

# Analysis of Survival Data

Jimin Lee

University of North Carolina-Asheville

February 9, 2011

## Introduction

- Survival Data
- Malignant Melanoma Data
- Distribution of Survival Data

## Survival Data

- Data I
- Data II

## Competing Risks Data

- Data III
- Semiparametric Additive Risks Model
- Estimation of Cumulative Incidence Function

## Simulation Study

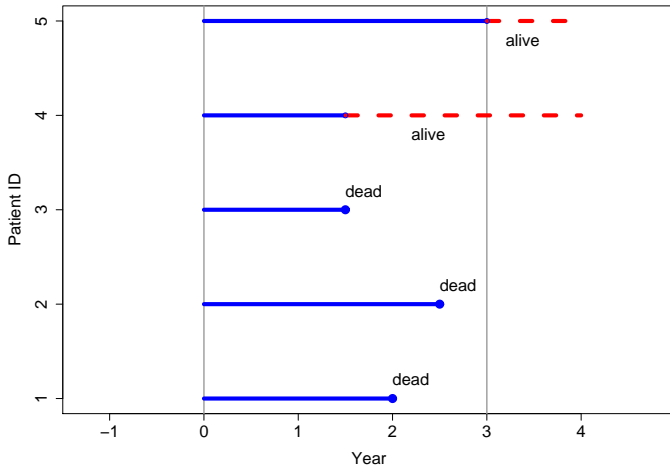
## Real Data Application

- Malignant Melanoma Data (Cntd.)

## Conclusion

- ▶ Time to death from a cancer
- ▶ Time to development of some disease
- ▶ Time to equipment breakdown
- ▶ Time to cessation of smoking
- ▶ Time to a certain event

# Survival Data



- ▶ Some patients may still be alive at the end of the study.
- ▶ Some individuals are lost to follow-up.
- ▶ The exact survival times of these subjects are unknown.
- ▶ Censored times

# Malignant Melanoma Data (Andersen *et al.*, 1993)

- ▶ 205 patients operated for malignant melanoma during 1962–1977
- ▶ 57 patients died from malignant melanoma, 14 patients died from other causes, and 134 were alive.
- ▶ covariates: tumor thickness, ulceration status, age, and gender
  
- ▶ predict the patient-specific cumulative incidence for death due to melanoma

Let  $T$  denote the survival time.

► Survival Function

$$\begin{aligned} S(t) &= P(T > t) \\ &= P(\text{surviving longer than time } t) \end{aligned}$$

► Hazard Function

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{1}{h} P(t \leq T < t + h | T \geq t)$$

the instantaneous failure risk after  $t$ , given survival at  $t$

Let  $T$  denote the survival time,  $C$  the censoring time.  
We have observed data  $(X_i, \Delta_i)$ ,  $i = 1, \dots, n$ ,

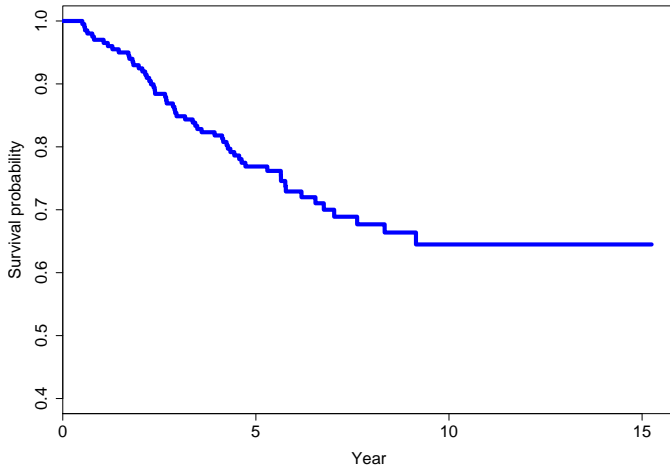
$$\begin{aligned} X_i &= \min(T_i, C_i), \\ \Delta_i &= \begin{cases} 1 & \text{if } T_i \leq C_i \text{ (observed failure)} \\ 0 & \text{if } T_i > C_i \text{ (observed censoring)} \end{cases} \end{aligned}$$

Kaplan-Meier Estimator:

$$\widehat{S}(t) = \prod_{X_{(i)} \leq t} \left( 1 - \frac{1}{n - i + 1} \right)^{\Delta_{(i)}}$$

Here  $X_{(1)} \leq \dots \leq X_{(n)}$  are the ordered  $X_1, \dots, X_n$  and  $\Delta_{(1)}, \dots, \Delta_{(n)}$  are the corresponding  $\Delta$ s

# Malignant Melanoma Data (Cntd.)



We have observed data  $(X_i, \Delta_i, \mathbf{z}_i)$ ,  $i = 1, \dots, n$ ,

where

$\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{ip})'$  is a vector of covariates,

such as treatment indicator, age or gender.

