Survival Analysis Part I

Luke Keele

January 10, 2022

Overview

- Medical research often focuses on time to event outcomes (survival).

Overview

- Medical research often focuses on time to event outcomes (survival).
- This type of outcome creates many complications.

Overview

- Medical research often focuses on time to event outcomes (survival).
- This type of outcome creates many complications.
- First, I will review data complications.

Overview

- Medical research often focuses on time to event outcomes (survival).
- This type of outcome creates many complications.
- First, I will review data complications.
- Second, I will review study design complications.

Overview

- Medical research often focuses on time to event outcomes (survival).
- This type of outcome creates many complications.
- First, I will review data complications.
- Second, I will review study design complications.
- Third, I will review basic survival analyses.

Overview

- Medical research often focuses on time to event outcomes (survival).
- This type of outcome creates many complications.
- First, I will review data complications.
- Second, I will review study design complications.
- Third, I will review basic survival analyses.
- Fourth, I will conclude with an example in Stata.

Data

- “Clean” vs “Dirty” data.

Data

- “Clean” vs “Dirty” data.
- Most tutorial examples: data are clean.

Data

- “Clean” vs “Dirty” data.
- Most tutorial examples: data are clean.
- Data from clinical registries are “dirty.”

Data

- "Clean" vs "Dirty" data.
- Most tutorial examples: data are clean.
- Data from clinical registries are "dirty."
- Survival analysis requires very specific data formatting.

Data

- “Clean” vs “Dirty” data.
- Most tutorial examples: data are clean.
- Data from clinical registries are “dirty.”
- Survival analysis requires very specific data formatting.
- Stata requires special formatting before it will give you any results for a survival analysis.

Data

- “Clean” vs “Dirty” data.
- Most tutorial examples: data are clean.
- Data from clinical registries are “dirty.”
- Survival analysis requires very specific data formatting.
- Stata requires special formatting before it will give you any results for a survival analysis.
- Such formatting is a key part of “cleaning” the data for analysis.

Data Notation

Typically, we need to create:

$$Y_i \quad = \quad \text{the duration until the event occurs, ie. 12 months}$$

Dates

- Registry data typically contain dates of events.

Dates

- Registry data typically contain dates of events.
- Handling dates in Stata is a key first step.

Dates

- Registry data typically contain dates of events.
- Handling dates in Stata is a key first step.
- Date variables need to be converted to date format using the date() command.

Dates

1. Date variable must be stored as a string.

Dates

1. Date variable must be stored as a string.
2. Date command requires you to know format of your date.

Dates

1. Date variable must be stored as a string.
2. Date command requires you to know format of your date.
3. Date might be recorded as 31jan2000 or 1/31/2000.

Dates

1. Date variable must be stored as a string.
2. Date command requires you to know format of your date.
3. Date might be recorded as 31jan2000 or 1/31/2000.
4. You must know the difference.

Dates

- The date command: date(string, “FMT”, 201X)

Dates

- The date command: date(string, “FMT”, 201X)
- Key parts FMT and 201X

Dates

- The date command: date(string, “FMT”, 201X)
- Key parts FMT and 201X
- gen date1 = date(date-string-var, “DMY”, 2019)

Dates

- The date command: date(string, “FMT”, 201X)
- Key parts FMT and 201X
- gen date1 = date(date-string-var, “DMY”, 2019)
- FMT is DMY for 31jan2000.

Dates

- The date command: date(string, “FMT”, 201X)
- Key parts FMT and 201X
- gen date1 = date(date-string-var, “DMY”, 2019)
- FMT is DMY for 31jan2000.
- FMT is MDY for 1/31/2000.

Dates

- The date command: date(string, “FMT”, 201X)
- Key parts FMT and 201X
- gen date1 = date(date-string-var, “DMY”, 2019)
- FMT is DMY for 31jan2000.
- FMT is MDY for 1/31/2000.
- Final part is topyear–most recent year in your data.

Dates

- Stata saves dates as number of days since Jan 1, 1960.

Dates

- Stata saves dates as number of days since Jan 1, 1960.
- The command:
- format date1 %td
- converts date from numeric form to date format.

Calculating Duration

- Stata: gen duration = date2 - date1
- This gives survival time in days.

Censoring

- Next, we must develop a censoring variable.

Censoring

- Next, we must develop a censoring variable.
- What is censoring?

Censoring

- Next, we must develop a censoring variable.
- What is censoring?
- Censoring loosely refers to missing outcome data.

Censoring

- Next, we must develop a censoring variable.
- What is censoring?
- Censoring loosely refers to missing outcome data.
- For some individuals, duration is known but survival time is unknown.

Censoring

- Next, we must develop a censoring variable.
- What is censoring?
- Censoring loosely refers to missing outcome data.
- For some individuals, duration is known but survival time is unknown.
- Most common is administrative censoring.

