ØRNULF BORGAN

Event History Analysis: An Overview and some Areas of Current Research

Event history analysis is a set of statistical concepts, models and methods for studying the occurrences of events over time for a number of subjects. The subjects are usually humans or animals, while the events may be deaths, onsets of a disease, divorces, etc. The aim of an investigation may be to study the effect of a medical treatment, to establish risk factors for a disease, to monitor a demographic phenomenon, or to make predictions on future occurrences of an event. Modern event history analysis has been developed over the last 30–40 years, motivated mainly by medical research, but also by problems in econometrics and technical reliability.

Professor Ørnulf Borgan
Department of Mathematics,
University of Oslo, Norway
E-mail: borgan@math.uio.no
CAS Group Leader 2005/2006

Single events

Traditionally, research in event history analysis has focused on situations where the interest is in a single event for each subject under study, which is commonly denoted *survival analysis*. A *survival time* is the time elapsed from an initial event to a well-defined end-point; e.g., time from birth to death, time from disease onset to death, or time from marriage to divorce. A special feature of survival data is *censoring:* all subjects will not experience the event of interest during the course of a study, and for some subjects it will only be known that their true survival times exceed certain censoring times.

Two basic concepts in survival analysis are the *survival function* and the *hazard rate*. The survival function $S(t)$ is the probability that a survival time will exceed t (in the study time scale: age, duration of marriage, etc.). Thus $S(t)$ describes the proportion of the population that has not yet experienced the event by time t. The *hazard rate* $h(t)$ is the instantaneous probability of the event per unit of time, i.e. $h(t)\mathrm{d}t$ is the probability that the event will happen between time t and time $t+\mathrm{d}t$ (for a small $\mathrm{d}t$) given that it has not happened earlier.

To illustrate these concepts, we consider data from Statistics Norway on divorce for couples married in 1960, 1970 and 1980. Figure 1 shows empirical hazard rates (rates of divorce) and empirical survival functions for marriages contracted in 1960, 1970, and 1980. The increase in divorce risk with marriage cohort is clearly seen. Furthermore, the hazard rates show an increase with duration of marriage until about 5 years, when a slight decline occurs. The survival functions show how the proportions still married are decreasing in the different marriage cohorts.

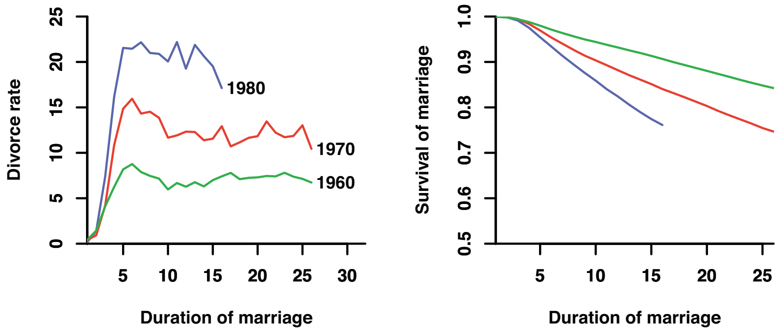

Figure 1. Rates of divorce per 1000 marriages per year (left panel) and empirical survival functions (right panel) for marriages contracted in Norway in 1960 (green lines), 1970 (red lines), and 1980 (blue lines). Based on data from Statistics Norway.

Hazard rates that are first increasing and then decreasing are found, e.g., for divorce rates and for the mortality rates of many cancers. It is tempting to interpret a decreasing hazard rate as a reduced risk at the *individual level*. However, in many cases it is more likely to be due to selection caused by *unobserved heterogeneity* between individuals. By this we mean that, for reasons unknown to us, there is variation in the hazard rates among the subjects (e.g. couples in the marriage example). Then the subjects with a high hazard rate will tend to experience the event earlier than those with a lower hazard rate. The population of subjects who have not yet experienced the event will therefore change over time, and it will eventually contain a large proportion of subjects with a low hazard rate yielding a decreasing hazard at the population level. The study of statistical models that may be used to better understand such effects of unobserved heterogeneity is an important topic of current research (e.g. Aalen *et al.* 2007).

Regression models

In most studies one would like to assess the effect of one or more *covariates* (or explanatory variables) on survival. As in other parts of statistics, regression models are called for at that point. We will discuss regression models for censored survival data by means of an example.