- ▶ Conditional Hazard Function

$$\lambda(t|\mathbf{z}) = \lim_{h \rightarrow 0} \frac{1}{h} P(t \leq T < t + h | T \geq t, \mathbf{z})$$

# Cox's Proportional Hazard Model (1972)

$$\lambda(t|\mathbf{z}) = \lambda_0(t) \exp(\mathbf{z}'\beta) = \lambda_0(t) \exp(z_1\beta_1 + \dots + z_p\beta_p),$$

where  $\lambda_0(t)$  is an arbitrary base hazard function and  $\beta$  is  $p \times 1$  vector of regression coefficients.



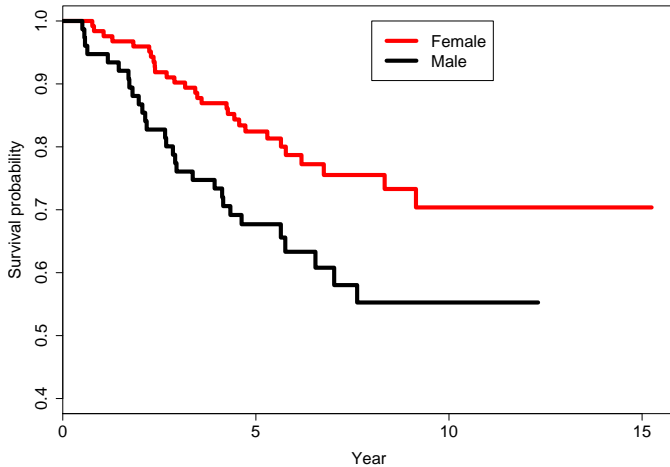
$$\frac{\lambda(t|z = 1 \text{ male})}{\lambda(t|z = 0 \text{ female})} = \frac{\lambda_0(t) \exp(1 \cdot \beta)}{\lambda_0(t) \exp(0 \cdot \beta)} = \exp(\beta)$$

The hazard rates are proportional.

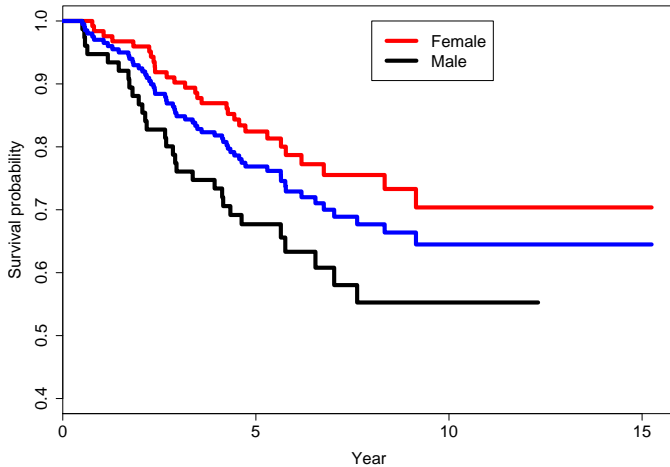
# Malignant Melanoma Data (Cntd.)

```
coef exp(coef) se(coef) z p sex 0.662 1.94 0.265 2.5 0.013
```

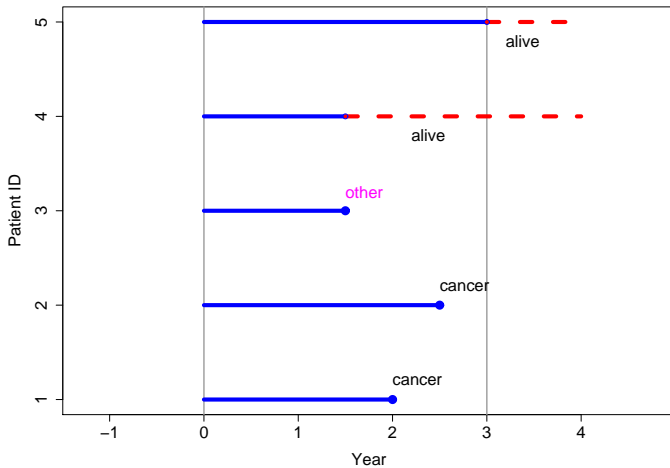
# Malignant Melanoma Data (Cntd.)



