Paper 375-2008

Advanced Statistical and Graphical features of SAS® PHREG

Lida Gharibvand, University of California, Riverside

George Fernandez, University of Nevada, Reno

ABSTRACT

Survival analysis involves the modeling of time-to-event data whereby death or failure is considered an "event". The graphic presentation of Cox proportional hazards model using SAS PHREG is a significant tool which facilitates effective data exploration in survival analysis. The SAS PROC PHREG can generate some of the useful survival analysis plots using the ODS graphics option in version 9.1.3. In this paper, we will demonstrate the advanced features of PHREG for investigating the cumulative martingale residual plots and for selecting best candidate models in model selection. In clinical trials, potential outlier individuals who 'died far too early' or 'lived far too long' are identified and compared to what the fitted model predicts. The cumulative residuals from PROC PHREG are used to investigate the model specification error of covariate and validate the proportion hazard function. Methods to identify outliers are commonly based on Cox regression residuals such as martingale and deviance residuals. We will use PROC GPLOT in SAS/GRAPH to generate these two residual plots and to detect influential outliers. We will outline a method to perform all possible subset model selection within user-defined subsets using AIC information criterion. Also we will discus the new and improved features of the BPHREG, an experimental upgrade to PHREG procedure that has some user-friendly options such as 'class', the 'hazards ratio', and 'strata' statements which can be used to fit Cox proportional hazards model more efficiently.

INTRODUCTION

Survival analysis is the phrase used to describe the analysis of data in the form of times from a well-defined "time origin" until the occurrence of some particular event or "end-point". In medical research, the time origin often corresponds to the recruitment of an individual into an experimental study, such as a clinical trial to compare two or more treatments. This in turn may coincide with the diagnosis of a particular condition, the commencement of a treatment regiment, or the occurrence of some adverse event. If the end point is the death of a patient, the resulting data are literally survival times. However, data of a similar form can be obtained when the end-point is not fatal, such as the relief of pain, or the recurrence of symptoms. In this case, the observations are often referred to as *time to event* data.

The analysis of survival data requires special techniques because the data are almost always incomplete and familiar parametric assumptions may be unjustifiable. Investigators follow subjects until they reach a pre-specified endpoint (for example, death). However, subjects sometimes withdraw from a study, or the study is completed before the endpoint is reached. In these cases, the survival times (also known as failure times) are *censored*; subjects survived to a certain time beyond which their status is unknown. The uncensored survival times are sometimes referred to as *event* times. Methods for survival analysis must account for both censored and uncensored data. This paper is focused on PROC PHREG and PROC BPHREG, but we can also use PROC LIFEREG, and PROC LIFETEST to fit other types of survival analysis. For more information, please refer to our previous paper (Survival Analysis Plots Using SAS ® ODS Graphics (Gharibvand and Fernandez 2007).

SAS PROC PHREG:

PROC PHREG is a semi-parametric procedure that fits the Cox proportional hazards model (SAS Institute, Inc. (2007b)). PROC BPHREG is an experimental upgrade to PHREG procedure that can be used to fit Bayesian Cox proportional hazards model (SAS Institute, Inc. (2007c)). Cox's semi-parametric model is widely used in survival analysis to model the effect of covariates on hazard rates. Cox's proportional hazards model assumes a parametric form for the effects of the covariates, but it allows an unspecified form for the underlying survivor function. The partial likelihood of Cox model also allows time-dependent covariates. A covariate is time-dependent if its value for any given individual can change over time. The validity of the proportional hazards model can be tested by testing for interaction between time-dependent covariates and the response time.

ODS SAS STATISTICAL GRAPHICS

In SAS 9.1, a number of SAS/STAT ® procedures have been modified to use an experimental extension to the Output Delivery System (ODS) that enables them to create statistical graphics as automatically as tables. This facility is referred to as ODS Statistical Graphics (or ODS Graphics for short), and it is invoked when you provide the experimental ODS GRAPHICS statement prior to your procedure statements. Any procedures that use ODS Graphics then create graphics, either by default or when you specify procedure options for requesting specific graphs. ODS graphics is experimental in SAS 9.1, and is expected to achieve production status in SAS 9.2 (SAS Institute, Inc. (2005)). With ODS Graphics, a procedure creates the graphs that are most commonly needed for a particular analysis. Using ODS Graphics eliminates the need to save numerical results in an output data set, manipulate them with a DATA step program, and display them with a graphics procedure. One requirement for using ODS GHRAPHICS is that ODS output files must be saved in RTF or HTML format.

