User's Guide to the SAS Programs

Introduction and Background

The published paper [1] describes the application of the piecewise exponential (PE) regression model for performing overall survival (OS) and conditional survival (CS) analysis. It indicates how to formulate the model, derive the associated OS and CS probabilities together with their confidence limits, and perform tests of significance using Wald χ^2 procedures. The methodology is illustrated using an example published by Moertel *et al*. [2,3] evaluating long-term survival following treatment for colon cancer. Code was written in SAS to implement this methodology and perform the analyses. It is now being made available to the wider research community so that it can be applied to similar problems.

This document provides a commentary and User's Guide to that code. Details are provided on how to structure the dataset for the PE model, proceed through the various steps to manipulate the resulting parameter estimates to derive the ln OS and CS probabilities, and perform the tests of significance described in the paper. At the end, tips are provided on how to adapt this code to any new problem.

Colon Cancer Dataset

The dataset was downloaded from the GitHub public data-sharing website [4] and includes the following.

`colon.csv` .... a comma-delimited file with the raw data.

`C colon cancer.PDF` .... the accompanying documentation file.

The data file contains all the variables from Moertel's original analysis. Note that the original file used the code 'NA' for missing values. They were changed to the standard '.' used by SAS.

Of particular interest are the following variables.

`rx` .... the original treatment assignment in the clinical trial. This is a character variable with values, **`obs,`** **`lev`**, and **`Lev+5FU`**. The first two groups were subsequently pooled together in our re-analysis (see below).

`etype` .... distinguishes the two outcomes reported in the original publication. **`etype=1`** corresponds to recurrence, while **`etype=2`** corresponds to death. Death was analyzed in the paper by using the **`IF etype=2;`** statement in the **`DATA`** step.

`time` .... the elapsed time (in days) until either the event or censoring occurred.

`status`........ censoring indicator, where **`status=1`** signifies an event while **`status=0`** indicates censored.

`age,sex` .... covariates used in our re-analysis. Both are numeric variables. **`sex=0`** is for females while **`sex=1`** is for males. In our re-analysis, **`age`** was centered about its overall mean of 60, while **`sex`** was recoded to a (-1, +1) variable (see below).

`PE Model with LIFEREG.txt` Program

This is the first of two SAS programs which implement the analysis. It manipulates the input data into the format necessary for LIFEREG, runs LIFEREG, and generates the $\{\mu_{aj}\}$ and its associated covariance structure. They are subsequently read into the second program and manipulated in IML.

The first part of the program uses the SAS procedure IMPORT to read data from the .csv file and convert it to the SAS dataset COLON. The following variable was created for the treatment assignment as used in the re-analysis.

`group` .... the treatment assignment in the clinical trial, but with **`Obs`** and **`Lev`** pooled together into one group. This is a character variable with values, **`Lev+5FU`** and **`Pool`**.

At the end of the program, it is saved to the permanent SAS dataset, **`colon.sas7bdat`** , so that it can be accessed in later programs.

The paper references the monograph by Allison [5] which describes how to restructure the input dataset for a PE analysis in LIFEREG. This is performed in the second part of the program. A dataset is created with one observation for each interval during which the individual is at risk. The last interval is the one in which s/he either has the event or is censored by the end of his/her follow-up. The time variable is the elapsed time from the beginning of the interval until the event or censoring occurs; the censoring indicator in each interval is defined accordingly. Baseline covariates are repeated across the multiple observations per subject.

As described in the paper, 7 intermediate cutpoints were used in the analysis including 6, 12, 18, 24, 36, 48, and 72 corresponding to 7 of the follow-up intervals. These cutpoints were converted to days by assuming 365 days in a year. They are specified in the analysis using the **CPOINTS** macro variable. Note, however, that there is an 8th interval extending from month 72 through the end of follow-up. The last value in **CPOINTS** sets the upper boundary for this last time interval. It only needs to be a number greater than or equal to the maximum amount of follow-up. This is 3229 days in the Moertel dataset; however, I just used 10000 days because it was convenient.

CPOINTS is subsequently used in the **DO** loop. It creates the dataset with the structure described in Allison's monograph. The variable **INTERVAL** Identifies the distinct intervals and corresponds to the subscript j. The **IF LAST = 1 THEN LEAVE;** statement jumps out of the **DO** loop when the final interval for that ID is determined.

This dataset is subsequently used in PROC LIFEREG with the following considerations.

1. The **ODS OUTPUT** statement identifies two datasets. **PARAM** is used for the estimates of the $\{\mu_{aj}\}$ while **COVB** is used for its associated covariance structure.
2. **GROUP** and **INTERVAL** are declared as **CLASS** variables.
3. The **MODEL** statement specifies the **GROUP*INTERVAL** corresponds to the group × time interaction model specified in the paper. This order is important. The IML program expects the $\{\mu_{aj}\}$ to be sorted by group indexed by a and then by interval indexed by j.
4. The **NOINT** turns off the default intercept term. **D=EXPONENTIAL** specifies the exponential distribution.

