BI476: Biostatistics - Case Studies

Lec08: Semiparametric Survival Data Analysis

Maoying,Wu
ricket.woo@gmail.com

Dept. of Bioinformatics & Biostatistics
Shanghai Jiao Tong University

Spring, 2018

Outline

Next Section ...

1. Semiparametric: Cox proportional hazard regression
   - Cox Model inference
2. Cox PH Model Assessment
3. Competing Risk

Semiparametric modeling

- A semiparametric model contains two portions:
  - Nonparametric portion: The underlying survival distribution.
  - Parametric portion: The way in which covariates affect that underlying distribution.
- Two popular semiparametric models for survival data regression
  - Proportional hazard (PH) framework
  - Accelerated failure time (AFT) framework

Semiparametric Models: Introduction

- For the AFT model:

$$Y_i = x_i^T \beta + W_i, \text{ where } Y_i = \log T_i$$

- The parametric aspect of the model lies in the $x_i^T \beta$ portion, while the nonparametric aspect involves assuming that $W_i \overset{\text{i.i.d}}{\sim} F$, where F is some generic, unspecified distribution.
- That is, β should be inferred without depending on a specific distribution F.

Cox proportional hazard regression

- Proposed by David Cox (1972)
- The model is specified in terms of the hazard function instead of the survival function
- Assumption of changes in the concomitant variable corresponding to multiplicative changes in the hazard function.
- Additive changes in the log of hazard function.

$$h(t|X) = \exp(X\beta)h_0(t)$$

where

- X is a vector of concomitant, covariate or regressor information $X = (x_1, x_2, \ldots, x_p)$
- β is the column vector of parameters
- $\exp(X\beta)$ is the **parametric portion**.
- $h_0(t)$ (**nonparametric portion**) is the time-specific baseline hazard function, referred to as a homogeneous or baseline hazard function.

$$\frac{h(t|X_1)}{h(t|X_2)} = \frac{\exp(X_1\beta)}{\exp(X_2\beta)}$$

Special case: Two observations

- Assume that we have only two observations, T_1 and T_2, with hazard functions $\lambda_1(t)$ and $\lambda_2(t)$
- Suppose that the first failure occurs at time t, how likely is the failure subject 1?
- **Proposition**

$$\mathbb{P}(T_1 = t | T_{(1)} = t) = \frac{\lambda_1(t)}{\lambda_1(t) + \lambda_2(t)}$$

- Under the PH assumption, we have

$$\mathbb{P}(T_1 < T_2) = \frac{\exp(x_1^T \beta)}{\exp(x_1^T \beta) + \exp(x_2^T \beta)}$$

  Now the baseline hazard, $\lambda_0(t)$, is canceled out.
- This can be extended to multiple subjects **without censoring**:

$$\mathbb{P}(T_1 < T_2 < \cdots < T_J) = \prod_{j=1}^{J} \frac{\exp x_j^T \beta}{\sum_{k=j}^{J} \exp(x_k^T \beta)}$$

PH model with censoring

- The probability that subject j fails at time t given that one of the subjects from the risk set $R(t)$ failed at time t is

$$\frac{\exp(x_j^T \beta)}{\sum_{k \in R(t)} \exp(x_k^T \beta)}$$

- Therefore, the full likelihood is

$$L(\beta) = \prod_j \frac{\exp(x_j^T \beta)}{\sum_{k \in R(t)} \exp(x_k^T \beta)}$$

- This is not exactly a likelihood, but a **partial likelihood**.
- But the partial likelihood, **Cox partial likelihood** still yields a score with mean zero and a variance given by the negative Hessian matrix.

Cox proportional hazard regression estimator

The estimation is based on the principle of **maximum partial likelihood** through working out the score vector and Hessian matrix and then applying an **iterative Newton-Raphson procedure**.

$$L(\beta) = \prod_j \frac{\exp(x_j^T \beta)}{\sum_{k \in R(t_j)} \exp(x_k^T \beta)}$$

- The denominator can be written as

$$\sum_{k=1}^{n} Y_k(t_j) \exp(x_k^T \beta)$$

where

$$Y_i(t) = \begin{cases} 1 & \text{subject } i \text{ is at risk at time } t \\ 0 & \text{otherwise} \end{cases}$$

Cox's PH Model Inference

Let

$$w_j = \exp(\mathbf{x}^T \beta), W_i = \sum_{j \in R_i} w_j$$

then

$$L(\beta) = \prod_i \left\{ \frac{w_i}{\sum_{R_i} w_j} \right\}^{d_i}$$

where w_i is the relative probability of failure for subject i.

