1 Representativity

2 Notation and basic concepts

Let T represent the time to arrive to a predefined station. We assume that T has a cumulative distribution function $P(t) = Pr(T \leq t)$ and probability density function $p(t) = \frac{d}{dt}P(T)$. The survival function represents the complement of P, and is defined as $S(t) = 1 - P(t)$. Alternatively, the arrival time can be represented iun terms of the hazard function, which can be interpreted as the instantaneous arrival time t, conditional on survival (not arrival) to that time:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{Pr(T \in [t, t + \Delta t) | T \geq t)}{\Delta t} \\ &= \frac{p(t)}{S(t)}. \end{aligned} \tag{1}$$

We can assume that the hazard function corresponds to a predefined parametric family of distributions, e.g. Gompertz, Weibull or Exponential. They come with a predefined increasing or decreasing with time of the hazard function. Alternatively, we can model the hazard function nonparametricaly, i.e. model-free, in that case the behaviour of the hazard function is completely specified by the data analyzed.

The hazard function by itself can be used to model and analyse the time to an event form homogeneous data, i.e. independent and identicaly distributed (i.i.d.) data. In some ocasions the i.i.d. assumption is not suitable because the data are grouped or we have several observations for every individual, so it is necessary to do cluster or stratified analysis. Also, it is also convenient to relate the hazard function to some specific covariate characteristics for each individual. The model described in the folowing section allows us to analyse data under this assumptions.

3 Frailty models

The frailty models (FM), introduced by Clayton and Cuzick (1985), generalizes the porportional hazards model (Cox, 1972) for survival time data describing the hazard function as

$$h_{ij}(t|\boldsymbol{x}_i, \nu_i) = h_0(t) \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta})\nu_i, \tag{2}$$

where i indexes the group of individuals, j indexes individuals, $\boldsymbol{x}_{ij}$ is a p-dimensional covariate vector, $\boldsymbol{\beta}$ is a p-dimensional unknown parameter vector, h_0 is the base-line hazard function, and ν_i is a multiplicative random effect. The parameter ν_i measures the "frailties" shared by all the members of the same group. This parameter plays an important role differenciating the hazard function behaviour across groups. The incorporation of this parameter in the model allows us to carry out a "stratified" analysis of the data, recognizing common intra group behaviour and differences between groups, simultaneously.

In this model, the base-line hazard function and the parameter vector are common across groups. The multiplicative random effect is usually reparameterized as $\theta_i = \log(\nu_i)$ so that the new random effect parameter can be incorporated in the linear combination of the exponential multiplicative component

of the model. The reparameterized model is written as

$$h_{ij}(t|\boldsymbol{x}_i, \nu_i) = h_0(t) \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \theta_i). \tag{3}$$

We can interpret this model in the following way. Every simgle observation follows a common base-line hazard function, measuring their exposure to risk failure. But the failure risk for every one of them is modified by a multiplicative component determinated by local covariate characteristics and also by a local-shared common factor for the group that they belong. Comparison across groups can be realised comparing their random effects.

In the Metro Map analysis we have repeated observations for several individuals and also we have information regarding their answers for the different maps (cities). We can assume that individuals across "groups" are independent, so we can carry out independenpendt analysis for each one of them (three). Repeated observations for a single idividual are modeled with "frailties", i.e. we are considering i indexes in the model description for each individual and j-indexes for the number of responses of each individual. Also, we consider the map of the city as one single covariate for each model, defining $x_{ij} \in \{1, 2, ..., 6\}$ acording to the city and question answered. The survival time is considered actually as **arrival** time, i.e. the 'event-histpry' variable is defined as the arrival to the 'correct' terminal or station in the corresponding question.

At the end, we will have three different base-line survival functions, which can be used to compare the behaviuour concerning arrival times across groups.

Main references

Clayton, D. G. and Cruzick, J. (1985) "Multivariate generalization of the proportional hazards model (with discussion)." *Journal of the Royal Statistical Society, Series A*, **148**: 82-117.

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 & Hall.

Therneau, T. M. (1999) A package for survival analysis in S. *Technical Report*, Mayo Foundation.

Therneau, T.M., Grambsch, P. M. and Pankratz, V. S. (2003) "Penalized survival models and frailty." *Journal of Computational and Graphical Statistics*, **12**(1): 156-175.