

Applications of PROC PHREG Enhancements in Version 6.10

Ying So
SAS Institute Inc., Cary, NC
Email: sasycs@unx.sas.com

Abstract

There are two major enhancements to the PHREG procedure in Release 6.10. One is the counting process style of input, and the other is the availability of various residuals and diagnostic measures. The counting process formulation allows PROC PHREG to fit a superset of the Cox model, known as the multiplicative hazards model. This extension includes multiple events per subjects, left truncation on the observation of survival time, and time-dependent covariates. Residuals and regression diagnostics are useful in assessing the adequacy of the model and in identifying influential observations. This article also discusses two applications that involve robust sandwich variance estimates.

Introduction

The enhancements to PROC PHREG in Release 6.10 are documented in *SASSTAT® Software: Changes and Enhancements, Release 6.10*.

In the counting process formulation, data for each subject are identified by three processes N , Y , and Z , where N is the counting process, Y is the censoring process, and Z is the covariate process. Here, $N(t)$ indicates the number of events the subject experiences over the time interval $(0, t]$; $Y(t)$ indicates whether the subject is at risk at time t (1 if at risk and 0 otherwise); and $Z(t)$ is a vector of explanatory variables for the subject at time t . The sample path of N is a step function with jumps of size +1 at the event times, and $N(0) = 0$. Unless $Z(t)$ changes continuously with time, the data for each subject can be represented by multiple observations, each identifying a semi-closed time interval $(t1, t2]$, the values of the explanatory variables over that interval, and the event status at $t2$. The subject remains at risk during the interval $(t1, t2]$, and an event may occur at $t2$. Values of the explanatory variables for the subject remain unchanged in the interval. This counting process style of input is available by the specification

```
MODEL (t1,t2) <*censor(list)> = variables </options>;
```

This MODEL statement identifies a pair of failure time variables, an optional censoring variable with an optional list of censoring values, and the explanatory variables. Preceding the equal sign are the names of failure time variables that signify the beginning and end of a semi-closed interval $(t1, t2]$ during which the subject is at risk. The names are enclosed in a pair of parentheses and can be followed by an asterisk, the name of the censoring variable, and a list of values enclosed in parentheses. If the censoring variable takes on one of these values (separated by blanks and commas if there is more than one), the time $t2$ is considered to be censored. The variables following the equal sign are the explanatory variables for the model. Various applications of the counting process formulation are given in the next section.

After a model has been fitted to the data, the adequacy of the model needs to be assessed. This includes checking the assumption of the model (for example, the proportional hazards assumption), detecting whether certain explanatory variables need to be transformed, and identifying outliers and influential observations. Plots based on residuals and other regression diagnostics are useful for assessing model adequacy. PROC PHREG now provides Schoenfeld residuals and score residuals in addition to the martingale residuals and deviance residuals. You can also request various diagnostics for identifying influential observations.

Not only can score residuals be used to assess the leverage of each observation in the parameter estimation, they also play an important role in the construction of sandwich variance estimators. The Wei, Lin, and Weissfeld (1989) approach of analyzing multivariate failure requires the computation of such a sandwich variance estimator. The Barlow (1994) analysis of case-cohort design also involves robust sandwich variance estimates. To illustrate their methodologies, the AIDS virus data are analyzed as multivariate failure times and a set of nasal sinus cancer data are fitted to the case-cohort model.

Applications of the Counting Processing Formulation

Multiple Failure Outcomes for Each Subject

The multiplicative hazards model can be used in the analysis of repeated failure outcomes of the same kind. Consider the CGD study (Fleming and Harrington 1991) in which patients in one of the two treatment groups (placebo or gamma interferon) might experience recurrent infections. The data are saved in a SAS® data set named CGD, which contain the variables SUBJECT (patient sequence number), T1 (start of the risk interval), T2 (end of the risk interval), DELTA (1= infection occurred at T2, 2 otherwise), and Z1 (0= placebo, 1= gamma interferon). A partial listing of the data set is shown in the following output:

CGD Data					
OBS	SUBJECT	T1	T2	DELTA	TRT
1	1	0	293	2	1
2	2	0	255	2	0
3	3	0	213	2	0
4	4	0	203	2	0
5	5	0	219	1	1
6	5	220	373	1	1
7	5	374	414	2	1
8	6	0	8	1	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
200	126	0	146	1	0
201	126	147	316	2	0
202	127	0	316	2	0
203	128	0	315	2	1

Subject 1 was followed to day 293 with no infection. Subject 5 had two infections, one at day 219 and the other at day 373, and was followed to day 414. Subject 1 has only one observation, while the data for subject 5 are represented by three observations, each identifying an at-risk interval.

The following SAS statements invoke PROC PHREG to fit the multiplicative hazards model with a single treatment indicator variable (Z1). The RISKLIMIT option was specified to print the 95% confidence interval of the hazard ratio $\exp(\beta)$.

```
proc phreg data=cgd;
  model (t2,t1)*delta(2)=z1/risklimits;
run;
```

The printed output looks very much like that for the Cox model, except that both T1 and T2 are listed as dependent variables. Results of this analysis indicate that gamma interferon reduces the rate of infection in the CGD patients.