   Taken together, this ensures that the parameters correspond to the $\{\mu_{aj}\}$ terms in the PE model as shown in equation (4) in the manuscript.

5. **age** and **sex** are specified as covariates in the **MODEL** statement. In the earlier **DATA** step, **age** is centered as its mean of 60, while **sex** is recoded to a (-1, +1) variable. The $\{\mu_{aj}\}$ are effectively adjusted to the means of these two covariates.

6. Note that although **sex** is categorical, it is not declared as a **CLASS** variable. SAS uses idiosyncratic coding for categorical variables declared in the **CLASS** statement – the last level is used as the" reference level". If **sex** is left as a (0,1) variable and declared in the **CLASS** statement, the $\{\mu_{aj}\}$ would the means adjusted to the last value of **sex**, i.e., for males, and not averaged over the two sexes. Recoding **sex** to a (-1, +1) variable and using it as a covariate in the **model** statement is an easy work-around when there are only two levels (see below).

7. The output datasets **PARAM** and **COVB** both contain information for a number of parameters incidental to our use. They include the variance terms as well as the β coefficients for the covariates. Extraneous rows and columns are removed from these datasets at the bottom of the program.

8. **PARAM** and **COVB** are saved to permanent SAS datasets so that they can be read into the second program.

`IML Code to Analyze Moertel Cancer Data.txt` Program

This is the second of two SAS programs. The estimates of the $\{\mu_{aj}\}$ and its associated covariance structure from the first program are read into this program. The four steps for deriving the OS probabilities are performed. Then, the user defines the $\boldsymbol{L}$ matrix in equation (7) to derive the CS probabilities. Point estimates and confidence intervals are reported for both.

Here are some considerations.

1. First, PROC IML is invoked. Then, the **%INCLUDE** statement is used to read in the IML macros located in the file, **`IML Macros for OS and CS Analysis.txt`**

2. Next. **BEGIN** is invoked to initialize various matrices. After it is invoked, the user can change the following parameters:

`_ALPHA_`............a scalar with the α level used for confidence intervals. It defaults to 0.05 for two-sided 95% confidence intervals.

`_PRINT_`............a scalar with the level of detail in the printout. The 0 value turns off most output, e.g., for a simulation study, while higher values provide progressively more. It defaults to 1.

`_MUNAMES_`........a column vector of labels for the elements $\{\mu_{aj}\}$. This is well worthwhile doing to make sure that you are interpreting the output correctly. The IML program expects the $\{\mu_{aj}\}$ to be sorted by group and interval in that order. It defaults to generic names.

In the analysis, the default values for `_ALPHA_` and `_PRINT_` are used, while `_MUNAMES_` is redefined to names specific to the Moertel re-analysis.

3. `USE` and `READ` statements are then used to access the data files created in LIFEREG and read the corresponding quantities into IML variables. `PARAM` is read into `_MU_` while `COVB` is read into `_COV_MU_`.

4. Next, the 4 steps described in the paper to generate the ln OS probabilities and associated covariance matrix are performed. Calls are made to the different modules defined in the `IML Macros for OS and CS Analysis.txt` file.

   A critical step is #3 for accumulating the hazards over the discrete intervals, i.e., $\hat{\Lambda}_{aj} = \sum_{k=1}^{j} \hat{\lambda}_{ak}\Delta_k$, where $\Delta_k = \tau_k - \tau_{k-1}$. The vector `_DELTA_` defines the $\{\Delta_k\}$ terms, i.e., the elapsed times between the cutpoints. The elapsed times in `_DELTA_` must be consistent with the `CPOINTS` macro used in LIFEREG. `CPOINTS` defines the elapsed time from Time 0 to the cutpoints, while `_DELTA_` defines the elapsed times from one cutpoint to the next.

   Thus, `DELTA_ = {182.5 182.5 182.5 182.5 365 365 730 1095};` indicates that the elapsed time to the first 4 cutpoints at months 6, 12, 18 and 24 is 182.5 days; 365 days to those at months 36 and 48; 730 days to that at month 72; and 1095 days (3 years) up to month 108.

5. Calls are then made to the `LTRIANG` module to create the lower triangular matrix and `LINEAR` to perform the linear transformation.

6. Point estimates and standard errors for the OS probabilities, together with their confidence intervals at the α level of significance, are then derived. If **`_PRINT_`** is ≥ 1, they are reported.

7. Next, hypothesis testing is performed. Each test is performed by defining the appropriate contrast matrix $\boldsymbol{C}$, denoted by **`_C_`** in the code, and invoking the **`WALDTEST`** module using the **`RUN`** statement. If **`_PRINT_`** is ≥ 1, the results are reported. For example, **`_C_ = I(8) || (-1)#I(8);`** evaluates whether there is an overall between-group difference across the 8 intervals.

