General Multiplicative Model for Grouped Data

\[
\log \lambda_{jk} = \alpha_j + x_{jk} \beta \\
\]

where

\[
\alpha_j = \text{Nuisance parameters (strata - age, etc)} \\
x_{jk} = \text{Exposure(s) of Interest (may be a vector)} \\
\frac{\lambda_{jk}}{\lambda_{jk'}} = \exp\{(x_{jk} - x_{jk'})\beta\} \text{ depends only on exposure}
\]

Poisson model: \( \log E[d_{jk}] = \log t_{jk} + \alpha_j + x_{jk} \beta. \)

Note that \( \log t_{jk} \) is a known constant – does not need to be estimated (i.e. it is NOT a parameter, nor does it have a parameter associated with it.) It is usually called the Offset.

In SAS

```
PROC GENMOD ;
  Model deaths = age exposure / dist = poisson offset = log_year ;
```

where \text{log\_year} is a variable whose value for the \( j,k \) cell is \( \log t_{jk}. \)

In R/Splus

```
glm(deaths ~ age + exposure + offset(log.year), family = poisson, ...)
```

\[
G^2 = 2 \left( \sum_j \sum_k d_{jk} \log \left( \frac{d_{jk}}{\hat{d}_{jk}} \right) + (\hat{d}_{jk} - d_{jk}) \right)
\]

Since we nearly always have \( \sum_j \hat{d}_{jk} = \sum_j d_{jk}, \) this is the usual \( G^2 = 2 \sum_{\text{All Cells}} \text{Obs log} \left( \frac{\text{Obs}}{\text{Exp}} \right) \)

Calendar and Age Specific Rates

In some settings, there may be interest in how age specific risk changes over time, or in adjusting for changes age specific risk changes over time. These changes may result from changes in environment, diet, lifestyle, etc.

This requires that we stratify in two dimensions as in the grid at right. A given subject begins follow-up at a particular age and calendar time indicated by the dots. Follow-up ends at a later time when both calendar time and age are increased by the total follow-up time. At the end of followup, the subject either experiences an event or not, as indicated by a D or a W.
For each age/calendar cell we can compute the person-years of exposure for each subject. We summarize the total person-years exposure and total number of events in each cell, and proceed as we have previously.

Most generally, we would consider each cell as a separate stratum and allocate a parameter to each (say $\theta_{jl}$ where $j$ denotes age category and $l$ denotes calendar category). Another way to model the underlying hazard would be to assume that age-specific rates are proportional across calendar time, in which case the parameterization would take the form $\theta_{jl}\lambda(t)$.

### Survival Analysis

Thus far we have considered subjects grouped by age/time/exposure cohorts. There are several drawbacks to this approach. First, the grouping is somewhat arbitrary. The number/length of intervals may have a significant effect on the result. Second, as the number of potential covariates increases, the number potential cells increases rapidly and the number of subjects contributing to each cell decreases leading to instability. Special methods are available which allow us to let the intervals get smaller. I will approach this problem from a survival analysis viewpoint.

First, we suppose that all subjects start at a common time. Subjects who withdraw (“W”) prior to experiencing an event (death/disease) are considered censored. Subjects who experience an event are considered failures. For each subject we observe a pair, $(t_i, \delta_i)$ where $t_i =$ time of end of observation, and $\delta_i$ is the indicator of failure (or censoring indicator).

$$\delta_i = \begin{cases} 
1 & \text{event} \\
0 & \text{no event} 
\end{cases}$$

The underlying assumption is that all subjects will eventually experience an event if followed long enough. Thus each subject has a (possibly unobserved) event time $y_i$. Each subject also has a potential censoring time, $g_i$, which is assumed to be independent of the failure time. The observed time to end of followup is denoted by $t_i = \min(y_i, g_i)$.

If $y_i$ and $g_i$ are dependent, then $g_i$ contains information about $y_i$, and we have informative censoring. Since we observe only one of $y_i$ and $g_i$, it is not possible to test the independence assumption, nor to estimate or account for it without making some additional (untestable) assumptions. In this setting, we will always assume that independence holds.
Letting $Y$ be the time to failure,

\[ S(t) = \Pr\{Y > t\} = \text{Survivor Function} \]

\[ = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(u) du} \]

where $\lambda(t)$ is the hazard, and $\Lambda(t)$ is the cumulative hazard.