SURVIVAL DATA USED TO DEMONSTRATE THE ADVANCED FEATURES OF PHREG

The Mayo liver disease example of Lin, Wei, and Ying (1993) is used here to demonstrate the features. The data consists of 418 patients with Primary Biliary Cirrhosis (PBC), among which 161 had died as of the date of data listing. The data set contains the following variables:

Time follow-up time in years
Censor event indicator with value 1 for death time and value 0 for censored time
Age age in years from birth to study registration
Alb serum albumin level in gm/dl
Bili serum Bilirubin level in mg/dl
Edema Edema with value 0 for presence of no Edema, Edema with value 0.5 for untreated or successfully treated, and Edema with value 1 for unsuccessfully treated Edema
Protime prothrombin time in seconds

ALL POSSIBLE MODEL SELECTION WITHIN USER DEFINED SUBSET

When a researcher is interested in determining the important prognosis factors from a large list of candidate variables, he or she will use the variable selection which is a standard exploratory exercise in survival analysis. There are four ways to select models based on the PHREG procedure: forward selection, backward elimination, stepwise selection, and best subset selection. The forward selection and the backward elimination methodologies are straightforward. The stepwise selection method is highly unstable in terms of regression coefficients estimates, standard errors, and confidence intervals. The likelihood score statistic is the basis for the best subset selection methodology in which a specified number of best models containing several variables, up to the single model containing all of the explanatory variables are determined. For example, if the number of possible explanatory variables is a relatively small number such as p=10, then the number of possible models to compare is $K = 2^{10} = 1024$. Assuming a more reasonable number such as p=20 would result in one million possible models. Shtatland et. al, (2005) advocate a method for building models based on combination of stepwise regression, Akaike information criteria, and the best subset selection. Specifically, their proposed method avoids the complicated process of choosing the “right” critical p-value in stepwise regression. Because of the millions of possible models to compare, a direct comparison would be impossible. Shtatland, et al (2005) offer a reasonable solution to use the stepwise selection method with SLENTRY and SLSTAY close to 1 (e.g., SLENTRY = 0.99 and SLSTAY = 0.995). Therefore, the sequence of models starting with the null model and ending with the full model including all the explanatory variables can be determined. The models in this sequence will be ordered in the way minimizing the AIC value at every step. As a result, it is natural to call this sequence the “stepwise sequence” (Shtatland et. al, 2005). The paper presented by Shtatland et.al (2005) and the code they provided was based on SAS version 8.2. However, we found some errors when we ran this code in SAS version 9.1. Thus, in this paper, we are presenting an updated SAS code, which will be compatible with version 9.12.

```
proc phreg data=pbc;

     model time*censor(0)= age  logbili  edema   logalb
logprotime edema0 edema05/ selection=stepwise slentry=0.99
slstay=0.995;

      run;
```

Code 1: **Sequential stepwise regression code**

Step	Criterion	Without Covariates	With Covariates
1	AIC	2659.489	2632.123
2	AIC	2659.489	2617.642
3	**AIC**	**2659.489**	**2617.143**
4	AIC	2659.489	2617.952
5	AIC	2659.489	2619.195
6	AIC	2659.489	2621.184
7	AIC	2659.489	2623.172

Table 1: **Sequential stepwise regression output**

The minimum AIC is obtained in step three with three variables model. Therefore, we can perform all possible subset selection between 2 and 4 subsets (3-1 and 3+1). The SAS code for performing all subset selection between subset 2 and 4 using all combination is given in the code box 2.

```
ods rtf select all;
ods output BestSubsets=Best_subsets;
 ods rtf exclude all;
 proc phreg data=pbc;
 model time*censor(0)=age logbili  edema logalb logprotime edema0 edema05 /
selection=score START= 2 STOP= 4 ;
run;
ods rtf exclude all;
proc print data=Best_subsets;
run;
OPTIONS MPRINT SYMBOLGEN MLOGIC;

%MACRO SCORE;
proc sql noprint;
select (nobs -delobs) into: num
from dictionary.tables
where libname ='WORK'
and memname = "BEST_SUBSETS";
%let num=&num;
quit;
%do i=1 %to &num;
data _null_ ;
set Best_Subsets;
if _N_ = &i;
call symput('list', VariablesInModel);
run;

ods output FitStatistics=Fit2;
ods rtf exclude all;
proc PHREG data=pbc;
model time*censor(0) = &list;
run;

ods rtf select all;
data icaic(keep=aic); set fit2;
       if criterion='AIC';
       aic=withcovariates;
       run;
       data ic(keep=model  aic ); set icaic;
       model="&list"; Run;
       run;
```

```
Proc append base=subaic data=ic force;
run;

%end;
%MEND;
%SCORE;
```

Code 2: All subset selection between subset 2 and 4 using all combinations

The results of all possible subset selection within user defined subset out of so many subsets (91) are presented in the table 2. We are only showing result of ten top models in this table 2. The best model includes four variables: logbili, logalb, logprotime, and edema0.

AIC	Model
2617.95	**logbili logalb logprotime edema0**
2631.19	logbili logalb logprotime
2631.85	age logbili logalb logprotime
2632.21	logbili edema logalb logprotime
2633.19	logbili logalb logprotime edema05
2664.69	logbili logalb edema0
2665.58	logbili edema logalb edema0
2666.55	age logbili logalb edema0
2666.68	logbili logalb edema0 edema05
2677.41	logbili edema logalb

Table 2: Top ten models selected based on minimum AIC from the subset selection between subset 2 and 4.

