Syllabus

1. Introduction
   - Survival data
   - Censoring mechanism
   - Application in medical field
2. Concepts and definitions
   - Survival function
   - Hazard function
3. Non-parametric approach
   - Life table
   - Kaplan-Meier survival estimate
   - Hazard function
   - Median and percentile survival time
4. Hypothesis testing
   - Overview – hypothesis, test statistics, p-values
   - Log-rank
   - Wilcoxon
   - Gehan test
5. Study design and sample size estimation
   - Overview
   - Survival sample size estimation
   - Accrual time and Study duration
6. **Semiparametric model – proportional hazard model**
   - **Partial likelihood**
   - **Inference**
   - Time varying covariates
   - Prediction
   - Stratification
7. PH model checking
   - Model checking
   - Residuals
8. Parametric model
   - Parametric proportional hazard model
   - Accelerate failure model
9. Other topics
   - Competing risk
   - Recurrent events
   - Non-proportional hazard ratio
   - Interval censoring

Survival Data With Covariates

- Why do we want to include covariates
  - To understand the impact of covariates on survival function
    - For both categorical and continuous variables
    - To estimate the magnitude of the impact
    - To quantify the confidence of the impact
  - To control confounding factors
  - To predict survival

- Covariates can be
  - Continuous
  - Discrete
  - Time varying

Example – Covariates in the Ovarian Data Set

Dataset available in R survival package

- futime: survival or censoring time (day)
- fustat: censoring status (censor=0)
- age: in years
- resid.ds: residual disease present (1=no,2=yes)
- rx: treatment group
- ecog.ps: ECOG performance status (1 is better, see reference)

	futime	fustat	age	resid.ds	rx	ecog.ps
1	59	1	72.3315	2	1	1
2	115	1	74.4932	2	1	1
3	156	1	66.4658	2	1	2
4	421	0	53.3644	2	2	1
5	431	1	50.3397	2	1	1
6	448	0	56.4301	1	1	2
7	464	1	56.9370	2	2	2
8	475	1	59.8548	2	2	2
9	477	0	64.1753	2	1	1
10	563	1	55.1781	1	2	2
11	638	1	56.7562	1	1	2
12	744	0	50.1096	1	2	1
13	769	0	59.6301	2	2	2
14	770	0	57.0521	2	2	1
15	803	0	39.2712	1	1	1
16	855	0	43.1233	1	1	2
17	1040	0	38.8932	2	1	2
18	1106	0	44.6000	1	1	1
19	1129	0	53.9068	1	2	1
20	1206	0	44.2055	2	2	1
21	1227	0	59.5890	1	2	2
22	268	1	74.5041	2	1	2
23	329	1	43.1370	2	1	1
24	353	1	63.2192	1	2	2
25	365	1	64.4247	2	2	1
26	377	0	58.3096	1	2	1

The Proportional Hazards (PH) Regression Model

- PH model - AKA as Cox proportional hazards model (Cox, JRSSB 1972)

- Survival data (T, Δ, Z), the hazard function can be written as the following based on the Cox model

$$h(t|Z = z) = h_0(t)e^{\beta' z}$$

  - where $h_0(t)$ is a baseline hazard function,
    Z can a vector of p covariates,
    β is a vector of p coefficients

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Baseline Hazard Function $h_0(t)$

A couple of notes on baseline hazard function $h_0(t)$

1. $h(t|Z = z) = h_0(t)e^{\beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p}$

   $h_0(t)$ is equivalent to the intercept of the regression model

2. Hazard ratio

$$h(t|Z = \phi_1) = h_0(t)e^{\beta' \phi_1}$$
$$h(t|Z = \phi_2) = h_0(t)e^{\beta' \phi_2}$$

$$\frac{h(t|Z = \phi_1)}{h(t|Z = \phi_2)} = e^{\beta'(\phi_1 - \phi_2)}$$

The hazard ratio does not depend upon the baseline hazard function

Understanding the Coefficient

Let Z be a univariate covariate and take values of 0,1

$$h(t|Z=1) = h_0(t)e^{\beta} \qquad h(t|Z=0) = h_0(t)$$

$$\frac{h(t|Z=1)}{h(t|Z=0)} = e^{\beta}$$

$$\beta = \log\frac{h(t|Z=1)}{h(t|Z=0)}$$

Interpretation

- β is the log hazard ratio
- e^{β} is the hazard ratio
- $1 - e^{\beta}$ represents risk reduction if $Z = 0$ represents the control arm.

Understanding the Coefficient

Let Z be a continuous univariate covariate

$$\frac{h(t|Z = z + 1)}{h(t|Z = z)} = e^{\beta}$$

$$\beta = \log \frac{h(t|Z = z + 1)}{h(t|Z = z)}$$

β is the log of the hazard ratio with unit change in Z

Example: Ovarian Data

- Let us consider rx and ecog in the Cox model
  - Rx – treatment indicator
  - Ecog – ECOG performance status (1 is better than 2)