With these, we can compute the absolute probability of failure for subject i at time t_j:

$$\pi_{ij} = Y_i(t_j) \frac{w_i}{W_j}$$

Score function

The partial log-likelihood is:

$$\begin{aligned} l &= \sum_i d_i \log w_i - \sum_i d_i \log W_i \\ &= \sum_i d_i \eta_i - \sum_i d_i \log W_i \end{aligned}$$

Note that W_i contains many η terms in addition to η_i.

Then the score equation is obtained by maximizing the log-likelihood function:

$$\frac{\partial l}{\partial \beta} = 0$$

we start by

$$\frac{\partial l}{\partial \eta_k} = d_k - \sum_i \pi_{ki} d_i$$

which can be written as the matrix-form:

$$\mathbf{u}(\eta) = \mathbf{d} - \mathbf{P}\mathbf{d}$$

where $\mathbf{P} \in \mathbb{R}^{n \times n} = (\pi_{ij})_{n \times n}$

Then we can write:

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial \eta} \frac{\partial \eta}{\partial \beta} = \mathbf{X}^{\mathbf{T}}(\mathbf{d} - \mathbf{P}\mathbf{d})$$

Hessian function

The hessian function can be obtained by:

$$\begin{aligned} \frac{\partial^2 l}{\partial \eta_k^2} &= -\sum_i d_i \pi_{ki}(1-\pi_{ki}) \\ \frac{\partial^2 l}{\partial \eta_k \partial \eta_j} &= \sum_i d_i \pi_{ki}\pi_{ji} \end{aligned}$$

and the Hessian matrix can be computed as

$$\begin{aligned} H(\beta) &= -\mathbf{X^T W X} \\ &= -\sum_j \sum_k \pi_{kj}(x_k - E_j x)(x_k - E_j x)^T \end{aligned}$$

Newton-Raphson Updates

$$\hat{\beta}^{(m+1)} = (\mathbf{X^T W X})^{-1}\mathbf{X^T}(\mathbf{d} - \mathbf{Pd}) + \hat{\beta}^{(m)}$$

```
for (i in 1:m){
        eta <- X %*% b
        haz <- as.numeric(exp(eta)) # w[i]
        rsk <- rev(cumsum(rev(haz))) # W[i]
        P <- outer(haz, rsk, '/')
        P[upper.tri(P)] <- 0
        W <- - P %*% diag(d) %*% t(1-P)
        diag(W) <- diag(P %*% diag(P) %*% t(1-P))
        b <- solve(t(X) %*% W %*% X) %*% t(X) %*% (d - P%
            *%d) + b
}
```

The above R code assumes that the data has been sorted by time on study, and assumes that no ties are present.

Fit Cox Regression Model

```
# fit Cox
fit.Cox <- coxph(Surv(Time, Status==0) ~ TRT, data=carcinoma)
summary(fit.Cox)


## Call:
## coxph(formula = Surv(Time, Status == 0) ~ TRT, data = carcinoma)
##
##   n= 31, number of events= 14
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## TRTS+CT+IT -1.0855    0.3377   0.7156 -1.517    0.129
## TRTS+IT    -0.5482    0.5780   0.6082 -0.901    0.367
##
##            exp(coef) exp(-coef) lower .95 upper .95
## TRTS+CT+IT    0.3377      2.961   0.08308     1.373
## TRTS+IT       0.5780      1.730   0.17547     1.904
##
## Concordance= 0.599  (se = 0.08 )
## Rsquare= 0.077   (max possible= 0.933 )
## Likelihood ratio test= 2.49  on 2 df,   p=0.3
## Wald test            = 2.41  on 2 df,   p=0.3
## Score (logrank) test = 2.57  on 2 df,   p=0.3
```