PROC PHREG ODS STATISTICAL GRAPHICS FEATURES

There is only one ODS GRAPHICS PLOT, cumulative residual plot for detecting model specification error, available in Version 9.1.3. As Lin et al. (1993) suggests, a Cox model for the survival time of the PCB patients with covariates Bili, log(Protime), log(Albumin), Age and Edema was fitted using the following code 3.

```
ODS RTF FILE='path\filename.rtf'style=statistical;
ODS LISTING CLOSE;
ODS GRAPHICS on;
DATA pbc;
PROC PHREG DATA=pbc;
MODEL Time*Censor(0)= Bili logAlb logProtime  Edema Age;
logProtime=log(Protime); logAlb=log(Alb);
ASSESS VAR=(Bili) /  RESAMPLE;
RUN;
```

```
OUTPUT OUT=outp XBETA=xb RESMART=mart RESDEV=dev RESSCH =ressch LMAX=lmax
RESSCO=ressco;
ODS GRAPHICS oFF;
ODS RTF CLOSE;
ODS LISTING;
QUIT;
```

Code 3: Fitting the survival time model

The ASSESS statement creates a plot of the cumulative Martingale residuals against the values of the covariate Bili, which is specified in the VAR= option. The RESAMPLE option computes the p-value of a Kolmogorov-type supremum test based on a sample of 1,000 simulated residual patterns.

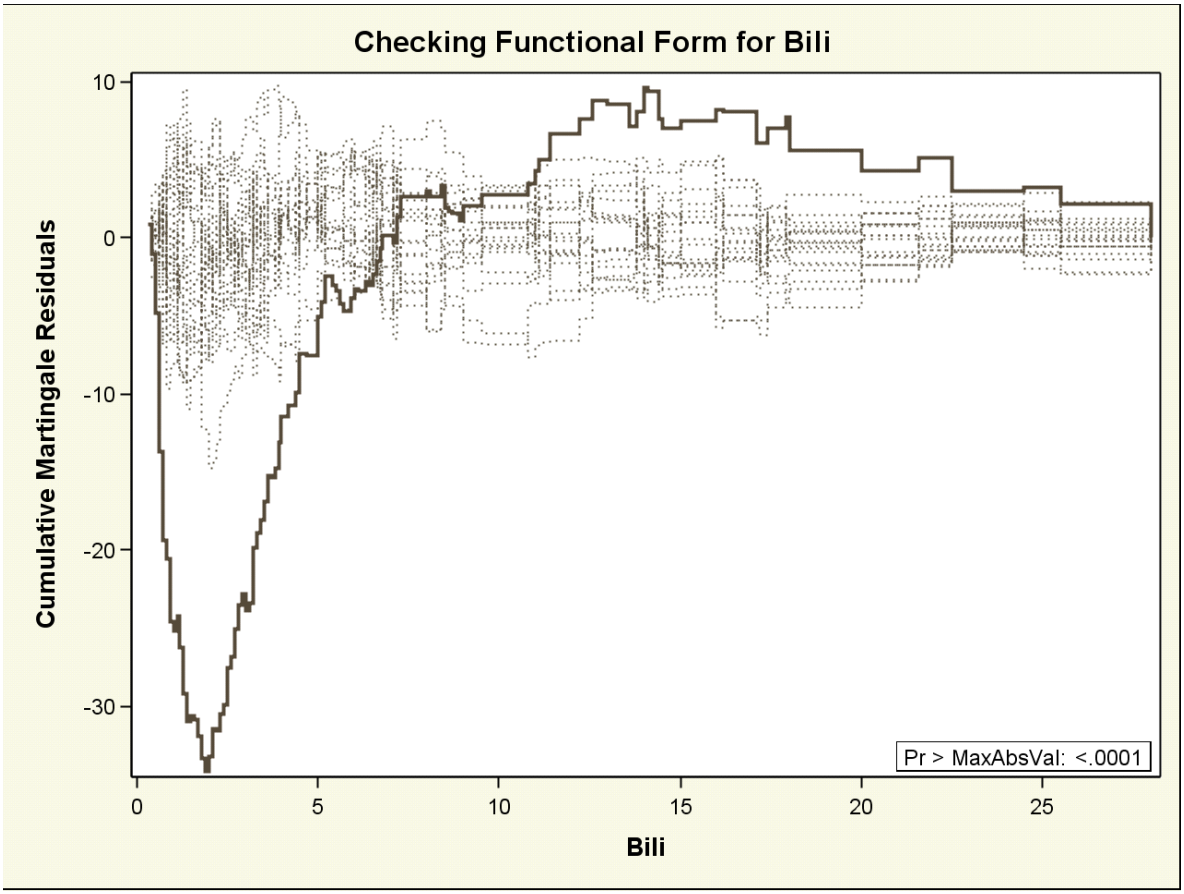

Figure 1: Cumulative Martingale residuals vs Bili (Experimental)

The plot in Figure 1 displays the observed cumulative Martingale residual process for Bili together with 20 simulated realizations from the null distribution. This graphical display is requested by specifying the experimental ODS GRAPHICS statement and the experimental ASSESS statement.

