Paper 273-28
Analysis of Data from Recurrent Events

Gordon Johnston and Ying So
SAS Institute Inc.
Cary, North Carolina, USA

Abstract

Time-to-event data have long been important in many applied fields. Many models and analysis methods have been developed for this type of data, in which each sample unit experiences at most a single end-of-life event.

In contrast, many applications involve repeated events, where a subject or sample unit may experience any number of events over a lifetime. There is a growing interest in the analysis of recurrent events data, also called repeated events data and recurrence data. This type of data arises in many fields. For example, the repair history of manufactured items can be modeled as recurrent events. In medical studies, the times of recurrent disease episodes in patients can also be modeled as recurrent events.

This paper describes methods for the analysis of recurrent events data. Nonparametric methods involving extensive use of graphics for the analysis of such data are discussed in a new book by Nelson (2003). These methods are illustrated using the SAS/QC® RELIABILITY procedure.

The use of the SAS/STAT® GENMOD and PHREG procedures to fit regression models to recurrent events data is also illustrated.

Examples are presented from the fields of medical studies and product reliability.

Introduction

Recurrence data consists of the times to any number of repeated events for each sample unit, for example, times of recurrent episodes of a disease in patients or times of repair of a manufactured product. The sample units are considered to be statistically independent, but the times between events within a sample unit are not necessarily independent nor identically distributed. The data are usually censored in the sense that sample units have different ends of histories. The *Mean Cumulative Function* (MCF) for the number (or associated cost) of events contains the information usually of interest in the analysis of recurrence data. Denoted $M(t)$, it is the population mean cumulative number (or cost) of events up to time t.

Nelson (2003) describes a simple nonparametric estimator of $M(t)$, denoted here $\hat{M}(t)$, analogous to the Nelson-Aalen estimator for the cumulative hazard function of lifetime data. Plots of $\hat{M}(t)$ and confidence limits versus t yield information such as the number of events (e.g., repairs or disease episodes) expected by time t, whether the rate of occurrence of events is increasing, decreasing, or constant, and whether two groups differ significantly in expected number of events. Additional insights may be gained into any anomalies in the data through MCF plots. You can compute and plot $\hat{M}(t)$ and confidence limits using the RELIABILITY procedure, for a single group or for different groups to assess the differences between the groups.

For illustration, the Chronic Granulomatous Disease (CGD) data described in Fleming and Harrington (1991) are considered. The data are from a clinical trial involving recurrent infections in patients with CGD. Several covariates were recorded, but the focus of the study is treatment with gamma interferon vs. placebo. Data on replacement of defrost controls in refrigerators from Nelson (2003) are also considered.

Lawless and Nadeau (1995) and Lin et al. (2000) describe *proportional rates* and *proportional means* regression models for recurrence data. These are semiparametric models analogous to the proportional hazards model for lifetime data. The proportional means model is, for each observation,

$$M(t) = M_0(t)\exp(X'\beta)$$

where X' is a vector of time invariant covariates and $M_0(t)$ is a baseline MCF. The proportional rates model is a similar model that allows for time dependent covariates, and is not discussed further here. Correct covariances for the estimators of the regression parameters $\hat{\beta}$, accounting for the dependence structure of the recurrence times, can be computed using a robust (or sandwich) estimator of the covariances. PROC PHREG can compute both the regression estimates $\hat{\beta}$ and their robust covariance estimates for both the proportional rates and means models. For illustration, PHREG is used to fit a regression model for gamma interferon vs. placebo for the CGD data. PHREG is also used to perform an analysis similar to the Lawless and Nadeau (1995) regression model analysis for the Nelson (1995) diesel locomotive engine valve seat repair data.

In the case where the data can be modeled as a Poisson process, the GENMOD procedure can be used to estimate parameters in a Poisson regression model describing the recurrence data. The Poisson model is illustrated with a regression model for the CGD data for gamma interferon vs. placebo.