The PHREG Procedure					
Data Set: WORK.CGD					
Dependent Variable: T1					
Dependent Variable: T2					
Censoring Variable: DELTA					
Censoring Value(s): 2					
Ties Handling: BRESLOW					
Summary of the Number of Event and Censored Values					
Total	Event	Censored	Percent Censored		
203	76	127	62.56		
Testing Global Null Hypothesis: BRESLOW					
Criterion	Without Covariates	With Covariates	Model Chi-Square		
-2 LOG L	684.332	664.108	20.224 with 1 DF	(p=0.0001)	
Score	.	.	19.534 with 1 DF	(p=0.0001)	
Wald	.	.	17.707 with 1 DF	(p=0.0001)	
Analysis of Maximum Likelihood Estimates					
Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
TRT	1	-1.098564	0.26107	17.70708	0.0001
Conditional Risk Ratio and 95% Confidence Limits					
Variable		Risk Ratio	Lower	Upper	
TRT		0.333	0.200	0.556	

Time-Dependent Covariates

The Cox model is a special case of the multiplicative hazards model, where $Y_i(t) = 1$ until the first event or censoring, and $Y_i(t) = 0$ thereafter. The covariate process $Z(t)$ allows the covariates to be updated with time. One of the most discussed examples is the Stanford heart transplant data (Crowley and Hu 1977), which concern the survival of patients who did or did not undergo a heart transplant operation. To receive a transplant, a patient had to survive long enough for a suitable donor to be found, so that the occurrence and timing of the transplant had to be taken into account to avoid a time-to-transplant bias. The observed failure time of interest is the time a patient entered into the program until the patient died or was censored. The data are saved in a SAS data set named HEART. A partial listing of the data* is shown in the following output. The measurements are ID (patient's id), TIME (failure time), STATUS (censoring indicator, 1=dead, 0=censored), SURGERY (previous surgery, 1=yes, 0=no), AGE (patient's age at entry in years minus 48), and WAITTIME (time from entry to transplant).

* To avoid ambiguity in the ties between death times and transplant times, a patient who died on the day of transplant is assumed to die after the transplant, and the day's deaths are assumed to come before the day's transplants. This is accomplished by adding .2 to the value of TIME when TIME=WAITTIME and by adding .1 to all values of WAITTIME.

Data HEART: Stanford Heart Transplant Data					
ID	TIME	STATUS	SURGERY	AGE	WAITTIME
1	49	1	0	-17.1554	.
2	5	1	0	3.8357	.
3	15	1	0	6.2971	0.1
4	38	1	0	-7.7372	35.1
5	17	1	0	-27.2142	.
6	2	1	0	6.5681	.
7	674	1	0	2.8693	50.1
8	39	1	0	-2.6502	.
9	84	1	0	-0.8378	.
10	57	1	0	-5.4976	11.1

Suppose you want to fit a Cox model with explanatory variables AGE, SURGERY, XP (transplant status, 1=yes, 0=no), AGE_XP (AGE and XP interaction), and SUR_XP (SURGERY and XP interaction). XP is a time-dependent variable, since its value changes from 0 before the transplant to 1 after the transplant. The interaction terms AGE_XP and SUR_XP are also time-dependent. To use the counting process style of input, the data for a patient with a transplant are represented by two observations, one identifying the interval from entry to transplant and the other identifying the interval from transplant to death or censored. For a patient who never receives a transplant, the risk interval is from entry to death or censored. A new censoring variable DEAD is created from STATUS to indicate the event status at the end of the risk interval. XP, AGE_XP, and SUR_XP have a value of 0 before the transplant, and XP has a value of 1 after the transplant. The following SAS statements create the data set HEART2 from the data set HEART and display the data for the first 10 patients:

```
data heart2;
  keep id t1 t2 dead age surgery xp
      age_xp sur_xp;
  set heart;
  if (waittime = .) then do;
    t1=0;
    t2=time;
    dead=status;
    xp=0;
    age_xp=0;
    sur_xp=0;
    dead=status;
    output;
  end;
  else do;
    t1=0;
    t2=waittime;
    dead=0;
    xp=0;
    age_xp=0;
    sur_xp=0;
    output;
    t1=waittime;
    t2=time;
    dead=status;
    xp=1;
    age_xp=age;
    sur_xp=surgery;
    output;
  end;
run;
```

```
proc print data=heart2;
  where id <= 10;
  id id;
  title 'Data HEART2: Counting Process Style';
run;
```

Data HEART2: Counting Process Style							
ID	T1	T2	DEAD	SURGERY	AGE	XP	AGE_XP
1	0.0	49.0	1	0	-17.1554	0	0.00000
2	0.0	5.0	1	0	3.8357	0	0.00000
3	0.0	0.1	0	0	6.2971	0	0.00000
3	0.1	15.0	1	0	6.2971	1	6.29706
4	0.0	35.1	0	0	-7.7372	0	0.00000
4	35.1	38.0	1	0	-7.7372	1	-7.73717
5	0.0	17.0	1	0	-27.2142	0	0.00000
6	0.0	2.0	1	0	6.5681	0	0.00000
7	0.0	50.1	0	0	2.8693	0	0.00000
7	50.1	674.0	1	0	2.8693	1	2.86927
8	0.0	39.0	1	0	-2.6502	0	0.00000
9	0.0	84.0	1	0	-0.8378	0	0.00000
10	0.0	11.1	0	0	-5.4976	0	0.00000
10	11.1	57.0	1	0	-5.4976	1	-5.49760