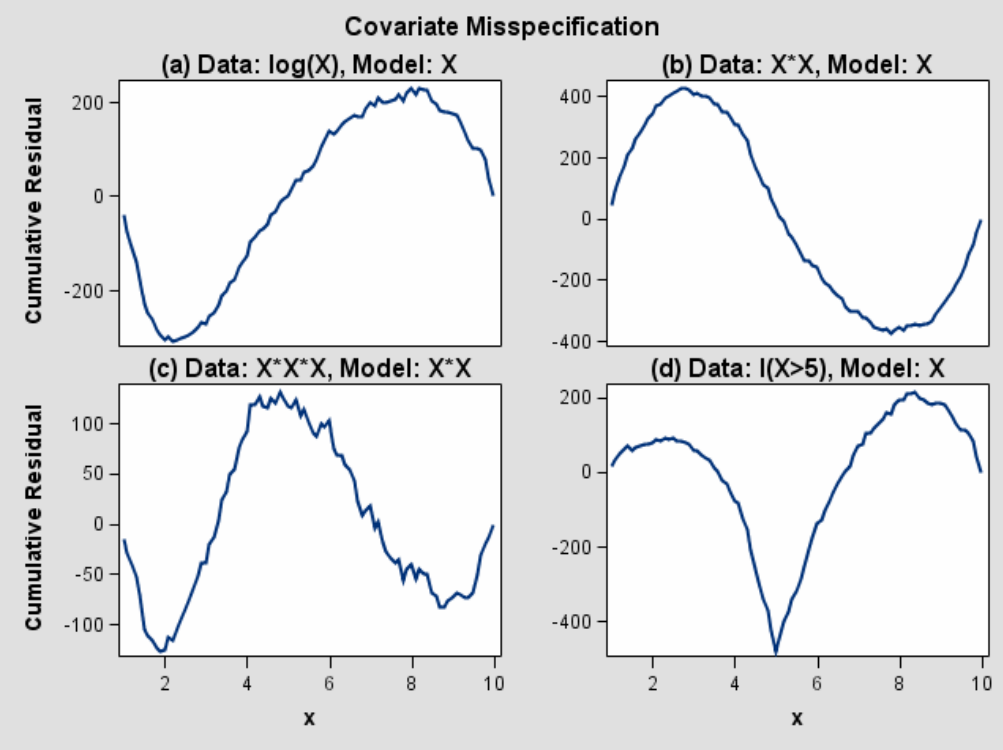

Figure 2: **Typical cumulative residual plot patterns**

The curve of observed cumulative Martingale residuals in Figure 1 most resembles the behavior of the curve in Figure 2(a), indicating that log(Bili) might be a more appropriate term in the model than Bili. Next, the analysis of the natural history of the PBC is repeated with log(Bili) replacing Bili, and the functional form of log(Bili) is assessed (Figure 3). When we compare Figures 1 and 3, we can see that Figure 3 shows a better function form when log(Bili) was used. Therefore, the model fit is improved when Bili is replaced by log(Bili).

