University of Illinois | Department of Economics
Spring 2017 | Econ 574 | Roger Koenker

Lecture 24
Competing Risks and Opportunities

A common situation in survival analysis is that there are several competing risks. We might imagine a latent variable model for $\{T_j : 1, \ldots, m\}$ denoting failure times due to causes $j = 1, \ldots, m$. We observe

$$(Y, J) = (\min_j\{T_j\}, \arg\min_j\{T_j\})$$

This formulation has some philosophical difficulties: Joe dies after being hit by a bus at age 35, of alcoholism at 55, by drowning at 75, ... It seems to require what is called a "many worlds" interpretation in modern physics. In this view it seems desirable to analyze the joint distribution of the vector T. But in the immortal words of the Rolling Stones, "you can't always get what you want, but if you try real hard – you get what you need."

Identifiability of Competing Risks

Cox (1962, p. 112) notes that given data on (Y, J) we cannot identify the form of the underlying dependence in the vector T. In particular, we may imagine estimating the conditional densities:

$$\begin{aligned} f_1(t) &= \lim_{h\to 0} \frac{P(T_1 \in (t, t+h], T_2 > t)}{h} \\ f_2(t) &= \lim_{h\to 0} \frac{P(T_2 \in (t, t+h], T_1 > t)}{h} \end{aligned}$$

but for an arbitrary joint distribution of T_1 and T_2 we can always find a specification of independent T_1 and T_2 with the same f_1 and f_2.

Do we really need a sample of cats, who die nine times, to estimate a dependent competing risk model? No, under some conditions we can make some progress without cats. Heckman and Honoré (1989) consider a variant of the Cox PH model. They suppose that

$$S_i(t|x) = \exp\{-\Lambda_i(t)\phi_i(x)\} \quad i = 1, \ldots, m$$

solving for the cumulative hazards

$$\Lambda_i(t) = -\log S_i(t|x)(\phi_i(x))^{-1}$$

and evaluating at a random event time T_i we have

$$\Lambda_i(T_i) = -\log U_i/\phi_i(x)$$

where $U_i \sim U[0, 1]$. This formulation yields a simple way to simulate from this model by computing

$$T_i = \Lambda_i^{-1}(-\log U_i/\phi_i(x))$$

The U_i's can be taken as independent, or alternatively can be generated as dependent. Note that random vectors $U \in [0, 1]^m$ with uniform marginals are characterized by their distribution function $C(u_i, \ldots, u_m)$, or copula function. In the independent case this is just the uniform distribution on $[0, 1]^m$. The form of the PH model imposes enough structure so that under some further regularity conditions we can identify some (restricted) forms of dependence. Further details are given in Heckman and Honoré (1989) and Abbring and van den Berg (2003).

An alternative approach originating, with Prentice et al (1978), that seems to have become more common in Biostatistics, is to focus on cause specific hazard and incidence functions. These are always identifiable. Briefly, consider what is called the crude risk of failure from cause 1 represented by the random variable

$$T_1^* = \begin{cases} Y & \text{if } J = 1 \\ +\infty & \text{otherwise,} \end{cases}$$

which has df

$$F_1(t) = P(T_1^* < t) = P(Y < t, J = 1).$$

This is estimable. In the counting process notation of L23, let $N_i(t) = I(Y_i \leq t, \delta_i = 1, J = 1)$ and $R_i(t) = I(Y_i \geq t)$ where δ_i is the usual censoring indicator $I(Y_i \leq C)$. A nonparametric estimator of F_1 is given by

$$\hat{F}_1(t) = \int_0^t \frac{\hat{S}(u)}{\bar{R}(u)} I(\bar{Y}(u) > 0) d\bar{N}(u)$$

where $\hat{S}(u)$ is the Kaplan-Meier estimator based on $\{(Y_i, \delta_i) : i = 1, \ldots, n\}$ the Kaplan-Meier estimator pooling all the risk categories into one.

Note that this is again using the fact that

$$F(t) = \int_0^t dF(s) = \int_0^t (1 - F(s-)) d\Lambda(s)$$

except that we have $\hat{S}(s)$ for $(1 - F(s-))$ and we need to account for the possibility that $\bar{Y}(u)$ can take the value 0.

