# Survival analysis

Wenda Zhou

June 21, 2017

# Survival analysis

Study of data under censoring.

## Censoring

A variable is said to be censored if its value is only partially known.

# Censoring

Can be seen as a type of missing data.

## Right censoring

Wish to understand survival time after cancer diagnosis.
Not all diagnosis made at the same time: after 5 years, status still unknown for some people.
Know that the value is greater than.

# Censoring

Can be seen as a type of missing data.

## Right censoring

Wish to understand survival time after cancer diagnosis.
Not all diagnosis made at the same time: after 5 years, status still unknown for some people.
Know that the value is greater than.

## Left censoring

Wish to understand reliability of car.
First inspection after 200 days – however, some cars may fail before.
Know that the value is less than.

# Censoring

Can be seen as a type of missing data.

## Right censoring

Wish to understand survival time after cancer diagnosis.
Not all diagnosis made at the same time: after 5 years, status still unknown for some people.
Know that the value is greater than.

## Left censoring

Wish to understand reliability of car.
First inspection after 200 days – however, some cars may fail before.
Know that the value is less than.

## Interval censoring

Observe animal every 7 days, wish to determine time of hibernation.
Know that the value is in an interval.

# Survival function

The object of interest in survival analysis is the survival function. Let $T$ denote the time to event:

$$S(t) = 1 - F(t) = \mathrm{P}(T \geq t) \tag{1}$$

# Survival function

The object of interest in survival analysis is the survival function. Let $T$ denote the time to event:

$$S(t) = 1 - F(t) = \mathrm{P}(T \geq t) \tag{1}$$

Usually the object of interest.

# Kaplan-Meier estimator of the survival function

Suppose that we have observations $t_1, \ldots, t_n$. How should we estimate the survival function?

## No censoring

If there is no censoring, can just consider the complement of the ecdf:

$$\hat{S}(t) = \frac{\#\{i : t_i > t\}}{n} \tag{2}$$

# Kaplan-Meier estimator of the survival function

Suppose that we have observations $t_1, \ldots, t_n$. How should we estimate the survival function?

## No censoring

If there is no censoring, can just consider the complement of the ecdf:

$$\hat{S}(t) = \frac{\#\{i : t_i > t\}}{n} \tag{2}$$

However, if there is censoring, cannot fully know the number of observations such that $t_i > t$.

# KM estimator

## What does not work

### Idea 1: complete case analysis

Longer time to events are more likely to be censored. Hence complete case analysis is to pessimistic.

# KM estimator

## What does not work

### Idea 1: complete case analysis

Longer time to events are more likely to be censored. Hence complete case analysis is to pessimistic.

### Idea 2

Count censored observations as having survived: too optimistic.

# KM estimator

## Life tables

| Age | Number dying | Number at start | Probability of dying |
|---|---|---|---|
| 0 - 1 | 596 | 100,000 | 0.005958 |
| 1 - 2 | 42 | 99,404 | 0.000422 |
| 2 - 3 | 25 | 99,362 | 0.000255 |

# KM estimator

What can we estimate? Probability of surviving until next year.

$$\mathrm{P}(T \geq t+1 \mid T \geq t) = \frac{\#\text{ alive at time } t \text{ who survive until } t+1}{\#\text{ alive at time } t} \tag{3}$$

Can we get to $\mathrm{P}(T \geq t)$ from the above?

# KM estimator

What can we estimate? Probability of surviving until next year.

$$\mathrm{P}(T \geq t+1 \mid T \geq t) = \frac{\#\text{ alive at time } t \text{ who survive until } t+1}{\#\text{ alive at time } t} \tag{3}$$

Can we get to $\mathrm{P}(T \geq t)$ from the above?

$$\mathrm{P}(T \geq t) = \mathrm{P}(T \geq t \mid T \geq t-1)\,\mathrm{P}(T \geq t-1 \mid T \geq t-2)\cdots \tag{4}$$

# Hazard function

Can define the hazard function, which is the limit:

$$\lambda(x) = \lim_{\delta \to 0} \frac{1}{\delta} \mathrm{P}(T \geq t + \delta \mid T \geq t) = \frac{f_T(t)}{S(t)} \tag{5}$$

# Hazard function

Can define the hazard function, which is the limit:

$$\lambda(x) = \lim_{\delta \to 0} \frac{1}{\delta} \mathrm{P}(T \geq t + \delta \mid T \geq t) = \frac{f_T(t)}{S(t)} \tag{5}$$

How to get from the hazard function to the survival function? Let the cumulative hazard function be $\Lambda(t)$ defined by:

$$\Lambda(t) = \int_0^t \lambda(t')\, dt' \tag{6}$$

then

$$S(t) = e^{-\Lambda(t)} \tag{7}$$

# KM estimator

The Kaplan-Meier estimator estimates the hazard function, and produces the survival function from that.

# KM estimator

The Kaplan-Meier estimator estimates the hazard function, and produces the survival function from that.
It is an unbiased estimator of the survival function under mild conditions.

# Survival and covariates

We will often wish to understand how some covariates affect survival.

## Examples

- Understand how the drug dose affects survival
- Understand how design of car affects reliability

# Survival and covariates

We will often wish to understand how some covariates affect survival.

## Examples

- Understand how the drug dose affects survival
- Understand how design of car affects reliability

Potentially interested in estimation and testing.

# Cox proportional hazards

"Linear regression for survival".
Model the hazard rate as for a unit with covariates $x_1, \ldots, x_p$ as:

$$\lambda(t) = \lambda_0(t) \exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\} \tag{8}$$

where $\lambda_0(t)$ is an unknown base hazard rate.