Nonparametric Estimation of $M(t)$

CGD Data

Of the 128 patients in the CGD study, 65 were randomized into the placebo group and 63 into the gamma interferon group. The MCF plots of the gamma interferon and placebo groups in Figure 1 were created by the following SAS statements using PROC RELIABILITY.

```
proc reliability data = CGD;
   unitid Id;
   mcfplot Tstop*Status(0)=Treatment /
      overlay
      font=Arial
      vaxis = 0 to 2.5 by .5;
      i=step
      noconf;
   run;
```

Here, the variable Tstop in the input data set CGD represents the time in days of each infection if the variable Status is equal to one, or the end-of-history if Status is equal to zero. The variable Id identifies individual patients. Treatment identifies the two groups, either placebo or gamma interferon.

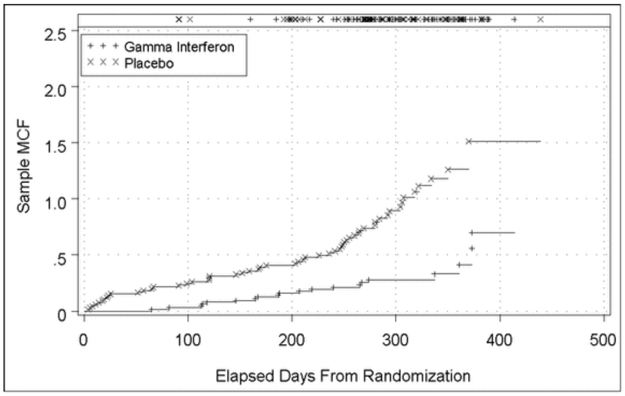

Figure 1. MCFs for Placebo and Gamma Interferon Groups

You can see clearly from the plots the higher infection rate of the placebo group. The points plotted in the strip along the top of the figure are the values of the end-of-history times. Figure 2 shows a plot of the difference between the gamma interferon and placebo MCFs created using PROC RELIABILITY. ALso, 95% pointwise confidence limits are plotted to help assess the significance of the difference between the two groups. Since the confidence limits do not enclose zero at later times in the study, there is evidence of a significant difference.

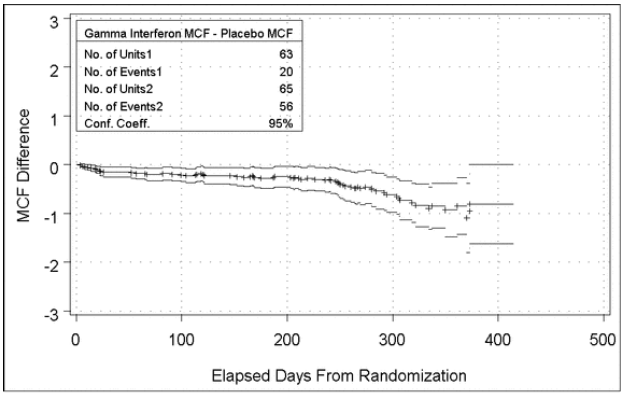

Figure 2. MCF Differences for Placebo and Gamma Interferon Groups

You can plot the MCFs using log scales to help assess the suitability of the proportional means model with an indicator variable for the two groups, perhaps leading to further analysis with the type of regression models that PROC PHREG can fit. If the proportional means model is appropriate, MCFs for the two groups will have the same shape and be approximately parallel. Figure 3 shows the MCFs plotted using a log scale for both axes. We use PROC PHREG to fit a proportional means model in the next section.

