Survival Analysis and Event History

Andy Grogan-Kaylor

13 May 2021

Introduction

> "Survival analysis is a key technique in data-driven decision-making, which is now central to public interest because of COVID-19. Applying the correct technique for the specific question at hand is crucial for credible public health inferences. If you are interested in assessing how a risk factor or a potential treatment affects the progression of a disease—such as how long a patient takes to recover—then survival analysis techniques come into play. Survival analysis deeply respects the ultimate source of its data, often the disease experience or even the life and death of human patients. It seeks to exploit every last drop of information that this experience can render for saving lives—in particular, not only whether patients survived, but how long, and why. And it strives to do so with minimal assumptions, so that the data are truly driving the decision."

—SAS Corporation

Key Concepts

WHO CARES how we measure time? Isn't it self-evident?

- Implementations differ; formulas are our friends
- $h(t) = x1 + x2 +$ etc....: formula (effect on hazard (instantaneous rate of occurrence))

The "Hospital Bed Problem"

- Imagine a *Hypothetical Hospital*
- Imagine that there are 52 patients *total.*
- 51 of the patients are *long term patients*, who each stay for *1 year.*
- 1 of the patients is a *short term patient*, who stays for *1 week.*

Is this a hospital that serves mostly long-term, or short term patients?

```
. clear all

. set obs 52 // 52 hypothetical obervations
Number of observations (_N) was 0, now 52.
```

```
. generate id = _n // set id = to observation #

. generate weeks = 52

. replace weeks = 1 if id == 52
(1 real change made)

. twoway (scatter id weeks if weeks == 52, msize(small)) /// staying 52 weeks
> (scatter id weeks if weeks == 1, msize(small)), /// staying 1 week
> title("Hypothetical Hospital") ///
> legend(on order(1 "long term" 2 "short term")) ///
> xtitle("week of discharge") ///
> ylabel(1(1)52, labels labsize(tiny) angle(horizontal) noticks nogrid) ///
> scheme(michigan)

. graph export hospital_bed_problem.png, width(1000) replace
file
    /Users/agrogan/Desktop/newstuff/categorical/survival-analysis-and-event-history/hospital_bed_p
    > roblem.png saved as PNG format
```

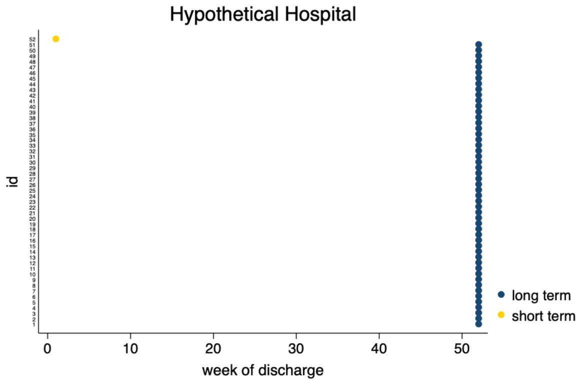
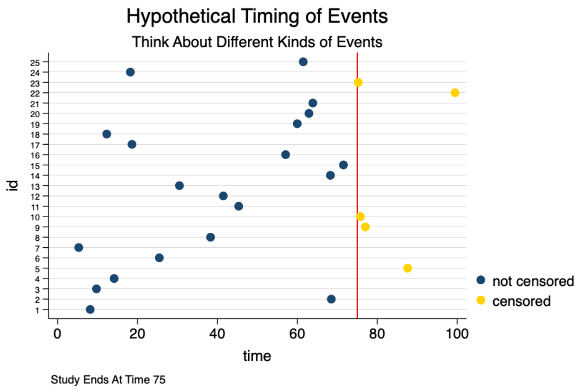