# Malignant Melanoma Data (Cntd.)



# Competing Risks Data



- ▶ The lifetimes of patients must be classified: those who died from cause of interest, or those who died from other causes.
- ▶ 57 patients died from malignant melanoma, 14 patients from other causes

Suppose that there are  $K$  distinct failure types.

Let  $T_{ij}$  be the  $j^{\text{th}}$  latent failure time for the  $i^{\text{th}}$  subject, where  $j = 1, \dots, K$  and  $i = 1, \dots, n$ .

One can only observe  $(T_i, \delta_i, \epsilon_i, \mathbf{z}_i)$ , where

$$T_i = \min(\bar{T}_i, C_i), \quad \bar{T}_i = \min\{T_{ij} : j = 1, \dots, K\}$$

$$\delta_i = \begin{cases} 1 & \text{if } \bar{T}_i \leq C_i \\ 0 & \text{if } \bar{T}_i > C_i \end{cases}$$

$$\epsilon_i = k \text{ if } \bar{T}_i = T_{ik}$$

$$\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{ip})' \text{ a vector of covariates}$$

- ▶ Cause-specific Hazard Function

$$\lambda_j(t|\mathbf{z}) = \lim_{h \rightarrow 0} \frac{1}{h} P(t \leq \bar{T} < t + h, \epsilon = j | \bar{T} \geq t, \mathbf{z})$$

the instantaneous rate of occurrence of the  $j$ th failure in the presence of all causes given  $\mathbf{z}$

- ▶ Cumulative Incidence Function

$$F_j(t|\mathbf{z}) = P(\bar{T} \leq t, \epsilon = j | \mathbf{z})$$

the probability of a subject failing from cause  $j$  in the presence of all causes given  $\mathbf{z}$

$$\begin{aligned} F_j(t|\mathbf{z}) &= \int_0^t S(u|\mathbf{z}) \lambda_j(u|\mathbf{z}) du \\ S(t|\mathbf{z}) &= \exp \left( - \sum_{l=1}^K \int_0^t \lambda_l(u|\mathbf{z}) du \right) \end{aligned} \tag{1}$$

# Proportional Hazard Model

$$\lambda_j(t|\mathbf{z}) = \lambda_{0j}(t) \exp\left(\mathbf{z}'\beta_j\right), \quad j = 1, \dots, K$$

- ▶ Kalbfleisch & Prentice (2002), Cheng, Fine & Wei (1998)
- ▶ The influence of the covariate is restricted to be constant.
- ▶ Strong assumption of proportionality

$$\lambda_j(t|\mathbf{x}, \mathbf{z}) = \mathbf{x}'\alpha_j(t) + \mathbf{z}'\beta_j,$$

where  $\alpha_j(t)$  is a  $p$ -dimensional function, and  $\beta_j$  a  $q$ -dimensional regression vector.

- ▶ Some covariates can be investigated nonparametrically to yield time-varying effects.

# Estimation of Cumulative Incidence Function

We estimate

$$\widehat{F}_1(t|\mathbf{x}_0, \mathbf{z}_0) = \int_0^t \widehat{S}(u|\mathbf{x}_0, \mathbf{z}_0) d\widehat{\Lambda}_1(u|\mathbf{x}_0, \mathbf{z}_0),$$

where

$$\begin{aligned}\widehat{S}(t|\mathbf{x}_0, \mathbf{z}_0) &= \exp\left(-\sum_{j=1}^K \widehat{\Lambda}_j(t|\mathbf{x}_0, \mathbf{z}_0)\right), \\ \widehat{\Lambda}_j(t|\mathbf{x}_0, \mathbf{z}_0) &= \int_0^t \mathbf{x}_0' d\widehat{A}_j(u) + t\mathbf{z}_0' \widehat{\beta}_j.\end{aligned}$$