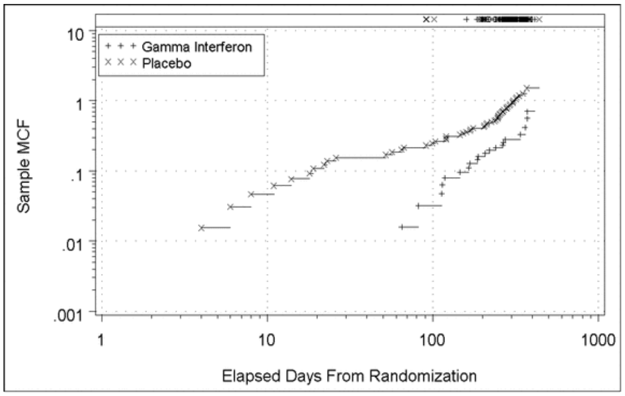

Figure 3. MCFs on Log-Log Scale

Appliance Defrost Control Data

The data in the preceding example contains exact times of recurrence of disease episodes (to the nearest day). In this example, the age of refrigerators at the time of replacement of defrost controls under warranty is grouped into intervals. The data, from Nelson (2003), contains data on 22,914 units, grouped into 29 one-month intervals. Figure 4 is an MCF plot of the data.

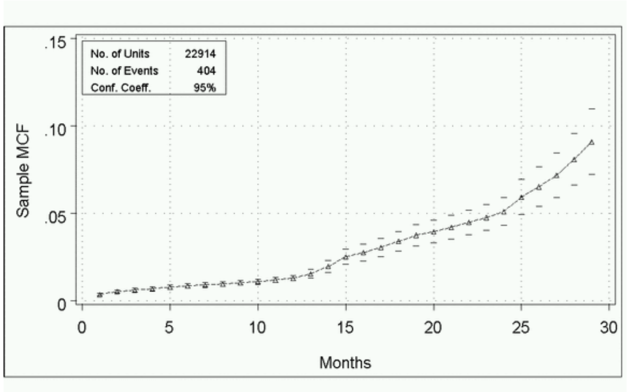

Figure 4. MCF for the Defrost Control Data

The data are in a SAS data set called DEFROST, and are shown in Table 7. The variables Lower and Upper are the lower and upper endpoints of the one-month intervals, Recurrences is the number of replacements in a monthly interval, and Censored is the number censored (removed from the study) during a monthly interval.

Table 1. Defrost Control Data

Obs	Lower	Upper	Recurrences	Censored
1	0	1	83	0
2	1	2	35	0
3	2	3	23	0
4	3	4	15	0
5	4	5	22	0
6	5	6	16	3
7	6	7	13	36
8	7	8	12	24
9	8	9	15	29
10	9	10	15	37
11	10	11	24	40
12	11	12	12	20041
13	12	13	7	14
14	13	14	11	17
15	14	15	15	13
16	15	16	6	28
17	16	17	8	22
18	17	18	9	27
19	18	19	9	64
20	19	20	5	94
21	20	21	6	119
22	21	22	6	118
23	22	23	6	138
24	23	24	5	1188
25	24	25	7	17
26	25	26	5	28
27	26	27	5	99
28	27	28	6	128
29	28	29	3	590

The following SAS statements use PROC RELIABILITY to compute and plot the MCF shown in Figure 4.

```
proc reliability data=DEFROST;
   mcfplot ( interval    = Lower Upper
             recurrences = Recurrences
             censor      = Censored ) / plotsymbol = triangle
                                        i=join lfit=3
                                        vaxis = 0 to .15 by .05
                                        inborder
                                        height = 5 inheight = 5
                                        font = Arial;
   inset / height = 3.5;
run;
```

Examination of the plot in Figure 4 yields useful insight into the nature of the population. The slope of the MCF, which is the failure rate, increases noticeably at 12 and 24 months, with fairly straight lines in between. Further study reveals that the original warranty expired at 12 months, with an option to purchase an extended warranty. The units for which extended warranties were purchased tended to be more trouble prone, hence with a higher failure rate. Similarly, another extended warranty option was available at 24 months. The MCF between 0 and 12 months is more representative of the overall population failure rate.

Proportional Means Model for $M(t)$

Regression Model for the CGD Data

You can fit the proportional means model using PROC PHREG. PROC PHREG computes the regression parameter estimates by maximizing a partial likelihood.