Figure 1: Illustration of Hospital Bed Problem

How To Measure Length of Stay (1)

```
. clear all

. set obs 25 // 25 hypothetical obervations
Number of observations (_N) was 0, now 25.

. generate id = _n // set id = to observation #

. generate time = runiform(1, 100) // random times

. generate censored = time > 75 // censored if time > 75

. twoway (scatter id time if censored == 0) ///
> (scatter id time if censored == 1), ///
> title("Hypothetical Timing of Events") ///
> subtitle("Think About Different Kinds of Events") ///
> note("Study Ends At Time 75") ///
> legend(on order(1 "not censored" 2 "censored")) ///
> xline(75, lcolor("red")) /// censoring line at 75
> ylabel(1(1)25, labsize(vsmall) angle(horizontal)) /// lines from 1 to 25
> scheme(michigan)

. graph export timing_of_events.png, width(1000) replace
```

```
file
    /Users/agrogan/Desktop/newstuff/categorical/survival-analysis-and-event-history/timing_of_even
    > ts.png saved as PNG format
```



Study Ends At Time 75

Figure 2: Timing Of Events

Animated

See times-events-and-censoring.html

How To Measure Length of Stay (2)

Event happened within a specified time (yes/no)

$$\ln(\frac{P(\text{event})}{1 - P(\text{event})}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i$$

- Statistically accurate, but we lose information on *when* the event happened.
- Statistically *less efficient*.

Time until Event

$$\text{time until event} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i$$

- What to do with events that haven't happened yet? (Censoring)
- Code as `missing`. Loss of information if using complete cases. Possible bias.
- Code as 0. Possible bias. They might happen at some point.
- Code as `time of censoring`. Possible bias. They might never happen. They might happen much later.

Hazard (Risk) of Event Occurence

A more heuristic definition:

$$h(t) = \lim_{\delta \to 0} \frac{\text{probability of having an event before time } t + \delta}{\delta}$$

This definition per Johnson & Shih (2007)

A more formal definition:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T > t)}{\Delta t}$$

This definition per Ragnar Frisch Centre for Economic Research (2020)

A Policy Example (Welfare Reform, 1996)

From LaDonna Pavetti (1995)

- time in months
- new entrants (percent)
- all current recipients at a point in time (percent)

```
. clear all

. use Pavetti.dta
(Written by R.              )

. list, abbreviate(25) // list out the data

     +-----------------------------------------------------+
     |    time   new_entrants   all_current_recipients     |
     |-----------------------------------------------------|
  1. |    1-12           27.4                      4.5     |
  2. |   13-24           14.8                      4.8     |
  3. |   25-36             10                      4.9     |
  4. |   37-48            7.7                        5     |
  5. |   49-60            5.5                      4.5     |
     |-----------------------------------------------------|
  6. | Over 60           34.6                     76.3     |
     +-----------------------------------------------------+

. graph bar (asis) all_current_recipients, /// this particular set of options was difficult to figur
> e out!
> asyvars ///
> over(time) ///
> title("All Current Recipients") ///
> sub("By Months On Caseload") ///
> ytitle("percent") ///
> scheme(michigan)

. graph export all_current_recipients.png, width(1000) replace
file
    /Users/agrogan/Desktop/newstuff/categorical/survival-analysis-and-event-history/all_current_re
    > cipients.png saved as PNG format
```

Welfare Reform (2)

```
. graph bar (asis) new_entrants, ///
> asyvars ///
> over(time) ///
> title("New Recipients") ///
> sub("By Months On Caseload") ///
> ytitle("percent") ///
> scheme(michigan)

. graph export new_recipients.png, width(1000) replace
file
    /Users/agrogan/Desktop/newstuff/categorical/survival-analysis-and-event-history/new_recipients
    > .png saved as PNG format
```