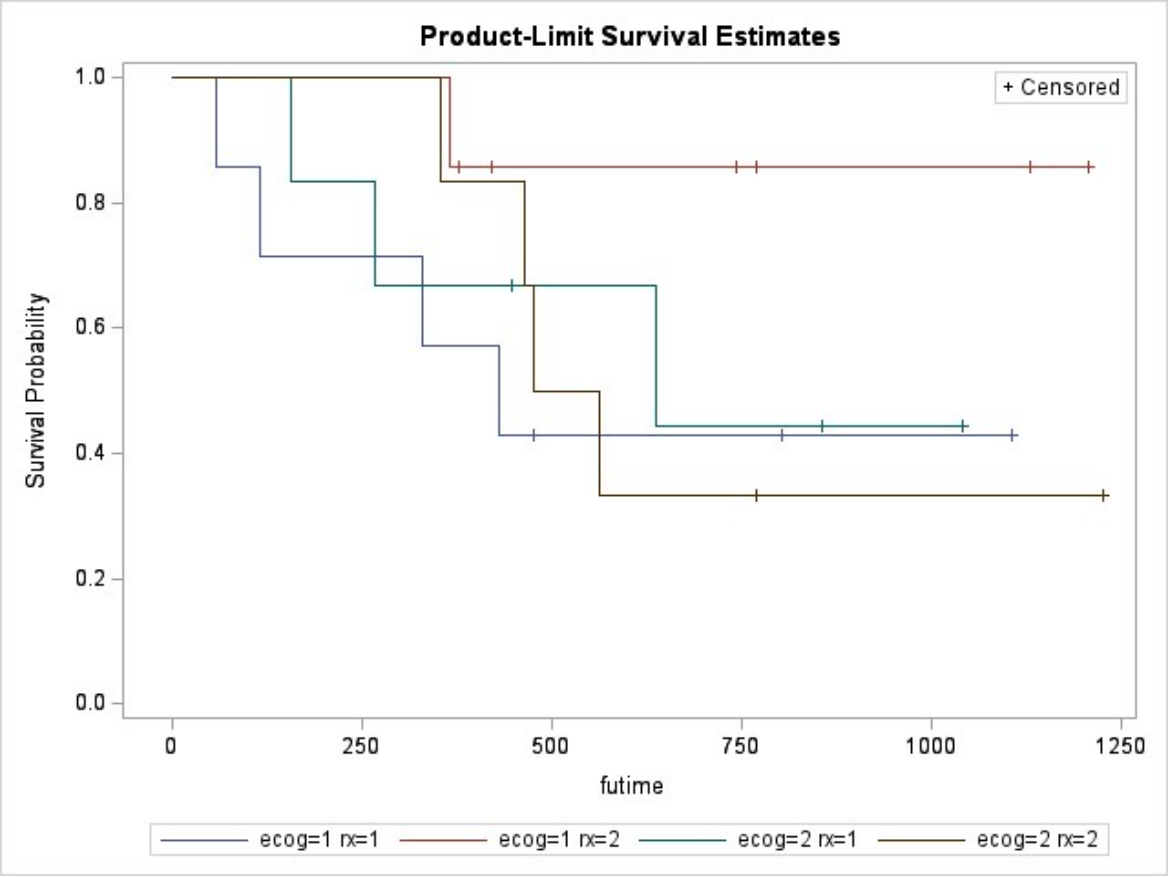
- Models

$$\log h(t|rx, ecog) = \log h_0(t) + \beta_1 rx + \beta_2 ecog$$

  - $rx = 1\,, ecog = 1, \quad \log h(t|rx, ecog) = \log h_0(t) + \beta_1 + \beta_2$
  - $rx = 1\,, ecog = 2, \quad \log h(t|rx, ecog) = \log h_0(t) + \beta_1 + 2\beta_2$
  - $rx = 2\,, ecog = 1, \quad \log h(t|rx, ecog) = \log h_0(t) + 2\beta_1 + \beta_2$
  - $rx = 2\,, ecog = 2, \quad \log h(t|rx, ecog) = \log h_0(t) + 2\beta_1 + 2\beta_2$

- The log hazard ratio between
  - Stratum $rx = 1\,, ecog = 1$ and
  - Stratum $rx = 2\,, ecog = 2$
  - is $-(\beta_1 + \beta_2)$

Example: Ovarian Cancer



Why Cox PH Model

- Since Cox proposed the model in 1972

$$h(t|Z = z) = h_0(t)e^{\beta' z}$$

- It has been the most popular and commonly used regression model for survival analyses
  - Easy to use
  - Flexible for covariate
  - Well-understood operating characteristics
  - Can be interpreted even assumption does not hold

- Nonetheless, it has limitations
  - Not easy to handle tied events
  - Small bias when randomization is not 1:1
  - Low power when PH assumption is violated

Type I error always inflate a little bit. 0.025 → 0.028.

- Let us move on to the inference

Notations

- Survival data in n subjects
  - (T_i, Δ_i, Z_i) $i = 1, 2, \dots, n$
  - Z_i a vector of covariates
  - $T_i = \min(X_i, C_i)$
  - $\Delta_i = I(X_i \leq C_i)$
- Number of events in both groups: r in J distinct event time
  - Ordered distinct event time: $t_{(1)} < t_{(2)} < \cdots < t_{(J)}$
  - $j = 1, 2, 3, \dots, J \leq r$
  - $z_{(j)}$ - covariates of the subject who experienced the j^{th} event
- Define risk set as $R(t) = \{\text{set of subjects } I(T_i \geq t)\}$

Partial Likelihood

- Assuming no tie at the distinct event time
- The likelihood at the j^{th} event for the subject who experienced the event at $t_{(j)}$ given $R(t_{(j)})$ can be written as

$$L_j(\beta) = P_r(j^{th} \text{ event at } t_{(j)} | R(t_{(j)}))$$

$$= \frac{h(t_{(j)}|z_{(j)})}{\sum_{l \in R(t_{(j)})} h(t_{(j)}|z_l)}$$

$$= \frac{h_0(t_{(j)}) e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} h_0(t_{(j)}) e^{\beta' z_l}}$$

$$= \frac{e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta' z_l}}$$

Partial Likelihood

- The partial likelihood can be written as

$$L_P(\beta) = \prod_{j=1}^{J} \frac{e^{\beta' z_{(j)}}}{\sum_{l \in R(t_{(j)})} e^{\beta' z_l}}$$

events index

$$= \prod_{i=1}^{n} \left\{ \frac{e^{\beta' z_i}}{\sum_{l \in R(t_i)} e^{\beta' z_l}} \right\}^{\Delta_i}$$

Subject index

$\Delta_i = \begin{cases} 1 \rightarrow \text{event} \\ 0 \rightarrow \text{censor} \end{cases}$

When i^{th} subject meet a event

$e^{\beta' z_i} = e^{\beta' z_{(j)}}$

$t_i = t_j$.