The baseline MCF $M_0(t)$ is computed using the nonparametric estimate described previously. If no regression parameters are specified, the $M_0(t)$ PHREG computes is identical to the MCF PROC RELIABILITY computes.

For example, the following SAS statements fit a regression model for the CGD data, with the indicator variable X as a covariate, $X = 1$: gamma interferon, $X = 2$: placebo.

```
data IN2;
   Treatment=1;
   output;
   Treatment=2;
   output;
   run;

proc phreg data=CGD covs(aggregate) covm;
  model (Tstart,Tstop)*Status(0)=Treatment;
  baseline covariates=IN2 out=OUT2 cmf=_all_ / nomean;
  id Id;
  run;
```

For recurrent events data, the counting process style of input is required. For instance, a unit which has three recurrences and a censoring time has four observations, each observation has a start time, a stop time, and an indicator of whether the stop time is a recurrent event time or a censored time. The start time is 0 or the last recurrence time.

Here, Tstart and Tstop are variables in the input data set CGD that specify the times (after randomization into groups) of infection onset or an end-of-history. Tstop is the time at which the current infection occurred for a patient, and Tstart is the time of the previous infection. For the first infection for a patient, Tstart=0. Status=0 identifies Tstop as the censoring, or end-of-history time. The BASELINE statement specifies the output data set OUT2 containing the baseline mean function $M_0(t)$ and the MCFs at values of the covariates in the input data set IN2. IN2 contains values for the gamma interferon and placebo covariate.

The plot of the MCFs for the treatment and control groups shown in Figure 5 was created using the MCF estimates in the output data set OUT2 and ODS graphics. The SAS statements used to create the plot is shown in Appendix A.

Table 2 shows the PHREG output table for testing the hypothesis $\beta = 0$. In this case, this is a test for the treatment effect. Both the score test and the Wald test using the robust sandwich estimates are highly significant. The resulting parameter estimates with robust standard errors are also shown in Table 2.

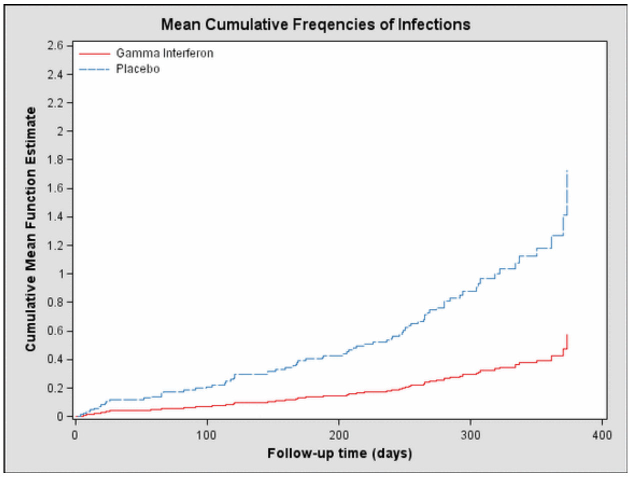

Figure 5. Regression Model for Treatment

Table 2. Regression Model for Treatment

```
                        The PHREG Procedure

             Testing Global Null Hypothesis: BETA=0

Test                        Chi-Square       DF      Pr > ChiSq

Likelihood Ratio               20.1671        1          <.0001
Score  (Model-Based)           19.4760        1          <.0001
Score  (Sandwich)              10.2372        1          0.0014
Wald (Model-Based)             17.6590        1          <.0001
Wald (Sandwich)                12.4313        1          0.0004

           Analysis of Maximum Likelihood Estimates
                with Sandwich Variance Estimate

                         Parameter       Standard      StdErr
 Variable       DF        Estimate          Error       Ratio

 treatment       1         1.09708        0.31116       1.192

                                    Hazard
Chi-Square      Pr > ChiSq           Ratio

   12.4313          0.0004           2.995
```

