PROPORTIONAL HAZARDS ANALYSIS OF LEFT-TRUNCATED SURVIVAL DATA

Dennis Heisey
University of Wisconsin-Madison

ABSTRACT

Failure time data are typically recorded as the times from some origin until some endpoint, such as the curing of a disease, or death. The influence of covariates or prognosticators on time until endpoint is often of interest. The proportional hazards model (Cox 1972) has become a very popular approach for evaluating the influence of covariates on failure times. PROC LIFEREG and PROC PHREG are examples of procedures that perform proportional hazards modelling of time until endpoint data. Such procedures assume that all subjects effectively enter the study at the same time as the origin of the process that gives rise to the endpoint of interest. This is not an appropriate assumption for some studies. For example, in studies of infection incubation times, it is not uncommon for subjects to enter the study well after initial infection (the origin) if initial infection is asymptomatic. Some potential subjects may reach their endpoints before study entry and will not be included in the study if, for example, they are too ill to be interviewed for covariate information; in some cases the researchers may not even be aware of the existence of such potential subjects. Such potential subjects are said to be left-truncated. A consequence of left-truncation is that the observed sample is not a random sample of times until endpoint; the observations are conditional on the subject having not reached the endpoint until after study entry. Analyses must accommodate this; direct application of procedures intended for untruncated data may return seriously misleading results. This paper discusses some of the issues in analyzing left-truncated data, and presents an example analysis suitable for grouped data, implemented with PROC PROBIT.

INTRODUCTION

Survival analysis often focuses on analyzing the times from some origin until the occurrence of some endpoint ("failure"). In many situations, the origin of the process giving rise to the endpoint and the time at which the subject enters the study coincide. For example, the origin might be the appearance of symptoms of some disease, and the endpoint might be the curing of the disease. Subjects seek medical attention when symptoms appear and immediately enter the study.

In other situations study entry may occur well after the natural origin of interest. For example, suppose we are interested in the incubation time of a disease from infection until the appearance of symptoms. For a disease such as AIDS (acquired immunodeficiency syndrome), initial infection may go unnoticed, and the subject may not be tested and determined to be HIV (human immunodeficiency virus) positive until well after infection. In such studies, even if the date of infection could be determined retrospectively (for example, in the case of blood transfusion recipients), it is not appropriate to analyze such data with techniques developed for data with study entry at the origin (Jewell 1990). This is because the observed failure times are not a random sample of all failure times; they are conditional on the subjects having not reached endpoint before study entry. Such data sets are said to be left-truncated.

Another situation where left-truncated data commonly arises is when the origin is some fixed calender time, such as the date that a polluting factory starts production. Perhaps subjects are recruited into the study from the nearby countryside over some period of time after startup. Such data are left-truncated if potential subjects that reach endpoint before they enter the study are prevented from entering the study. In addition to medical applications, left-truncation is an important problem in animal ecology (Pollock, Winterstein, Bunck, and Curtis, 1989). Animals may die before they can be captured and radiomarked for monitoring, hence the marked population returns a biased sample of ages at death.

Regression analysis of failure times examine how covariates influence the times until failure. The proportional hazards model (Cox 1972) has become especially popular for such regression analyses. The proportional hazards model assumes some function of the covariates, usually log-linear, acts in a multiplicative fashion on a baseline hazard function. If the origin corresponds with the time of study entry, procedures such as PROC LIFEREG and PROC PHREG can be used for regression analysis of failure time data. In some cases, it may be reasonable to assume a parametric model for the baseline; PROC LIFEREG permits proportional hazards modelling assuming a Weibull baseline hazard. PHREG takes a nonparametric approach and makes no assumptions about the functional nature of the baseline; it is especially useful for exploratory analyses.