We estimate the regression coefficients  $A_j(t) = \int_0^t \alpha_j(u) du$  and  $\beta_j$  by

$$\begin{aligned}\widehat{A}_j(t) &= \int_0^t X^-(u) \left( dN_j(u) - Z(u)\widehat{\beta}_j du \right), \\ \widehat{\beta}_j &= \left[ \int_0^\infty Z(t)' H(t) Z(t) dt \right]^{-1} \int_0^\infty Z(t)' H(t) dN_j(t),\end{aligned}$$

where

$$\begin{aligned}H(t) &= I_n - X(t)X^-(t) \\ X^-(t) &= (X(t)'X(t))^{-1}X(t)'\end{aligned}$$

with treating all the failure times  $T_i$  with  $\epsilon_i \neq j$  as censored observation.

## Theorem

$\sqrt{n} \left( \widehat{F}_1(t|\mathbf{x}_0, \mathbf{z}_0) - F_1(t|\mathbf{x}_0, \mathbf{z}_0) \right)$  is asymptotically equivalent to a sum of martingale  $U_1(t|\mathbf{x}_0, \mathbf{z}_0) = \sqrt{n} \sum_{i=1}^n \Phi_i(t|\mathbf{x}_0, \mathbf{z}_0)$ , and it converges in distribution to a zero-mean Gaussian process.

Here

$$\begin{aligned}\Phi_i(t|\mathbf{x}_0, \mathbf{z}_0) &= \int_0^t S(u|\mathbf{x}_0, \mathbf{z}_0) \mathbf{x}_0' (X(u)' X(u))^{-1} \mathbf{x}_i dM_{i1}(u) \\ &\quad - \int_0^t S(u|\mathbf{x}_0, \mathbf{z}_0) \mathbf{x}_0' X^{-1}(u) Z(u) du C^{-1} D_{i1} \\ &\quad - \int_0^t S(u|\mathbf{x}_0, \mathbf{z}_0) du \mathbf{z}_0' C^{-1} D_{i1} \\ &\quad - \sum_{j=1}^K \left( \int_0^t H_1(t, u) \mathbf{x}_0' (X(u)' X(u))^{-1} \mathbf{x}_i dM_{ij}(u) \right. \\ &\quad \quad - \int_0^t H_1(t, u) \mathbf{x}_0' X^{-1}(u) Z(u) du C^{-1} D_{ij} \\ &\quad \quad \left. - \int_0^t H_1(t, u) du \mathbf{z}_0' C^{-1} D_{ij} \right)\end{aligned}$$

and

$$M_{ij}(t) = N_{ij}(t) - \int_0^t Y_i(u) (\mathbf{x}_i' \alpha_j(u) + \mathbf{z}_i' \beta_j) du,$$

$$C = \int_0^\infty Z(t)' H(t) Z(t) dt,$$

$$D_{ij} = \int_0^\infty \left( \mathbf{z}_i - Z(t)' X(t) (X(t)' X(t))^{-1} \mathbf{x}_i \right) dM_{ij}(t),$$

$$H_1(t, u) = F_1(t | \mathbf{x}_0, \mathbf{z}_0) - F_1(u | \mathbf{x}_0, \mathbf{z}_0)$$

# Simulation Study

- ▶  $\mathbf{x} = (1, x_2)$ , the second component  $x_2$  Bernoulli distribution with  $p = 0.5$
- ▶  $z$  uniformly distributed on  $[0, 1]$
- ▶  $\mathbf{x}_0 = (1, 1)$  and  $z_0 = 0.5$
- ▶ The  $K = 2$  latent failure times were taken to be conditionally independent with cause-specific hazard functions given by

$$\lambda_k(t|\mathbf{x}, z) = \alpha_{k1}(t) + x_2\alpha_{k2}(t) + z\beta_k$$

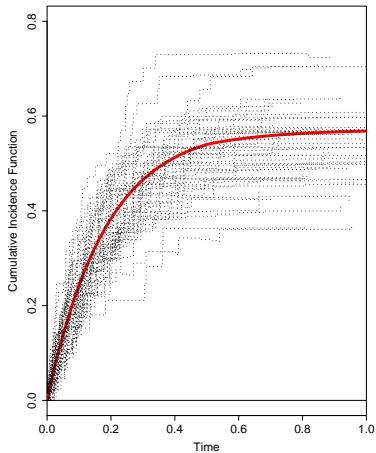
where  $\alpha_{11}(t) = \alpha_{21}(t) = \alpha_{22}(t) = 1.0$  for  $t > 0$ ,

$$\alpha_{12}(t) = \begin{cases} 2.0 & \text{for } 0 \leq t \leq 0.5 \\ 1.0 & \text{for } t > 0.5 \end{cases}$$