Recover the above formula

Example – Construct Partial Likelihood

Survival data

- Group 0: 6+,7,9+,10,11+,13,16+,17+, 20+ $n_0 = 9$
- Group 1: 4, 5+, 8+,11+,12,15,17+,22+,23+ $n_1 = 9$
- $Z = 0,1$
- 7 Distinct event time: 4, 7, 10, 12, 13, 15,

$$L_P(\beta) = \frac{e^{\beta}}{9e^{\beta}+9} \cdot \frac{1}{7e^{\beta}+8} \cdot \frac{1}{6e^{\beta}+6} \cdot \frac{e^{\beta}}{5e^{\beta}+4} \cdot \frac{1}{4e^{\beta}+4} \cdot \frac{e^{\beta}}{4e^{\beta}+3}$$

Risk sets

Group 0	6+,7,9+, 10,11+,13, 16+,17+, 20+	7,9+,10, 11+,13, 16+,17+, 20+	10, 11+, 13,16+, 17+, 20+	13,16+, 17+, 20+	13,16+, 17+, 20+	16+, 17+, 20+
Group 1	4, 5+, 8+,11+, 12,15,17+, 22+,23+	8+,11+, 12,15,17+, 22+,23+	11+,12, 15,17+, 22+,23+	12,15, 17+,22+, 23+	15,17+, 22+,23+	15,17+,22+, 23+
Event Time	4	7	10	12	13	15

$e^{\beta z} = \begin{cases} e^{\beta} & \text{Group 1} \\ 1 & \text{Group 0} \end{cases}$

Why Partial Likelihood

- For censored data, likelihood consists observed data (with events) and partially observed data (censored)
- For observed events (complete)

$$L_O(\beta) = \prod_{i=1}^{n} f(T_i)^{\Delta_i} \longrightarrow \text{censor indicator}$$

$$= \prod_{i=1}^{n} (h(T_i)S(T_i))^{\Delta_i}$$

subject who $\Delta_i = 0$ (censored, not contribute to this part)

- For incompletely observed data (incomplete)

$$L_C(\beta) = \prod_{i=1}^{n} S(T_i)^{1-\Delta_i}$$

Subject who has a $\Delta_i = 1$ (complete) observe contribute to this part.

- The complete likelihood

$$L(\beta) = L_O(\beta)L_C(\beta)$$

$$= \prod_{i=1}^{n} h(T_i)^{\Delta_i} S(T_i)$$

$\Delta_i = 1$ $\quad h(T_i)S(T_i)$

$\Delta_i = 0$ $\quad S(T_i)$

Why Partial Likelihood

- Rewrite the full likelihood

$$L(\beta) = \prod_{i=1}^{n} h(t_i|z_i)^{\Delta_i} S(T_i)$$

$$= \prod_{i=1}^{n} \left\{ \frac{h(t_i|z_i)}{\sum_{l \in R(t_i)} h(t_i|z_l)} \right\}^{\Delta_i} \left\{ \sum_{l \in R(t_i)} h(t_i|z_l) \right\}^{\Delta_i} S(T_i)$$

$$= L_P(\beta) \prod_{i=1}^{n} \left\{ \sum_{l \in R(t_i)} h(t_i|z_l) \right\}^{\Delta_i} S(T_i)$$

- Cox argued that there is very little information on β beyond $L_P(\beta)$ in the full likelihood

Log-Partial Likelihood

$$l_P(\beta) = \log \prod_{i=1}^{n} \left\{ \frac{e^{\beta' z_i}}{\sum_{l \in R(t_i)} e^{\beta' z_l}} \right\}^{\Delta_i}$$

$$= \sum_{i=1}^{n} \Delta_i \left\{ \beta' z_i - \log \left\{ \sum_{l \in R(t_i)} e^{\beta' z_l} \right\} \right\}$$

$$= \sum_{i=1}^{n} l_i(\beta)$$

Note, $l_i(\beta)$ are not independent!

Score Function

- The score function

$$U(\beta) = \frac{\partial}{\partial \beta} l_P(\beta)$$

chain rule.

$$= \sum_{i=1}^{n} \Delta_i \left\{ z_i - \frac{\sum_{l \in R(t_i)} z_l e^{\beta' z_l}}{\sum_{l \in R(t_i)} e^{\beta' z_l}} \right\}$$

$$= \sum_{i=1}^{n} \Delta_i \{ z_i - \bar{Z}_i \}$$

for eig $\longrightarrow$ link with log-rank test.

Where $\bar{Z}_i = \frac{\sum_{l \in R(t_i)} z_l e^{\beta' z_l}}{\sum_{l \in R(t_i)} e^{\beta' z_l}}$, weighted average of the Z in the risk set $R(t_i)$

Information Definition

$$I(\beta) = -\left\{\frac{\partial^2}{\partial\beta^2} l_P(\beta)\right\}$$

$$= \sum_{i=1}^{n} \Delta_i \left\{\frac{\sum_{l\in R(t_i)} z_l \otimes z_l e^{\beta' z_l}}{\sum_{l\in R(t_i)} e^{\beta' z_l}} - \left(\frac{\sum_{l\in R(t_i)} z_l e^{\beta' z_l}}{\sum_{l\in R(t_i)} e^{\beta' z_l}}\right)^2\right\}$$

$$= \sum_{i=1}^{n} \Delta_i \left\{\frac{S_2}{S_0} - \frac{S_1 S_1'}{S_0{}^2}\right\}$$

where

$$S_0 = \sum_{l\in R(t_i)} e^{\beta' z_l}, \qquad S_1 = \bar{Z}_i = \sum_{l\in R(t_i)} z_l e^{\beta' z_l}, \qquad \text{and} \quad S_2 = \sum_{l\in R(t_i)} z_l z_l' e^{\beta' z_l}$$

Note, S_0 is a scaler, S_1 is a*Kx1* vector, and S_2 is a *KxK* matrix

K x K information matrix.