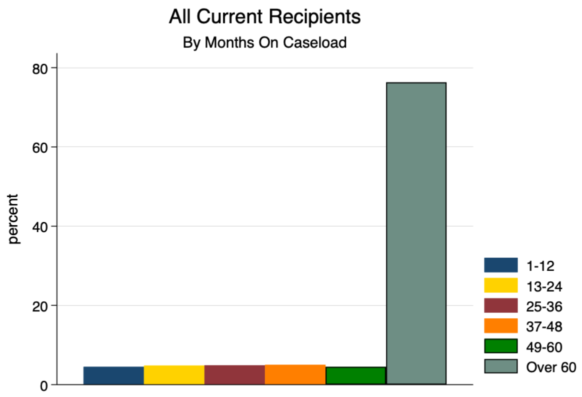


Figure 3: All Current Recipients by Months on Caseload

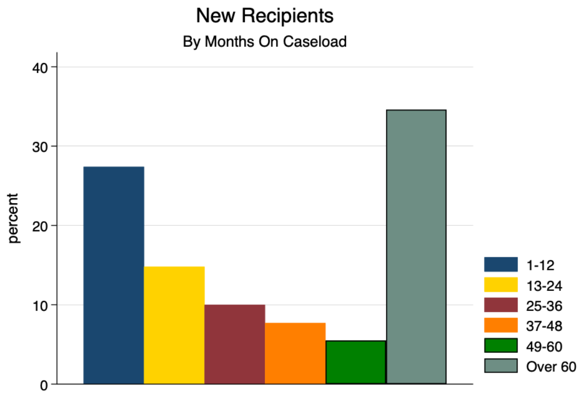
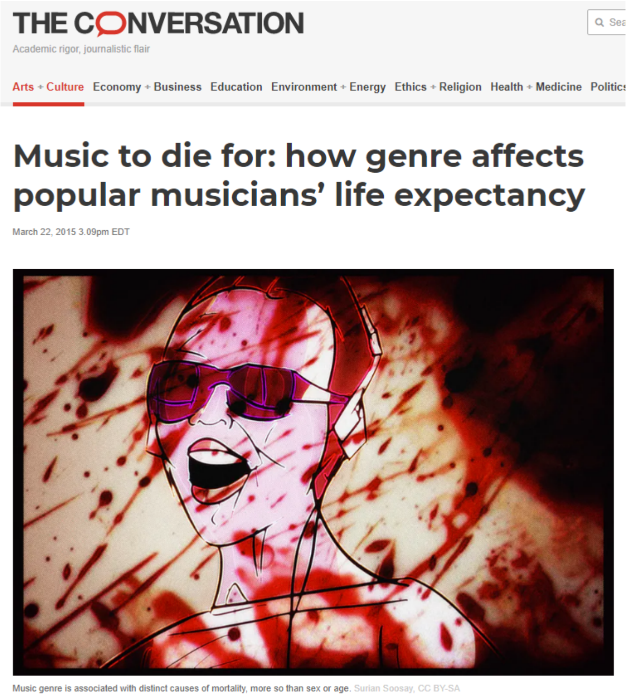


Figure 4: New Recipients by Months on Caseload

Musicians and Mortality (1)



Music to die for: how genre affects popular musicians' life expectancy

March 22, 2015 3.09pm EDT

Music genre is associated with distinct causes of mortality, more so than sex or age. Surian Soosay, CC BY-SA

Figure 5: Music To Die For

Musicians and Mortality (2)

Cox Proportional Hazards Model

Formula

$h(t)$ the rate of occurrence.

$$h(t) = \lim_{\delta \to \infty} \frac{\text{probability of having an event before time } t + \delta}{\delta}$$

This definition per Johnson & Shih (2007).

$$h(t) = h_0(t) e^{\beta_1 x 1 + \beta_2 x_2 + etc.}$$

> We don't directly estimate the hazard, but estimate the effect of covariates on the hazard.

> The event (birth, death, program entry, program departure) is coded as 1, so we are estimating the association of the covariates with event occurrence.

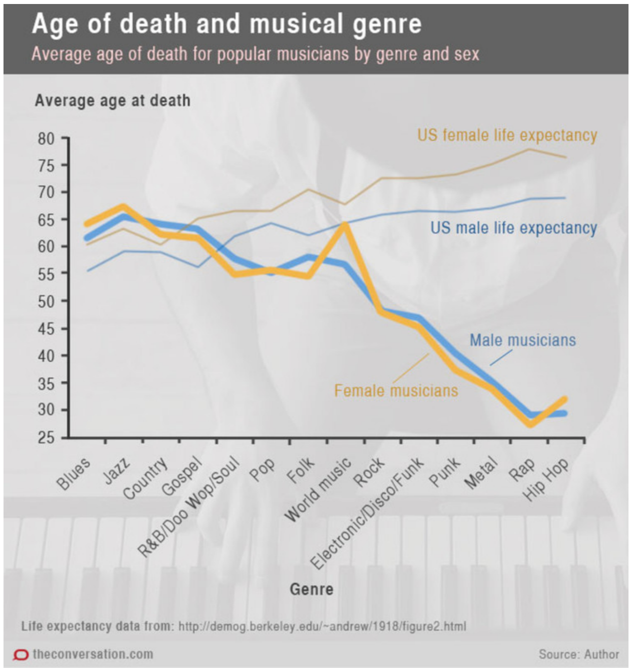


Figure 6: Musician Mortality

Cox Proportional Hazards Model in Stata

Using a data set referenced frequently in Stata `help` and Stata YouTube Videos

```
. clear all

. webuse drugtr // demonstration data set from Stata
(Patient survival in drug trial)
```