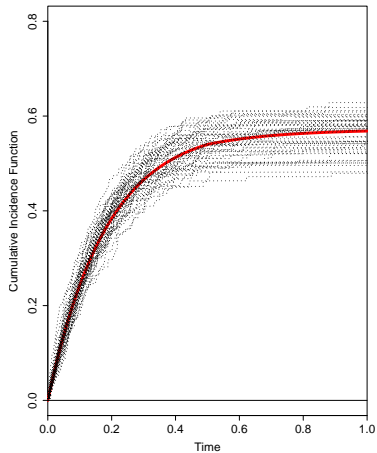
and  $\beta_1 = 0.3, \beta_2 = 0.6$

- ▶ The censoring  $C$  exponentially distributed

# Simulation Study



(a) 100



(b) 500

		Sample size	Censoring		
			10%	30%	50%
Cause 1	250	0.0307	0.0309	0.0310	
	500	0.0221	0.0221	0.0222	
	750	0.0177	0.0179	0.0180	
Cause 2	250	0.0284	0.0282	0.0278	
	500	0.0203	0.0201	0.0199	
	750	0.0163	0.0164	0.0162	

**Table:** Mean Absolute Deviation  $\Sigma_t |\hat{F}_j(t) - F_j(t)|$  of predicted cumulative incidence function

# Malignant Melanoma Data (Cntd.)

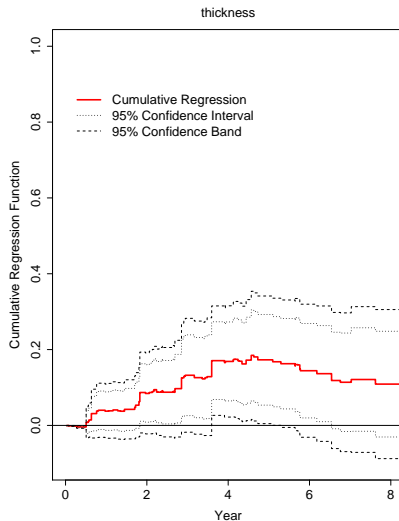
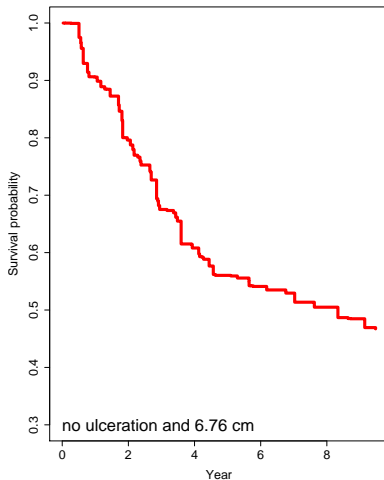
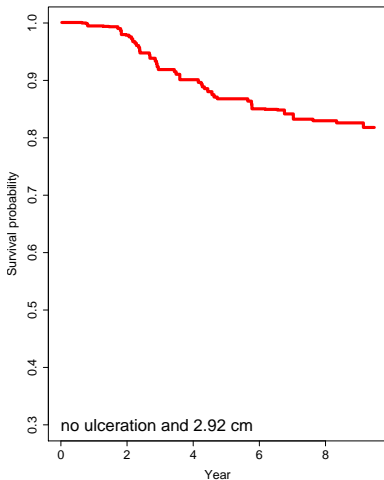


Figure: the effect of tumor thickness

# Malignant Melanoma Data (Cntd.)

Survival probability for a 52-year-old female patient,



# Areas of Application

- ▶ Medicine, Public Health
- ▶ Life Insurance
- ▶ Quality Control & Reliability in Manufacturing

Thank You.