MLE

- MLE $\hat{\beta}$ can be obtained from the score function

$$U(\hat{\beta}) = 0$$

$$Var(\hat{\beta}) = I(\hat{\beta})^{-1}$$

Inference

- Estimation and confidence intervals
- Hypothesis testing,
  - Hazard ratio <1 represents risk reduction
  - $\beta < 0$ mean risk reduction

$$H_0: \beta \geq 0 \quad vs \quad H_A: \beta < 0$$

- Association
- Prediction

Confidence Intervals

- 95% CI confidence interval for the coefficient estimates

$$\hat{\beta} \pm z_{0.975} se(\hat{\beta}) = \hat{\beta} \pm 1.96 se(\hat{\beta})$$

$\hat{\beta}$ is a unit log hazard ratio, to covert to unit hazard ratio, lower and upper bound of the 95% CI for the unit hazard ratio is

$$\left[e^{\hat{\beta} - 1.96 se(\hat{\beta})}, e^{\hat{\beta} + 1.96 se(\hat{\beta})} \right]$$

$e^{LB(\hat{\beta})}$, $e^{UB(\hat{\beta})}$

Wald Test

- For components of $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$,

$H_0: \quad \beta_1 = \beta_2 = \cdots = \beta_p = 0$

- A Wald test for one covariate is

$$Z_i = \frac{(\hat{\beta}_i - \beta_i)}{\sqrt{I(\hat{\beta}_i)^{-1}}} \sim N(0,1)$$

$$for\ i = 1, \ldots, p$$

A composite test can be written as

$$\chi_p^2 = \hat{\beta}' I(\hat{\beta})^{-1} \hat{\beta}$$

Score Test

Under null H_0: $\beta = 0$, score test

$$U(0)/\sqrt{I(0)} \sim N(0,1)$$

$$U(0) = \sum_{i=1}^{n} \Delta_i \left\{ z_i - \frac{\sum_{l \in R(t_i)} z_l e^{\beta' z_l}}{\sum_{l \in R(t_i)} e^{\beta' z_l}} \right\}$$

$$= \sum_{i=1}^{n} \Delta_i \left\{ z_i - \frac{\sum_{l \in R(t_i)} z_l}{\sum_{l \in R(t_i)} 1} \right\}$$

$$= \sum_{i=1}^{n} \Delta_i \left\{ z_i - \frac{\sum_{l \in R(t_i)} z_l}{n_i} \right\}$$

Likelihood Ratio Test

- Let full covariates be

$$Z^{K\prime} = (Z_1, Z_2, \ldots, Z_k, Z_{k+1}, \ldots, Z_K)$$

- Subset of covariates for $k < K$

$$Z^{k\prime} = (Z_1, Z_2, \ldots, Z_k)$$

- The full model is

$$L_P(\beta^K | Z^K) = \prod_{i=1}^{n} \left\{ \frac{exp(\beta_1 Z_{i1} + \beta_2 Z_{i2} + \cdots + \beta_K Z_{iK})}{\sum_{l \in R(t_i)} exp(\beta_1 Z_{l1} + \beta_2 Z_{l2} + \cdots + \beta_K Z_{lK})} \right\}^{\Delta_i}$$

- The sub-model is

$$L_P(\beta^k | Z^k) = \prod_{i=1}^{n} \left\{ \frac{exp(\beta_1 Z_{i1} + \beta_2 Z_{i2} + \cdots + \beta_k Z_{ik})}{\sum_{l \in R(t_i)} exp(\beta_1 Z_{l1} + \beta_2 Z_{l2} + \cdots + \beta_K Z_{lk})} \right\}^{\Delta_i}$$

Likelihood Ratio Test

- To test

$$H_0: \beta_{k+1} = \beta_{k+2} = \cdots = \beta_K = 0$$

- The likelihood Ratio test is

$$\Lambda = \frac{L_P\left(\hat{\beta}^K | Z^K\right)}{L_P\left(\hat{\beta}^k | Z^k\right)}$$

$$-2\log\Lambda = -2\left\{\log L_P\left(\hat{\beta}^K | Z^K\right) - \log L_P\left(\hat{\beta}^k | Z^k\right)\right\} \sim \chi^2_{(K-k)}$$

Example – Melanoma Data

- Variable names
  - Time – survival time in days
  - Status – 1=died from melanoma, 2=alive, 3=dead from other causes
  - Sex – 1=male, 0=female
  - Age – age in years
  - Year – operation
  - Thickness – tumor thickness in mm
  - Ulcer – 1=presence; 0=absence

