PH 716 Applied Survival Analysis

Part VII: Stratified Cox PH Model

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

2024/Apr/02 20:29:22

Recall the Cox PH model

- Observed $\widetilde{T}_i = \tilde{t}_i$ and $\Delta_i = \delta_i$
- T_i are independent across i, given covariates $x_{i1}, \ldots, x_{ip}$
- 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, 10=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
par(mfrow=c(2,3))
plot(cox.zph(fit.ex71, transform="identity", terms=F, global=F))

# Score tests for PH assumption for each covariate
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 sth stratum, $s = 1, \ldots, S$

- Observed $\widetilde{T}_{si} = \tilde{t}_{si}$ and $\Delta_{si} = \delta_{si}$, $i = 1, \ldots, n_s$, $s = 1, \ldots, S$
- T_{is} are independent from each other, given covariates $x_{si1}, \ldots, x_{sip}$
- 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 sth stratum
    - Assuming PH within each stratum BUT NOT across strata
  - No relationship assumed among $\lambda_{10}(t), \ldots, \lambda_{S0}(t)$

Partial likelihood and log-partial likelihood

- Partial likelihood $pL(\boldsymbol{\beta}) = pL_1(\boldsymbol{\beta}) \times \cdots \times pL_S(\boldsymbol{\beta})$
  - $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$
  - $pL_s(\boldsymbol{\beta})$: the partial likelihood in stratum s (solely based on data for those subjects in stratum s)
- Log-partial likelihood $p\ell(\boldsymbol{\beta}) = p\ell_1(\boldsymbol{\beta}) + \cdots + p\ell_S(\boldsymbol{\beta})$
  - $p\ell_s(\boldsymbol{\beta}) = \ln pL_s(\boldsymbol{\beta})$: 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
par(mfrow=c(2,2))
plot(cox.zph(fit.ex72, transform="identity", terms=F, global=F))

# Score tests for PH assumption for each covariate
cox.zph(fit.ex72, transform="identity", terms=F, global=F)
```

Survival function

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

Plotting survival functions for Ex. 7.2

```
# Baseline cumulative 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.ex72.1 <- 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.ex72.1,
  conf.type = 'log-log'
)
plot(
  baseline.surv.curve,
  xlab="Time", ylab="Estimated Survival Probability",
  lty=1:nrow(newdata.ex72.1), col=1:nrow(newdata.ex72.1), lwd=2,
)
legend(
  "topright",
  c(
    "Baseline survival function for large=F",
    "Baseline survival function for large=T"
  ),
  lty=1:nrow(newdata.ex72.1), col=1:nrow(newdata.ex72.1), lwd=2
)
## Or
survminer::ggsurvplot(
  baseline.surv.curve,
  data = veteran,
  conf.int = F,
  censor = F,
  legend = c(.75,.95), # legend position
  legend.title = '',
  legend.labs = c(
    "Baseline survival function for large=F",
    "Baseline survival function for large=T"
  )
)

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

# Plot survival functions with given values of covariates
plot(
  predicted.surv.curv,
  xlab="Time", ylab="Estimated Survival Probability",
```

```
  conf.int = .95,
  lty=1:nrow(newdata.ex72.2), col=1:nrow(newdata.ex72.2), 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.ex72.2), col=1:nrow(newdata.ex72.2), lwd=2
)
## Or
survminer::ggsurvplot(
  predicted.surv.curv,
  data = veteran,
  conf.int = T,
  censor = F,
  legend = c(.75,.95), # legend position
  legend.title = '',
  legend.labs = c(
    "trt=1, age=40, prior=0, large=F",
    "trt=2, age=50, prior=10, large=T"
  )
)
```

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