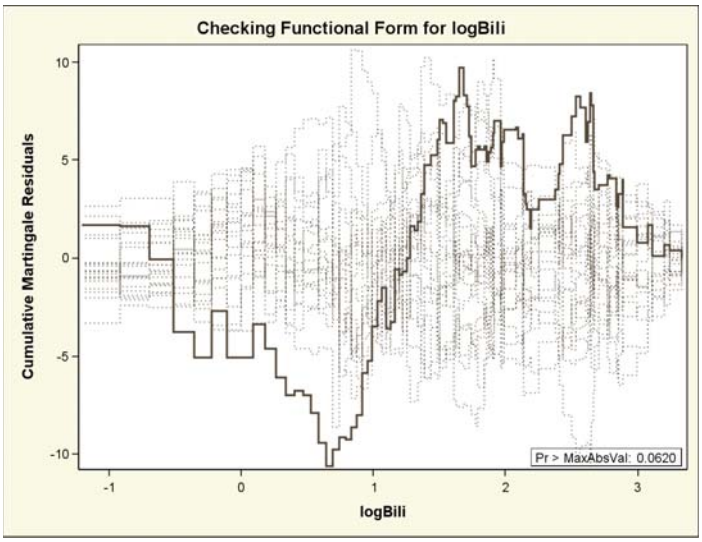

Figure 3: **Cumulative Martingale residuals vs log(Bili)**

OTHER EXPLORATORY PLOTS: OUTLIER DETECTION

All releases of PROC PHREG have options for three different residual statistics that are computed for each individual in the sample: Cox-Snell residuals (LOGSURV), Martingale residuals (RESMART), and deviance residuals (RESDEV) (Code 3). In addition, LMAX, relative influence of observations on the overall fit of the model can also be generated in the OUPUT. This diagnostic is useful in assessing the relative influence (sensitivity) of the fit of the model to each observation. The residual analysis for detecting outliers and influential data are lacking in the SAS ODS GRAPHICS for survival analysis. However, PHREG (and BPHREG) have options for outputting for Martingale, deviance residuals, and LMAX. By using SAS GPLOT procedure, we can perform residual analysis and identify the outlier and the influential one. LMAX statistic is useful to detect the influential observations. See Code 4 & 5.

MARTINGALE RESIDUAL PLOT

Martingale residuals are obtained by transforming Cox-Snell residuals, and deviance residuals are considered as a transformed Martingale residuals. While Cox-Snell residuals were useful for assessing the fit of the parametric models, they are not very informative for Cox models estimated by partial likelihood.

```
PROC GPLOT DATA=outp ;
TITLE1    "Martingale residuals plot";
PLOT  mart*xb /CFRAME=white OVERLAY VAXIS=axis1 HAXIS=axis2 FRAME VREF=0
VMINOR=0    HMINOR=0  CAXIS = BLACK NAME='plot3';
AXIS1 LABEL=(A=90 R=0 F="<ttf> Arial ""Martingale Residual")WIDTH=2;
AXIS2 LABEL=("Linear Predictor") VALUE=none WIDTH=2; RUN;  QUIT;
```

Code 4: Martingale residuals

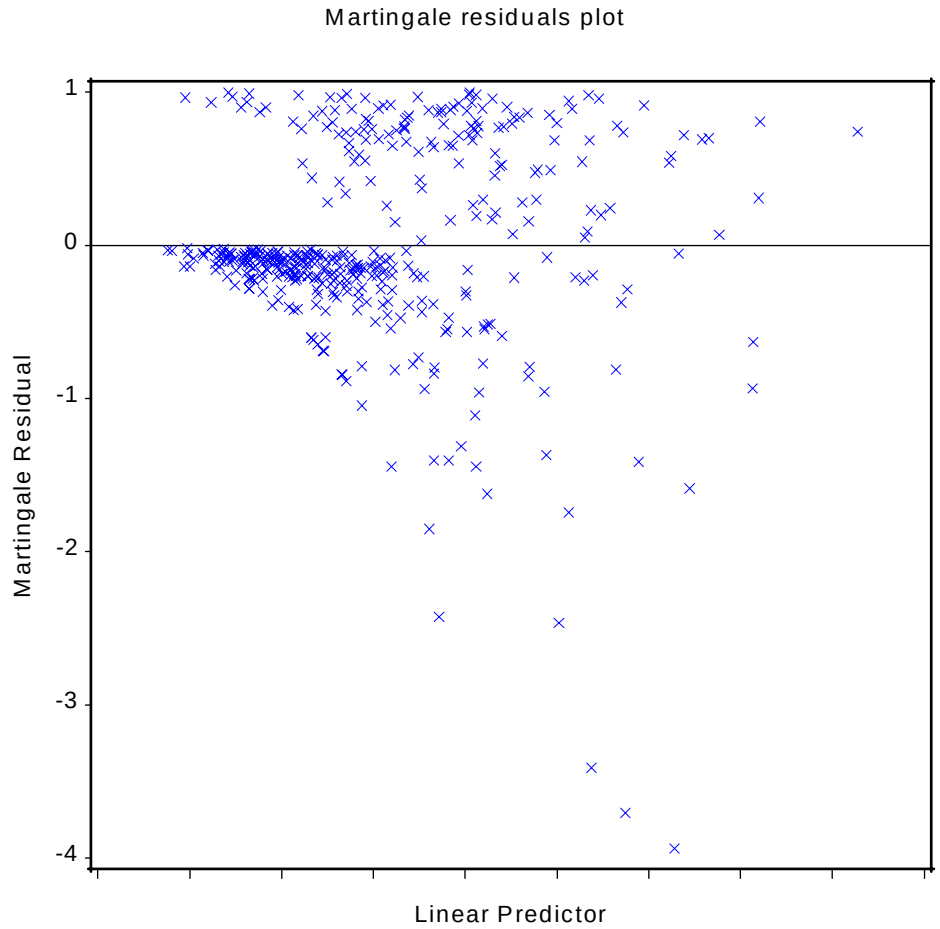

Figure 4: Martingale residuals using PROC GPLOT

Although Martingale residuals share many of the properties possessed by residuals encountered in other situations, such as in linear regression analysis, they are not symmetrically distributed about zero, even when the fitted model is correct (Figure 4). This skewness makes plots based on the residuals difficult to interpret. The deviance residuals, which were introduced by Therneau et al. (1990), are much more symmetrically distributed about zero (Figure 5).