You can include other covariates and plot MCFs for specific values of the covariates. The following statements fit a regression model for treatment and age and create the output data set OUT3 for the plot in Figure 6 of MCFs for treatment group, age 30, and placebo group, age 1.

```
data IN3;
   Treatment=1; Age=30;
   output;
   Treatment=2; Age=1;
   output;
   run;

proc phreg data=CGD covs(aggregate) covm;
  model (Tstart,Tstop)*Status(0)=Treatment Age;
  baseline covariates=IN3 out=OUT3 cmf=_all_ / nomean;
  id Id;
  run;
```

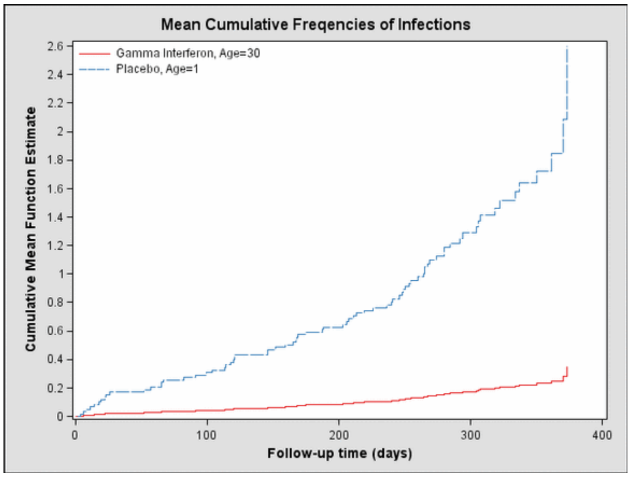

Figure 6. Regression Model for Treatment and Age

Regression Model for the Valve Seat Data

This example illustrates how to check an underlying assumption in recurrence data analysis, that the end-of-history times are independent of the event processes. Lawless and Nadeau (1995) proposed to check the assumption by fitting a proportional means model with end-of history as the covariate, and checking that the resulting estimates of $M(t)$ do not change with end-of-history. The data for the example is the Nelson (1995) valve seat data, shown below.

```
data valve;
   input Id Days Value@@;
   cards;
251 761 -1   252 759 -1   327  98 1
327 667 -1   328 326 1    328 653 1
328 653 1    328 667 -1   329 665 -1
330  84 1    330 667 -1   331 87 1
331 663 -1   389 646 1    389 653 -1
390  92 1    390 653 -1   391 651 -1
392 258 1    392 328 1    392 377 1
392 621 1    392 650 -1   393  61 1
393 539 1    393 648 -1   394 254 1
394 276 1    394 298 1    394 640 1
394 644 -1   395 76 1     395 538 1
395 642 -1   396 635 1    396 641 -1
397 349 1    397 404 1    397 561 1
397 649 -1   398 631 -1   399 596 -1
400 120 1    400 479 1    400 614 -1
401 323 1    401 449 1    401 582 -1
402 139 1    402 139 1    402 589 -1
403 593 -1   404 573 1    404 589 -1
405 165 1    405 408 1    405 604 1
405 606 -1   406 249 1    406 594 -1
407 344 1    407 497 1    407 613 -1
408 265 1    408 586 1    408 595 -1
409 166 1    409 206 1    409 348 1
409 389 -1   410 601 -1   411 410 1
411 581 1    411 601 -1   412 611 -1
413 608 -1   414 587 -1   415 367 1
415 603 -1   416 202 1    416 563 1
416 570 1    416 585 -1   417 587 -1
418 578 -1   419 578 -1   420 586 -1
421 585 -1   422 582 -1
;
```

Here, Id identifies the particular unit, Days is the time in days of valve seat replacement, and Value indicates a replacement (1) or end-of-history (-1).

Let τ_i denote the end-of-history time for the unit associated with observation i. Two models of the form $M_i(t) = M_0(t)\exp(x_i'\beta)$ were considered:

- $x_i = 0$ if $\tau_i <= 610$, and $x_i = 1$ if $\tau_i > 610$
- $x_i = \tau_i$

In both cases, $\beta = 0$ if if the τ_i's are independent of the event processes.