Administrative Censoring

- Survival analysis requires a stopping point data collection.

Administrative Censoring

- Survival analysis requires a stopping point data collection.
- Some patients may not have experienced the event when the data is collected.

Administrative Censoring

- Survival analysis requires a stopping point data collection.
- Some patients may not have experienced the event when the data is collected.
- For these patients, we don't observe their survival time.

Administrative Censoring

- Survival analysis requires a stopping point data collection.
- Some patients may not have experienced the event when the data is collected.
- For these patients, we don't observe their survival time.
- We have a duration but survival time is missing, since the recorded time span is qualitatively different from patients that experienced the event.

Administrative Censoring

- Outcome: disease free survival after adj. chemo.

Administrative Censoring

- Outcome: disease free survival after adj. chemo.
- Data collection stops on Dec 1, 2021.

Administrative Censoring

- Outcome: disease free survival after adj. chemo.
- Data collection stops on Dec 1, 2021.
- All patients that have yet to recur when data collection stops are censored.

Administrative Censoring

- Outcome: disease free survival after adj. chemo.
- Data collection stops on Dec 1, 2021.
- All patients that have yet to recur when data collection stops are censored.
- We observe their duration but not the time to event, since the event has yet to occur.

Administrative Censoring

- Outcome: disease free survival after adj. chemo.
- Data collection stops on Dec 1, 2021.
- All patients that have yet to recur when data collection stops are censored.
- We observe their duration but not the time to event, since the event has yet to occur.
- The survival time is missing.

Administrative Censoring

- We must record observations that are censored.

Administrative Censoring

- We must record observations that are censored.
- Create new variable: $C = 1$ is an uncensored observation, $C = 0$ is a censored observation.

Competing Events Censoring

- A competing event prevents event of interest.

Competing Events Censoring

- A competing event prevents event of interest.
- Example: time to transplant.

Competing Events Censoring

- A competing event prevents event of interest.
- Example: time to transplant.
- Patient dies before transplant.

Competing Events Censoring

- A competing event prevents event of interest.
- Example: time to transplant.
- Patient dies before transplant.
- These patient should be considered censored.

Competing Events Censoring

- A competing event prevents event of interest.
- Example: time to transplant.
- Patient dies before transplant.
- These patient should be considered censored.
- Competing events are study specific.

Competing Events Censoring

- A competing event prevents event of interest.
- Example: time to transplant.
- Patient dies before transplant.
- These patient should be considered censored.
- Competing events are study specific.
- Need to be defined by the researcher.

Competing Events Censoring

- A competing event prevents event of interest.
- Example: time to transplant.
- Patient dies before transplant.
- These patient should be considered censored.
- Competing events are study specific.
- Need to be defined by the researcher.
- Widespread competing events makes interpretation of results difficult.

Effects of Censoring

- Censoring reduces power.

Effects of Censoring

- Censoring reduces power.
- If censoring is systematic, estimates from multivariate models can be biased.

Follow-up Time

- Follow-up time is the time from an event of interest: i.e. time since randomization.

Follow-up Time

- Follow-up time is the time from an event of interest: i.e. time since randomization.
- If follow-up time isn't long enough, censoring rates will be high.

Follow-up Time

- Follow-up time is the time from an event of interest: i.e. time since randomization.
- If follow-up time isn't long enough, censoring rates will be high.
- What is the appropriate follow-up time?

Follow-up Time

- Follow-up time is the time from an event of interest: i.e. time since randomization.
- If follow-up time isn't long enough, censoring rates will be high.
- What is the appropriate follow-up time?
- Will be study specific: follow up time in a transplant context is very different from cancer recurrence.

Follow-up Time

- Follow-up time is the time from an event of interest: i.e. time since randomization.
- If follow-up time isn't long enough, censoring rates will be high.
- What is the appropriate follow-up time?
- Will be study specific: follow up time in a transplant context is very different from cancer recurrence.

Baseline Time Point

- All subjects should be untreated at $t = 0$.

Baseline Time Point

- All subjects should be untreated at $t = 0$.
- E.g.: Effect of lipitor on time to cardiac event.

Baseline Time Point

- All subjects should be untreated at $t = 0$.
- E.g.: Effect of lipitor on time to cardiac event.
- No one should be on lipitor at $t = 0$.

Example Data

id	duration	censor
1	4	1
2	2	1
3	5	1
4	6	0

Stata stset

- Stata requires all survival data to be stset.
- Syntax: stset surv_var, failure(censor)

Stata stset

- Stata requires all survival data to be stset.
- Syntax: stset surv_var, failure(censor)
- Stata creates: _t0, _t1, _d, and _st
- time span, censoring, and relevant.

Survival Curves

- Survival curves are basic summary statistics.

Survival Curves

- Survival curves are basic summary statistics.
- Kaplan-Meier method is the most common.

