# BSTT536: Survival Data Analysis

Instructor: Hua Yun Chen, PhD

Division of Epidemiology and Biostatistics
School of Public Health
University of Illinois at Chicago

Table of Content

Generalization of PPH: Stratified PPH

Generalization of PPH: Time-dependent covariates

Estimation and Inference

Examples

# Stratified proportional hazards model

1. If we suspect the effect of a variable is not proportional, including it in the proportional hazards model may affect the estimation of the effect of other variables.
2. For example, the center of a multicenter clinic trial. patients population in different centers may be heterogeneous in terms of their baseline survival.
3. One way to approach this problem is to use the stratified proportional hazard model.

# Comparison to the proportional hazards model

1. The stratified proportional hazards model assumes

$$h(t \mid w, z) = h(t \mid w) \exp(\beta z),$$

   where $h(t \mid w)$ is unspecified.
2. The proportional hazards model assumes further that

$$h(t \mid w) = h_0(t) \exp(\alpha W),$$

   or a similar model.
3. A further simplification would assume that $W$ has no effect on survival, i.e.,

$$h(t \mid w) = h_0(t).$$

# Stratified analysis

1. Under the stratified proportional hazards model, analysis may be performed stratum-by-stratum using the proportional hazards model in each stratum.
2. Problem with this approach is that individual stratum can have a very small sample size.
3. Parameters shared across strata can be estimated by pooling subjects across different strata to increase the power of detection.

# Partial likelihood

1. Suppose that, for a given stratum $w = k$, the partial likelihood is

$$L_k(\beta) = \prod_{i=1}^{n_k} \left\{ \frac{\exp(\beta Z_{ik})}{\sum_{X_{jk} \geq X_{ik}} \exp(\beta Z_{jk})} \right\}^{\delta_{ik}},$$

   where $(X_{ik}, \delta_{ik}, Z_{ik}), i = 1, \cdots, n_k$ are the observed data in stratum $k$.

2. the pooled partial likelihood is

$$L(\beta) = \prod_{k=1}^{K} L_k(\beta).$$

3. Estimation and inference can be based on the pooled likelihood.

# Fit the stratified proportional hazards model in SAS

1. The basic SAS statement for fitting a stratified PH model is

```
Proc phreg;
        model X*d(0)=Z;
        strata W;
run;
```

2. In contrast, the basic SAS statement for fitting a PH model is

```
Proc phreg;
        class W;
        model X*d(0)=Z W;
run;
```

# Pros and cons for using stratified PH model versus PH model

1. Pros: Stratified PH model requires less assumptions and is therefore more robust.
2. Cons: Stratified PH model is less powerful in detecting the covariate effect when the porportional hazards model holds for the stratification variable.
3. Interpretation of the relative risk parameter remains unchanged.

# Time-dependent covariates in Cox regression model

1. Time-dependent covariates: $Z(t)$, $t \geq 0$.
2. Internal time-dependent covariates:can be affected by the subject's survival status. Such as the measure of a bio-marker.
3. External time-dependent covariates: not be affected by the subject's survival status. Such as the weather.
4. Synthetic time dependent covariates: such as a time-independent covariate multiplied by time.

# Cox regression model with time-dependent covariates

1. Hazards density model

$$h(t|z) = h_0(t)\exp\{\beta Z(t)\}.$$

2. The hazard ratio depends on time

$$\frac{h(t|Z_1)}{h(t|Z_2)} = \exp[\beta\{Z_1(t) - Z_2(t)\}].$$

3. Interpretation of $\exp(\beta)$: If $Z_1(t) - Z_2(t) = 1$, then

$$\frac{h(t|Z_1)}{h(t|Z_2)} = \exp(\beta).$$

   At the given time point $t$, one unit increase in $Z(t)$ results in $\exp(\beta)$ times increase in the hazard of failure at time $t$.

# Estimation and inference with time-dependent covariates

1. Observed $\{X_i, \delta_i, Z_i(t), t \leq X_i\}$, $i = 1, \cdots, n$. The partial likelihood is

$$\prod_{i=1}^{n} \left[ \frac{\exp\{\beta Z_i(X_i)\}}{\sum_{\{X_j \geq X_i\}} \exp\{\beta Z_j(X_i)} \right]^{\delta_i}.$$

2. Maximize the partial likelihood to obtain the parameter estimate for $\beta$.
3. The variance can be estimated by the inverse of the minus second derivative matrix of the log-partial likelihood.

# Baseline hazard and survival function estimation

1. Breslow estimator for the jump at failure time $X_k$.

$$\hat{h}_k = \frac{d_k}{\sum_{\{X_j \geq X_k\}} \exp\{\beta Z_j(X_k)\}}.$$

2. Baseline cumulative hazard estimator,

$$\hat{H}_0(t) = \sum_{\{k|X_k \leq t\}} \hat{h}_k.$$

3. Baseline survival function estimator,

$$\hat{S}_0(t) = \prod_{\{k|X_k \leq t\}} (1 - \hat{h}_k).$$

# Other hazard and survival function estimation

1. For a subject with covariate $\{Z(t), t > 0\}$, the predicted cumulative hazard is

$$\hat{H}(t|Z) = \sum_{\{k|X_k \leq t\}} \hat{h}_k \exp\{\hat{\beta} Z(X_k)\}.$$

2. Baseline survival function estimator,

$$\hat{S}(t|Z) = \prod_{\{k|T_k \leq t\}} \left[1 - \hat{h}_k \exp\{\beta Z(X_k)\}\right].$$

3. An alternative one is

$$\hat{S}(t|Z) = \exp\{-\hat{H}(t|Z)\}.$$

# Requirements on Data Structure in Cox Regression Model

1. The ideal situation is to continuously record the covariate values of a subject until the subject is failed or censored. This is however burdensome.
2. For a subject censored at time $t$, we need the covariate values of this subject at all the failure times observed in the sample up to before time $t$.
3. For a subject failed at time $t$, we need the covariate values of this subject at all the failure times observed in the sample up to and include time $t$.
4. In general, whenever a failure occurs in the sample, all at risk subjects are needed to have their covariate value recorded at the time point.