For left-truncated data, it is not appropriate to simply extend the data back to the origin and treat it as if it were uncensored. This does not accurately reflect the conditional nature of the data, and the resulting likelihood, either parametric or nonparametric, will be incorrect. For untruncated data, a right-censored subject contributes $S(c)$ to the likelihood, where c is the censoring time, and $S(c)$ is the probability of surviving to c. A subject that fails at t contributes $f(t)$, where $f(t)$ is the probability density function corresponding to $S(t)$. However, in the presence of left-truncation, the distribution of the failure time t is conditional on being greater than the time of study entry, say t'. Thus, the likelihood contribution for a right-censored subject that enters the study at time t' is $S(c)/S(t')$, while the contribution is $f(t)/S(t')$ for a subject that fails at t.

Relatively little attention is given to the problem of left-truncated data in some of the more popular survival analysis texts (Kalbfleisch and Prentice, 1980; Miller 1981; Lawless, 1982; Cox and Oakes, 1984). This paper illustrates some of the issues involved in the analysis of left-truncated data, primarily in the context of the nonparametric proportional hazards model. Unlike the Cox partial likelihood approach, the analysis illustrated here is especially suited for heavily tied (grouped) failure times. As implemented with PROC PROBIT, it may not be especially practical for data sets with many distinct failure times unless the data are grouped.

DISCRETE PROPORTIONAL HAZARDS MODEL

The classical Cox proportional hazards model (Cox, 1972) assumes that the hazard function for subject k can be modelled as

$$h_k(t|Z_k) = h_0(t) \exp(Z_k\beta) \quad ,$$

where t is time, $h_0(t)$ is an arbitrary unspecified baseline hazard function, Z_k is a vector of covariates, and β is a vector of regression coefficients. The covariates Z_k may vary over time, but this is notationally suppressed.

The probability that subject k survives day i, given that it is alive at the beginning of day i, is related to the hazard function as

$$s_{ik} = \exp - \int_{i-1}^{i} h_k(t|Z_k)\, dt$$

and expressing the baseline daily survival as

$$s_{0i} = \exp - \int_{i-1}^{i} h_0(t)\, dt$$

allows the daily survival probability s_{ik} to be expressed as

$$s_{ik} = s_{0i}^{\exp(Z_k\beta)}$$

It is convenient to use the reparametrization $s_{0i}=\exp(-\exp \alpha_i)$, which gives the model

$$\begin{aligned} s_{ik} &= \exp(-\exp(\alpha_i + Z_k\beta)) \\ &= \exp(-\exp(\eta_{ik})) \end{aligned}$$

This transformation is known as the complementary log-log link or the discrete time proportional hazards model (Prentice and Gloeckler, 1978; Kalbfleisch and Prentice 1980; Agresti 1990).

This discrete time proportional hazards model should not be confused with the discrete time model used by the TIES=DISCRETE option of PROC PHREG. The complementary log-log model is appropriate for the situation where discreteness arises from grouping the data from a continuous process. This occurs very frequently, for example, when records are summarized by day or by week. The discrete time model used by PROC PHREG is actually not a proportional hazards model, but rather a proportional logit model. The proportional logit model is appropriate for truly discrete trials, such as testing the functioning of an on/off switch. In the case where discreteness arises due to grouping of the data, the proportional logit model does not give consistent estimates of the proportional hazards parameters β (Kalbfleisch and Prentice, 1980, p78).

A subject that is recruited to the study on day i' and is last observed on day j contributes

Pr(fails on j|survives through i'-1) =

$$\frac{(1 - s_{jk}) \prod_{i=1}^{j-1} s_{ik}}{\prod_{i=1}^{i'-1} s_{ik}} = (1 - s_{jk}) \prod_{i=i'}^{j-1} s_{ik}$$

to the likelihood if it failed on day j and it contributes

Pr(survives through j|survives through i'-1) =

$$\frac{\prod_{i=1}^{j} s_{ik}}{\prod_{i=1}^{i'-1} s_{ik}} = \prod_{i=i'}^{j} s_{ik}$$

if it survived day j (i.e., was right-censored). When the log likelihood contributions are collected over all subjects, the log likelihood becomes

$$\log L = \sum_{i=1}^{I} \left[\sum_{k \in F_i} \ln(1 - s_{ik}) + \sum_{k \in U_i} \ln s_{ik} \right]$$

where I is the duration of the study, F_i is the set of subjects in the study that failed on day i, and U_i is the set of subjects in the study that survived day i. The log likelihood can be written as

$$\sum_{i=1}^{I} \left[\sum_{k \in F_i} \ln[1 - \exp(-\exp(\eta_{ik}))] + \sum_{k \in U_i} -\exp(\eta_{ik}) \right].$$

