Applied Survival Analysis Using R Chapter 9: Multiple Survival Outcomes and Competing Risks

Qi Guo

Department of Mathematical Sciences
The University of Texas at Dallas

April, 17 2019

Outline

1. Clustered Survival Times
2. Marginal Survival Models
3. Frailty Survival Models
4. Cause-Specific Hazards

Problem

- The type of survival data commonly we have considered has, as *an endpoint*, *a single cause of death*, and the survival times of each case have been assumed to be *independent*.
- Problem:
  - How about the independence assumption no longer holds in clustered data?
  - How about the event may repeat indefinitely, and then we would have multiple times per person?
  - How about only the first of several outcomes is observable?

Example

mutation carrier in female relatives among families

Determine if a female was a mutation carrier, and find the relationship between mutation and breast cancer, this subset consists of 1,960 families with two or more female relatives; for those with three or more female relatives, two were selected at random.

```
> ashkenazi[ashkenazi$famID %in% c(1, 9, 94), ]
     famID     brcancer     age     mutant
1        1            0      73          0
2        1            0      40          0
7        9            0      89          0
8        9            1      60          0
87      94            1      44          1
88      94            0      45          1
```

- And we know the covariate is “`mutant`” ,the censoring variable is “`brcancer`”.

Model

- Suppose that there is only one covariate, and its estimate is $\hat{\beta}$,*ignoring* the cluster structure first.
- Denote the estimate of its variance (from the Cox model) by $\hat{V}$, and the standard error of the estimate is then $\hat{V}^{1/2} = \sqrt{\hat{V}}$. and assume all subjects are independent.
- To obtain a correction due to the clustering structure, define a *score residual* for subject j in cluster i:

$$s_{ij} = \delta_{ij}[z_{ij} - \bar{z}(t_{ij})] - \sum_{t_u \le t_{ij}} [z_i - \bar{z}(t_{ij})] e^{z_i \beta} \left[\hat{H}_0(t_u) - \hat{H}_0(t_{u-1}) \right] \quad (1)$$

Model

- The first part of this residual is the *Schoenfeld residual* in Chapter 7
- Formulate a quantity C defined by:

$$C = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \sum_{m=1}^{n_i} s_{ij} s_{im} \tag{2}$$

- We can define a cluster-adjusted variance by $V^* = \hat{V}^2 \cdot C$, and a standard error for $\hat{\beta}$ by $\sqrt{V^*}$.
- If there are q covariates, $se(\beta) = [diag(V^*)]^{1/2}$, and the score residuals s_{ij} are $1 \times q$ matrices, $C = \sum_{i=1}^{G} \sum_{j=1}^{n_i} \sum_{m=1}^{n_i} \underset{\sim}{s'_{ij}} \underset{\sim}{s_{im}}$, and the cluster-adjusted covariance matrix is given by $V^* = \hat{V} C \hat{V}$.

Generalize to clustered survival data

- The independent survival data, with the *i*th observation given by (t_i, δ_i, z_i), the likelihood function:

$$L(\beta; z_i) = \prod_{i=1}^{n} f(t_i, \beta)^{\delta_i} S(t_i, \beta)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i, \beta)^{\delta_i} S(t_i, \beta) \qquad (3)$$

  or in baseline cumulative hazard:

$$L(\beta; z_i) = \prod_{i=1}^{n} [h_0(t_i) e^{z_i\beta}]^{\delta_i} \cdot e^{-H_0(t_i)} e^{z_i\beta} \qquad (4)$$

  where $H_0(t_i) = -\int_0^{t_i} h_0(v)dv$ is the baseline cumulative hazard.
- Now suppose that the survival times are organized into clusters