Note also that if there is no censoring – so $\hat{S}$ is the empirical df of the Y_i – then this expression for $\hat{F}_1(t)$ is just the edf of T_1^*.

Once you have $\hat{F}_i(t)$ you can compute quantiles as in Peng and Fine (2007), and given discrete treatment variables one can plot $\hat{F}_1$'s for the samples with and without treatment to visualize QTE's.

Postscript on the Roy Model

To end the course on an economic note I would like to conclude by briefly discussing a model of occupational choice that is closely related to the competing risk model. The crucial reference is Heckman and Honoré (1990).

Roy (1951) proposed the following model of occupational choice: Agents have skills of two types, fishing/hunting, fighting/loving, etc. (S_1, S_2) which are associated with wages π_1, π_2, the agents choose the occupation that maximizes earnings. Skills are distributed in the population according to the df $F(s_1, s_2)$, which is assumed to have density $f(s_1, s_2)$. In fact Roy assumed normality, but this seems to be something that warrants empirical investigation.

Let p denote the proportion of the population choosing occupation 1,

$$p = \int_0^\infty \int_0^{\pi_1 s_1/\pi_2} f(s_1, s_2) ds_1 ds_2$$

The marginal density of S_1 is,

$$f_1(s) = \int_0^\infty f(s, s_2) ds_2$$

which should be distinguished from the density of s_1 for those employed in occupation 1,

$$g_1(s) = p^{-1} \int_0^{\pi_1 s/\pi_2} f(s, s_2) ds_2$$

The density of earnings in the whole economy is, changing variables $s_i \to \pi_i s_i \equiv w_i$,

$$g(w) = pg_1(w) + (1-p)g_2(w).$$

Many questions ensue: does inequality increase as a result of self selection into occupations, who is self-selected into occupations, as wages change, how do average skill levels change?

Roy assumed lognormality of skills, but some of the results he derived are shown to be valid by Heckman and Honoré under the much weaker assumption of log concavity of the distribution of

$$\log(\pi_1 S_1) - \log(\pi_2 S_2)$$

or somewhat more generally under the assumption of log concavity of

$$(w_1, w_2) = (\log(\pi_1 S_1), \log(\pi_2 S_2))$$

Under normality, moments of w and various conditional versions are easily computed and consequently one can investigate the effects of comparative static "experiments" like what happens to income inequality when π_1 changes,

Theorem: For log-normal distributed skills, self selection reduces income inequality as measured by the variance of log earnings relative to random assignment.

Identifiability of the Log Normal Roy Model?

Can we identify the parameters of the normal theory Roy model? As in the Cox model the answer is "yes" provided that we have covariates that shift the mean skill levels. More surprising is

Theorem (Basu and J.K. Ghosh): Suppose $(X_1, X_2) \sim \mathcal{N}(u, \Sigma)$ and $Z = \min_k\{X_i\}$ and $I = \arg\min_i\{X_i\}$ then u, Σ are identified from (Z, I).

This is quite esoteric and attempting to generalize it leads back to the non-identifiability results of Cox. In the general case we can always find an independent skill distribution that rationalizes the observed data. Heckman and Honoré note however that if we have varying skill prices, π, then these independent configurations would have to be consistent over different π and this can be used to identify the joint distribution.

Abbring, J.H. and G.J. van den Berg (2003) "The identifiability of the mixed proportional hazards competing risks model", *JRSS(B)*, 65, 701-710.

Basu, A.P. and J.K. Ghosh (1978) Identifiability of the multinormal and other distributions under competing risks model, *J. of Multivariate Analysis*, 8, 413-429.

Cox, D.R. (1962) *Renewal Theory*, Methuen.

Heckman, J. and B. Honoré (1989) "Identifiability of the computing risk model," *Biometrika*, 76, 325-330.

Heckman, J. and B. Honoré (1990) "The empirical content of the Roy model," *Econometrica*, 58, 1121-49.

Peng, L. and J.P. Fine (2007) "Nonparametric quantile inference with competing risks data," *Biometrika*, 94, 735-744.

Prentice, R. et.al. (1978) "The analysis of failure times in the presence of competing risks," *Biometrics*, 34, 541-544.

Roy, A.D. (1951) Some thoughts on the distribution of Earnings, *Oxford Economic Papers*, 3, 135-146.