Fit Cox Regression Model with Covariate

```
# fit Cox
fit.Cox.age <- coxph(Surv(Time, Status==0) ~ TRT + Age, data=carcinoma)
summary(fit.Cox.age)


## Call:
## coxph(formula = Surv(Time, Status == 0) ~ TRT + Age, data = carcinoma)
##
##   n= 31, number of events= 14
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## TRTS+CT+IT -0.86214   0.42226  0.71845 -1.200  0.23014
## TRTS+IT    -0.29422   0.74511  0.61201 -0.481  0.63070
## Age         0.11251   1.11908  0.04125  2.728  0.00638 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## TRTS+CT+IT    0.4223     2.3682    0.1033     1.726
## TRTS+IT       0.7451     1.3421    0.2245     2.473
## Age           1.1191     0.8936    1.0322     1.213
##
## Concordance= 0.769  (se = 0.086 )
## Rsquare= 0.314   (max possible= 0.933 )
## Likelihood ratio test= 11.68  on 3 df,   p=0.009
```

Cox model inference

- Wald test
- Score test
- Likelihood ratio test (LRT)

Wald test

- Wald inference is based on the asymptotic result:

$$\hat{\beta} \dot{\sim} N(\beta, (X^T W X)^{-1}),$$

- The confidence interval are constructed using $\hat{\beta} \pm z_{\alpha/2} \sqrt{I_{jj}^{-1}}$

Likelihood ratio test

Let $\hat{\beta}_0$ denote the fit of the first model and $\hat{\beta}_1$ denote the fit of the second model, the likelihood ratio test (which is only requires fitting two nested models) is based on

$$2(l(\hat{\beta}_1) - l(\hat{\beta}_0)) \dot{\sim} \chi^2_k$$

where k is the difference of number of parameters for the two models.

Score test

The score test statistic for testing $H_0 : \beta = 0$ is

$$\begin{aligned} u(0) &= \sum_j (x_j - E_j x) \\ &= \sum_j \left(d_{1j} - d_j \frac{n_{1j}}{n_j} \right) \end{aligned}$$

The score test is in some sense equivalent to the log-rank test, although the variances are calculated differently and therefore do not produce the exact same p-value.

coxph function

- The function for fitting Cox proportional hazards models in `survival` package.
- The syntax is similar to other model-fitting functions in R:

```
fit <- coxph(Surv(time, status) ~ treatment + stage +
    hepato, pbc)
```

- Several functions on the `coxph` object:
  - `coef(fit)`: return the MLE of the coefficient vector.
  - `vcov(fit)`: return the inverse of the information matrix $(\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X})^{-1}$
  - `model.matrix(fit)`: return X.
  - The Wald, score, LRT tests are testing the global hypotheses $H_0 : \beta = 0$
  - `anova(fit0, fit1)`
  - `loglik(fit), AIC(fit), BIC(fit)`
  - `predict(fit)`

Next Section ...

1. Semiparametric: Cox proportional hazard regression
   - Cox Model inference
2. Cox PH Model Assessment
3. Competing Risk

Assesssment of the Cox Model

The Cox (PH) model:

$$\lambda(t|Z(t)) = \lambda_0(t)\exp(Z(t)\beta)$$

Assumption of this model:

(1) the regression effect of β is constant over time (PH assumption)

(2) linear combination of the covariates (including possibly higher order terms, interactions)

(3) the link function is exponential

Using residuals for checking assumptions

In the regression analysis, the residuals are often used to check the assumptions. However, residuals for survival data are somewhat different from those for the other type of data models, mainly due to the existence of censoring:

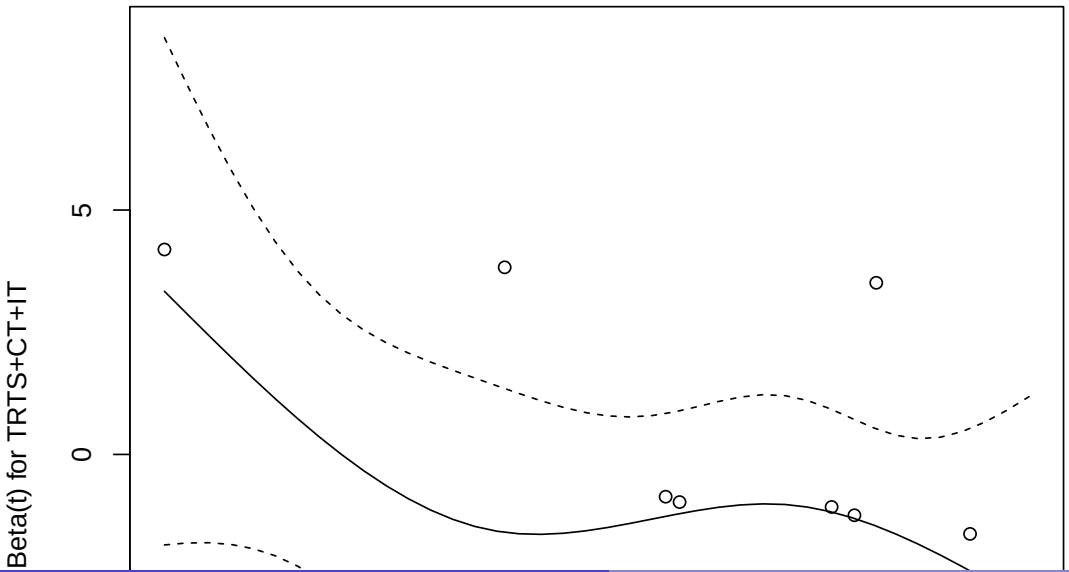
- Generalized (Cox-Snell) residuals
- **Schoenfeld residuals**
- Martingale residuals

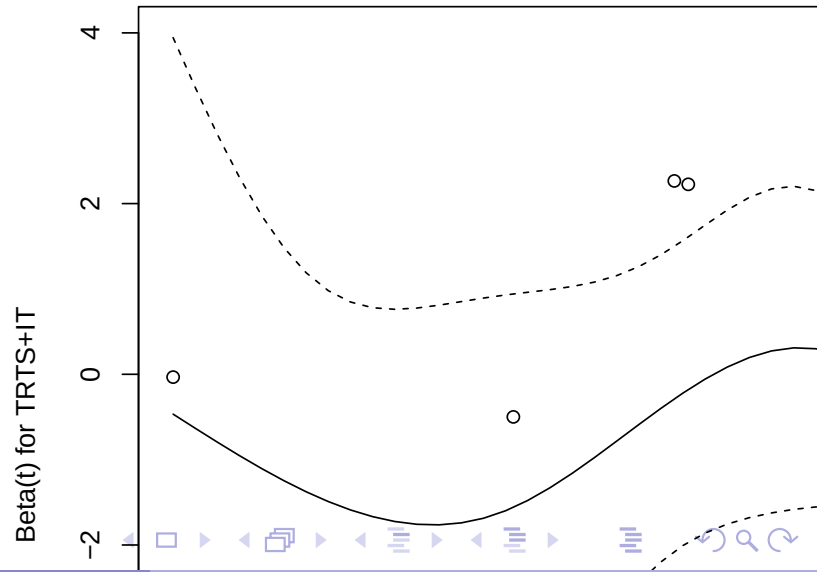
cox.zph function

```
check <- cox.zph(fit.Cox, transform="log", global=TRUE)
print(check)


##                rho chisq     p
## TRTS+CT+IT -0.385 2.278 0.131
## TRTS+IT     0.124 0.206 0.650
## GLOBAL         NA 3.581 0.167


plot(check)
```





Summary

Test of proportionality. The `cox.zph` function will test proportionality of all the predictors in the model by creating interactions with time using the transformation of time specified in the transform option. In this example we are testing proportionality by looking at the interactions with **log(time)**. The column **rho** is the Pearson product-moment correlation between the scaled Schoenfeld residuals and **log(time)** for each covariate. The last row contains the global test for all the interactions tested at once. A *p*-value less than 0.05 indicates a violation of the proportionality assumption.

Next Section ...

1. Semiparametric: Cox proportional hazard regression
   - Cox Model inference
2. Cox PH Model Assessment
3. Competing Risk

Stem cell transplant for acute leukemia

In this clinical trial, 117 acute leukemia patients received stem cell transplant. The aim of the analysis was to estimate the cumulative incidence of relapse in the presence of transplant-related death, which deals with competing events. The effect on relapse of predictive factors and covariates such as Sex, Disease (lymphoblastic or myeloblastic leukemia), Phase at transplant (Relapse, CR1, CR2, CR3), Source of stem cells (bone marrow and peripheral blood (BM+PB), or peripheral blood (PB), and Age will be evaluated.