Generalize to clustered survival data

- Assign each individual in a cluster a common factor known as a *frailty* or, alternatively, as a *random effect*, and denote the frailty for all individuals in the *i*th cluster by ω_i, then the hazard function for the *j*th subject in the *i*th cluster as follows:

$$h_{ij}(t_{ij}) = h_0(t_{ij}) \cdot \omega_i e^{z_{ij}\beta} \tag{5}$$

  and ω_i is vary from one cluster to another, and a common model that governs this variability is a *gamma distribution*

$$g(\omega_i, \theta) = \frac{\omega^{\frac{1}{\theta}-1} e^{-\frac{\omega}{\theta}}}{\Gamma(\frac{1}{\theta})\theta^{\frac{1}{\theta}}} \tag{6}$$

Generalize to clustered survival data

- When we take frailties ω_i into consideration, the the joint likelihood for the jth subject in the ith cluster would be:

$$L_{ij}(\beta, \theta; \omega_i, t_{ij}, \delta_{ij}, z_{ij}) = g(\omega_i, \theta) \cdot [h_0(t_{ij})\omega_i e^{z_{ij}\beta}]^{\delta_{ij}} \cdot e^{-H_0(t_{ij})\omega_i e^{z_{ij}\beta}} \quad (7)$$

  and the full likelihood would be:

$$L(\beta, \theta) = \prod_{i=1}^{G} \prod_{j=1}^{n_i} L_{ij}(\beta, \theta; \omega_i, t_{ij}, \delta_{ij}, z_{ij}) \quad (8)$$

- The frailties are *latent* variables, we cannot directly observe. Thus, to obtain estimates of β and θ we need to use a multistage procedure called the *EM (expectation-maximization) algorithm*.

Example

- Using standard Cox proportional hazards model to predict the age of onset of breast cancer depending on mutant.

```
> result.coxph <- coxph(Surv(age, brcancer) ~ mutant,
data=ashkenazi)
> summary(result.coxph)
 n= 3920, number of events= 473
            coef    exp(coef)    se(coef)    z       Pr(>|z|)
mutant    1.1907    3.2895       0.1984    6.002    1.95e-09 ***
```

- The log partial likelihood from this model is obtained as follows:

```
result.coxph$loglik
[1] -3579.707   -3566.745
```

- -3579.707 is no covarites and the -3566.745 is model with "`mutant`" included as a predictor, the likelihood ratio test statistic is twice the difference, $G^2 = 2(3579.707 - 3566.745)$ $= 25.924$,compared to a chi-square distribution with 1 df.

Model with cluster

```
> result.coxph.cluster <- coxph(Surv(age, brcancer) ~ mutant +
cluster(famID), data=ashkenazi)
> summary(result.coxph.cluster)
n= 3920, number of events= 473
         coef   exp(coef)  se(coef)  robust.se   z    Pr(>|z|)
mutant  1.1907   3.2895     0.1984     0.2023  5.886  3.96e-09 ***
```

- The “`robust se`”, this estimate is only slightly higher than the one from the standard Cox model, indicating that the effect of clustering within first-degree relatives is small.

```
> result.coxph.frail <- coxph(Surv(age, brcancer) ~ mutant +
frailty(famID), data=ashkenazi)
> summary(result.coxph.frail)
      n= 3920, number of events= 473
               coef  se(coef)    se2    Chisq    DF     p
mutant        1.272   0.2317   0.2004   30.13   1.0  4.0e-08
frailty(famID)                         221.50 211.6  3.1e-01
```

New facility in R

- The “frailty” option, is the “coxme” package,must be separately downloaded and installed.

```
> library(coxme)
> result.coxme <- coxme(Surv(age, brcancer) ~ mutant + (1|famID),
data=ashkenazi)
> summary(result.coxme)
Cox mixed-effects model fit by maximum likelihood
  Data: ashkenazi events,
  n = 473, 3920
  Iterations= 10 63
                    NULL     Integrated     Fitted
Log-likelihood  -3579.707    -3564.622   -3411.522
                  Chisq    df        p          AIC       BIC
Integrated loglik 30.17   2.0   2.8100e-07   26.17    17.85
Penalized loglik 336.37  150.1  2.2204e-16   36.16   -588.13
Model:  Surv(age, brcancer) ~ mutant + (1 | famID)
Fixed coefficients
           coef      exp(coef)     se(coef)     z       p
mutant   1.236609    3.443914    0.2205358   5.61   2.1e-08
Random effects
 Group Variable   Std Dev    Variance
 famID Intercept 0.5912135  0.3495334
```

