

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

- 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.

```
1 > ashkenazi[ashkenazi$famID %in% c(1, 9, 94), ]
2   famID   brcancer   age   mutant
3 1      1          0    73        0
4 2      1          0    40        0
5 7      9          0    89        0
6 8      9          1    60        0
7 87     94          1    44        1
8 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 :

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

Model

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

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

- 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:

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

or in baseline cumulative hazard:

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

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:

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

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

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

Generalize to clustered survival data

- When we take frailties ω_i into consideration, the the joint likelihood for the j th subject in the i th 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.

```

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

```

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

```

1 result.coxph$loglik
2 [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

```

1 > result.coxph.cluster <- coxph(Surv(age, brcancer) ~ mutant +
2 cluster(famID), data=ashkenazi)
3 > summary(result.coxph.cluster)
4 n= 3920, number of events= 473
5      coef    exp(coef)    se(coef)    robust.se      z      Pr(>|z|)
6 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.

```

1 > result.coxph.frail <- coxph(Surv(age, brcancer) ~ mutant +
2 frailty(famID), data=ashkenazi)
3 > summary(result.coxph.frail)
4      n= 3920, number of events= 473
5      coef    se(coef)    se2      Chisq      DF      p
6 mutant      1.272    0.2317    0.2004     30.13     1.0    4.0e-08
7 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.

```

1 > library(coxme)
2 > result.coxme <- coxme(Surv(age, brcancer) ~ mutant + (1|famID),
3 data=ashkenazi)
4 > summary(result.coxme)
5 Cox mixed-effects model fit by maximum likelihood
6 Data: ashkenazi events,
7 n = 473, 3920
8 Iterations= 10 63
9
10 NULL Integrated Fitted
10 Log-likelihood -3579.707 -3564.622 -3411.522
11 Chisq df p AIC BIC
12 Integrated loglik 30.17 2.0 2.8100e-07 26.17 17.85
13 Penalized loglik 336.37 150.1 2.2204e-16 36.16 -588.13
14 Model: Surv(age, brcancer) ~ mutant + (1 | famID)
15 Fixed coefficients
16 coef exp(coef) se(coef) z p
17 mutant 1.236609 3.443914 0.2205358 5.61 2.1e-08
18 Random effects
19 Group Variable Std Dev Variance
20 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

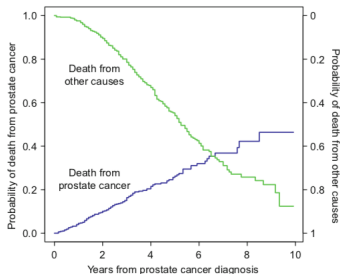
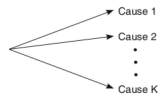


Fig. 9.2 Subject can die of only one of K causes



- 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*

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

- And the hazards is:

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

- It's easy to get:

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

Formula

- Suppose now that we have D distinct ordered failure times t_1, t_2, \dots, t_D , estimate the hazard at the i th time t_i using $\hat{h}(t_i) = d_i/n_i$
- The cause-specific hazard for the k th 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 $\hat{S}(t_{i-1})\hat{h}_k(t_i)$
- So the *cumulative incidence function* is:

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

Example

- First compute the overall survival distribution.

```

1 > tt <- c(2,7,5,3,4,6)
2 > status <- c(1,2,1,2,0,0)
3 > status.any <- as.numeric(status >= 1)
4 > result.any <- survfit(Surv(tt, status.any) ~ 1)
5 > result.any$surv
6 [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	CL1	CL2
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:

```

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

```

- Plot

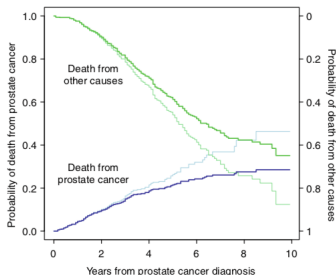
```

1 > plot(Rci2 ~ times, type="s", ylim=c(0,1),lwd=2, color="green",
2 xlab="Time in years". ylab="Survival probability")
3 > lines(ci1~times, type="s", lwd=2, col="blue")
4 > lines(surv.other.km ~ time.km, type="s",
5 col="lightgreen", lwd=1)
6 > lines(cumDist.prost.km ~ time.km, type="s",
7 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.