The following SAS statements fit the time-dependent covariates model using the counting process style of input:

```
proc phreg data=heart2;
  model (t1,t2)*dead(0)= age surgery xp
                        age_xp sur_xp;
run;
```

Testing Global Null Hypothesis: BETA=0					
Criterion	Without Covariates	With Covariates	Model Chi-Square		
-2 LOG L	596.628	584.213	12.416 with 5 DF	(p=0.0295)	
Score	.	.	12.006 with 5 DF	(p=0.0347)	
Wald	.	.	11.612 with 5 DF	(p=0.0405)	

Analysis of Maximum Likelihood Estimates					
Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Risk Ratio
AGE	1	0.013850	0.01812	0.58430	1.014
SURGERY	1	-0.547464	0.61090	0.80311	0.578
STATUS	1	0.122391	0.32730	0.13983	1.130
AGE_XP	1	0.034449	0.02722	1.60137	1.035
SUR_XP	1	-0.289794	0.75818	0.14610	0.748

PROC PHREG can also handle time-dependent explanatory variables through the use of programming statements. The following invocation of PROC PHREG gives the same results:

```
proc phreg data= heart;
  model time*status(0)= age surgery xp
                        age_xp sur_xp;
  if (waittime = . or time < waittime)
    then do;
      xp=0;
      age_xp=0;
      sur_xp=0;
    end;
  else do;
    xp=1;
    age_xp=age;
    sur_xp=surgery;
  end;
```

```

end;
title 'Handling Time-Dependent Covariates '
      'via Programming Statements';
run;

```

These two approaches of fitting the Cox model with time-dependent explanatory variables are not equally efficient with the PHREG procedure. The counting process approach requires preprocessing the data into records of risk intervals, but the values of the time-dependent variables do not have to be recomputed for every subject in the risk set for each event time. The amount of savings in computing time could be fairly substantial, especially when there is a large number of follow-up observations for each time-dependent variable. On the other hand, the convenience of using programming statements could be attractive. Moreover, time-dependent variables that change continuously with time can be handled by programming statements, as in the application of testing the proportional hazards assumption (refer to the introductory example in the PROC PHREG documentation in SAS Technical Report P-229, *SAS/STAT Software: Changes and Enhancements, Release 6.07*).

Left Truncation on Observation of Survival Time

Left truncation arises when individuals come under observation only some known time after the natural origin of the phenomenon under study. For example, in an occupational health study of nurses using age as the time metameter, the left truncation time is the age at which an individual takes up the nursing profession; had that person died before becoming a nurse, he or she would not have been recorded. The risk set for an event time does not include individuals whose left truncation times exceed the given event time. Thus, any contribution to the likelihood must be conditional on the truncation limit having been exceeded. The counting process style of input lends itself to the left truncation setting by allowing the specification of an interval under which the individual is at risk. The data of an individual consists of one observation, with the left truncation time as the start of the risk interval and the observed time as the end of the risk interval.

For the nurse health study, let AGE_BIR be the age of the nurse and let AGE_ENT be the age* at which he or she became a nurse. Let STATUS represent the censoring status (for example, 1=dead, 2=alive) and let Z1-Z10 be explanatory variables thought to be related to survival. The following SAS statements invoke PROC PHREG to fit the Cox model:

```

proc phreg data=nurses;
  model (age_ent, age_bir) * status(2) = z1-z10;
run;

```

Although you can also use a time-dependent covariate to trick PROC PHREG into producing correct risk sets

* If you want to include the age at entry in the risk interval, you should subtract a small amount from the age at entry.

for the left truncation situation, this approach is not recommended because it is considerably less efficient. However, if you have access to SAS/STAT software only prior to Release 6.10, you can designate one of the exploratory variables at the time-dependent covariate, as follows:

```

proc phreg data=nurses;
  model age_bir * status(2) = z1-z10;
  if age_bir < age_ent then z1= .;
run;

```

Here, variable Z1 becomes a time-dependent explanatory variable. Individuals whose entry times are less than an event time are excluded from the risk set for that event time because they have missing values for the time-dependent variable.

Other Applications

In the counting process formulation, data for each individual are broken into distinct intervals of risk. These intervals are not necessarily contiguous, allowing the specification of disjoint intervals of risk. For example, in the nurse health study, an individual may leave his or her nursing job to raise a family and later come back when his or her children are older. During this absence, this individual is no longer at risk and therefore will not be in any risk sets for that period of time. The counting process style of input can conveniently accommodate this situation by having different observations representing the periods the individual is working as a nurse.

