Censoring

- missing data
- right censoring : observations in progress are stopped by a reason other than death, e.g. end of investigation, retirement, policy surrender, policy expiry
- left censoring : observations are before start of investigation, e.g. calculating exact duration of sickness without exact date of getting sick
- interval censoring : observations do not provide exact point of time, e.g. only calendar year of death but not exact date of death
- random censoring : time of censoring is unknown, e.g. voluntary retirement, policy surrender
- informative censoring : future lifetime and time of censoring are dependent, e.g. those surrender their life policies tend to have better health than those who continue their policies

Censoring

- Type I censoring : time of censoring is known in advance, e.g. end of investigation, normal retirement
- Type II censoring : observations in progress are stopped when a pre-determined number of deaths is reached, e.g. a medical trial is ended after 50 lives under a particular treatment die

Notation & Assumptions

- a population of size N
- all aged x or have some common properties
- lives are independent
- $t_1 < t_2 < ... < t_k$ are times with at least one death
- number of deaths at t_j is d_j
- total number of deaths is $m = d_1 + d_2 + ... + d_k$
- t_0 is time 0 and t_{k+1} is end of investigation
- c_j is number of lives censored between t_j and t_{j+1}
- $t_{j,1} < t_{j,2} < ... < t_{j,c_j}$ are times of censoring
- if a life is censored at exactly t_j, treat it as if it occurred very shortly after t_j
- total number of censored lives is $N - m$
- t_j^- is just before t_j
- n_j is number of lives who are alive and at risk at t_j^-
- set $n_{k+1} = 0$
- $n_j = d_j + c_j + n_{j+1}$
- assume censoring is non-informative

Kaplan-Meier Estimation

- from maximum likelihood
- a step function with a jump at each t_j
- $\hat{\lambda}_j = \frac{d_j}{n_j}$

  for $t_j \leq t < t_{j+1}$

$$\hat{F}(t) = 1 - \prod_{i=1}^{j} \left(1 - \hat{\lambda}_i\right)$$

Greenwood's Formula

– for $t_j \leq t < t_{j+1}$

$$\text{Var}\left(\tilde{F}(t)\right) \approx \left(1-\hat{F}(t)\right)^2 \sum_{i=1}^{j} \frac{d_i}{n_i\left(n_i-d_i\right)}$$

Nelson-Aalen Estimation

– $\hat{\lambda}_j = d_j / n_j$

for $t_j \leq t < t_{j+1}$

$$\hat{F}(t) = 1 - \exp\left(-\sum_{i=1}^{j} \hat{\lambda}_i \right)$$

Greenwood's Formula

– for $t_j \leq t < t_{j+1}$

$$\text{Var}\left(\tilde{\Lambda}_t\right) = \text{Var}\left(\sum_{i=1}^{j} \tilde{\lambda}_i\right) \approx \sum_{i=1}^{j} \frac{d_i\left(n_i - d_i\right)}{n_i^3}$$

Kaplan-Meier vs Nelson-Aalen

$$\hat{F}(t) = 1 - \prod_{i=1}^{j}\left(1-\hat{\lambda}_i\right) \approx 1 - \prod_{i=1}^{j}\exp\left(-\hat{\lambda}_i\right)$$

$$= 1 - \exp\left(-\sum_{i=1}^{j}\hat{\lambda}_i\right)$$

Heterogeneity

- lives have different characteristics
  e.g. age, sex, smokers, occupation
- split the population into homogeneous subgroups
- smaller sample size reduces statistical significance
- information may be limited or inaccurate
- strike a balance

Heterogeneity

- suppose we are selling life policies
- if we estimate mortality rate from a heterogeneous population, we get an ‘average’ rate
- if we use this average rate to calculate premiums, healthier lives are overcharged while those with poor health are undercharged
- if there is another insurer who prices correctly, healthier lives will leave us and go to that insurer while those with poor health will come to us

Regression

- an alternative is to deal with those different characteristics directly with regression
- in regression these characteristics are treated as covariates
- a covariate can be quantitative

  e.g. age, height, weight
- a covariate can be qualitative

  e.g. 0 for male and 1 for female, 1 to 5 for increasing order of illness
- Cox model

Cox Model

- force of mortality or hazard function of *i*th life at time *t* :

$$\lambda_i(t) = \lambda_0(t)\exp\left(\vec{\beta}\,\vec{z}_i^{\mathrm{T}}\right) = \lambda_0(t)\exp\left(\sum_{j=1}^{p}\beta_j x_{i,j}\right)$$

- $\lambda_0(t)$ is baseline hazard at time *t*
- $\vec{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of regression parameters
- $\vec{z}_i = (x_{i,1}, x_{i,2}, ..., x_{i,p})$ is a vector of covariates of *i*th life
- *p* is number of parameters

Cox Model

– ratio of life 1's hazard function to life 2's is constant at all t :