```

1 > prostateSurvival.T2 <- prostateSurvival[prostateSurvival$stage
2 == "T2", ]
3 > attach(prostateSurvival.T2)
4 > result.prostate <- coxph(Surv(survTime, status.prost) ~ grade +
5 ageGroup)
6 > summary(result.prostate)
7
8      coef      exp(coef)    se(coef)      z      Pr(>|z|)
9 gradepoor      1.2199      3.3867      0.1004    12.154    < 2e-16 ***
10 ageGroup70-74  -0.2860      0.7513      0.2595    -1.102     0.2704
11 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”

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

where the conditional event is given by:

$$E = \{ \{ T_k > t \} \text{ or } \{ T_{k'} \leq t \text{ and } k' \neq k \} \} \quad (14)$$

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”.

```

1 > cov.matrix <- model.matrix(~ grade + ageGroup)
2 > head(cov.matrix)
3 (Intercept)  gradepoor  ageGroup70-74  ageGroup75-79  ageGroup80+
4 1          1          0          1          0          0
5 2          1          1          0          1          0
6 3          1          1          0          1          0
7 4          1          1          0          0          1
8 5          1          0          0          1          0
9 6          1          0          0          1          0
10 > 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.

```

1 > library(cmprsk)
2 > result.prostate.crr <- crr(survTime, status, cov1=cov.
3 matrix[, -1], failcode=1)
4
5      coef      exp(coef)    se(coef)      z      Pr(>|z|)
6 gradepoor      1.132      3.102      0.101    11.20    0.00000
7 ageGroup70-74  -0.272      0.762      0.253    -1.08    0.28000
8 ageGroup75-79   0.367      1.443      0.219     1.67    0.09400
9 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”,

```

1 > tmat <- trans.comprisk(2, names = c("event-free", "prostate",
2 "other"))
3 > tmat
4 from      to
5 event-free  NA      1      2
6 prostate    NA     NA     NA
7 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:

```

1 > prostate.long <- msprep(time = cbind(NA, survTime, survTime),
2   status = cbind(NA, status.prost, status.other),
3   keep = data.frame(grade, ageGroup), trans = tmat)
4 > head(prostate.long)
5 > events(prostate.long)
6 $Frequencies
7           to
8 from      event-free  prostate  other  no event  total entering
9 event-free      0         410    1345     4165      5920
10 prostate       0          0         0         0         0
11 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:

```
1 > summary(coxph(Surv(time, status) ~ grade + ageGroup,  
2 data=prostate.long, subset=trans==1))  
3 > summary(coxph(Surv(time, status) ~ grade + ageGroup,  
4 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

```

1 > summary(coxph(Surv(time, status) ~ grade*factor(trans) +
2   ageGroup + strata(trans), data=prostate.long))
3   n= 11840, number of events= 1755
4           coef      exp(coef)    se(coef)      z      Pr(>|z|)
5 gradepoor      1.239         3.451      0.100  12.391    < 2e-16 ***
6 factor(trans)2      NA          NA      0.000      NA          NA
7 ageGroup70-74    0.026         1.027      0.112    0.235    0.81431
8 ageGroup75-79    0.333         1.395      0.104    3.201    0.00137 **
9 ageGroup80+      0.833         2.301      0.099    8.394    < 2e-16 ***
10 gradepoor:
11 factor(trans)2 -0.963         0.382      0.116   -8.327    < 2e-16 ***
12 ---
13 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

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 $\exp(-0.963) = 0.382$ times the hazard of death from prostate cancer.
- How increasing age affects the risk of dying from prostate cancer and of other causes?

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