DEVIANCE RESIDUAL

Deviance residuals behave much like residuals from OLS regression: they are symmetrically distributed around 0 and have an approximate standard deviation of 1.0. They are negative for observations that have longer survival times than expected and positive for observations with survival times that are smaller than expected. Very high or very low values suggest that the observation may be an outlier in need of special attention. You can plot the residuals against the covariates, and any unusual patterns may suggest features of the data that have not been adequately fitted by the model. Be aware, however, that censoring can produce striking patterns that don't necessarily imply any problem with the model.

```
DATA INF;
SET outp(where =(dev ne .));
id=_n_;
LENGTH   text $12 function $8;
RETAIN XSYS '2' YSYS '2' size 1;
X=xb  ;   Y=dev;
IF abs(dev) > 2.5  THEN DO; function= 'LABEL'; position= '8'; TEXT=ID;
END;
RUN;

GOPTIONS RESET=all  COLORS=(Black, RED,BLUE,YELLOW,GREEN,MAGENTA,CYAN)
dev=EMF target=EMF  XMAX=7 YMAX=7 HTEXT=14pt FTEXT="<ttf> Arial";

PROC GPLOT DATA=outp ;
TITLE1    "Deviance residuals plot";
TITLE2    "Outlier and influential diagnostics  ";
BUBBLE  dev*xb=lmax /CFRAME=white ANNOTATE=inf  VAXIS=axis1 HAXIS=axis2
FRAME VREF=-2.5 0  2.5 VMINOR=0 HMINOR=0 CAXIS= black NAME='plot3'
BCOLOR=red BSIZE=12;
AXIS1 LABEL=(A=90 R=0 F="Arial ""Deviance Residual")WIDTH=2; RUN;
QUIT;

ODS GRAPHICS OFF;
ODS RTF CLOSE;
ODS LISTING;
```

Code 5: Deviance residuals using PROC GPLOT

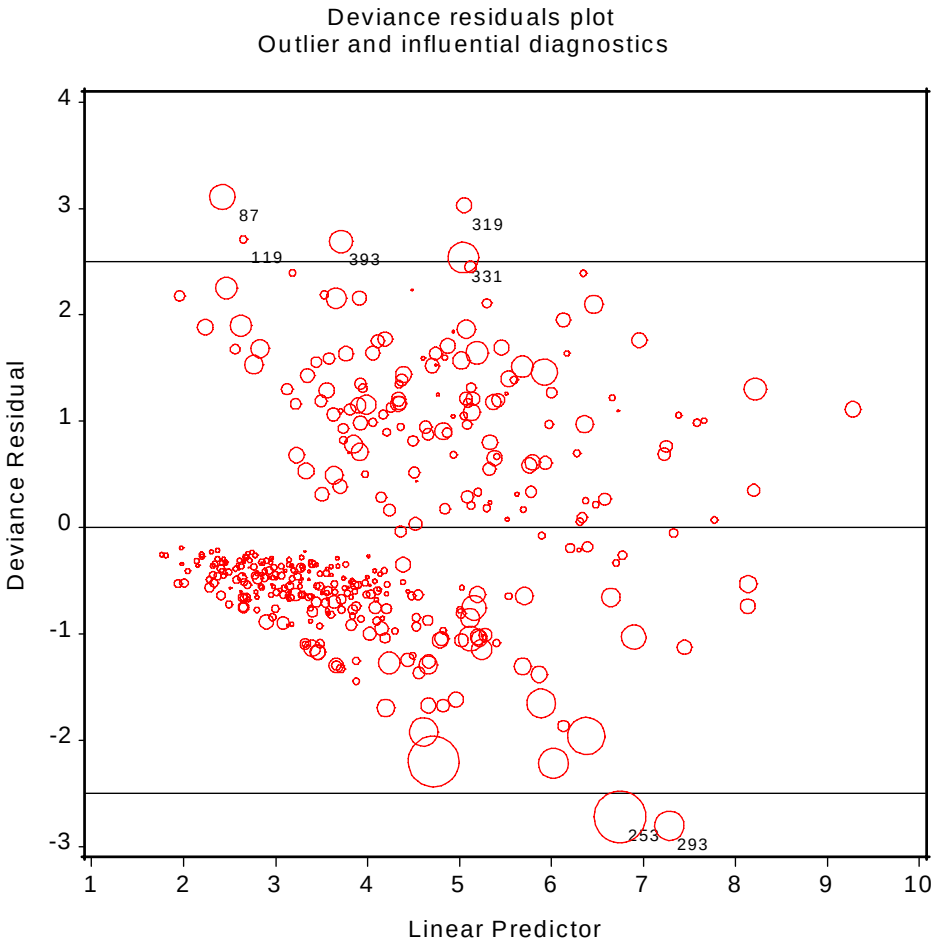

Figure 5: Deviance residuals using PROC GPLOT

Figure 5 shows deviance residuals with influential data and outliers. As we can see, numbers 87, 119, 292, 319, 331, 253, and 293 are outliers. The diameter of bubble is proportional to LMAX statistic. Observations with relatively large bubbles can be considered as the influential observations. The big bubbles inside the lines are not outliers, but are highly influential.