In order for PROC PHREG to handle more than one recurrence at a particular time within a unit, identical recurrence times within a subject have to be perturbed so that they appear nonidentical. In this example, the perturbation is carried out by subtracting a multiple of 10^{-10}. For instance, if there are three identical times, say t, within a unit, two of them are perturbed to become $t - 10^{-10}$ and $t - 2 * 10^{-10}$. The following PROC SORT and DATA step perturb the identical recurrence times. Also, two variables X1 and X2 are created to be used later in regression analysis. Variable X1 is the indicator variable of whether the censoring time of a unit is greater than 610. Variable X2 is the censoring time of the unit.

```
proc sort data=VALVE out=VALVE2;
   by Id Descending Days Value;
   run;
data VALVE2(drop=Lastdays Lastvalue Tau);
   retain Tau;
   set VALVE2;
   by Id;
   Lastdays=lag1(Days);
   Lastvalue=lag(Value);
   if first.Id and Value=-1 then Tau=Days;
   else do;
      if Days>=Lastdays and Value=1
      then Days= Lastdays - 1e-10;
   end;
   X1= (Tau > 610);
   X2= Tau;
   run;
```

For recurrent events data, the counting process style of input described in the previous example is required. The following PROC SORT and DATA step create such data.

```
proc sort data=VALVE2;
   by Id Days Descending Value;
   run;

data VALVE2(drop=Tmp);
  retain Tmp;
  set VALVE2;
  by Id;
  if first.Id then Tstart=0;
  else Tstart=Tmp;
  Tmp=Days;
  run;
```

The following SAS statements fit the two models

- regressing on X1 (indicator that censoring time of a unit is greater or equal to 610)
- regressing on X2 (the censoring time of the unit)

```
proc phreg data=VALVE2 covs(aggregate);
  model (Tstart,Days)*Value(-1)=X1 /ties=breslow;
  id Id;
  title
   "X1 indicates whether the censoring is larger than 610";
  run;

proc phreg data=VALVE2 covs(aggregate);
  model (Tstart,Days)*Value(-1)=X2 /ties=breslow;
  title "X2 is the censoring time of the unit";
  id Id;
  run;
```

Table 3 shows the table of tests of the hypothesis $\beta = 0$ and the parameter estimates for the first model, and Table 4 for the second model.

Table 3. Model for X1

```
                      The PHREG Procedure

             Testing Global Null Hypothesis: BETA=0

Test                        Chi-Square       DF     Pr > ChiSq

Likelihood Ratio                0.0629        1         0.8020
Score (Model-Based)             0.0629        1         0.8020
Score (Sandwich)                0.0459        1         0.8303
Wald (Model-Based)              0.0628        1         0.8021
Wald (Sandwich)                 0.0456        1         0.8309

            Analysis of Maximum Likelihood Estimates

                       Parameter       Standard      StdErr
 Variable     DF        Estimate          Error       Ratio

 x1            1         0.07806        0.36556       1.174

                                 Hazard
Chi-Square    Pr > ChiSq          Ratio

    0.0456        0.8309          1.081
```

Table 4. Model for X2

```
                      The PHREG Procedure

             Testing Global Null Hypothesis: BETA=0

Test                        Chi-Square       DF     Pr > ChiSq

Likelihood Ratio                4.7428        1         0.0294
Score (Model-Based)             5.1080        1         0.0238
Score (Sandwich)                1.5412        1         0.2144
Wald (Model-Based)              5.4508        1         0.0196
Wald (Sandwich)                 9.4956        1         0.0021

            Analysis of Maximum Likelihood Estimates

                       Parameter       Standard      StdErr
 Variable     DF        Estimate          Error       Ratio

 x2            1        -0.00572        0.00186       0.758

                                 Hazard
Chi-Square    Pr > ChiSq          Ratio

    9.4956        0.0021          0.994
```

The results are somewhat contradictory, with the first model providing no evidence against $\beta = 0$, but the second indicating a nonzero β. Upon further analysis, Lawless and Nadeau (1995) remove unit 409. Unit 409 has the smallest censoring time but also three recurrences, indicating a somewhat greater failure rate than the rest of the units. The unit is removed and the analyses repeated. Table 5 and Table 6 contain the results for the revised data, and both now provide no evidence against $\beta = 0$.