How do we estimate $S(t)$? (or $\Lambda$ or $\lambda$?) Suppose that we divide the time axis into intervals and for each interval estimate $\hat{\lambda}_j$, then

\[ \hat{\Lambda}_j = \sum_{l=1}^j \hat{\lambda}_l \Delta s_l \]

where $\Delta s_l$ is the length of the $l^{th}$ interval. While not necessary, we will assume that $\Delta s_l = \Delta s$, a common length. Then,

\[ \hat{S}_j = e^{-\hat{\Lambda}_j} = \prod_{l=1}^j e^{-\hat{\lambda}_l \Delta s} \approx \prod_{l=1}^j (1 - \hat{\lambda}_l \Delta s) \]

Now let $\Delta s \to 0$. Since we have a finite number of subjects, and hence a finite number of failure times, eventually, we will have at most one failure time per interval. (Most intervals will have no failures.) For intervals with no failures, $\hat{\lambda}_l = 0$, and these intervals do not contribute to the estimate of $S(t)$. Otherwise, since nearly all subjects in an interval will be followed for the entire interval, the exposure time in an interval can be approximated by $n_l \Delta s$, where $n_l$ is the number of subjects at risk during the $l^{th}$ interval. We have

\[ \hat{\lambda}_l = \frac{d_l}{n_l \Delta s} \]

and

\[ \hat{S}_j = \prod_{l=1}^j (1 - \frac{d_l}{n_l \Delta s} \Delta s) = \prod_{l=1}^j (1 - \frac{d_l}{n_l}) \]

or,

\[ \hat{S}(t) = \prod_{t_l < t} (1 - \frac{d_l}{n_l}) \]

where

\[ d_l = \# \text{ deaths at time } t_l \]
\[ n_l = \# \text{ at risk just prior to time } t_l \]

Convention: if observations are tied, assume that failures precede censored observations. The estimator $\hat{S}(t)$ is called the Kaplan-Meier or Product-Limit survival estimate.

Note: We may randomly break ties, by adding small perturbations to the $t_l$, (assuming that all censored observations occur after the last of the failure times) and achieve the same estimate. For example, if there observations are tied at time $t_l$, the contribution to $\hat{S}$ from time $t_l$ is

\[ (1 - \frac{1}{n_l})(1 - \frac{1}{n_l - 1})(1 - \frac{1}{n_l - 2}) = \frac{n_l - 1}{n_l} \frac{1}{n_l - 1} \frac{n_l - 2}{n_l - 2} = \frac{n_l - 3}{n_l} = 1 - \frac{3}{n_l} = 1 - \frac{d_l}{n_l} \]
Consider the following special case in which we have a common censoring time, so that all failures precede all censored observations. Then, the number at risk at any failure time
\[ = \text{total sample size} - \text{number of prior events} \]
and
\[
\hat{S}(t) = \left(1 - \frac{d_1}{n}\right) \left(1 - \frac{d_2}{n-d_1}\right) \left(1 - \frac{d_3}{n-d_1-d_2}\right) \cdots \left(1 - \frac{d_{j-1}}{n-d_1-d_2-\cdots-d_{j-1}}\right) \left(1 - \frac{d_j}{n-d_1-d_2-\cdots-d_{j-1}}\right) \\
\]
where \( t_j < t < t_{j+1} \).

\[
= \frac{n-d_1-d_2-\cdots-d_j}{n} \\
= 1 - \frac{d_1 + d_2 + \cdots + d_j}{n} \\
= 1 - \frac{\text{# dead prior to time } t}{\text{total # subjects}}
\]
is the natural estimate of \( S(t) \). Note that at each failure time, the size of the jump in the survival curve is \( \frac{d_j}{n} \).

With arbitrary censoring, \( \hat{S}_j = \hat{S}_{j-1} (1 - \frac{d_j}{n_j}) = \hat{S}_{j-1} - \frac{d_j}{n_j} \). The size of the jumps in the Kaplan-Meier estimate are \( \frac{d_j}{n_j} \).
At some point, when the size of the risk set is small, estimates of $S(t)$ become unreliable. A single failure can result in a large jump in $\hat{S}(t)$.

Variance of $\hat{S}(t)$. We first write $\log \hat{S}(t) = \sum_{t_i < t} \log(1 - \frac{d_i}{n_i})$. The $\frac{d_i}{n_i}$ are not independent, but they are uncorrelated, so the variance of the sum is the sum of the variances. If $d_i$ is binomial, we may estimate its variance by $d_i(n_i - d_i)/n_i$ and

$$\text{Var}(\log \hat{S}(t)) \approx \sum_{t_i < t} \frac{1}{(n_i - d_i)^2} d_i(n_i - d_i)/n_i = \sum_{t_i < t} \frac{d_i}{n_i(n_i - d_i)}$$

which is usually known as Greenwood’s formula.

An $\alpha$-confidence interval takes the form: $\hat{S}(t)e^{\pm \sqrt{\text{Var}(\log \hat{S}(t))} Z_{\alpha/2}}$. Survivor function estimates can be obtained using:

- SAS - PROC LIFETEST
- R/Splus - surv.fit

compute Kaplan-Meier estimates, use Greenwood’s Formula.