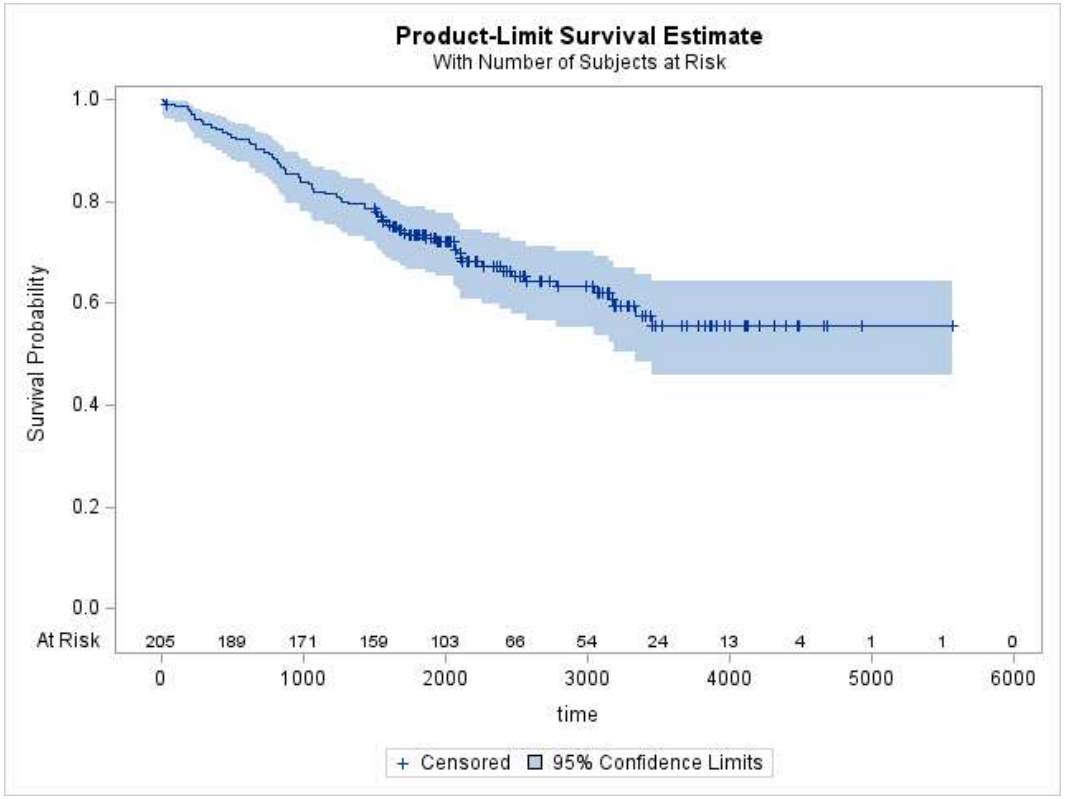
```
> head(Melanoma)
  time status sex age year thickness ulcer
1   10      3   1  76 1972      6.76     1
2   30      3   1  56 1968      0.65     0
3   35      2   1  41 1977      1.34     0
4   99      3   0  71 1968      2.90     0
5  185      1   1  52 1965     12.08     1
6  204      1   1  28 1971      4.84     1
>
```

- Source
  - P. K. Andersen, O. Borgan, R. D. Gill and N. Keiding (1993) Statistical Models based on Counting Processes. Springer.
  - Library(MASS)

SAS Code

```
ods graphics on;
proc lifetest data=example method=KM plots=survival (cl atrisk=0 to 6000 by
500) outsurv=survival;
  time time*status(2);
  run;
ods graphics off;
proc phreg data=example;
  model time*status(2)=age sex thickness;
  output out=Outp xbeta=Xb resmart=Mart resdev=Dev;
run;
```

Example - Melanoma



Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
205	71	134	65.37

Example - Melanoma

Interpretation:

Overall tests – reject the null that there is no difference in age, sex, and tumor thickness

- Likelihood Ratio
- Score
- Wald

Ward tests shows

- The risk of death increased 2% with 1 year old
- The risk in male is 67% higher than that of female
- The risk increased 15% with 1 unit increase in tumor thickness

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	700.985	666.615
AIC	700.985	672.615
SBC	700.985	679.403

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.3703	3	<.0001
Score	41.8566	3	<.0001
Wald	38.2646	3	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	0.02221	0.00795	7.8071	0.0052	1.022	age
sex	1	0.51242	0.23877	4.6056	0.0319	1.669	sex
thickness	1	0.13499	0.03048	19.6188	<.0001	1.145	thickness

Example - Melanoma

Model fit statistics

- A general form of information criteria (IC) is IC(c) = - 2logL(M) + c*p
  where p is the number of covariates and c is a penalty parameter.

Akaike Information Criterion

$$AIC = -2\log L + c * p$$

Schwarz Bayesian (Information) Criterion

$$SBC = -2\log L + p\log\left(\sum_j f_j \Delta_j\right)$$

Δ_j- event indicator

f_j - frequency

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	700.985	666.615
AIC	700.985	672.615
SBC	700.985	679.403

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.3703	3	<.0001
Score	41.8566	3	<.0001
Wald	38.2646	3	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	0.02221	0.00795	7.8071	0.0052	1.022	age
sex	1	0.51242	0.23877	4.6056	0.0319	1.669	sex
thickness	1	0.13499	0.03048	19.6188	<.0001	1.145	thickness

Compare to the Log-rank Test

2x2 table at the j^{th} event time

Group	Events occurred at $t_{(j)}$	Number of subjects Survival at $t_{(j)}^{+}$	Number of subject at risk at $t_{(j)}^{-}$
0	d_{0j}	$n_{0j} - d_{0j}$	n_{0j}
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
Total	d_j	$n_j - d_j$	n_j

The Log-rank Test

- The total deviation from null

$$L = \sum_{i=1}^{k} (d_{0i} - e_{0i})$$

$e_{0i} = d_i \cdot \frac{n_{1i}}{n_i}$

- Variance

$$\text{var}(L) = \text{var}\left(\sum_{i=1}^{k} (d_{0i} - e_{0i})\right) \approx \sum_{i=1}^{k} var(d_{0i})$$

