Survival analysis

Murthi

Why not use regression?

- Censoring
- If there is a lot of censoring, regression will give biased estimates
- Why not treat event as a binary variable and use logit?
- Time has more information and using logit will make estimates inefficient.

What is it?

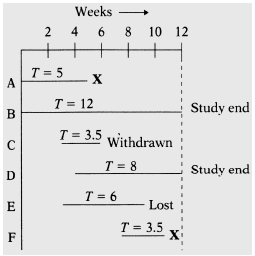
Class of methods for studying the occurrence and timing of events

Also known as

- Failure time analysis,
- Reliability analysis (in engineering),
- Event history analysis (in sociology),
- Duration analysis or transition analysis (in economics).
- It is used heavily in the insurance industry to set insurance premiums.

Censoring

- Subject does not experience event of interest
- Incomplete follow-up
  - Lost to follow-up
  - Withdraws from study
  - Dies (if not being studied)
- **Left** or **right** censored



What is an event?

- It is a qualitative change situated on a time scale.
- We know when a person moves from one state to another and the time between two changes in events.
- For example, a marriage is a transition from unmarried state to married state.
- A job promotion is a transition from one position to a higher position.
- We need to predict when a change will occur in time and what factors affect the time to transition (or duration).

- *Right censoring*: If a person has lived till age 50 then we know his lifetime T is >50 but we do not know how much longer he/she is likely to live. Right censoring is more commonly seen.
- *Left censoring*: If you are studying the age at which a person gets arrested for the first time and you start studying data on youth older than 15 years. The all we can say for a person who got arrested before 15 years is that T<15.
- *Interval censoring*: when there is both left- and right-censoring, a < T < b.

- *Type 1 censoring*: if censoring time is fixed (that is, it is under the control of the investigator) and all observations had the same censoring time
  - e.g., We stopped observing deaths after 3 years.
- *Type II censoring*: When observations are terminated after a pre-specified number of events have occurred. We do not see this kind of censoring in social sciences.
  - e.g., A researcher decides to stop an experiment after 50 out of his hundred rats died.

How to use survival analysis

- Time to death
- Time in remission after treatment
- Time before a customer leaves the firm
- Time before an insurance claim is made
- Time before a brand switch is made
- Time before a firm defaults
- Time before a customer churns
- Time to click or purchase on a website

- In all these cases we want to know factors that affect the duration

- *Random censoring*: When observations are terminated for reasons not under the control of the researcher.
- e.g., some participants leave the study
- Random censoring can also occur when there is a single termination time but the entry times vary randomly across individuals. People came in at different times into the study. In this case one solution is to introduce entry time as a covariate in the model.
- Try to avoid random censoring as much as possible. If there is informative censoring then the parameters will be biased. For instance, PhD students who drop out of the program are also likely to be weaker students, and families that drop out of a marriage study are more likely to have marriage problems.

Survivor Function

- Survivor function, S(t) defines the probability of surviving longer than time *t*
  - What is the probability that a person would live to age 90?
  - In a sample, count all people who are over 90 years and divide by the sample size to get this probability.
  - What is the probability that a customer will stay for 3 years with a firm?

Survival Analysis

- Model **time to failure** or event
- Able to account for censoring
- Can **compare survival** between 2+ groups
- Assess **relationship between covariates and survival time**

- T is considered to be a random variable following a parametric distribution function.
- The c.d.f. of T is denoted by $F(t)= Pr(T \leq t)$.
- Survivor function : $S(t) = 1-F(t) = Pr (T>t)$
- S(0)=1. and as T increases, S(t) can never increase. It can have a variety of shapes.

Hazard function

- Hazard function is the probability that an event will happen in the next instant given that it has not happened until now.
- P(t|t>T) : it is a conditional probability
- If a person survives until age 90, what is the chance that something will happen in the next period?
- Take a sample of 90 years olds, count all people who passed away in the next time period (e.g., a week)

Constant hazard over time model

- log h(t) = a
- => S(t) = exp(-at)
- => f(t) = a.exp(-at)
- pdf is an exponential distribution of time

- log h(t) = a + bt
- This implies that time has a Gompertz distribution

- log h(t) = a + b log(t)
- time has a Weibull distribution

Hazard function

- $h(t) = \lim_{\Delta t \to 0} Pr[t \leq T < t+\Delta t \mid T>t] / \Delta t$
- h(t) is the probability conditional on individual surviving to time t (T>t).
- $f(t) = \lim_{\Delta t \to 0} Pr[t<=T<t+\Delta t] / \Delta t$
- $h(t) = f(t) / S(t) = f(t) / [1-F(t)]$
- $S(t) = \exp\{- \int_0^t h(u)\, du\}$

- Note : hazard function is not strictly a probability as it can be greater than 1.0 but it has lower bound at 0.
- Survivor and hazard functions can be converted between each other

Univariate method: Kaplan-Meier survival curves

- Generates the characteristic “stair step” survival curves
- Does **not** account for confounding or effect modification by other covariates

- Hazard function can also be interpreted as the number of events in a given interval of time.
- Suppose the hazard for getting influenza at some point of time (measured in months) is 0.015,
- it means that it will take 1/0.015 months or 66.66 months before I can expect to get the flu.
- If my hazard for death is 0.011 per year then I can expect to live to 90.9 years.
- I am assuming in the above calculation that the hazard rate remains constant over time, which it may not.