This log likelihood is very similar to the log likelihood for logistic regression for binomial data, and it is in fact identical to the log likelihood for the complementary log-log regression model for binomial data (Thompson 1981). PROC PROBIT (also PROC LOGISTIC) performs regression analysis of binomial data; the GOMPERTZ distribution option specifies the complementary log-log model. Thus, with the proper preparation of data, PROC PROBIT with the D=GOMPERTZ option can be used to find the maximum likelihood estimates for the left-truncated discrete proportional hazards model.

Consider a day on which no failures are observed; the contribution to the likelihood for such a day is

$$\sum_{k \in U_i} - \exp(\alpha_i + Z_k\beta) \ .$$

This term is maximized at 0 (or equivalently $\alpha_i = -\infty$), so such days when no failures occurred may just as well be "thrown out". For nonparametric estimation (i.e., no assumptions about the structure of the α_i's), we only need to deal with days on which there was at least one failure, hereafter referred to as "event days".

Such an argument is the basic motivation behind Cox's partial likelihood approach (1972). That is, without some assumptions about the nature of the baseline, there is no information about the baseline between failures; for all we know, the true baseline hazard between failures is zero. Therefore, the only information in the data is in the neighborhood of the failure times, and the data from other times are of no consequence. If we had assumed some parametric structure for the baseline hazard $h_0(t)$, this would not be the case, days on which no failures occurred could not just be ignored.

Thus, for each event day, it is necessary to determine which subjects were in the study; these subjects compose the risk set on trial. The proportional hazards model is then estimated by fitting the binary regression model to the number of failures out of these risk sets. This is illustrated in the next section.

EXAMPLE

In this example, time is measured in days. The data set must include the day of study entry as well as the day of last observation. The fate variable indicates whether the subject failed on the last day (fate = 1), or survived (fate = 0). The data set would also include the covariates, say cov1-covn. These covariates could be quantitative or qualitative "class" variables, as specified with the class option in PROBIT. These are the input arguments for the macro %TRUNC_PH, listed in the Appendix.

The macro first scans the data set and identifies all event days. Then, for each subject, a record is output for each event day for which it was present in the study. This constructs the so-called risk sets. For an event day for which a subject was present, the date is output, as well as the fate for that subject on that particular day. Fate will always be zero for a subject unless the particular event day was the day on which the subject was observed to fail, in which case fate=1.

The discrete proportional hazards model for continuous covariates cov1, cov2, and cov3 is then estimated by specifying

```
PROC PROBIT;
CLASS DATE;
MODEL FATE/ONE = DATE COV1 COV2 COV3 /
D = GOMPERTZ;
```

The variable one is always equal to 1. Date plays a special role; the estimates resulting for date plus the intercept are estimates of the baseline parameters α_i. For nonparametric modelling, date must always be included in the model. The resulting coefficient estimates and standard errors for cov1-cov3 are the appropriate maximum likelihood estimates for the discrete proportional hazards model.

For illustration, I reanalyzed the black duck survival data presented by Pollock, Winterstein and Conroy (1989) in their Table 1; I derived entry times from their Figure 1. They originally analyzed the data with PROC PHGLM. Here I present the results from the PROC PROBIT approach, as well as the results I obtained with PROC PHREG using TIES=EXACT. (Even though this study resulted in left-truncated data, parameter estimates for the proportional hazards model could still be obtained with PROC PHREG because all subjects entered before the first failure occurred and hence all of the risk sets were the same with either approach. This will not generally be the case with left-truncated data.)

