

# PH 716 Applied Survival Analysis

## Part VII: Stratified Cox PH Model

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2024/Mar/24 16:41:13

### Recall the Cox PH model

- Observed  $\tilde{T}_i = \tilde{t}_i$  and  $\Delta_i = \delta_i$
- $T_i$  are independent across  $i$ , given covariates  $x_{i1}, \dots, x_{ip}$
- The independent and non-informative censoring
- $\lambda_{T_i}(t) = \lambda_0(t) \exp(\sum_{j=1}^p x_{ij}\beta_j)$ , or equiv.  $\ln \lambda_{T_i}(t) = \ln \lambda_0(t) + \sum_{j=1}^p x_{ij}\beta_j$ 
  - $\lambda_0(t)$ : the common baseline hazard

### Ex. 7.1: veterans' administration lung cancer study

- Randomized trial of two treatment regimens for lung cancer.
  - **trt**: 1=standard vs. 2=test.
  - **celltype**: type of cancer (squamous, smallcell, adeno, large).
  - **time**: survival time in days from the start of the study.
  - **status**: Indicator of whether the patient died (event occurred) or was censored at the end of the study.
  - **karno**: Karnofsky score, a measure of the patient's functional status, assessed on a scale from 0 to 100.
  - **age**: enrollment age.
  - **prior**: Indicator of whether the patient had received therapy before the study (0=no, 1=yes).

```
options(digits=4)
library(survival)
sapply(veteran, class) # print out data types of columns
veteran$trt = as.factor(veteran$trt)
veteran$prior = as.factor(veteran$prior)
fit.ex71 <- coxph(Surv(time, status) ~trt+celltype+age+prior, data=veteran, x=T)

# Schoenfeld residual plots
plot(cox.zph(fit.ex71, transform="identity", terms=F, global=F))
```

- Indicating a potential violation of PH assumption due to **celltype**

### Stratified Cox PH Model relaxing the assumption of common baseline hazard

- $n_s$  subjects in the  $s$ th stratum,  $s = 1, \dots, S$
- Observed  $\tilde{T}_{si} = \tilde{t}_{si}$  and  $\Delta_{si} = \delta_{si}$ ,  $i = 1, \dots, n_s$ ,  $s = 1, \dots, S$
- $T_{is}$  are independent from each other, given covariates  $x_{si1}, \dots, x_{sip}$

- The independent and non-informative censoring
- $\lambda_{T_{si}}(t) = \lambda_{s0}(t) \exp(\sum_{j=1}^p x_{sij} \beta_j)$ , or equiv.  $\ln \lambda_{T_{si}}(t) = \ln \lambda_{s0}(t) + \sum_{j=1}^p x_{sij} \beta_j$ 
  - $\lambda_{s0}(t)$ : the baseline hazard for the  $s$ th stratum
    - \* Assuming PH within each stratum BUT NOT across strata
  - No relationship assumed among  $\lambda_{10}(t), \dots, \lambda_{S0}(t)$

## Partial likelihood and log-partial likelihood

- Partial likelihood  $pL(\beta) = pL_1(\beta) \times \dots \times pL_S(\beta)$ 
  - $\beta = [\beta_1, \dots, \beta_p]^\top$
  - $pL_s(\beta)$ : the partial likelihood in stratum  $s$  (solely based on data for those subjects in stratum  $s$ )
- Log-partial likelihood  $p\ell(\beta) = p\ell_1(\beta) + \dots + p\ell_S(\beta)$ 
  - $p\ell_s(\beta) = \ln pL_s(\beta)$ : the log-partial likelihood in stratum  $s$  (solely based on data for those subjects in stratum  $s$ )

## Ex 7.2: Revisit the veterans' administration lung cancer study

```
options(digits=4)
library(survival)
veteran$trt = as.factor(veteran$trt)
veteran$prior = as.factor(veteran$prior)
veteran$large = as.factor(veteran$celltype == 'large')
fit.ex72 <- coxph(Surv(time, status) ~trt+age+prior+strata(large), data=veteran, x=T)

# Schoenfeld residual plots
plot(cox.zph(fit.ex72, transform="identity", terms=F, global=F))
```

## Survival function

- $\hat{S}_{T_{si}}(t) = \exp\{-\hat{\Lambda}_{s0}(t)\}^{\exp(\sum_{j=1}^p x_{sij} \hat{\beta}_j)} = \hat{S}_{s0}(t)^{\exp(\sum_{j=1}^p x_{sij} \hat{\beta}_j)}$ 
  - $\hat{S}_{s0}(t) = \exp\{-\hat{\Lambda}_{s0}(t)\}$

## Plotting survival functions for Ex. 7.2

```
# Baseline hazards and baseline survivals
baseline <- basehaz(fit.ex72, centered = FALSE)
names(baseline)[1] = 'cum.haz' # clarify the first column
baseline$surv = exp(-baseline$cum.haz)
baseline

# Plot baseline survival functions
newdata.ex72 <- data.frame(
  trt = factor(c(1,1)),
  age = c(0, 0),
  prior = factor(c(0,0)),
  large = factor(c(FALSE, TRUE))
)
baseline.surv.curve = survfit(
  fit.ex72,
  newdata=newdata.ex72,
  conf.type = 'log-log'
)
```

```

plot(
  baseline.surv.curve,
  xlab="Time", ylab="Estimated Survival Probability",
  lty=1:nrow(newdata.ex72), col=1:nrow(newdata.ex72), lwd=2,
)
legend(
  "topright",
  c(
    "Baseline survival for large=F",
    "Baseline survival for large=T"
  ),
  lty=1:nrow(newdata.ex72), col=1:nrow(newdata.ex72), lwd=2
)

# Predict the survival probability at specific times
newdata.ex72 <- data.frame(
  trt = factor(c(1,2)),
  age = c(40, 50),
  prior = factor(c(0,10)),
  large = factor(c(FALSE, TRUE))
)
newdata.ex72
cox.predicted.survival = survfit(
  fit.ex72,
  newdata=newdata.ex72,
  conf.type = 'log-log'
)
summary(cox.predicted.survival, times=c(20,30)) # Survival at times 20 and 30

# Plot the survival function with given values of covariates
plot(
  cox.predicted.survival,
  xlab="Time", ylab="Estimated Survival Probability",
  conf.int = .95,
  lty=1:nrow(newdata.ex72), col=1:nrow(newdata.ex72), lwd=2,
)
legend(
  "topright",
  c(
    "trt=1, age=40, prior=0, large=F",
    "trt=2, age=50, prior=10, large=T"
  ),
  lty=1:nrow(newdata.ex72), col=1:nrow(newdata.ex72), lwd=2
)

```

## Pros and cons

- Merely stratifying on categorical covariates
- Cannot estimate the effect of stratified factor using standard methods
- Requiring a large sample size and number of events within each stratum
  - May result in inaccuracy if stratifying too finely

- May speed up estimation considerably for large data sets
  - Since only within-stratum comparisons are made, sums and integrals will be totaled over much smaller numbers of subjects
  - So, for an extremely huge data set, stratification may be preferred even if the PH assumption is known to hold