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\lambda_0(t)\exp\left(\vec{\beta}\,\vec{z}_1^{\mathrm{T}}\right)}{\lambda_0(t)\exp\left(\vec{\beta}\,\vec{z}_2^{\mathrm{T}}\right)} = \exp\left(\vec{\beta}\,\vec{z}_1^{\mathrm{T}} - \vec{\beta}\,\vec{z}_2^{\mathrm{T}}\right) = \exp\left(\sum_{j=1}^{p}\beta_j\left(x_{1,j} - x_{2,j}\right)\right)$$

– also called proportional hazards model

Cox Model

- covariates are time-independent here
- it may be acceptable for variables such as age or weight IF their values do not vary greatly over the investigation, e.g. age does not change much after 6 months, OR main effect of a covariate depends on its value at a specific point of time, e.g. weight at start of a disease
- if covariates are time-dependent, use extended Cox model

Cox Model

- baseline hazard determines shape of hazard function
- exponential term shows differences between lives
- if only interested in the latter, ignore the former and still can estimate regression parameters
- this is a semi-parametric approach, only looking at part of hazard function
- identify relative levels of mortality
- flexible because no need to assume any shape at start
- understand data better before deciding baseline hazard

Partial Likelihood Function

- t_j is time at which (only) one death is observed
- N_j is set of lives who are alive and at risk just before t_j
- $\vec{z}_j^*$ are covariates for death at t_j
- partial likelihood function is :

$$L = \prod_{j=1}^{k} \frac{\exp\left(\vec{\beta}\, \vec{z}_j^{*\mathrm{T}}\right)}{\sum_{i \in N_j} \exp\left(\vec{\beta}\, \vec{z}_i^{\mathrm{T}}\right)}$$

- it is partial because time of death and time of censoring are ignored
- if a life is censored at exactly t_j, treat it as if it occurred very shortly after t_j

Approximate Partial Likelihood Function

- more than one death at a point of time
- $d_j \geq 1$ is the number of deaths at t_j
- $\vec{z}^*_{j,l}$ are covariates for lth death at t_j
- $\vec{s}_j = \sum_{l=1}^{d_j} \vec{z}^*_{j,l}$
- approximate partial likelihood function is :

$$L = \prod_{j=1}^{k} \frac{\exp\left(\vec{\beta}\, \vec{s}_j^{\mathrm{T}}\right)}{\left(\sum_{i \in N_j} \exp\left(\vec{\beta}\, \vec{z}_i^{\mathrm{T}}\right)\right)^{d_j}}$$

Maximum Likelihood

- maximum likelihood estimates :

$$\frac{\partial}{\partial \beta_1} \ln L \bigg|_{\vec{\beta}=\hat{\vec{\beta}}} = 0, \quad \frac{\partial}{\partial \beta_2} \ln L \bigg|_{\vec{\beta}=\hat{\vec{\beta}}} = 0, \quad \cdots, \quad \frac{\partial}{\partial \beta_p} \ln L \bigg|_{\vec{\beta}=\hat{\vec{\beta}}} = 0$$

- cell (i , j) of information matrix $\vec{I}$:

$$-\frac{\partial^2}{\partial \beta_i \partial \beta_j} \ln L \bigg|_{\vec{\beta}=\hat{\vec{\beta}}}$$

- variance-covariance matrix : $\vec{C} = \vec{I}^{-1}$
- diagonal cell (i , i) gives variance of $\tilde{\beta}_i$
- cell (i , j) gives covariance b/n $\tilde{\beta}_i$ and $\tilde{\beta}_j$
- asymptotically, regression parameter estimators are normal and unbiased
- test statistic $\hat{\beta}_i \Big/ \sqrt{\text{Var}\left(\tilde{\beta}_i\right)}$ for null hypothesis $H_0 : \beta_i = 0$
- rejected at 5% significance level if > 1.96 or < –1.96 (two-sided)

Likelihood Ratio Statistic

- identify important covariates and discard less important ones
- say, start with two models : one with p covariates and one with $p + q$ covariates
- test whether addition of extra q covariates has significant effects
- null hypothesis is $H_0 : \beta_{p+1} = \beta_{p+2} = ... = \beta_{p+q} = 0$
- likelihood ratio statistic is :

$$2 \ln \frac{L_{p+q}\Big|_{\vec{\beta}_{p+q}=\hat{\vec{\beta}}_{p+q}}}{L_p\Big|_{\vec{\beta}_p=\hat{\vec{\beta}}_p}}$$

- asymptotically, it has a χ^2_q distribution
- rejected at 5% significance level if $>$ 95th percentile of χ^2_q (one-sided)

Model Building

- start with a null model that has no covariates and then add covariates one by one
- start with a full model that includes all potential covariates and then exclude those that have no significant effects
- it may be useful to incorporate interactions between covariates