Kaplan-Meier Estimation with Competing Risks

- A patient may potentially experience *multiple events*, only the first-occurring of which can be observed, eg: diagnosis with prostate cancer until death from that Cause 1 to Cause 2, but for a particular patient we can only observe the time to the first event.
- One way:Select each as the primary event, and to treat the other as a censoring event.
- However,Obtain unbiased estimates of survival curves, this simplistic method would require the usually *false assumption* that the two causes of death are independent.

Kaplan-Meier Estimation with Competing Risks

Fig. 9.1 Kaplan-Meier estimates of the probabilities of death from prostate cancer and from other causes

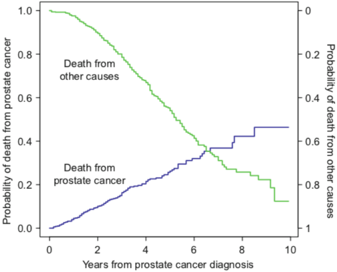
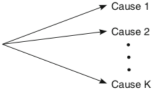


Fig. 9.2 Subject can die of only one of K causes



- Skip the R codes, and the result shows above, but such an exercise would require the assumption that the causes be independent. This assumption cannot be tested from the data,

Cause-Specific Hazards and Cumulative Incidence Functions

- Suppose that there are K distinct causes of death, and each subject can experience at most one of the K causes of death
- With competing risks, it is helpful to define, for each cause of interest, a function known as the *cumulative risk function*

$$F_j(t) = Pr(T \leq t, C = j) = \int_0^t h_j(u)S(u)du \tag{9}$$

- And the hazards is:

$$h_j(t) = \lim_{\delta \to 0} \left(\frac{Pr(t < T < t + \delta, C = j | T > t)}{\Delta} \right) \tag{10}$$

- It's easy to get:

$$h(t) = \sum_{j=1}^{K} h_j(t) \tag{11}$$

Formula

- Suppose now that we have D distinct ordered failure times t_1 ,t_2,..., t_D, estimate the hazard at the ith time ti using $\hat{h}(t_i) = d_i/n_i$
- The cause-specific hazard for the kth hazard may be written in a similar form as $\hat{h}_k(t_i) = d_{ik}/n_i$
- The probability of failure from any cause at time t_i is the product of $\hat{S}(t_{i-1})$, the probability of being alive just before t_i, and $\hat{h}(t_i)$, the risk of dying at t_i. Similarly, the probability of failure due to cause k at that time is $\hat{S}(t_{i-1})\hat{h}(t_i)$
- So the *cumulative incidence function* is:

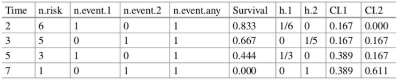
$$\hat{F}_k(t) = \sum_{t_i \leq t} \hat{S}(t_{i-1})\hat{h}(t_i) \tag{12}$$

Example

- First compute the overall survival distribution.

```
> tt <- c(2,7,5,3,4,6)
> status <- c(1,2,1,2,0,0)
> status.any <- as.numeric(status >= 1)
> result.any <- survfit(Surv(tt, status.any) ~ 1)
> result.any$surv
[1] 0.8333333 0.6666667 0.6666667 0.4444444 0.4444444 0.0000000
```

- Compute the cumulative incidence functions as in the following table:



Time	n.risk	n.event.1	n.event.2	n.event.any	Survival	h.1	h.2	CI.1	CI.2
2	6	1	0	1	0.833	1/6	0	0.167	0.000
3	5	0	1	1	0.667	0	1/5	0.167	0.167
5	3	1	0	1	0.444	1/3	0	0.389	0.167
7	1	0	1	1	0.000	0	1	0.389	0.611