Setup of Data

```
. stset // show st setup of data
-> stset studytime, failure(died)

Survival-time data settings

         Failure event: died!=0 & died<.
Observed time interval: (0, studytime]
     Exit on or before: failure

--------------------------------------------------------------------------

         48  total observations
          0  exclusions
--------------------------------------------------------------------------
         48  observations remaining, representing
         31  failures in single-record/single-failure data
        744  total analysis time at risk and under observation
                                                At risk from t =         0
                                     Earliest observed entry t =         0
                                          Last observed exit t =        39

. describe // show variables in data

Contains data from https://www.stata-press.com/data/r17/drugtr.dta
 Observations:            48                  Patient survival in drug trial
    Variables:             8                  3 Mar 2020 02:12
------------------------------------------------------------------------------------------------------
```

```
Variable      Storage   Display    Value
    name         type    format    label      Variable label
─────────────────────────────────────────────────────────────────────────
studytime       byte    %8.0g                 Months to death or end of exp.
died            byte    %8.0g                 1 if patient died
drug            byte    %8.0g                 Drug type (0=placebo)
age             byte    %8.0g                 Patient´s age at start of exp.
_st             byte    %8.0g                 1 if record is to be used; 0 otherwise
_d              byte    %8.0g                 1 if failure; 0 if censored
_t              byte    %10.0g                Analysis time when record ends
_t0             byte    %10.0g                Analysis time when record begins
─────────────────────────────────────────────────────────────────────────
Sorted by:
```

Kaplan-Meier Survivor Function (per Gabriela Ortiz, Stata)

$$S(t) = Pr(T > t)$$

```
. sts graph, scheme(michigan) // Kaplan-Meier Survivor Function

        Failure _d: died
  Analysis time _t: studytime

. graph export survival0.png, width(1000) replace
file /Users/agrogan/Desktop/newstuff/categorical/survival-analysis-and-event-history/survival0.png
    saved as PNG format
```

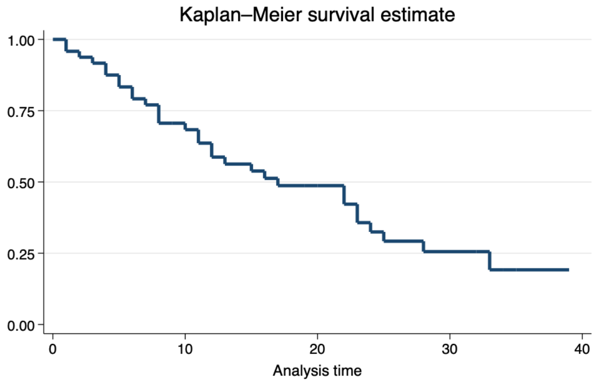


Figure 7: Kaplan-Meier Survivor Function

Cox Proportional Hazards Model

```
. stcox age drug // run Cox Proportional Hazards Model

        Failure _d: died
  Analysis time _t: studytime

Iteration 0:   log likelihood = -99.911448
Iteration 1:   log likelihood = -83.551879
Iteration 2:   log likelihood = -83.324009
Iteration 3:   log likelihood = -83.323546
Refining estimates:
Iteration 0:   log likelihood = -83.323546

Cox regression with Breslow method for ties

No. of subjects =  48                                   Number of obs =     48
No. of failures =  31
Time at risk    = 744
                                                        LR chi2(2)    =  33.18
```

```
Log likelihood = -83.323546                             Prob > chi2    = 0.0000

------------------------------------------------------------------------------
          _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   1.120325   .0417711     3.05   0.002     1.041375     1.20526
        drug |   .1048772   .0477017    -4.96   0.000     .0430057    .2557622
------------------------------------------------------------------------------
```