Fifty radiomarked ducks were included in the study. Deaths occurred on 16 dates, and a total of 18 ducks were observed to die (there were 2 pairs of ties). The covariates were the duck's age and its condition (the ratio of body weight to wing length). The regression parameter estimates and standard errors for the discrete proportional hazards model fitted with PROC PROBIT are shown below. The results from PROC PHREG with TIES=EXACT are shown in parentheses:

Model	Age Estimate	Condition Estimate
Age	0.140 (0.141)	
SE	0.502 (0.502)	
Cond.		-0.964 (-0.941)
SE		0.763 (0.760)
Both	0.445 (0.434)	-1.178 (-1.145)
SE	0.532 (0.532)	0.787 (0.784)

The resulting likelihood ratio chi squares for the covariate models against the baseline models are:

Model	PROBIT	PHREG	DF
Age	0.08	0.08	1
Cond.	1.65	1.58	1
Both	2.31	2.25	2

The results for both approaches are quite similar, and variations of similar size were observed between the other options for TIES (specifically DISCRETE and BRESLOW) within PROC PHREG. None of the factors appear to be important, either by the likelihood ratio tests shown here, or the Wald tests given in the output. Pollock et al. (1989) report a significant slope when only condition is in the model [$\hat{\beta}$=1.68, SE($\hat{\beta}$)=.8]. I could not duplicate this result, and survival plots stratified by condition suggest that if condition does influence the hazard, increasing condition depresses the hazard, consistent with a negative coefficient.

DISCUSSION

Improper treatment of left-truncated survival data can have serious consequences. This is not just an issue for regression analysis; it must also be addressed when estimating survival functions in the absence of covariates (Wang, 1991). When no covariates are included in the

approach outlined above (i.e., just date), the survival curve resulting from the estimated α_i's is the generalization of the Kaplan-Meier (1958) estimate to the left-truncation case. Interestingly, in the continuous time model used by PHREG, the resulting survival estimates for the no covariate case are generalizations of the Aalen-Nelson-Aultschuler estimator rather than Kaplan-Meier (e.g., Breslow, 1972, 1974; Johnasen, 1983). Thus the PROC PROBIT approach can be used to obtain Kaplan-Meier estimates when it is not appropriate to do so with PROC LIFETEST because of left-truncation. Such survival curves, constructed for strata of the covariate values, are important for diagnosing the appropriateness of the proportional hazards assumption (e.g., Kalbfleisch and Prentice, 1980).

Left-truncation, either in the continuous or discrete case, introduces no additional complications over the untruncated model in dealing with time varying covariates or stratified analyses. When a covariate for a subject changes, it is treated as if the subject with the previous covariate value leaves the study (i.e., becomes right-censored), and a new subject with the new value enters the study. This is repeated whenever the value of the covariate for a subject changes. %TRUNC_PH illustrates stratified analyses.

Regression diagnostics need to be generalized for the left-truncated case. A popular proportional hazards residual is defined as

$$\delta_k - ln\,\hat{S}(t_k) \ ,$$

where δ_k is the censoring indicator for subject k and t_k is the time of the last observation of subject k (Lawless, 1982). In the case of left-truncatien, a number of arguments would support generalizing this to

$$\delta_k - ln\,\hat{S}(t_k|t'_k) \ ,$$

where t'_k is the time of study entry for subject k (Therneau, Grambsch, and Fleming, 1990). %TRUNC_PH illustrates an ad hoc adaptation of this continuous time residual for the discrete case.

A feature like "ABSORB" in PROC GLM would be desirable for class variables in PROC PROBIT; this would increase the allowable number of distinct failure times for the grouped approach. It would be very desirable to have a continuous time procedure like PHREG extended to allow for left-truncation under various tie handling options; this could be implemented within the more general framework of Cox-Anderson-Gill models (Andersen and Gill, 1982) which permits subjects to enter and leave the study repeatedly and arbitrarily. A parametric procedure similar to LIFEREG that accommodates left-truncation would also be very useful. All of these extensions are relatively straightforward and would no doubt find wide application.