PROC BPHREG- NEW SAS EXPERIMENTAL PROC

In PHREG 9.1.3, there is no CLASS statement and the ASSESS statement fails to produce ODS Graphics in presence of STRATA variable. However, a new experimental PHREG procedure is available to run Bayesian PHREG (BPHREG). In the BPHREG statement we can include CLASS statement and ASSESS statement. Users can download BPHREG at http://www.sas.com/apps/demosdownloads/setupintro.jsp. Then click on SAS/STAT Software and then SAS/STAT Bayesian Procedures to download.

CHECKING FOR PROPORTIONAL HAZARD FUNCTIONS:

```
proc bphreg data=Liver;
class edema ;
   model Time*Status(0)=logBilirubin  logprotime  Albumin Age Edema;
   logBilirubin=log(Bilirubin);
   logProtime=log(Protime);
        assess ph/ crpanel resample seed=19;
   run;
```

Code 6: Proportional hazard functions

Using **Assess** statement and **ph** option, we can perform the test for proportional assumption (Table 3) and the ODS exploratory graphics for testing ph assumption (Figure 6.)

Supremum Test for Proportional's Hazards Assumption				
Variable	**Maximum Absolute Value**	**Replications**	**Seed**	**Pr > MaxAbsVal**
logBilirubin	1.1170	1000	19	0.1230
logprotime	**1.6681**	**1000**	**19**	**0.0010**
Albumin	0.8043	1000	19	0.5140
Age	0.6897	1000	19	0.5370
Edema0	2.1694	1000	19	0.0210
Edema0D5	1.7343	1000	19	0.1070

Table 3: Supremum test for proportional hazards assumption

Based on table 3, we can see that the logprotime seriously violates the ph assumption. Thus, the remedial measure for correcting the ph assumption violation is performing stratified analysis. Using the "strata" statement (strata protime(**10**, **11**, **12**)) in PROC BPHREG we can stratify the protime variable in three groups. Thus, the strata statement is creating three stratas based on protime values, where group 1 is less than or equal to 10, group2 is between 10 and 11, and group3 is greater than 11. Thus, a stratified PH COX regression model can be performed using the Strata statement in BPHREG (Code 7)

The parameter estimates of the final stratified PH COX regression model are given in table 4. Using the strata option, the AIC value is reduced from 1572.328 (un-stratified model) to 1124.272 (stratified model). We only showed here the results of how we fixed the effects of problematic variables that failed the PH assumption.

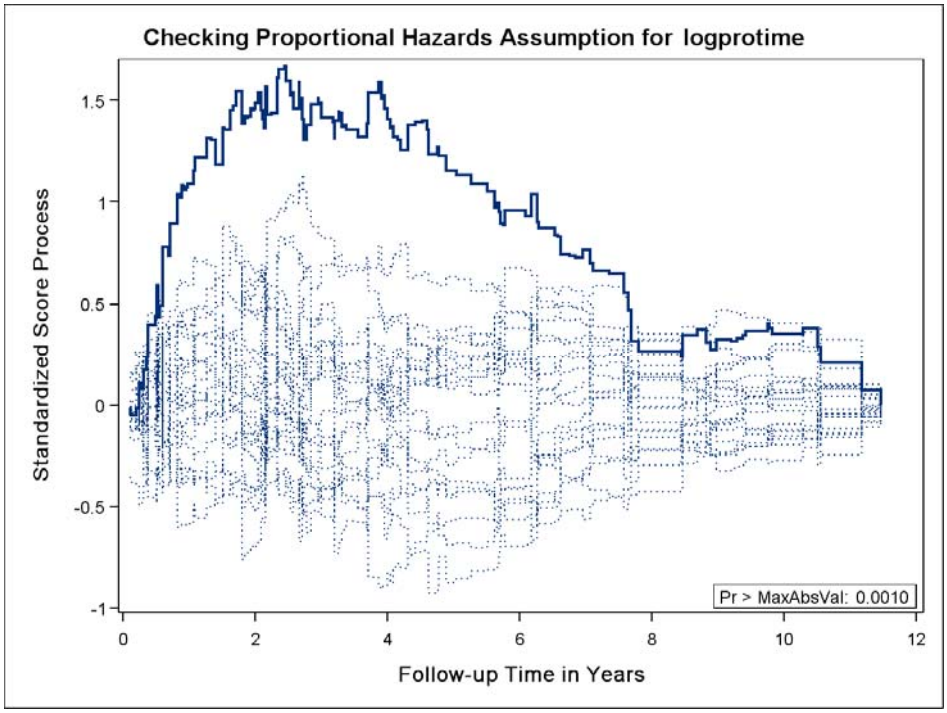

Figure 6: Explore plot for checking proportional hazard functions

```
proc bphreg data=Liver;
class edema ;
strata protime(10, 11, 12 );
   model Time*Status(0)=logBilirubin  Albumin Age Edema ;

   logBilirubin=log(Bilirubin);

      run;
```