Importing the data

```
bmt <- read.csv("data/bmtcrr.csv", header=TRUE)
head(bmt)
```

```
##   Sex   D   Phase Age Status Source  ftime
## 1   M ALL Relapse  48      2  BM+PB   0.67
## 2   F AML     CR2  23      1  BM+PB   9.50
## 3   M ALL     CR3   7      0  BM+PB 131.77
## 4   F ALL     CR2  26      2  BM+PB  24.03
## 5   F ALL     CR2  36      2  BM+PB   1.47
## 6   M ALL Relapse  17      2  BM+PB   2.23
```

Type-specific hazard

- Assume that we have multiple failure types
- Let T denote the time until failure, and J indicate the type of failure
- The type-specific hazard is then defined as

$$\lambda_j(t) = \lim_{h \to 0} \frac{\mathcal{P}(t \leq T < t + h, J = j | T \geq t)}{h}$$

- Then the overall hazard is

$$\lambda(t) = \sum_j \lambda_j(t)$$

- Therefore the overall survival becomes

$$S(t) = \exp\left\{ - \int_0^t \lambda(s) ds \right\}$$

Recoding the covariates

```
attach(bmt)


## The following objects are masked from bmt (pos = 3):
##
##     Age, D, ftime, Phase, Sex, Source, Status

## The following objects are masked from bmt (pos = 4):
##
##     Age, D, ftime, Phase, Sex, Source, Status

## The following objects are masked from bmt (pos = 5):
##
##     Age, D, ftime, Phase, Sex, Source, Status

## The following objects are masked from bmt (pos = 6):
##
##     Age, D, ftime, Phase, Sex, Source, Status

## The following objects are masked from bmt (pos = 7):
##
##     Age, D, ftime, Phase, Sex, Source, Status

## The following objects are masked from bmt (pos = 8):
##
##     Age, D, ftime, Phase, Sex, Source, Status

## The following objects are masked from bmt (pos = 9):
```

Competing risk modeling

The main function to fit regression models for competing risks data is `crr()`, which is contained in the `cmprsk` package. In the simplest form it requires

- a vector of follow-up times,
- a vector of status with a code for each failure type or censoring,
- a matrix of fixed covariates.

By default, the censoring code for status is set by the optional argument `cencode=0`, and the code that denotes the failure type of interest is set by the optional argument `failcode=1`.

In our example, transplant-related death, which is the competing event, is coded with 2.

Regresison model for relapse

```
mod1 <- crr(ftime, Status, x)
summary(mod1)


## Competing Risks Regression
##
## Call:
## crr(ftime = ftime, fstatus = Status, cov1 = x)
##
##               coef exp(coef) se(coef)      z p-value
## Age        -0.0185     0.982   0.0119 -1.554  0.1200
## Sex:F      -0.0352     0.965   0.2900 -0.122  0.9000
## D:AML      -0.4723     0.624   0.3054 -1.547  0.1200
## Phase:CR1 -1.1018     0.332   0.3764 -2.927  0.0034
## Phase:CR2 -1.0200     0.361   0.3558 -2.867  0.0041
## Phase:CR3 -0.7314     0.481   0.5766 -1.268  0.2000
## Source:PB  0.9211     2.512   0.5530  1.666  0.0960
##
##           exp(coef) exp(-coef)  2.5% 97.5%
## Age           0.982      1.019 0.959 1.005
## Sex:F         0.965      1.036 0.547 1.704
## D:AML         0.624      1.604 0.343 1.134
## Phase:CR1     0.332      3.009 0.159 0.695
## Phase:CR2     0.361      2.773 0.180 0.724
## Phase:CR3     0.481      2.078 0.155 1.490
```

Analysis of Results

- The first part of the output shows for each term in the design matrix
  - the estimated coefficient $\hat{\beta}_j$
  - the relative risk $\exp(\hat{\beta}_j)$
  - the standard error
  - the z-value
  - the corresponding P-value
- For Sex,
  - $\hat{\beta}_{female} = -0.0352$
  - relative risk 0.965 is the ratio of risk for female w.r.t. the male, with all other covariates equal.
  - p-value of 0.9000 indicates no significant effect.
  - for age, the relative risk of 0.982 is the relative risk for a 1 year increase in age.