REFERENCES

Agresti, A. (1990). Categorical Data Analysis. New York : Wiley.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. Annals of Statistics 10, 1100-1120.

Breslow, N. (1972). Comments in Cox (1972).

Breslow, N. (1974). Covariance analysis of censored survival data. Biometrics 30, 89-99.

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

Cox, D. R., and Oakes, D. (1984). Analysis of Survival Data. London: Chapman and Hall.

Jewell, N. (1990). Some statistical issues in studies of the epidemiology of AIDS. Statistics in Medicine 9, 1387-1416.

Johansen, S. (1983). An extension of Cox's regression model. International Statistical Review 51, 165-174.

Kalbfleisch, J. D., and Prentice, R. L. (1980). The Statistical Analysis of Failure Time Data. New York : Wiley.

Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457-481.

Lawless, J. F. (1982). Statistical Models and Methods for Lifetime Data. New York: John Wiley.

Pollock, K. H., Winterstein, S. R., Bunck, C. M., and Curtis, P. D. (1989). Survival analysis in telemetry studies: The staggered entry design. Journal of Wildlife Management 53, 7-15.

Pollock, K. H., Winterstein, S. R., and Conroy, M. J. (1989). Estimation and analysis of survival distributions for radio-tagged animals. Biometrics 45, 99-109.

Prentice, R. L., and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. Biometrics 34, 57-67.

Miller, R. G. (1981). Survival Analysis. New York: John Wiley.

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

Thompson, R. (1981). Survival data and GLIM. Applied Statistics 30, 310.

Wang, M. (1991). Nonparameteric estimation from cross-sectional survival data. Journal of the American Statistical Association. 86, 130-143.