# Stanford Heart Transplant Data (partial)

| id | date of birth | date of accept | date of transplant | date last seen | dead |
|---|---|---|---|---|---|
| 1 | 1/10/37 | 11/15/67 | NA | 1/3/68 | 1 |
| 16 | 5/16/19 | 10/26/68 | 11/22/68 | 8/29/69 | 1 |
| 39 | 11/12/19 | 5/20/70 | 5/21/70 | 7/11/70 | 1 |

| id | prior surgery | Number of mismatch | HLA-A2 | mismatch score | reject |
|---|---|---|---|---|---|
| 1 | 0 | NA | NA | NA | - |
| 16 | 0 | 2 | 0 | 1.12 | 1 |
| 39 | 0 | NA | NA | NA | - |

# Stanford Heart Transplant Data Summary

1. Events: Acceptance, transplant, death.
2. Even Times:
   - Time from acceptance to transplant,
   - Time from transplant to death,
   - Time from acceptance to death.
3. Which event time to use in the analysis depends on the research questions.

# Model Stanford Heart Transplant Data Model

1. Major question: How does transplant change the survival of a patient ?
2. Event time: Time between acceptance to death.
3. Transplant is a time-dependent covariate

$$Z(t) = \begin{cases} 0 & \text{if the patient not yet received transplant at time } t, \\ 1 & \text{if the patient received transplant at time } t. \end{cases}$$

4. Cox regression model

$$h(t|Z) = h_0(t) \exp\{\beta Z(t)\}.$$

age at acceptance, previous surgery, etc. may also be added to the covariate list.

# Model Stanford Heart Transplant Data Analysis

1. The time-dependent covariate is programmed in the PROC PHREG through the waiting time variable.
2. Age at acceptance was not found to significantly affect the survival of the patients. It is excluded from the second model.
3. The second model includes Xstatus (the transplant status), Xage (age at the transplant), and Xscore (the mismatch score at the transplant). All are time-dependent covariates and are programmed in the PROC PHREG through the waiting time variable.

# Fit Cox regression model with time-dependent covariates using SAS: I

Model 1: Cox regression model with the transplant status as a time dependent covariate (`Xstat`) and age at the acceptance to the waiting list as a time-independent covariate (`AccAge`).

```
proc phreg data= Heart;
    model Time*Status(0)= Xtrans AccAge;
    if (WaitTime = . or Time < WaitTime) then Xtrans=0.;
    else Xtrans= 1.0;
run;
```

Age at acceptance was not found to significantly affect the survival of the patients. It is excluded from the next model.

# Fit Cox regression model with time-dependent covariates using SAS: II

Model 2: Cox regression model with the transplant status (`Xstatus`) and the age at transplant (`Xage`) as time-dependent covariates, and the mismatch score as time-independent covariate.

```
proc phreg data= Heart;
    model Time*Status(0)= Xtrans XAge Score;
    where NotTyped .NE. 'y';
    if (WaitTime = . or Time < WaitTime) then do;
    Xtrans=0.; XAge=0.;
    end;
    else do;
    Xtrans= 1.0; XAge= XplAge;
    end;
run;
```

Those who misses mismatch score are excluded from the analysis.

# Time-dependent covariates in repeated measurements form

1. Data were from an experiment to study the dosing effect of a tumor-promoting agent.
2. Rodents were randomly assigned to three dose groups.
3. After the first death, the rodents are examined **every week** for the number of papillomas.
4. Failure times are (27,34,37,41,43,45,46,47,49,50,51,53,65,67,71) in days.
5. Data include ID, Survival time of the subject, censoring status, dose group, and P1-P15 representing the results of the examination on the number of papillomas at the 15 weeks where deaths were observed in those weeks .

# Analysis of the data

1. Event time: From the treatment of the agent to the death from cancer.
2. Model:

$$h(t|dose, Npap) = h_0(t) \exp\{\beta_1 dose + \beta_2 Npap(t)\}.$$

   where `Npap(t)` is the number of papilomas at time $t$.
3. The number of papillomas is a time-dependent covartiates. This variable is not continuously observed, nor it is observed at the exact time point where failure occurred.
4. In order to apply the Cox regression model, we need to interpolate the missing values.

# Using SAS to fit the model with repeatedly measured time-dependent covariate

```
proc phreg data=Tumor;
    model Time*Dead(0)=Dose NPap;
    array pp* P1-P14; array tt* t1-t15;
    t1 = 27; t2 = 34; t3 = 37; t4 = 41; t5 = 43; t6 = 45;
    t7 = 46; t8 = 47; t9 = 49; t10= 50; t11= 51; t12= 53;
    t13= 65; t14= 67; t15= 71;
    if Time < tt[1] then NPap=0;
    else if time >= tt[15] then NPap=P15;
    else do i=1 to dim(pp);
    if tt[i] <= Time < tt[i+1] then NPap= pp[i];
    end;
run;
```