Example

- Returning to the prostate cancer example of Fig. 9.1,now estimate the competing risks cumulative incidence functions as follows:

```
> sf <- survfit(Surv(survTime, status, type="mstate") ~ 1,
data=prostateSurvival.highrisk)
> tt <- sf$time
> CIs <- sf$pstate
> ci1 <- CIs[,1]
> ci2 <- CIs[,2]
> times <- tt/12
> Rci2 <- 1 - ci2
```

- Plot

```
> plot(Rci2 ~ times, type="s", ylim=c(0,1),lwd=2, color="green",
xlab="Time in years". ylab="Survival probability")
> lines(ci1~times, type="s", lwd=2, col="blue")
> lines(surv.other.km ~ time.km, type="s",
col="lightgreen", lwd=1)
> lines(cumDist.prost.km ~ time.km, type="s",
col="lightblue", lwd=1)
```

Example

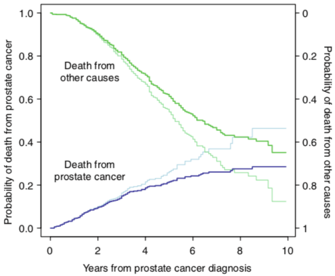


Fig. 9.4 Cumulative incidence of death from prostate cancer and from other causes, compared to the Kaplan-Meier estimates

- These curves represent estimates of the actual probabilities that a patient will die of a *particular cause*, rather than hypothetical probabilities that he would die of one cause in the absence of the other.

Regression Methods for Cause-Specific Hazards

- Modeling covariate information for competing risks too difficult to define precisely the hazard function on which the covariates should operate.
- Study the effects of the, *remaining covariates (grade and age)* on prostate cancer death, treating other causes of death as censoring indicator.

```
> prostateSurvival.T2 <- prostateSurvival[prostateSurvival$stage
=="T2",]
> attach(prostateSurvival.T2)
> result.prostate <- coxph(Surv(survTime, status.prost) ~ grade +
ageGroup)
> summary(result.prostate)
                  coef    exp(coef)    se(coef)     z      Pr(>|z|)
gradepoor        1.2199    3.3867       0.1004   12.154   < 2e-16 ***
ageGroup70-74   -0.2860    0.7513       0.2595   -1.102   0.2704
ageGroup75-79    0.4027    1.4958       0.2257    1.784   0.0744
ageGroup80+      0.9728    2.6454       0.2148    4.529   5.92e-06 ***
```

Regression Methods for Cause-Specific Hazards

- Conclusion:Patients having poorly differentiated disease (grade = poor) have much worse prognosis than do patients with moderately differentiated disease (the reference group here), with a log-hazard ratio of 1.2199.
- Define a "sub-distribution hazard"

$$\bar{h}_k(t) = \lim_{\delta \to 0} \frac{pr(t < T_k < t + \delta | E)}{\delta} \tag{13}$$

where the conditional event is given by:

$$E = \{\{T_k > t\} or \{T_{k'} \leq t \; and \; k' \neq k\}\} \tag{14}$$

Example

- When computing these sub-distribution hazards, the risk set includes not only those currently alive and at risk for the *k*th event type, but also those who *died earlier of other causes*.
- This method is implemented in the “`crr`” function in the R package “`cmprsk`”.

```
> cov.matrix <- model.matrix(~ grade + ageGroup)
> head(cov.matrix)
(Intercept)  gradepoor  ageGroup70-74  ageGroup75-79  ageGroup80+
1      1         0           1               0             0
2      1         1           0               1             0
3      1         1           0               1             0
4      1         1           0               0             1
5      1         0           0               1             0
6      1         0           0               1             0
> cov.matrix.use <- cov.matrix[,-1] # drop the first column
```