APPENDIX

```
%macro trunc_ph(data=z3q5w1x,in=,out=,fate=,
class=,covars=,strata=);
```

```
/*
in=time of entry
out=time of last observation
- in & out cannot exceed 9999
fate=1 for failure, =0 for censoring
class=class (qualitative) covariates
covars=quantitative covariates
strata=strata variable (0 thru n)
multiple class or covars variables can be
specified as class=%str(x1 x2 x3)

RESERVED DATASETS: z3q5w1x, x1w5q3z, alof_day,
resids, survival
RESERVED VARIABLE NAMES: strata (except as the
argument for strata=), _s_in_, _s_out_, alofd#
(where # is a number), _dday_, _outcom_, _start_,
_newset_, _tt_, _one_,day_surv, _strunc_, _tot_,
_resid_, _subj_.
*/

%let cv_n=0;
%do %while(%scan(&covars,&cv_n+1)^=);
  %let cv_n=%eval(&cv_n+1);
  %let cov&cv_n=%scan(&covars,&cv_n);
%end;
%let cl_n=0;
%do %while(%scan(&class,&cl_n+1)^=);
  %let cl_n=%eval(&cl_n+1);
  %let class&cl_n=%scan(&class,&cl_n);
%end;
%let allv=;
%if %length(&strata) GT 0 %then %let allv = &strata,;
%do i=1 %to &cv_n; %let allv = &allv &&cov&i,; %end;
%do i=1 %to &cl_n; %let allv = &allv &&class&i,; %end;
%let allv = &allv &in, &out, &fate;
data z3q5w1x; set &data;
  if (nmiss(of &allv) GT 0) then delete;
  %if %length(&strata) EQ 0 %then %do;
    strata = 0;
    %let strata = strata;
  %end;
  _s_out_ = &strata * 10000 + &out;
  _s_in_ = &strata * 10000 + &in;
  keep &in &out &fate &class &covars &strata _s_in_
_s_out_;
proc sort data=z3q5w1x; by &out;
data z3q5w1x; set z3q5w1x end=longest;
  if longest then call symput ('longest', &out);
 run;
data alof_day; set z3q5w1x; /* now find alof days */
  if (&fate EQ 1); keep _s_out_;
proc sort data=alof_day; by _s_out_;
data alof_day; set alof_day; /* remove redundancies */
  by _s_out_; if (first._s_out_);
data _null_; set alof_day end=lastalof;
  retain nalofs 0;
  nalofs = nalofs + 1;  /* count the alof days */
  if lastalof then call symput ('nalofs', nalofs);
  drop nalofs;
 run;
proc transpose data=alof_day out=alof_day prefix=alofd;
data alof_day; set alof_day;
  if (_n_ EQ 1);
  drop _name_; /* created a row vector of alof days */
proc sort data=z3q5w1x; by _s_in_; /* expand risk sets */
data z3q5w1x; if (_n_ EQ 1) then set alof_day;
  set z3q5w1x;
  retain _newset_ 1;
  retain _subj_ 0;
  array alofd {&nalofs};
  _subj_ = _subj_ + 1;
  _start_ = _newset_;
  do _tt_ = _start_ to &nalofs;
    if (alofd [_tt_] LE _s_in_) then _newset_ = _tt_;
    if (alofd [_tt_] GE _s_in_) AND
    (alofd [_tt_] LE _s_out_) then
    do;
      _dday_ = alofd [_tt_];
      _outcom_ = 0;
      if (alofd [_tt_] EQ _s_out_) then _outcom_ = &fate;
      _one_ = 1;
      output;
    end;
    else if (alofd [_tt_] GT _s_out_) then goto outloop;
  end; outloop:
  keep _subj_ &in &out _dday_ _outcom_ &class &covars
&strata _one_;
/* if there are few covariate patterns, sum _outcom_ & _one_
within _dday_ &class &covars now */
proc probit data=z3q5w1x; /* fit the model */
  class _dday_ &class;
  model _outcom_ / _one_ = _dday_ &class &covars /
  d=gompertz;
  output out=x1w5q3z xbeta=xbeta;
%if (%length(&covars) EQ 0) AND
(%length(&class) EQ 0) %then %do;
data survival; merge z3q5w1x x1w5q3z;
  if _outcom_; /* only alof dates */
  day_surv = exp (-exp (xbeta));
  _dday_ = _dday_ - 10000*&strata;
  keep day_surv _outcom_ _dday_ &strata day_surv &out;
proc sort data = survival; by &strata _dday_;
data survival; set survival;
  by &strata _dday_; if (first._dday_); /* skip ties */
  retain survival 1.0;
  retain _start_ 1;
  if (first.&strata) then do;
    survival = 1.0;
    _start_ = 1;
  end;
  do i = _start_ to _dday_;
    output;
  end;
  _start_ = _dday_;
  survival = survival * day_surv;
  if (last.&strata) then do;
    do i = _dday_ to &longest;
      output;
    end;
  end;
proc plot data=survival;
  title 'Survival function';
  plot survival * i = &strata;
%end;
%else %do;
data resids; merge z3q5w1x x1w5q3z;
  day_surv = exp (-exp (xbeta));
  by _subj_;
  retain _strunc_ 1.0;
  if (first._subj_) AND NOT (last._subj_) then
  _strunc_ = day_surv;
  else if NOT (last._subj_) then
  _strunc_ = _strunc_ * day_surv;
  else if (_outcom_) then do; /* survives to midpoint */
```

```
   _strunc_ = 1.0 - _strunc_ * sqrt (day_surv);
   output;
  end;
  else do;
   _strunc_ = _strunc_ * day_surv;
   output;
  end;
data resids; set resids;
  _tot_  = &out - &in;
  _resid_ = log (_strunc_);
  keep _tot_  _resid_ &strata;
proc plot data=resids;
  title 'Residuals vs. time on test';
  plot _resid_ * _tot_ = &strata;
run;
%end;
proc datasets;
  delete x1w5q3z z3q5w1x alof_day;


%mend trunc_ph;
```