From hypergeometric distribution

$$\text{var}(d_{0i}) = \frac{n_{0i} n_{1i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

$$\text{var}(L) = \sum_{i=1}^{k} \frac{n_{0i} n_{1i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

$$\frac{L}{\sqrt{\text{var}(L)}} \sim N(0,1)$$

Compared To the Log-rank Test

Consider a special case of two sample problem and no ties

Under null H_0: $\beta = 0$, score test

$$U(0)/\sqrt{I(0)} \sim N(0,1)$$

$$U(0) = \sum_{i=1}^{n} \Delta_i \left\{ z_i - \frac{\sum_{l \in R(t_i)} z_l}{\sum_{l \in R(t_i)} 1} \right\}$$

$$= \sum_{j=1}^{J} \left(z_j d_j - d_j \frac{n_{1j}}{n_j} \right)$$

$$= \sum_{j=1}^{J} \left(d_{1j} - d_j \frac{n_{1j}}{n_j} \right)$$

$$= \sum_{j=1}^{J} \left(d_{1j} - e_{1j} \right)$$

Compare to the Log-rank Test

When $d_j = 1$, no tie

$$I(\beta) = \sum_{i=1}^{n} \Delta_i \left\{ \frac{\sum_{l \in R(t_i)} z_l{}^2 e^{\beta' z_l}}{\sum_{l \in R(t_i)} e^{\beta' z_l}} - \left(\frac{\sum_{l \in R(t_i)} z_l e^{\beta' z_l}}{\sum_{l \in R(t_i)} e^{\beta' z_l}} \right)^2 \right\}$$

$$I(0) = \sum_{i=1}^{n} \Delta_i \left\{ \frac{\sum_{l \in R(t_i)} z_l{}^2}{\sum_{l \in R(t_i)} 1} - \left(\frac{\sum_{l \in R(t_i)} z_l}{\sum_{l \in R(t_i)} 1} \right)^2 \right\}$$

$$= \sum_{j=1}^{J} d_j \left\{ \frac{n_{1j}}{n_j} - \left(\frac{n_{1j}}{n_j} \right)^2 \right\} = \mathrm{var}(L)$$

There score test

$$\frac{U(0)}{\sqrt{I(0)}} = \frac{L}{\sqrt{\mathrm{var}(L)}} \sim N(0,1)$$

Compared To Log-rank

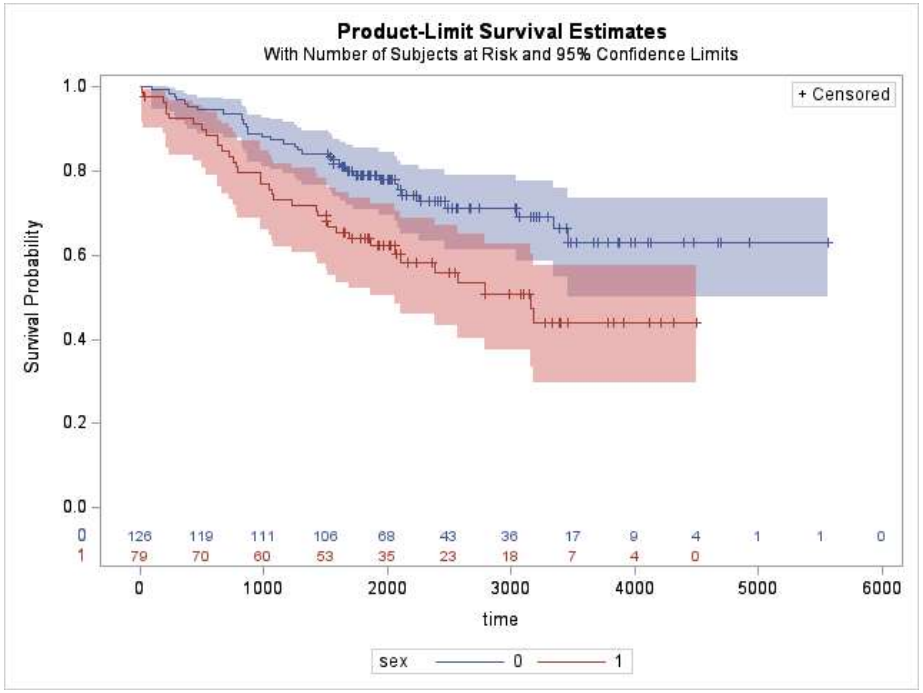
- Score test is exact the same test as the log-rank
- Both have the optimal power when hazard ratio is constant

- Advantages of PH model
  - Provide the estimate of hazard ratio and CI in addition to p-values
  - Adjust confounding factors in
    - Hypothesis test
    - Understanding association
  - Fit multiple covariates
  - Can be used for prediction

Example Melanoma

- Test difference between male and female
- Sex – 1=male, 0=female

- K-M survival curves suggest that females survival longer than males



Example Melanoma

- Test difference between male and female
- Sex – 1=male, 0=female
- Males' risk doubles female's

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	7.8965	1	0.0050
Wilcoxon	7.9688	1	0.0048
-2Log(LR)	7.4974	1	0.0062

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	700.985	693.475
AIC	700.985	695.475
SBC	700.985	697.738

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.5102	1	0.0061
Score	7.8953	1	0.0050
Wald	7.6190	1	0.0058

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
sex	1	0.65586	0.23761	7.6190	0.0058	1.927	sex

Example - Melanoma

- See the difference of the sex effect in different model?