The counting process formulation can also be used to accommodate time-dependent strata. For example, in the Stanford heart transplant study, you may believe that the underlying hazard rate of a patient is no longer the same after receiving a transplant, and therefore patients should be included in a different stratum after receiving a transplant, since different strata correspond to different baseline hazard rates. You should carefully consider the appropriate origin of failure times after the transplant, but the counting process style of input enables you to split the data of patients into different strata easily.

Model Checking in the Proportional Hazards Models

Many model checking procedures are based on residuals and diagnostic statistics, the behavior of which are approximately known when the fitted model is satisfactory. You can output these statistics by giving names under appropriate keywords in the OUTPUT statement.

The martingale residual is the difference between the observed number of events for an individual and the corresponding estimated expected number based on the fitted model. Martingale residuals have an expected value of 0 and are uncorrelated asymptotically. This property, which is also possessed by residuals in linear regression analysis, makes plotting the martingale

residuals a useful model-checking procedure. However, the martingale residuals are not symmetric about 0 even when the fitted model is correct. The skewness makes the martingale residual plots difficult to interpret. Deviance residuals are a transform of the martingale residuals. Deviance residuals are distributed symmetrically about 0 when the fitted model is appropriate.

The Schoenfeld residual for an explanatory variable at an event time is the difference between the explanatory variable and a weighted average of the covariates over a risk set at the given event time. Under the proportional hazards assumption, the Schoenfeld residuals have the sample path of a random walk and are therefore useful in assessing time trend or lack of proportionality.

Score residuals are a decomposition of the first partial derivatives of the log likelihood. Score residuals are useful in assessing the influence of each individual on the parameter estimates. The DFBETA statistics are score residuals weighted by the model-based covariance matrix.

The LD and LMAX statistics are useful in assessing the influence of each individual on the overall fit of the model.

Residual plots can be used in graphical assessments of the adequacy of the linear predictors and the validity of the proportional hazards assumption. Typically, martingale residuals and deviance residuals are plotted against the failure times, the rank order of the failure times, and each of the explanatory variables in the models. Schoenfeld residuals and score residuals are plotted against the failure times or the rank order of failure times. Any trends or changes in the variability are taken as an indication of departure of proportionality assumption. In practice, these plots are difficult to interpret. Nevertheless, these residuals are useful for detecting outlying covariate values.

Multivariate Failure Time Data

While the multiplicative hazards model can only handle repeated failure outcomes of the sample kind, the marginal analysis approach of Wei, Lin, and Weissfeld (1989) can be applied to failures that may be repetitions of the same kind of event or may be events of different natures.

Consider the AIDS virus data listed in Wei, Lin, and Weissfeld (1989). Blood samples were collected at weeks 4, 8, and 12 from each patient in three treatment groups (placebo, low dose, and high dose of ribavirin). For each serum sample, the failure time is the number of days when virus positivity was detected, and it is censored when the culture required a longer period of time to register for virus positive than was achievable in a laboratory. Some failure times were missing because the patient did not make the scheduled visit or the sample is inadequate for laboratory analysis. Investigators were interested in whether the drug prolonged the time to virus positivity.

The Wei, Lin and Weissfeld analysis regards the three failure times for each individual as a multivariate failure time and models each component time with a separate Cox proportional hazards model. No specific correlation structure is assumed for the multivariate failure times. For $j = 1, 2, 3$, let β_j be the vector of regression coefficient for the j th marginal model; let $\hat{\beta}_j$ be the maximum likelihood estimate of β_j . The estimators $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are generally correlated, and their covariances are given by the following sandwich estimators:

$$\widehat{\text{cov}}(\hat{\beta}_i, \hat{\beta}_j) = A_i R_i' R_j A_j, 1 \leq i, j \leq 3$$

where A_j is the model-based covariance matrix of $\hat{\beta}_j$ obtained by inverting the information matrix for the j th marginal model, and R_j is the matrix of score residuals for the j th marginal model. Inference on the parameters is based on the asymptotic normality of the parameter estimators.

When using PROC PHREG to fit the marginal models, you can request the DFBETA statistics, which are precisely the columns of the product matrix $R_j A_j$. By saving the DFBETA statistics into a data set, you can subsequently use SAS/IML® software to compute the sandwich covariance matrix. The analysis of the AIDS virus data is illustrated below.

The AIDS virus data are saved in the data set VIRUS. The variables are ID (sequence number of the patients), TRT (treatment, 1=placebo, 2=low dose, 3=high dose), and WEEK4, WEEK8, and WEEK12 (representing the failure times for samples collected at weeks 4, 8, and 12, respectively). Censored times are indicated by a negative sign.

```
data virus;
  retain id 0;
  input trt week4 week8 week12 @@;
  id + 1;
  cards;
1 9 6 7 1 4 5 10 1 6 7 6 1 10 . -21
1 15 8 . 1 3 . 6 1 4 7 3 1 9 12 12
1 9 19 -19 1 6 5 6 1 9 . 18 1 9 -20 -17
2 6 4 5 2 16 17 -21 2 31 -19 -21 2 -27 -19 .
2 7 16 -23 2 -28 7 -19 2 -28 3 16 2 15 12 16
2 18 -21 22 2 8 4 7 2 4 -21 7 3 21 9 -8
3 13 7 -21 3 16 6 20 3 3 8 6 3 21 . -25
3 7 19 3 3 11 13 -21 3 -27 -18 9 3 14 14 6
3 8 11 15 3 8 4 7 3 8 3 9 3 -19 10 -17
;
run;
```