Code 7: **Strata statement in BPHREG**

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
logBilirubin		1	0.81147	0.08744	86.1174	<.0001	2.251	
Albumin		1	-0.85582	0.21526	15.8070	<.0001	0.425	
Age		1	0.03567	0.00778	21.0112	<.0001	1.036	
Edema	0	1	-0.62515	0.30849	4.1066	0.0427	0.535	Edema 0
Edema	0.5	1	-0.42826	0.33760	1.6092	0.2046	0.652	Edema 0.5

Table 4: **Parameter estimates of the final model**

NEW 'HAZARDSRATIO' OPTION

The SAS PROC BPH has a new option for performing all possible comparisons between the levels of categorical variable at a given level of continuous variable. For example, let us assume that we want to compare the hazard ratio for three edema levels (0, 0.5, 1) at age 75 years which is (hazardratio edema / diff=all at(age=75)) statement. The output of all possible hazard ratio comparisons is presented in table 5.

Hazard Ratios for Edema			
Description	**Point Estimate**	**95% Wald Confidence Limits**	
Edema 0 vs 0.5 At Age=75	0.821	0.513	1.314
Edema 0 vs 1 At Age=75	0.535	0.292	0.980
Edema 0.5 vs 1 At Age=75	0.652	0.336	1.263

Table 5: **SAS output from hazard ratio statement**

SUMMARY

In this paper, we demonstrated some important tools and plots to conduct Cox's proportional hazard survival analysis. We presented key features such as cumulative Martingale residual plots, and outlier detection plots using PROCS, PHREG, BPHREG, and ODS Graphics. The use of ODS Graphics would enable us to make customized graphics with ease. Furthermore, we showed the solutions to the limitations in the PHREG PROC in producing the ODS Graphics in the presence of both STRATA and ASSESS statements. By using the experimental BPHREG procedure and the new CLASS statement, we showed how to produce the cumulative residual plots using the ODS Graphics. In conclusion, ODS Graphics is an extremely useful new

experimental feature in SAS that allows the creation of sophisticated statistical graphics. The graphs will maintain a professional appearance, and with the use of styles, will look consistent with other ODS output.

REFERENCES

Gharibvand, L. and Fernandez, G. (2007), “Survival Analysis Plots Using SAS ® ODS Graphics”, Western Users of SAS Software (WUSS) conference proceedings San Francisco
http://www.crda.unr.edu/crda/publications/ANL_Gharibvand_SurvivalAnalysis.pdf

Lin, D. and Wei, L. and Ying, Z. (1993), "Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals", *Biometrika*, 80, 557 - 572

SAS Institute, Inc. (2005), “TEMPLATE Procedure: Creating ODS Statistical Graphics Output (Experimental)” http://support.sas.com/rnd/base/topics/statgraph/proctemplate/a002774500.htm

SAS Institute, Inc. (2007a), SAS/STAT ® User’s Guide, SAS OnlineDoc® 9.1.3, Cary, NC: SAS Institute, Inc.

SAS Institute, Inc. (2007b), PHREG: SAS/STAT ® User’s Guide, SAS OnlineDoc® 9.1.3
http://support.sas.com/91doc/getDoc/statug.hlp/phreg_index.htm
Cary, NC: SAS Institute, Inc.

SAS Institute, Inc. (2007c), BPHREG: SAS/STAT ® User’s Guide, SAS OnlineDoc® 9.1.3
http://support.sas.com/rnd/app/papers/bayesian.pdf

Therneau, T.M. and Grambsch, P.M. and Fleming, T.R. (1990), “Martingale-based residuals for survival models”, Biometrika 1990 77(1):147-160

Shtatland, E.S. and Kleinman, K. and Cain, E.M. (2005), “MODEL BUILDING IN PROC PHREG WITH AUTOMATIC VARIABLE SELECTION AND INFORMATION CRITERIA”, online proceedings paper, SAS Users Global Forum 2005
http://www2.sas.com/proceedings/sugi30/206-30.pdf

CONTACT INFORMATION

Your comments are greatly appreciated and encouraged. Contact the authors at:

Lida Gharibvand
University of California, Riverside
Work Phone: (949) 230-5439
Email: lida.gharibvand@email.ucr.edu

George Fernandez
University of Nevada, Reno
Work Phone: (775) 784-4206
Email: gcjf@unr.edu