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	700.985	693.475
AIC	700.985	695.475
SBC	700.985	697.738

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.5102	1	0.0061
Score	7.8953	1	0.0050
Wald	7.6190	1	0.0058

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
sex	1	0.65586	0.23761	7.6190	0.0058	1.927	sex

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	700.985	666.615
AIC	700.985	672.615
SBC	700.985	679.403

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	34.3703	3	<.0001
Score	41.8566	3	<.0001
Wald	38.2646	3	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
age	1	0.02221	0.00795	7.8071	0.0052	1.022	age
sex	1	0.51242	0.23877	4.6056	0.0319	1.669	sex
thickness	1	0.13499	0.03048	19.6188	<.0001	1.145	thickness

Example - Melanoma

Variable		n	Mean (SD)	Min	Max
age	Male	79	53.90 (17.61)	12.00	95.00
	Female	126	51.56 (16.06)	4.00	89.00
thickness	Male	79	2.49 (2.75)	0.16	14.67
	Female	126	3.61 (3.12)	0.10	17.42

When There Are Ties

- The set-up of the partial likelihood are based on one event only at distinct survival time
- When there are ties, the following approaches are available to adjust ties
  - Discrete method – Cox's modification (1972)
  - Breslow method – 1972
  - Efron method
  - Exact method – Kalbfleisch and Prentice (1973)

- Reference

Breslow, NE (1972). Contribution to the discussion of Cox (1972). Journal of the Royal Statistical Society B, 34: 216-217.

Cox DR (1972). Regression models and life tables (with Discussion). Journal of the Royal Statistical Society B, 34:187-220.

Kalbfleisch JD and Prentice RL (1973). Marginal likelihoods based on Cox's regression and life model. Biometrika, 60: 267-278.

Notations

- Survival data in n subjects
  - $(T_i, \Delta_i, Z_i)\ \ i = 1, 2, \dots, n$
  - Z_i a vector of covariates
  - $T_i = \min(X_i, C_i)$
  - $\Delta_i = I(X_i \leq C_i)$

- Number of events in both groups: r in J distinct event time
  - Ordered distinct event time: $t_{(1)} < t_{(2)} < \cdots < t_{(J)}$
  - $j = 1, 2, 3, \dots, J \leq r$
  - d_j – the number of events at $t_{(j)}$
  - $i_{j1}, \dots, i_{jd_j}$ - subject index at $t_{(j)}$ who has an event

Notation

- Define combination risk set at $t_{(k)}$ as
  - $CR(t_{(j)}) = \{\text{sets } d_j \text{of subjects } I(T_i \geq t_{(j)})\}$

- Ex.: Risk set at $t_{(2)}$
  - $R(t_{(2)}) = \{1,2,3,4,5\}$
  - Event set at $t_{(2)} = \{4,5\}, d_2 = 2$ =number of event at $t_{(2)}$
  - $CR(t_{(2)}) = \{(1,2), (1,3), (1,4), (1,5), (2,3), (2,4), (2,5), (3,4), (3,5), (4,5)\}$

Discrete Method

- The partial likelihood

$$L_P(\beta) = \prod_{j=1}^{J} P_r\left(d_j \; events \middle| the \; number \; of \; sets \; in \; \{CR(t_{(j)})\}\right)$$

$$= \prod_{j=1}^{J} \frac{P_r\left(\left(d_j \; events \middle| R(t_{(j)})\right)\right)}{\sum_{l \in \{CR(t_{(j)})\}} P_r\left(\left(d_j \; events \; in \; set \; l \middle| R(t_{(j)})\right)\right)}$$

$$= \prod_{j=1}^{J} \frac{e^{\beta z_{j1}} e^{\beta z_{j2}} \ldots e^{\beta z_{jd_j}}}{\sum_{l \in \{CR(t_{(j)})\}} e^{\beta z_{l1}} e^{\beta z_{l2}} \ldots e^{\beta z_{ld_j}}}$$

$$= \prod_{j=1}^{J} \frac{\exp(\beta \sum_{c=1}^{d_j} z_{jc})}{\sum_{l \in \{CR(t_{(j)})\}} \exp(\beta \sum_{c=1}^{d_j} z_{lc})}$$

Exercise – Partial Likelihood With Ties

Survival data

- Group 0: 6+, 7, 9+,10, 11+, 13, 15, 17+, 20+ $n_0 = 9$
- Group 1: 4, 5+, 8+, 11+, 12, 15, 17+, 22+, 23+ $n_1 = 9$
- $Z = 0,1$
- 7 Distinct event time: 4, 7, 10, 12, 13, 15

$$L_P(\beta) = \frac{e^{\beta}}{9e^{\beta}+9} \cdot \frac{1}{7e^{\beta}+8} \cdot \frac{1}{6e^{\beta}+6} \cdot \frac{e^{\beta}}{5e^{\beta}+4} \cdot \frac{1}{4e^{\beta}+4} \cdot \frac{e^{\beta}}{6e^{2\beta}+12e^{\beta}+3}$$

Risk sets	Group 0	6+,7,9+, 10,11+,15,17+, 20+	7,9+,10, 11+,13,15,17+, 20+	10, 11+, 13,15,17+, 20+	13,15, 17+, 20+	13, 15, 17+, 20+	15,17+, 20+ (15,17), (15,20),(17,20)
Combination sets	Group 1	4, 5+, 8+,11+, 12,15,17+,22+,23+	8+,11+, 12,15,17+, 22+,23+	11+,12, 15,17+, 22+,23+	12,15, 17+,22+,23+	15,17+, 22+,23+	15,17+,22+, 23+ (15,17),(15,22),(15,23),(17,22), (17,23) ,(22,23)
							(15,15),(15,17), (15,22),(15,23), (17,15),(17,17), (17,22),(17,23), (20,15),(20,17), (20,22), (20,23)
	Event Time	5	7	10	12	13	15

Exact Method

- The ties are caused by the fact that the observation time interval is not fine enough
- If reduce the interval, the ties follow certain oder
- The exact method
  - Permutt all possible orders of the tied events d_j