There are three marginal models, one for each of the three occasions that the samples were collected. Instead of fitting one model at a time, you can fit all three models in one invocation of PROC PHREG. The trick is to employ a stratified analysis in such a way that the partial likelihood for a stratum is identical to the partial likelihood for the marginal model that corresponds to the stratum. The following DATA step reorganizes the AIDS virus data into the data set VIRUS1. Three observations are created for each individual, one for each observed failure time. Variable SAMPLE indicates the

sample number with values 1, 2, and 3 corresponding to weeks 4, 8, and 12, respectively, DAYS represents the observed time, and POSITIVE is the censoring variable with value 0 indicating a censored time. Six variables (Z11, Z21, Z12, Z22, Z13, and Z23) are created to represent the treatments. For the first failure time, Z11 is the indicator for the low dose and Z21 is the indicator for the high dose, and Z12, Z22, Z13, and Z23 are null. For the second failure time, Z12 is the low dose indicator and Z22 is the high dose indicator, and Z11, Z21, Z13, and Z23 are null. For the third failure time, Z13 is the low dose indicator and Z23 is the high dose indicator, and Z11, Z21, Z12, and Z22 are null. If you perform a stratified Cox regression analysis using SAMPLE as the stratification variable, the regression coefficients associated with Z1j and Z2j are the coefficients for the low dose and high dose, respectively, for the jth marginal model, j = 1, 2, 3.

```
data virus1 (drop=tmp week4 week8 week12);
  array wk[3] week4 week8 week12;
  set virus;
  do sample=1 to 3;
    tmp= wk[sample];
    if tmp ^= . then do;
      if tmp < 0 then do;
        days= -tmp;
        positive= 0;
      end;
      else do;
        days= tmp;
        positive= 1;
      end;
      z11= (trt=1) * (sample=1);
      z12= (trt=1) * (sample=2);
      z13= (trt=1) * (sample=3);
      z21= (trt=2) * (sample=1);
      z22= (trt=2) * (sample=2);
      z23= (trt=2) * (sample=3);
      output;
    end;
  end;
run;
```

The following SAS statements invoke PROC PHREG to fit separate marginal models to the multivariate failure times. Corresponding to the explanatory variables Z11, Z21, Z12, Z22, Z13, and Z23, the DFBETA statistics DB11, DB21, DB12, DB22, DB13, and DB23 are saved in the OUTPUT data set OUT1. However, not all entries in DFBETA statistics are relevant. For instance, the relevant entries for DB11 and DB12 are those that correspond to SAMPLE=1 (that is, the first marginal model); all other entries are zero. A convenient way to get rid of the zero entries in the DFBETA statistics is to use PROC MEANS to sum up the DFBETA statistics for each individual.

```
proc phreg data=virus1 outest=est1;
  strata sample;
  model days*positive(0)=z11 z21 z12 z22 z13 z23;
  output out= out1
         dfbeta=db11 db21 db12 db22 db13 db23 /
         order=data;
  id id;
run;
proc means data=out1 noprint;
  by id;
```

```
var db11 db21 db12 db22 db13 db23;
output out=out2
      sum=db11 db21 db12 db22 db13 db23;
run;
```

The parameter estimates for the marginal models are contained in the data set EST1, and the DFBETA statistics for the marginal models are contained in the data set OUT2. SAS/IML software is used to compute the sandwich variance estimators. Linear hypotheses about the model parameters can be tested. As an illustration, the global null hypothesis that ribavirin does not affect the time to virus positivity is tested.

```
proc iml;
  use est1;
  read all var{z11 z21 z12 z22 z13 z23} into b;
  vname={"Z11","Z21","Z12","Z22","Z13","Z23"};
  use out2;
  read all var{db11 db21 db12 db22 db13 db23}
    into al;
  b= t(b);
  v= al' * al;
  se=sqrt(vecdiag(v));
  reset noname;
  estprt=b||se;
  stitle={"Estimate", " Std Error"};
  vname={"Z11","Z21","Z12","Z22","Z13","Z23"};
  print,estprt[colname=stitle rowname=vname
               format=10.4];
  print,"Estimated Covariance Matrix",,
    v[colname=vname rowname=vname
      format=10.4],,
  /* global null hypothesis */
  chisq= t(b) * inv(v) * b;
  df= nrow(b);
  p= 1- probchi( chisq, df);
  print ,,
    "Testing Global Null Hypothesis: BETA=0",,
    "Wald Chi-Square = " chisq, "DF = " df,
    "p-value = "p[format=5.4],,
quit;
```

AIDS Virus Data		
	Estimate	Std Error
Z11	-1.3939	0.4953
Z21	-0.9383	0.3686
Z12	-0.6552	0.5354
Z22	0.0200	0.4087
Z13	-0.6151	0.5070
Z23	-0.3310	0.4790