Example

- Obtain estimates for the prostate cancer as follows, dropping the first (intercept) column of the covariate matrix.

```
> library(cmprsk)
> result.prostate.crr <- crr(survTime, status, cov1=cov.
matrix[,-1], failcode=1)
                    coef    exp(coef)   se(coef)   z       Pr(>|z|)
gradepoor          1.132     3.102       0.101   11.20    0.00000
ageGroup70-74     -0.272     0.762       0.253   -1.08    0.28000
ageGroup75-79      0.367     1.443       0.219    1.67    0.09400
ageGroup80+        0.799     2.224       0.208    3.85    0.00012
```

- The argument "`failcode=1`" refers to death from prostate cancer. For death from other causes, we use "`failcode=2`",

Comparing the Effects of Covariates on Different Causes of Death

- We know the risk of both causes of death increase with age. But does the effect of age differ for these two causes?
- Just split each patient's data into separate rows, one for each cause of death.
- Begin by setting up a "`transition`" matrix using the function "`trans.comprisk`",

```
> tmat <- trans.comprisk(2, names = c("event-free", "prostate",
"other"))
> tmat          to
from          event-free   prostate        other
  event-free      NA           1              2
  prostate        NA          NA             NA
  other           NA          NA             NA
```

- The matrix states that a patient's status can change from "`event-free`" to either "`prostate`" or "`other`", these latter two being causes of death.

Example

- Next, we use the function "`msprep`" to create the new data set, and examine the first few rows, and obtain a summary of the numbers of events of each type as follows:

```
> prostate.long <- msprep(time = cbind(NA, survTime, survTime),
status = cbind(NA, status.prost, status.other),
keep = data.frame(grade, ageGroup), trans = tmat)
> head(prostate.long)
> events(prostate.long)
$Frequencies
            to
from        event-free  prostate  other  no event  total entering
 event-free      0           410   1345      4165              5920
 prostate        0             0      0         0                 0
 other           0             0      0         0                 0
```

- These results indicate that there are 410 deaths due to prostate cancer, 1345 due to other causes,and 4165 censored observations, for 5920 total.

Summary

- Use separate commands, one for “`trans = 1`” (prostate cancer) and the other for “`trans = 2`” (other causes of death), as follows:

```
> summary(coxph(Surv(time, status) ~ grade + ageGroup,
data=prostate.long, subset=trans==1))
> summary(coxph(Surv(time, status) ~ grade + ageGroup,
data=prostate.long, subset=trans==2))
```

- The results are identical to what we obtained before.
- Expect that cancer grade affects prostate cancer death differently than it does death from other causes.

```
> summary(coxph(Surv(time, status) ~ grade*factor(trans) +
 ageGroup + strata(trans), data=prostate.long))
   n= 11840, number of events= 1755
                 coef   exp(coef)  se(coef)     z    Pr(>|z|)
gradepoor       1.239       3.451     0.100  12.391   < 2e-16 ***
factor(trans)2     NA          NA     0.000      NA        NA
ageGroup70-74   0.026       1.027     0.112   0.235   0.81431
ageGroup75-79   0.333       1.395     0.104   3.201   0.00137 **
ageGroup80+     0.833       2.301     0.099   8.394   < 2e-16 ***
gradepoor:
factor(trans)2 -0.963       0.382     0.116  -8.327   < 2e-16 ***
```

Conclusion

- Conclusion: The interaction between a grade of “`poor`” and cause “`2`” (other death). The estimate -0.963 represents the additional effect of *poor grade* on risk of death from *other causes* relative to its effect on prostate cancer death. And the hazard of death from other causes is $exp(0.963) = 0.381$ times the hazard of death from prostate cancer.
- How increasing age affects the risk of dying from prostate cancer and of other causes?

```
> summary(coxph(Surv(time, status) ~ (grade + ageGroup)*trans +
ageGroup + strata(trans), data=prostate.long))
```