Exercise –Exact Method

Survival data

- Group 0: 6+, 7, 9+,10, 11+, 13, 15, 17+, 20+ $n_0 = 9$
- Group 1: 4, 5+, 8+, 11+, 12, 15, 17+, 22+, 23+ $n_1 = 9$
- $Z = 0,1$
- 7 Distinct event time: 4, 7, 10, 12, 13, 15

$$L_P(\beta) = \frac{e^{\beta}}{9e^{\beta}+9} \cdot \frac{1}{7e^{\beta}+8} \cdot \frac{1}{6e^{\beta}+6} \cdot \frac{e^{\beta}}{5e^{\beta}+4} \cdot \frac{1}{4e^{\beta}+4} \cdot \left(\frac{1}{4e^{\beta}+3} \cdot \frac{e^{\beta}}{4e^{\beta}+2} + \frac{e^{\beta}}{4e^{\beta}+3} \cdot \frac{1}{4e^{\beta}+2}\right)$$

Breslow Method

- Proposed an approximation
- In partial likelihood, replace

$$\sum_{l \in \{CR(t_{(j)})\}} \exp(\beta \sum_{c=1}^{d_j} z_{lc})$$

in the discrete method with

$$\sum_{l \in \{R(t_{(j)})\}} \exp(\beta z_l)^{d_j}$$

Efron's Method

- Another approximation
- Preferred method when there are large number of ties

$$L_P(\beta) = \prod_{j=1}^{J} \frac{\exp(\beta \sum_{c=1}^{d_j} z_{jc})}{\prod_{c=1}^{d_j} \left(\sum_{l \in \{R(t_{(j)})\}} \exp(\beta z_l) - \frac{c-1}{d_j} \sum_{l \in \{D(t_{(j)})\}} \exp(\beta z_l) \right)}$$

Which Methods to Use?

- When the number of ties is small, all methods yield similar results
- When ties are large
  - Discrete – if the number of ties or sample size are large, the combination sets can be computation intensive
  - Efron's method is preferred
  - Breslow method may produce larger bias towards null

SAS Code For Tied Events

```
proc phreg data=example;
  class trt;
  model time*event(0)=trt/ties=discrete;
  run;
```

```
proc phreg data=example;
  class trt;
  model time*event(0)=trt/ties=exact;
  run;
```

```
proc phreg data=example;
  class trt;
  model time*event(0)=trt/ties=efron;
  run;
```

```
proc phreg data=example;
  class trt;
  model time*event(0)=trt/ties=breslow;
  run;
```

Example - Leukemia

Analysis of Maximum Likelihood Estimates - discrete								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
trt	6-MP	1	-1.62822	0.43313	14.1316	0.0002	0.196	trt 6-MP

Analysis of Maximum Likelihood Estimates - Exact								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
trt	6-MP	1	-1.59787	0.42162	14.3630	0.0002	0.202	trt 6-MP

Analysis of Maximum Likelihood Estimates - Efron								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
trt	6-MP	1	-1.57213	0.41240	14.5326	0.0001	0.208	trt 6-MP

Analysis of Maximum Likelihood Estimates - Breslow								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
trt	6-MP	1	-1.50919	0.40956	13.5783	0.0002	0.221	trt 6-MP

Homework 6

1. A trial is designed to evaluate a biomarker reduction post-treatment from baseline. The biomarker is a continuous variable. The reduction in the geometric mean is 33% with 95% CI (15%,47%) reported in literature. The team would like to know the sample size with 30% and 50% risk reduction. The statistician in the team suggests to calculate sample size for 80% and 90% power at the 2-sided significance levels of 0.05 and 0.1. Please show steps for the sample size calculation.
2. A randomized, controlled, and three-arm phase II study for Alzheimer's disease is designed using 2 active doses and a placebo control. The randomization ratio is 1:1:1. The primary endpoint will be clinical dementia rating scale sum of boxes (CDR-SB). Two recent phase III studies for Aducanumab showed a mean change from baseline to be 1.74 with standard error of 0.11 in the placebo arm and 548 subjects in Study Emerge 302; and a mean change from baseline of 1.56 with standard error also 0.11 calculated from 545 subjects. Please choose the significance level and power and calculate the sample for each arm and the total sample size. Please explain your rationale of the sample size calculation, including the assumptions used and if multiplicity adjustment should be considered.
3. Construct 95% CI for the hazard ratio from a PH model shown in the follow table for the risk reduction between two treatment groups

outcome

Analysis of Maximum Likelihood Estimates - discrete

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
trt	6-MP	1	-1.62822	0.43313	14.1316	0.0002	0.196	trt 6-MP

placebo

does1

does2

548

545

Homework 6

4. Obtain the PBC dataset from R survival package. Fit the following PH model

$$h(t,Z) = h_0(t)\exp(\beta_1 sex + \beta_2 edema + \beta_3 bili + \beta_4 albumin + \beta_5 copper + \beta_6 stage) \quad (1)$$

   a) Identify the information matrix
   b) Construct likelihood ratio test for hypothesis $H_0: \beta_4 = \beta_5 = \beta_6 = 0$ (show steps)
   c) Compare the p-values for sex between Model (1) and the log-rank test
   d) Suppose that the team plans to publish a paper about the study result. You need to describe the statistical methods in the method section of the paper. Please write all needed analyses in the method section.
   e) You are also responsible for the result section of the potential publication. Please describe the results, including interpretation of the hazard ratios, p-values, 2-sided 95% CIs of each covariate.

5. Derive PH model score test is the same as the log-rank test for an indicator covariate when no ties.

6. The observed survival data $(T_i, \Delta_i, Z_i)\ \ i = 1,2,3,4,5,6$ are

   (16,1,1), (20, 0,1), (12, 1,0), (14, 0,0), (11, 1, 0), (9, 1, 1).

   Please construct the partial likelihood.