Table 5. Revised Model for X1

```
                      The PHREG Procedure

             Testing Global Null Hypothesis: BETA=0

Test                        Chi-Square       DF     Pr > ChiSq

Likelihood Ratio                0.4091        1         0.5225
Score (Model-Based)             0.4080        1         0.5230
Score (Sandwich)                0.3059        1         0.5802
Wald (Model-Based)              0.4065        1         0.5237
Wald (Sandwich)                 0.3018        1         0.5827

            Analysis of Maximum Likelihood Estimates

                       Parameter       Standard      StdErr
 Variable     DF        Estimate          Error       Ratio

 x1            1         0.20743        0.37755       1.161

                                 Hazard
Chi-Square    Pr > ChiSq          Ratio

    0.3018        0.5827          1.231
```

Table 6. Revised Model for X2

```
                          The PHREG Procedure

            Testing Global Null Hypothesis: BETA=0

Test                        Chi-Square       DF      Pr > ChiSq

Likelihood Ratio                0.7249        1          0.3945
Score (Model-Based)             0.6803        1          0.4095
Score (Sandwich)                0.5348        1          0.4646
Wald (Model-Based)              0.6775        1          0.4104
Wald (Sandwich)                 0.6462        1          0.4215

           Analysis of Maximum Likelihood Estimates

                      Parameter      Standard     StdErr
 Variable     DF       Estimate         Error      Ratio

 x2            1       -0.00319       0.00397      1.024

                                 Hazard
Chi-Square     Pr > ChiSq         Ratio

    0.6462         0.4215         0.997
```

Regression Model for Poisson Process

If the process of event occurrence is a homogeneous Poisson process, the number of recurrences for a unit in the time period over which it is observed is a Poisson random variable with mean $M(t)$ of the form

$$M(t) = \lambda t$$

where λ is the rate of occurrence. A popular form for Poisson regression is

$$\log(\mu_i) = x_i'\beta + o_i$$

where μ_i is the Poisson mean, x_i is a vector of regression coefficients, and o_i is a constant *offset* for observation i. For the Poisson process, we have

$$\log(M(t)) = \log(\lambda) + \log(t)$$

so that λ can be estimated by Poisson regression on N_i, the number of recurrences for unit i, using the log of the censoring time as offset: $o_i = \log(t_i)$. Then $\lambda_i = \exp(x_i'\beta)$.

If the recurrence process is a Poisson process, a plot of the MCF will be a straight line. Examining Figure 1 and Figure 5 indicates that the Poisson assumption is of questionable validity for the CGD data. However, for illustration, PROC GENMOD is used to estimate the Poisson rates for the placebo and gamma interferon groups. The following statements use PROC GENMOD to fit a Poisson regression to the cgd data.

```
proc genmod data = CGD3;
   class Treatment;
   model Nevent = Treatment / dist = Poisson offset = Logt;
   estimate 'Placebo'   Intercept 1 Treatment 0 1 / exp;
   estimate 'Treatment' Intercept 1 Treatment 1 0 / exp;
run;
```

Here, the variable Nevent is the total number of recurrences of disease for a patient and Logt is the logarithm of the censoring time. The two ESTIMATE statements compute estimates of the log rates for the two groups, and the EXP option exponentiates to give estimates of the rates for the two groups. Output from the ESTIMATE statements is shown in Table 7. The rate of occurrence for the placebo group is .0030 and the rate for the gamma interferon group is .0011. Figure 7 shows plots of the resulting MCFs.