Graph Survival Curves

```
. stcurve, survival scheme(michigan) // survival curve

. graph export survival1.png, width(1000) replace
file /Users/agrogan/Desktop/newstuff/categorical/survival-analysis-and-event-history/survival1.png
    saved as PNG format
```

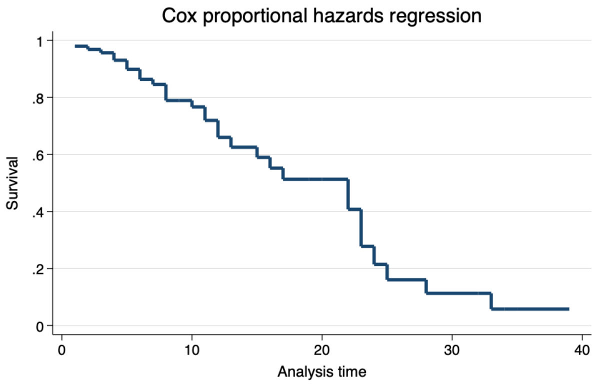


Figure 8: Survival Curve

```
. stcurve, survival at1(drug=0) at2(drug=1) scheme(michigan) // survival curve by group

. graph export survival2.png, width(1000) replace
file /Users/agrogan/Desktop/newstuff/categorical/survival-analysis-and-event-history/survival2.png
    saved as PNG format
```

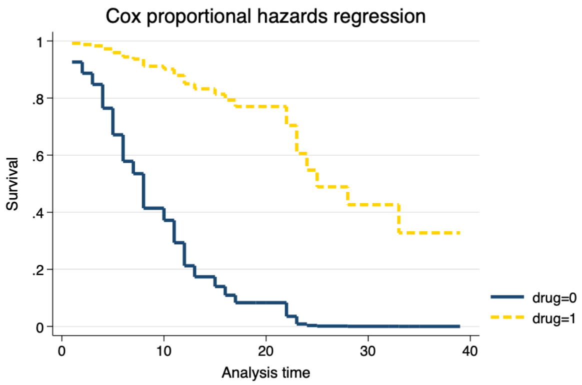


Figure 9: Survival Curve by Drug Group

Proportional Hazards Assumption

```
. estat phtest // formal test of PH assumption
Test of proportional-hazards assumption
Time function: Analysis time
────────────────────────────────────────────────
             │     chi2       df       Prob>chi2
─────────────┼──────────────────────────────────
 Global test │     0.43        2          0.8064
────────────────────────────────────────────────

. stphplot, by(drug) scheme(michigan) // graphical test of PH assumption
        Failure _d: died
  Analysis time _t: studytime

. graph export ph.png, width(1000) replace
file /Users/agrogan/Desktop/newstuff/categorical/survival-analysis-and-event-history/ph.png saved
    as PNG format
```

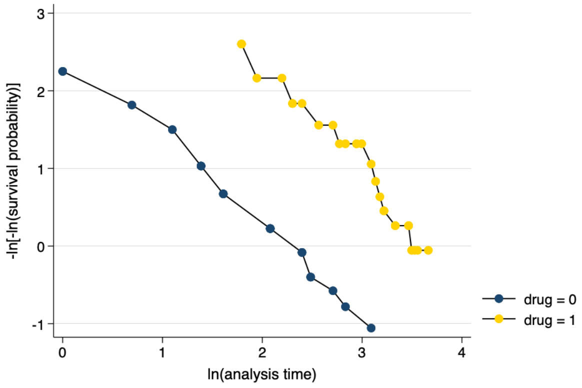


Figure 10: Graphical Assessment of Proportional Hazards Assumptions

References

Johnson, L. L., & Shih, J. H. (2007). CHAPTER 20 - An Introduction to Survival Analysis (J. I. Gallin & F. P. Ognibene, eds.). https://doi.org/https://doi.org/10.1016/B978-012369440-9/50024-4

Ragnar Frisch Centre for Economic Research (2020). Event History Analysis, Survival Analysis, Duration Analysis ,Transition Data Analysis, Hazard Rate Analysis. Oslo, Norway.