8. The CS probabilities are derived conditional on surviving the 1-year treatment period, i.e., the second cutpoint. The linear transformation matrix, $\boldsymbol{L} = [\mathbf{0}_{6\times1} \quad -\mathbf{1}_{6\times1} \quad \mathbf{I}_{6\times6}]$ is applied by defining **`_L_ = J(6,1,0) || J(6,1,-1) || I(6);`** and performing a **`LINEAR`** transformation.

9. Point estimates, standard errors, and confidence intervals at the α level of significance for the CS probabilities are derived. If **`_PRINT_`** is ≥ 1, they are reported.

10. Hypotheses on between-group differences in the CS probabilities both overall and at months 24 and 36 are tested by defining the appropriate **`_C_`** matrices and invoking the **`WALDTEST`** module. For example, **`_C_ = {0 1 0 0 0 0  0 -1 0 0 0 0}`** evaluates whether there is a between-group difference in the CS probabilities at month 24. If **`_PRINT_`** is ≥ 1, the results are reported.

Tips for Modifying the Code for a New Analysis

The two SAS programs can be modified fairly easily for any new problem. But, attention must be paid to the following issues.

1. Be consistent with the time metric. Although I use days, months and years interchangeably in the manuscript, days is used exclusively in the code.

2. Careful consideration must be made for the choice of cutpoints. As described in the paper, the number of cutpoints must strike a balance between fully characterizing the survival curves and estimating all the resulting $\{\mu_{aj}\}$ reliably. Check the SAS log from the **`PE Model with LIFEREG.txt`** program to ensure that there are no convergence issues.

3. Sensitivity analyses should be performed to assess robustness of the inferences to the choice of the partition. Nevertheless, the number of cutpoints is not generally the critical issue. Rather, identifying cutpoints that capture the dynamics of the changing survival curve is the key consideration.

4. The **CPOINTS** macro in the LIFEREG program is defined as the elapsed time from Time 0 to the cutpoints, while **_DELTA_** in the IML program defines the elapsed times from one cutpoint to the next. You need to make sure that you appreciate the difference and that they are consistent.

5. You need to be careful about the final interval. There are 7 time intervals corresponding to the 7 cutpoints. But, there is an 8th interval that extends from month 72 through the end of follow-up. Thus, the **CPOINTS** macro and **_DELTA_** both have 8 elements – one for each interval.

6. The maximum follow-up time in the Moertel dataset is 3329 days, i.e., slightly beyond 9 years. Specifying that the last element of the **CPOINTS** macro is greater than or equal to the maximum follow-up creates the 8th interval for the analysis, and ensures that all the available data are used in the LIFEREG analysis.

7. Note, however, that for reporting in the manuscript tables, I just used 9 years = 108 months = 3285 days for the 8th time point. This is more meaningful than the 3329 days actually observed.

8. Adjusting for covariates is a little tricky. The IML code expects the $\{\mu_{aj}\}$ to be adjusted to reference values of the covariates, generally the overall means. You therefore need to center continuous covariates at their reference values in a **DATA** step beforehand. Recode binary covariates to (-1, +1) variables as performed for **sex** and treat them as continuous.

9. This is more complicated when a categorical covariate has more than two levels, e.g., race or SES. In some SAS procedures, you can specify the **PARAM=EFFECT** option to the **CLASS** statement to implement this coding. Although this is available in PHREG, it is not available in LIFEREG at the time of writing.

10. Until it is, you need to create (-1, 0, +1) coding manually in a **DATA** step beforehand and treat the resulting variables as continuous. Basically, if there are l levels for a categorical

covariate, create $l - 1$ indicator variables corresponding to the first $l - 1$ levels. For the last level, the value for all the $l - 1$ indicators is -1 .

Read the "Parameterization of Model Effects" section of the "Shared Concepts and Topics" chapter in the SAS documentation for the details.

A good way to get started with this software is to run the sensitivity analysis described in the supplemental material. There, 12 cutpoints were identified including 3, 6, 9, 12, 15, 18, 21, 24, 30, 36, 48, and 72 months, with the final interval extending through the end of follow-up. Compare your results against the tables in the supplemental document.

References:

1. Rochon J. An integrated model for overall and conditional survival analysis in epidemiologic studies. *In press*.
2. Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Goodman PJ, et al. Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *N Engl J Med*. 1990;322:352-358.
3. Moertel CG, Fleming TR, Macdonald JS, Haller DG, Laurie JA, Tangen CM, et al. Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage III colon carcinoma: a final report. *Ann Intern Med*. 1995;122:321-326.
4. Rdatasets – An archive of datasets distributed with R. Chemotherapy for Stage B/C colon cancer. https://vincentarelbundock.github.io/Rdatasets/. Accessed March 13, 2018.
5. Allison, P.D. *Survival Analysis using SAS: A Practical G*uide. 2nd ed. Cary, NC: SAS Institute, Publishers; 2010.