In the period 1962–77, a total of 205 patients were operated for malignant melanoma (cancer of the skin) at Odense University Hospital. Based on survival data for these patients, one may assess the effect of covariates on the rate of cancer death. For the purpose of illustration, we restrict our attention to the covariates sex (coded as 0 for females and 1 for males) and tumour thickness (in mm), and refer to Andersen *et al.* (1993) for a detailed analysis. The most common regression model for censored survival data is Cox's model, which for the malignant melanoma example takes the form $h(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{sex} + \beta_2 \cdot \text{thickness})$. Once the function $h_0(t)$ and the parameters β_1 and β_2 have been estimated from data, one may use the fitted model to predict the survival of future patients. To illustrate, Figure 2 shows the predicted survival functions for male and female patients with tumour thicknesses of 1 mm and 5 mm.

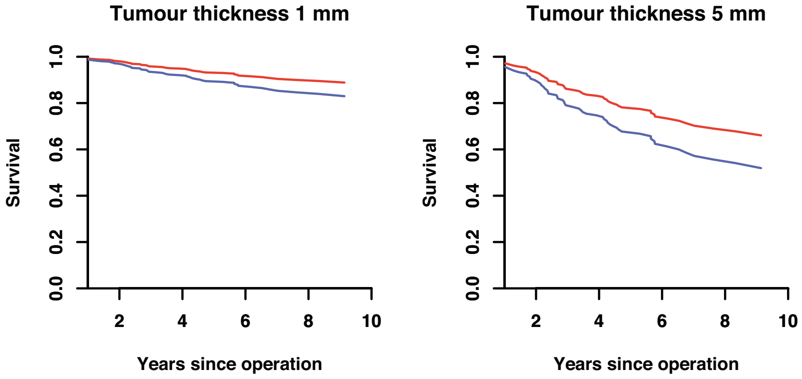

Figure 2. Predicted survival functions for female (red lines) and male (blue lines) patients with malignant melanoma. Left panel: tumour thickness 1mm; right panel: tumour thickness 5 mm. (Based on a fitted model with sex and the logarithm of tumour thickness as covariates.)

An alternative regression model is Aalen's model, which for the malignant melanoma example takes the form $h(t) = h_0(t) + \beta_1(t) \cdot \text{sex} + \beta_2(t) \cdot \text{thickness}$. One important part of any statistical analysis is to assess whether a model gives a reasonable fit to the data. For censored survival data in particular, one has to choose between Cox's and Aalen's models (and other regression models for censored survival data). How to assess the fit of a regression model and how to choose between alternative regression models is an important area of current research (e.g. Martinussen and Scheike, 2006; Gandy and Jensen, 2006).

Internal time-dependent covariates

Regression models for censored survival data may include covariates that depend on time. This causes no problems for *exogenous* covariates, but care has to be exercised when including *endogenous* time-dependent covariates in a regression analysis.

To illustrate this, we consider data from a clinical trial where 488 patients with liver cirrhosis at several Copenhagen hospitals were randomized to treatment with prednisone (a hormone) or placebo in the period 1962–69 and followed-up until 1974 (see Andersen *et al.* 1993, for details). A number of covariates were measured at randomization, and the prothrombin index (a measure of liver function) was recorded at follow-up visits to a doctor. The main aim of the study was to assess the effect of treatment. However, in order to understand better how the treatment is functioning, it is also of interest to study the effect of the prothrombin index. As earlier analyses of the data suggest that there is interaction between ascites (excess fluid in the abdomen) and treatment, for our illustrative purpose, we will restrict our attention to the 386 patients with no ascites.

Table 1. Estimated regression coefficients (with standard errors) for two Cox regression models for patients with liver cirrhosis. Model I: all covariates measured at randomization; model II: as model I, but with the prothrombin index measured at randomization replaced by its last recorded value.

Covariate	Model I	Model II
Treatment (0=placebo; 1=prednisone)	-0.28 (0.14)	-0.06 (0.14)
Sex (0=female; 1=male)	0.27 (0.16)	0.31 (0.15)
Age (in years)	0.041 (0.008)	0.043 (0.008)
Acetylcholinesterase concentration	-0.0019 (0.0007)	-0.0015 (0.0006)
Inflammation (0=absent; 1=present)	-0.47 (0.007)	-0.43 (0.15)
Baseline prothrombin (in percent of normal)	-0.014 (0.007)	
Current prothrombin (in percent of normal)		-0.054 (0.004)

Table 1 shows estimated regression coefficients (with standard errors) for two Cox regression models for the liver cirrhosis patients. In model I, all covariates are measured at randomization, while in model II, the last recorded value of the prothrombin index is used instead of the one measured at entry. Model I estimates the *total treatment effect*, and shows that mortality is reduced among those treated. However, since treatment mainly has an *indirect effect* operating via the time-dependent covariate current prothrombin, the *direct effect* of treatment estimated in model II is substantially lower. How to handle endogenous time-dependent covariates and how to define precisely concepts such as total, direct and indirect effects are challenging areas of current research (e.g. Fosen *et al.*, 2006).