Estimated Covariance Matrix						
	Z11	Z21	Z12	Z22	Z13	Z23
Z11	0.2454	0.0753	0.0507	0.0168	0.1073	0.0607
Z21	0.0753	0.1359	0.0265	0.0486	0.0458	0.0911
Z12	0.0507	0.0265	0.2867	0.1187	0.1333	0.0907
Z22	0.0168	0.0486	0.1187	0.1670	0.0763	0.0686
Z13	0.1073	0.0458	0.1333	0.0763	0.2570	0.1142
Z23	0.0607	0.0911	0.0907	0.0686	0.1142	0.2294

Testing Global Null Hypothesis: BETA=0	
Wald Chi-Square =	13.058034
DF =	6
p-value =	.0421

The drug appeared to have an effect (p=.0421) on the time to register for virus positive. Other inferences

about the model parameters that are of interest include testing whether the effects of the drug are the same across samples and estimating the common effect for the low dose and for the high dose (refer to the SAS program phrwls.sas in the SAS/STAT sample library).

Fitting Case-Cohort Design

When a cohort of individuals is followed in order to relate failure rates with covariate histories, the assembly of the covariate histories for individual cohort members requires a tremendous amount of resources and efforts. For a low rate of disease occurrence, a large cohort is needed to give meaningful results, and the cost would be prohibitively expensive. The case-cohort design proposed by Prentice (1986) requires covariate histories only for the cases (those experienced failure) and for members in a randomly selected subcohort. All members of the cohort are followed until failure or censoring, but only data for the cases and subcohort members are used in the analysis. The concentrated effort of the case-cohort design reduces cost and improves data quality.

In the analysis of a case-cohort study, the risk set at each event time consists of the case that fails at that particular event time and members of the subcohort who were at risk at the event time. A case outside the subcohort can only be in the risk set of his event time. By explicitly constructing the risk sets for the event times, you can use PROC PHREG to obtain the correct parameter estimates, but the standard error estimates are incorrect and will tend to underestimate the true standard errors. Prentice (1976) has pointed out that the components of the total scores for different event times are often correlated. He has provided a closed form expression for the covariance matrix that corrects for the correlated score contributions. However, the expression is difficult to evaluate and will not be pursued here.

Barlow (1994) proposes a robust variance estimator that allows for more complicated sampling mechanisms. The technique uses a jackknife approach of the variance of individual influence function. The approach used by Prentice (1978) is concerned with the contribution of the score function of each individual summed over the events times, while the approach of Barlow (1994) is concerned with the contributions at each event time summed over individuals.

As an illustration of Barlow's method, consider the cohort of the South Wales Nickel Worker study (Breslow and Day 1987, pp. 375-390). The full cohort consists of 679 workers, and the event of interest is the death from nasal sinus cancer (ICD=160). A 10% subcohort is randomly chosen. The data for the cases and the subcohort members are saved in the data set NICKEL (see the Appendix).

The data set NICKEL contains seven variables: ID (identification number), EXPOSURE (exposure level) YR.EMP (year employed), AGE.STRT (age at first

employment), AGE.STOP (age at death or withdrawal), NASAL (1=nasal sinus cancer, 0=otherwise), SUBCOHRT (1=subcohort member, 0=otherwise). There are 68 individuals in the subcohort. Fifty-six workers died from nasal sinus cancer, 9 of them are in the chosen subcohort.

Barlow (1994) has considered weights that reflect the subcohort membership. Controls in the subcohort are weighted inversely proportional to the sampling fraction. For a cohort of 10%, the weight is 10. Cases have a weight of 1 at the instant the individual fails regardless of subcohort membership. Cases inside the subcohort are controls until they become a case, and at such instant the weight changes. The analysis requires a staggered entry into the cohort since the workers started their employment at different times. Although PROC PHREG allows a staggered entry with the use of the counting process style of input, with all the given restrictions, it is simpler to construct the risk sets explicitly rather than tricking PROC PHREG into forming appropriate risk sets. The following SAS statements set up the appropriate risk sets for the event times:

```
proc sort data=nickel;
  by age_stop;
run;
data one (drop=evnttime);
  retain strat 0;
  set nickel;
  if nasal=1 then do;
    strat + 1;
    evnttime= age_stop;
    t=1;
    wt=1;
    output;
  do i=_n+1 to n;
    set nickel point=i nobs=n;
    if (age_strt lt evnttime le age_stop)
      and subcohort=1 then do;
      t=2;
      wt=10; /* 10% subcohort */
      output;
    end;
  end;
end;
run;
```

The data set ONE contains three additional variables: STRAT, T, and WT. The values of STRAT represent the distinct case-control sets. T has a value of 1 for a case and a value of 2 for a control. WT represents the weight of the observations.

In the following SAS statements, PROC PHREG is invoked to fit the case-cohort model with YR.EMP and EXPOSURE as the explanatory variables. STRAT is used as a stratifying variable, T is the failure time variable, and the control observations (T=2) are regarded as censored. WT is used as a FREQ variable. The parameter estimates are output to the data set EST1, and the DFBETA statistics are output to the data set OUT1. The DFBETA statistics are useful for computing the robust variance estimates for the parameter estimates.

```
proc phreg data=one outest=est1;
  model t*t(2)=yr_emp exposure;
  freq wt;
  strata strat;
  output out=out1 dfbeta=dfb1 dfb2;
  id id;
run;
```

Results of the PROC PHREG analysis are shown in the following table. Although the parameters estimates are correct, the standard error estimates may underestimate the true standard errors.