SAS

- PROC LIFETEST uses Kaplan Meier (KM) method for small datasets and life-table method for large datasets.
- to compute and plot the estimate of the distribution of the survival time.
- Produces life tables and graphs of survival curves
- For example, you want to compare the survival experiences of patients who receive different treatments for their disease.

```
Proc lifetest data=myel;
time dur*status(0);
run;
```

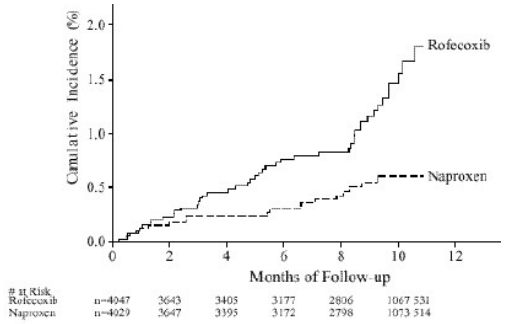
- Plots

```
proc lifetest data=myel plots=(s) graphics outsurv=a;
time dur*status(0);
symbol1 v=none;
run;
```

- Testing differences

```
proc lifetest data=myel plots=(s) graphics outsurv=a;
time dur*status(0);
strata treat;
symbol1 v=none color=black line=1;
symbol2 v=none color=black line=2;
run;
```

Time to Cardiovascular Adverse Event in VIGOR Trial



- Lifetable method

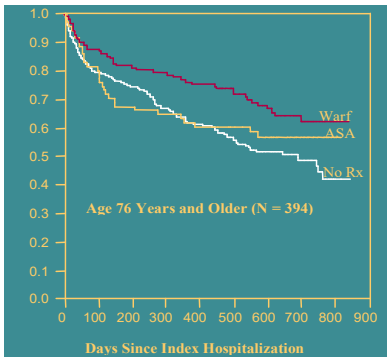
```
proc lifetest data=myel plots=(s) graphics outsurv=a
method=life;
time dur*status(0);
strata treat;
symbol1 v=none color=black line=1;
symbol2 v=none color=black line=2;
run;
```

Parametric hazard models: PROC LIFEREG

- $\text{Log}(T_i) = \beta X_i + \sigma\varepsilon_i$
- Log ensures that predicted values of T are positive.

```
PROC LIFEREG data =a1 outest=a;
model week*arrest(0)=fin age race wexp mar paro prio / dist=lnormal;
output out=b xbeta=lp;
run;
```

- exp(β) gives the ratio of the expected survival times for the two groups if the X variables is binary.
- for a continuous variable, 100(exp(β)-1) gives the percent increase in expected survival time for each unit increase in the variable.
- Weibull, exponential, gamma, log-logistic, log-normal distributions are available in SAS.



Multivariate methods: Cox proportional hazards (PH)

- Needed to assess effect of multiple covariates on survival
- Cox-proportional hazards is the most commonly used multivariate survival method
  - Easy to implement in SPSS, Stata, or SAS
  - Parametric approaches are an alternative, but they require stronger assumptions about h(t).
  - Does not require assumption of a distribution for survival function. Hence it is called a semiparametric method.

Cox proportional hazard model

- Conveniently separates baseline hazard function from covariates
  - Baseline hazard function over time $h_0(t)$
  - Covariates are time independent
- Nonparametric
- Quasi-likelihood function

$$h(t,\mathbf{X}) = h_0(t)\, e^{\sum \beta_i X_i}$$

- proc phreg data=a1;
- model week*arrest(0)= fin age race wexp mar paro prio;run;

- You will see no intercepts - characteristic of partial likelihood.
- Hazard ratio or risk ratio is exp(β).
- 100(exp(β)-1) is the percentage change in hazard of arrest due to a unit change in X variable.

Cox proportional hazards model, continued

- Can handle both continuous and categorical predictor variables
- Without knowing baseline hazard $h_o(t)$, can still calculate coefficients for each covariate, and therefore **hazard ratio**
- Assumes multiplicative risk—this is the proportional hazard assumption
  - Can be compensated in part with interaction terms

Differences between LIFEREG and PHREG

- Lifereg accomodates left censoring and interval censoring, PHREG only allows right censoring
- PROC LIFEREG you can test hypotheses about the shape of the hazard function. PHREG you cannot.
- If the shape of the survival distribution is known, LIFEREG gives efficient estimates.
- LIFEREG automatically creates dummy variables. PHREG does not.

Limitations of Cox PH model

- Covariates normally do not vary over time
  - True with respect to gender, ethnicity, or congenital condition
    - One can program time-dependent variables
    - When might you want this?
- Baseline hazard function, $h_o(t)$, is never specified, but Cox PH models known hazard functions
  - You can estimate $h_o(t)$ accurately if you need to estimate S(t).