Event History Analysis: Introductory Concepts

Categorical and Limited Dependent Variables

Paul A. Jargowsky

Events in Time

- Events
  - Deaths, seizures, heart attacks
  - Computer crashes, mechanical failures
  - Dropping out, getting pregnant
  - Crimes, arrests, **re-arrest** (Rossi et al. 1980)
- Can do simple logit, but information is lost
  - Some happen quickly, some slowly
  - Censoring in time
  - Covariates vary in time also

Single vs. Repeated Events

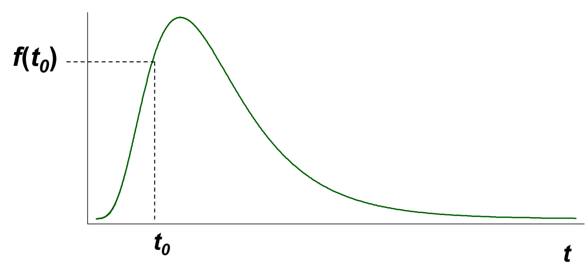
- Single events are easier to model (e.g., death, first marriage, first recurrence of cancer after a treatment).
- Repeated events complicate analysis but offer a more complete picture (e.g., job changes, marriages, childbirths).
- Start with single events only. Can always study time to first Y.

Simple vs. Complex Events

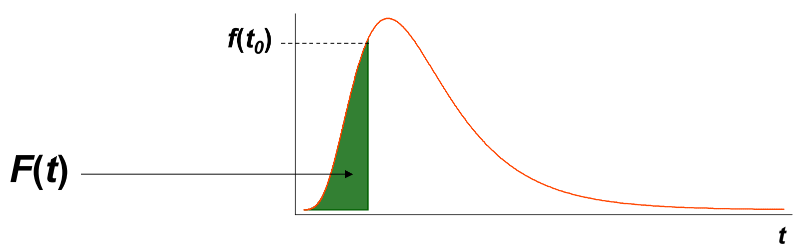
- Question of definition of DV.
- For example, death is a simple event (0/1).
- But there are different kinds of death:
  - Accident;
  - Disease;
  - Old Age.
- Complex events (with subtypes) can be analyzed, but much more complicated.
- We will only discuss simple events.

Framework and Terminology

- Failure Time/Survival Terminology
  - t = time to failure (a random variable)
  - t_0 = some specific time at which failure might occur
  - $f(t)$ the density function of t (the relative likelihood of a failure at time t)



Cumulative Probability of Failure

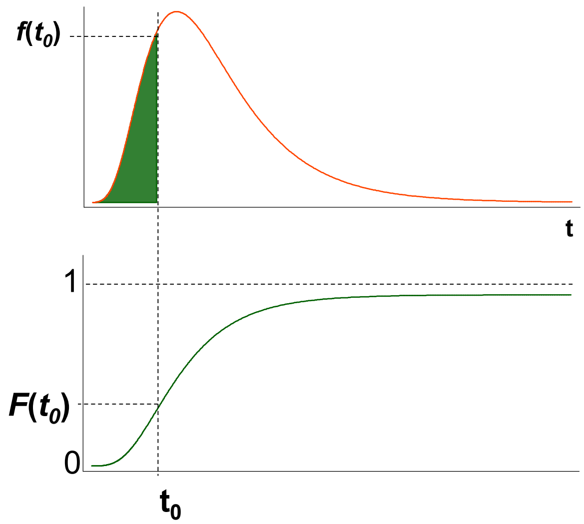


The cumulative probability of a failure up to time t_0 is the area under the density curve from the beginning (t=0) up to time t_0, and is designated as $F(t_0)$. *Generally F(t).*

$$F(t_0) = \int_0^{t_0} f(t)dt = \Pr(t \le t_0)$$

Relationship between f(t) and F(t)

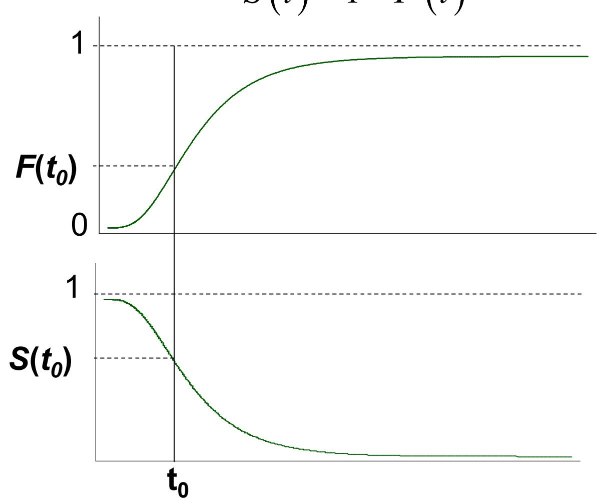
The area under the curve f(t) is the height of the curve F(t). F(t) is bounded by zero and one. The area under the curve f(t) sums to one.