Analysis of Maximum Likelihood Estimates					
Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
YR_EMP	1	-0.033956	0.02667	1.62096	0.2030
EXPOSURE	1	0.204011	0.02871	50.49008	0.0001

The next step is to compute the robust variance estimates. Since observations in the data set OUT1 correspond to the contributions at each failure time, the DFBETA statistics are weighted by WT and then summed over individuals.

```
data two;
  array a dfb1 dfb2;
  set out1;
  do over a;
    a=a*wt;
  end;
run;
proc summary data=two nway;
  class id;
  var dfb1 dfb2;
  output out=three sum=dfb1 dfb2;
run;
```

The sandwich estimate of the covariance matrix is computed as the sum of the cross-product matrix of the DFBETA vectors for the individuals. This can be carried out easily by SAS/IML software. Also computed are the Wald tests for individual variables based on the robust variance estimates.

```
proc iml;
  use est1;
  read all var{yr_emp exposure} into e;
  b= e[1,];
  b= t(b);

  use three;
  read all var{dfb1 dfb2} into dfb;
  v2= t(dfb) * dfb;
  sd2= sqrt(vecdiag(v2));

  t= b / sd2;
  p= 1- probchi(t # t, 1);
  cname={"Estimate", "Std Err", "t-statistic",
    "p-value"};
  rname={"AGE_EMP", "EXPOSURE"};
  variable= b || sd2 || t || p;

  print variable[colname=cname rowname=rname
    format=10.5];
quit;
```

Results of the analysis based on the robust variance estimates are given in the following table.

Analysis based on Robust Variance Estimates				
VARIABLE	Estimate	Std Err	t-statistic	p-value
YR_EMP	-0.03396	0.03558	-0.95437	0.33989
EXPOSURE	0.20401	0.05654	3.60841	0.00031

Note that the robust standard error estimates are larger than the naive standard error estimates obtained directly from PROC PHREG (.03558 vs 0.02667 for YR_EMP, .05654 vs .02871 for EXPOSURE).

Appendix

A 10% subcohort was randomly chosen from the full cohort of the South Wales Nickel Worker study (Breslow and Day 1987). The data for the cases and the subcohort members are saved in the data set NICKEL.

```
data nickel;
  input id exposure yr_emp age_strt age_stop
    nasal subcohort;
  cards;
    8 9.0 1911 47.907 69.679 0 1
    16 0.5 1924 59.915 65.583 1 0
    19 10.0 1915 52.520 58.490 1 0
    35 0.5 1909 47.852 82.827 0 1
    42 0.0 1923 34.748 83.049 0 1
    47 1.0 1911 45.934 47.114 1 0
    52 0.5 1915 39.499 86.688 0 1
    53 1.0 1923 33.536 72.761 1 0
    54 11.0 1914 44.849 77.808 1 0
    55 0.0 1923 33.260 81.014 0 1
    56 2.0 1922 29.746 51.585 0 1
    63 14.0 1908 56.216 58.065 1 0
    64 6.0 1907 45.746 64.812 0 1
    75 0.0 1922 28.838 76.592 0 1
    106 11.0 1908 53.935 58.072 1 0
    107 0.0 1923 36.592 79.847 0 1
    117 1.0 1916 42.594 66.337 0 1
    126 18.0 1908 50.746 57.279 0 1
    142 7.5 1910 50.331 57.449 1 0
    143 3.0 1920 46.746 77.303 0 1
    144 1.5 1913 49.741 67.229 1 0
    150 9.5 1910 62.315 64.732 1 0
    151 10.5 1908 58.882 77.745 1 0
    154 2.5 1922 44.096 70.000 0 1
    156 12.5 1914 46.746 58.837 0 1
    180 3.0 1913 41.746 48.117 0 1
    208 1.5 1913 47.219 67.211 0 1
    222 3.5 1913 58.162 70.910 1 0
    224 0.0 1917 30.746 37.645 0 1
    256 2.0 1920 53.746 68.235 0 1
    271 9.0 1915 49.727 64.990 1 0
    286 5.0 1910 47.746 63.832 0 1
    294 5.5 1920 48.948 62.096 1 0
    298 12.0 1911 48.425 54.700 1 0
    299 0.0 1921 35.746 68.653 1 0
    306 0.0 1922 36.746 37.300 0 1
    319 0.0 1922 33.509 66.320 1 0
    330 0.0 1910 52.668 70.079 1 1
    353 0.0 1924 32.746 48.763 0 1
    360 0.0 1913 50.340 86.241 0 1
    363 10.0 1910 50.825 57.789 1 0
    378 3.5 1913 40.746 63.974 1 0
```