Survival Curves

- Survival curves are basic summary statistics.
- Kaplan-Meier method is the most common.
- Displays probability of event for those at risk.

Survival Curves

- We have N observations.
- $n_t =$ the number of observations "at risk" for the event at time t.

Survival Curves

- We have N observations.
- $n_t =$ the number of observations “at risk” for the event at time t.
- $d_t =$ the number of observations which experience the event at time t

Survival Curves

- We have N observations.
- $n_t =$ the number of observations "at risk" for the event at time t.
- $d_t =$ the number of observations which experience the event at time t
- For any particular time $t = k$, we can get an estimate of the survival function $S(t)$ as the product of the conditional proportions of all survivors to that point

$$\widehat{S(t_k)} = \prod_{t \leq t_k} \frac{n_t - d_t}{n_t}$$

Survival Curves

- We have N observations.
- $n_t =$ the number of observations “at risk” for the event at time t.
- $d_t =$ the number of observations which experience the event at time t
- For any particular time $t = k$, we can get an estimate of the survival function $S(t)$ as the product of the conditional proportions of all survivors to that point

$$\widehat{S(t_k)} = \prod_{t \leq t_k} \frac{n_t - d_t}{n_t}$$

This is known as the “Kaplan–Meier” estimate of the survivor function.

Estimating Survival Curves - An Example

Time	No. at risk	No. failed	No. censored
2	6	1	0
4	5	2	0
5	3	0	1
7	2	1	0
8	1	0	1

Estimating Survival Curves

Time	No. at risk	No. failed	No. censored	p	$\hat{S}(t)$
2	6	1	0	5/6	5/6
4	5	2	0	3/5	1/2
5	3	0	1	1	1/2
7	2	1	0	1/2	1/4
8	1	0	1	1	1/4

Estimating Survival Curves

Time	No. at risk	No. failed	No. censored	p	$\hat{S}(t)$
2	6	1	0	5/6	5/6
4	5	2	0	3/5	1/2
5	3	0	1	1	1/2
7	2	1	0	1/2	1/4
8	1	0	1	1	1/4

Estimating Survival Curves

Time	No. at risk	No. failed	No. censored	p	$\hat{S}(t)$
2	6	1	0	5/6	5/6
4	5	2	0	3/5	1/2
5	3	0	1	1	1/2
7	2	1	0	1/2	1/4
8	1	0	1	1	1/4

$\frac{(6-1)}{6} = \frac{5}{6}$

$\frac{(5-2)}{5} = \frac{3}{5} \times \frac{5}{6} = \frac{1}{2}$

Survival Curves

- Typically, we plot stratified KM curves.

Survival Curves

- Typically, we plot stratified KM curves.
- Stratify by key covariate: treatment, sex, etc.

Survival Curves

- Typically, we plot stratified KM curves.
- Stratify by key covariate: treatment, sex, etc.
- Inference is now important: are the curves statistically different?

If we're interested in inference, or just want to know the uncertainty surrounding our estimates, we need some measure of the variability of these estimates. The most commonly–used of these is the "Greenwood" variance estimator:

$$Var[\widehat{S(t_k)}] = [\widehat{S(t_k)}]^2 \sum_{t \leq t_k} \frac{d_t}{n_t(n_t - d_t)}$$

Log-rank Test

We have two groups; *treatment* (=1) and *placebo* (=0), and we want to know if the survival curves are statistically different. Standard test is the log-rank test.

Log-rank Test

	Treatment	Placebo	Total
Event	d_{1t}	d_{0t}	d_t
No Event	$n_{1t} - d_{1t}$	$n_{0t} - d_{0t}$	$n_t - d_t$
Total	n_{1t}	n_{0t}	n_t

Normally, we'd do a χ^2 test here, using the observed and expected number of events per cell.

The same general intuition applies, except that we conduct a similar test for each time period t.

Log-rank Test

$$\hat{e}_{1t} = \frac{n_{1t}d_t}{n_t}$$

is the “expected” number of events in that time period.

Log-rank Test

$$\hat{Q} = \frac{\left[\sum_t (d_{1t} - \hat{e}_{1t})\right]^2}{\left[\frac{n_{1t} n_{0t} d_{0t} (n_t - d_t)}{n_t^2 (n_t - 1)}\right]}$$

The numerator of $\hat{Q}$ is the sum of the (squared) observed minus expected events. We use this to test the null hypothesis of no difference between the treatment and placebo groups.

$\hat{Q}$ is distributed as χ^2_1.

Unadjusted Survival Analysis

- The analysis thus far assumes treated and control groups are exchangeable.
- Only reason survival curves differ is treatment–not some baseline characteristic of the treated group.
- Next time we take up methods to control for confounders.

Conclusion

- Data cleaning is a key step in survival analysis.
- May require several consequential choices.