Table 7. Poisson Rate Estimates for the Poisson Model for the CGD Data

```
                    Contrast Estimate Results

                                  Standard
Label               Estimate         Error     Alpha

Placebo              -5.8015        0.1336      0.05
Exp(Placebo)          0.0030        0.0004      0.05
Treatment            -6.8540        0.2236      0.05
Exp(Treatment)        0.0011        0.0002      0.05

                             Chi-

    Confidence Limits       Square      Pr > ChiSq

   -6.0634     -5.5396      1884.8          <.0001
    0.0023      0.0039
   -7.2922     -6.4157      939.54          <.0001
    0.0007      0.0016
```

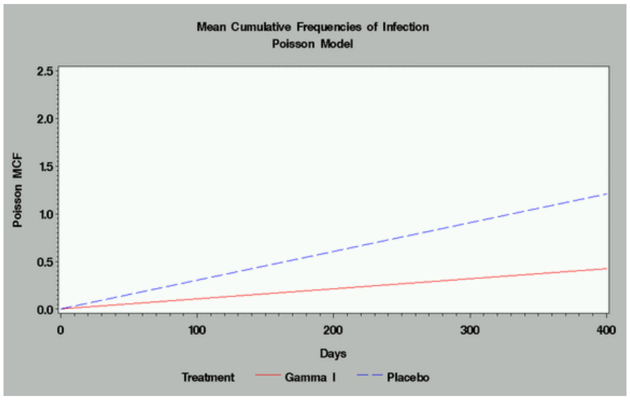

Figure 7. MCF for Poisson Model

Conclusions

Recurrent events data arise in many fields. Nelson (2003) presents numerous examples from diverse areas such as medicine, manufacturing, and social sciences. We have presented here several ways to analyze such data using features of SAS Release 9.1. If you need to analyze this kind of data, the techniques described here will be a valuable addition to your set of statistical tools.

Appendix A

The following SAS statements use PROC PHREG to create the output data set OUT2 and ODS graphics statements to create the plot of the mean cumulative function.

```
proc template;
   define statgraph cmf;
      dynamic _title;
      layout gridded;
         entrytitle
            "Mean Cumulative Frequencies of Infections";
         layout overlay /
            xaxisopts=(label="Follow-up time (days)")
            yaxisopts=(ticks={0 .2  .4  .6  .8 1.0 1.2 1.4
                               1.6 1.8 2.0 2.2 2.4 2.6}) ;
            stepplot y=cmf x=tstop / group=rx name="mean";
               discretelegend "mean" / hAlign=left
                                       vAlign=top
                                       across=1;
         endlayout;
       endlayout;
    end;
run;

ods graphics on;
ods html;

Title2 "Regression Analysis with Treatment as Covariate";
data in2;
   treatment=1;
   output;
   treatment=2;
   output;
   run;

proc phreg data=cgd covs(aggregate);
  model (tstart,tstop)*status(0)=treatment;
  baseline covariates=in2 out=out2 cmf=_all_ /nomean;
  id id;
  run;

data out2;
   set out2;
   if treatment=1 then rx='Gamma Interferon';
   else rx='Placebo';

data _null_;
   set out2;
   file print ods=(template="cmf");
   put _ods_;
   output;
   run;
```

References

Fleming, T. R. and Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, New York: Wiley.

Lawless, J. F. and Nadeau, C. (1995), "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics*, 37, 158–168.

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000), "Semiparametric Regression for the Mean and Rate Functions of Recurrent Events," *J. R. Statist. Soc.* B, 62, 711–730.

Nelson, W. B. (1995) "Confidence Limits for Recurrence Data—Applied to Cost or Number of Repairs," *Technometrics*, 37, 147–157.

Nelson, W. B. (2003), *Recurrent Events Analysis for Product Repairs, Disease Recurrences, and Other Applications*, The ASA-SIAM Series on Statistics and Applied Probability.

Contact Information

Gordon Johnston, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513.
Email gordon.johnston@sas.com

Ying So, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513.
Email ying.so@sas.com