Royal Statistical Society, Series B 34: 187-220.

8. Dorcey, Tim (1992). Personal communication. E-mail address: tcd@cornellc

9. Huffer, F.W. and McKeague, I.W. (1987) 'Survival analysis using additive risk models', technical report 396, Department of Statistics, Stanford University.

10. Huffer, F.W. and McKeague, I.W. (1991) 'Weighted least squares estimation for Aalen's additive risk model', Journal of the American Statistical Association 86: 114-129.

11. Kardaun, O. (1983) 'Statistical survival analysis of male larynx-cancer patients -- A case study', Statistica Neerlandica 37: 103-125.

12. Klein, John, P. and Moeschberger, Melvin L. (1996) Survival Analysis: Applied Methods and Examples, New York: Springer-Verlag (in press).

13. Mau, J. (1986) 'On a graphical method for the detection of time-dependent effects of covariates in survival data', Applied Statistics 35: 245-255.

14. Mau, J. (1988) 'A comparison of counting process models for complicated life histories', Applied Stochastic Models and Data Analysis 4: 283-298.

15. McKeague, I. W. (1986) 'Estimation for a semimartingale regression model using the method of sieves', Annals of Statistics 14: 579-589.

16. McKeague, I.W. and Utikal, K.J. (1988) 'Goodness-of-fit tests for additive hazards and proportional hazards models', technical report M-793, Department of Statistics, Florida State University.

17. SAS Institute Inc., SAS/GRAPH ®Software: Introduction, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1990. 122 pp.

18. SAS Institute Inc., SAS® Guide to Macro Processing, Version 6, Second Edition, Cary, NC: SAS Institute Inc., 1990. 319 pp.

19. SAS Institute Inc., SAS/IML® Software: Usage and Reference, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1989. 501 pp.