381 13.0 1914 49.651 78.087 1 1
 386 5.0 1917 32.573 78.375 0 1
 392 15.0 1912 43.227 57.585 1 0
 395 3.0 1919 38.083 70.806 1 0
 400 0.0 1922 30.877 77.067 0 1
 408 1.0 1908 51.746 74.481 0 1
 409 2.0 1912 59.746 70.807 0 1
 410 0.0 1923 33.746 51.661 0 1
 419 0.0 1922 31.707 70.808 0 1
 421 1.0 1914 49.746 69.782 1 0
 423 0.0 1906 59.746 74.522 0 1
 427 3.5 1915 45.910 67.417 0 1
 436 0.0 1922 34.868 73.351 0 1
 446 0.0 1923 46.175 65.017 1 0
 453 0.0 1923 41.744 64.183 1 0
 478 0.0 1924 38.909 73.926 0 1
 499 9.0 1910 44.746 67.003 0 1
 506 6.0 1911 44.786 67.523 1 1
 518 5.5 1903 53.599 80.361 1 0
 522 6.0 1906 48.746 63.276 0 1
 525 0.0 1920 34.934 65.439 0 1
 529 1.0 1923 32.611 50.293 1 0
 538 4.5 1912 39.000 59.129 0 1
 545 0.5 1920 33.995 71.436 0 1
 553 0.0 1923 37.637 70.495 0 1
 566 2.5 1912 57.645 66.448 1 0
 578 0.0 1910 50.252 74.504 1 0
 593 7.5 1914 51.808 61.821 1 0
 595 8.0 1911 49.309 79.560 1 0
 599 0.0 1916 48.315 81.403 1 0
 608 3.0 1922 33.744 72.634 0 1
 609 3.0 1915 57.907 70.455 1 0
 618 9.0 1913 51.746 53.735 0 1
 622 4.5 1905 50.746 71.555 0 1
 629 0.0 1911 39.235 75.016 0 1
 639 16.0 1909 52.334 78.145 1 0
 658 1.0 1918 53.987 68.503 1 0
 674 12.5 1904 61.569 63.065 1 0
 675 3.0 1922 35.746 53.831 0 1
 687 4.0 1914 49.077 78.433 1 1
 706 6.0 1913 59.805 74.937 1 0
 717 11.5 1908 58.400 59.101 1 1
 719 8.0 1915 53.509 56.030 1 0
 729 2.0 1922 41.746 70.333 1 1
 742 3.0 1903 51.746 57.360 0 1
 768 13.0 1907 52.583 59.548 1 0
 772 0.0 1922 38.746 55.275 0 1
 776 0.0 1912 57.746 66.881 1 0
 780 0.0 1923 47.441 56.795 1 0
 781 6.0 1915 58.457 69.926 1 0
 790 3.0 1914 49.621 59.341 1 0
 802 0.0 1903 57.746 86.094 0 1
 811 6.0 1909 51.835 63.367 1 0
 815 0.0 1916 39.746 53.174 0 1
 818 4.0 1909 44.657 71.521 1 1
 827 12.0 1911 59.268 64.510 1 1
 833 1.0 1915 57.637 62.443 1 0
 835 0.5 1919 36.926 54.868 0 1
 836 6.5 1912 54.746 68.004 0 1
 839 6.5 1914 45.246 62.334 1 0
 841 3.0 1903 60.746 66.607 0 1
 846 2.0 1923 38.320 77.066 0 1
 861 0.0 1923 35.937 67.860 0 1
 864 0.0 1923 38.715 80.444 0 1
 875 5.0 1914 60.279 76.501 1 1
 888 2.0 1913 43.746 67.363 0 1
 919 0.0 1913 42.091 82.159 0 1
 931 0.0 1916 64.746 70.292 0 1
 942 0.0 1906 50.039 51.461 1 0
 945 4.5 1913 70.746 83.289 0 1
 963 3.5 1914 47.746 51.695 0 1

964 3.0 1923 41.746 73.640 0 1
 969 14.0 1910 52.746 57.686 1 0
 ;

Reference

- Barlow, W.E. (1994), "Robust variance estimation of case-cohort design," *Biometrics*, 50, 1064-1072.
- Breslow, N.E. and Day, N.E. (1987), *Statistical Methods in Cancer Research*, Vol. II - The Design and Analysis of Cohort Studies, Lyon: International Agency for Research on Cancer.
- Crowley, J. and Hu, M. (1977), "Covariance analysis of heart transplant survival data," *Journal of American Statistical Association*, 72, 27-36.
- Fleming, T.R. and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*, New York: John Wiley & Sons, Inc.
- Prentice, R.L. (1986), "A case-cohort design in epidemiologic cohort studies and disease prevention trials," *Biometrika*, 73, 1-11.
- SAS Institute Inc. (1992), SAS Technical Report P-229, *SAS/STAT Software: Changes and Enhancements, Release 6.07*, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1994), *SAS/STAT Software: Changes and Enhancements, Release 6.10*, Cary, NC: SAS Institute Inc.
- Wei, L.J., Lin, D.Y., and Weissfeld, L. (1989), "Regression analysis of multivariate incomplete failure time data by modeling marginal distribution," *Journal of the American Statistical Association*, 84, 1065-1073.

SAS, SAS/IML, and SAS/STAT are trademarks or registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.