Event histories

Connecting several events for a subject as they occur over time yields *event histories*. As an example, consider the model in Figure 3 for leukaemia patients who have undergone bone marrow transplantation (Keiding *et al.*, 2001). All patients start in remission (i.e. without clinical symptoms of the disease) in state 0 at transplantation, and they will ultimately either relapse (R) or die in remission (D). As intermediate events, patients may experience acute (A), chronic (C), or both acute and chronic (AC) graft-versus host disease. People are typically interested in the probability of relapse free survival, i.e. the probability of being in one of the states 0, A, C or AC, and how this is affected by treatment and other covariates. One way of approaching this problem is to fit Cox or Aalen models for the transition rates between the states, and then 'piece together' these estimates to obtain estimates of the probability of relapse-free survival (e.g., Keiding *et al.*, 2001; Aalen *et al.*, 2001). One problem with this approach is that the 'local modelling' of transition rates may fail to be a good 'global fit' for the probability of relapse-free survival. An alternative currently being investigated is direct regression modelling of the probability of relapse-free survival (e.g. Andersen and Klein, 2007; Scheike and Zhang, 2007).

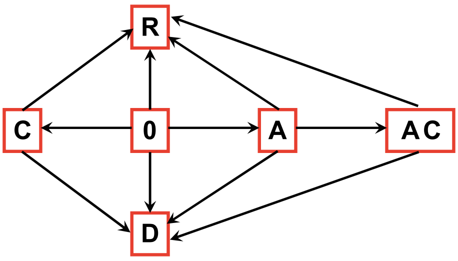

Figure 3: A multistate model for events after bone marrow transplantation (see text for details).

Concluding comments

The aims of this paper are to convey 'what event history analysis is all about' and to point out some areas of current research. A number of other areas could have been mentioned; important examples being sampling designs for event history data (e.g. Langholz and Goldstein, 1996) and the handling of the high dimensional data of modern genomics (e.g. van Houwelingen *et al.*, 2006).

References

Aalen, O.O., Borgan, Ø., and Fekjær, H. "Covariate adjustment of event histories estimated from Markov chains: The additive approach". *Biometrics* 57, 2001, 993–1001.

Aalen, O.O., Borgan, Ø., and Gjessing, H.K. *Event History Analysis: A Process Point of View*. To be published by Springer-Verlag, New York, 2007.

Andersen, P.K., Borgan, Ø., Gill, R.D., and Keiding, N. *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1993.

Andersen, P.K. and Klein, J.P. "Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies". *Scandinavian Journal of Statistics*, to appear 2007
(published online August 2006: doi: 10.1111/j.1467–9469.2006.00526.x).

Fosen, J., Ferkingstad, E., Borgan, Ø., and Aalen, O.O. "Dynamic path analysis—a new approach to analyzing time-dependent covariates". *Lifetime Data Analysis* 12, 2006, 143–167.

Gandy, A. Jensen U. "On goodness-of-fit tests for Aalen's additive risk model". *Scandinavian Journal of Statistics* 32, 2005, 425–445.

Keiding, N., Klein, J.P., and Horowitz, M.M. "Multi-state models and outcome prediction in bone marrow transplantation". *Statistics in Medicine* 20, 2001, 1871–1885.

Langholz, B. and Goldstein, L. "Risk set sampling in epidemiologic cohort studies". *Statistical Science* 11, 1996, 35–53.

Martinussen, T. and Scheike, T.H. *Dynamic Regression Models for Survival Data*. Springer-Verlag, New York, 2006.

Scheike, T.H. and Zhang, M.-J. "Directly modelling the regression effects for transition probabilities in multistate models". *Scandinavian Journal of Statistics*, to appear, 2007.

van Houwelingen, H.C., Bruinsma, T., Hart, A.A.M., van't Veer, L.J., and Wessels, L.F.A. "Cross-validated Cox regression on microarray gene expression data". *Statistics in Medicine* 25, 2006, 3201–3216.