Survival

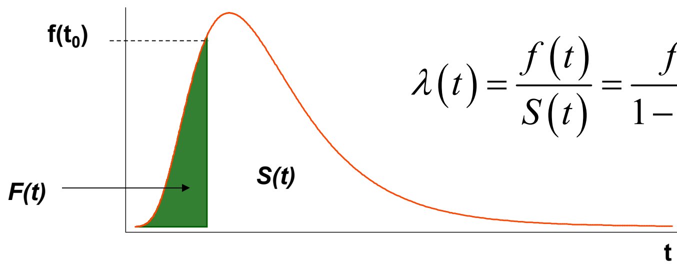
$$S(t) = 1 - F(t)$$

Survival is not failing. Terminology depends on the field of origin. Engineers more interested in when the machine breaks, doctors want to know if the patient lives.



The Hazard Rate

Key concept! The density function for failure, the cumulative probability of failure, and the cumulative probability of survival describe the population as a whole. *But what about the rate of failure today given that a person has survived so far?* This is the "**hazard rate**."



$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}$$

(Note: if the time periods are discrete, you must use S(t-1), since you can't both die and survive in the same time period.)

Example: A Matter of Life and Death

- Event=death, t=years.
  - f(90) = 0.01. Probability (at birth) that a person dies during the 90^{th} year of life (between 90.0000000…. and 90.9999999….)
  - F(89) = 0.95. Probability (at birth) of death before age 90.
  - S(89) = 1-F(89) = 1-0.95 = 0.05. Probability of surviving to the 90^{th} birthday or beyond
  - F(89)+S(89)=1.
- Hazard of death at exactly age 90 is…

$$\lambda(90) = \frac{f(90)}{S(89)} = \frac{0.01}{0.05} = 0.2$$

In other words, the death rate at age 90 for someone who has attained that age is 0.2 (or 20 percent). Can be greater than 1 in continuous time data, but usually not.

Hazard Rates and Time

- Hazard rates can be constant, such as the decay of radioactive atoms, probability of red on roulette wheel
- Hazard rates can have "duration dependence" (vary with time)
  - Positive duration dependence (hazard of machine failure rises over time)
  - Negative duration dependence (hazard of company failure, companies become established over time)
  - Both, e.g. U shape (hazard of death, first declines then is basically flat, then rises – "bathtub")

Hazard Rates vary depending on personal characteristics (the Xs)

Probably of death in given time period:

- $\lambda_{white}(t) \neq \lambda_{black}(t)$
- $\lambda_{rich}(t) \neq \lambda_{poor}(t)$
- $\lambda_{male}(t) \neq \lambda_{female}(t)$

If prostate cancer, two treatments:

- $\lambda_{surgery}(t) \neq \lambda_{seeds}(t)$

"Surgery" is the removal of the prostate. "Seeds"…

Seeds vs. Surgery

STANFORD COMPREHENSIVE CANCER CENTER

Prostate Seed Implants
Prostate seed implants can be a particularly suitable radiotherapy option for patients diagnosed with early stage prostate cancer.

How Prostate Seed Implants Work
About 100 radioactive seeds (Iodine-125) are injected into the prostate under anesthesia where they emit low levels of radiation for a few months. The procedure is usually performed on a one-time, outpatient basis and takes about two hours.

Alternative Models for the Hazard Rate

1. Define the risk: re-occurrence of cancer.
2. Define the time frame: 5 years post procedure.
3. Model the hazard (discrete probability of a recurrence.

$$Y_i = \begin{cases} 0 & \text{No cancer after 5 years} \\ 1 & \text{Cancer recurs within 5 years} \end{cases}$$

OLS: $\lambda_i = \beta_1 + \beta_2 Black_i + \beta_3 SES_i + \beta_3 Seeds_i + u_i$

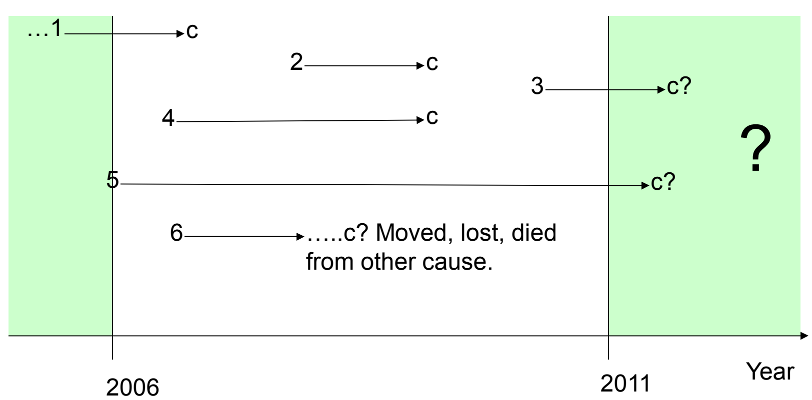
Does not work well. Why?

Logit Approach

$$\ln\left(\frac{\lambda_i}{1-\lambda_i}\right) = \beta_1 + \beta_2 Black_i + \beta_3 SES_i + \beta_3 Seeds_i + u_i$$

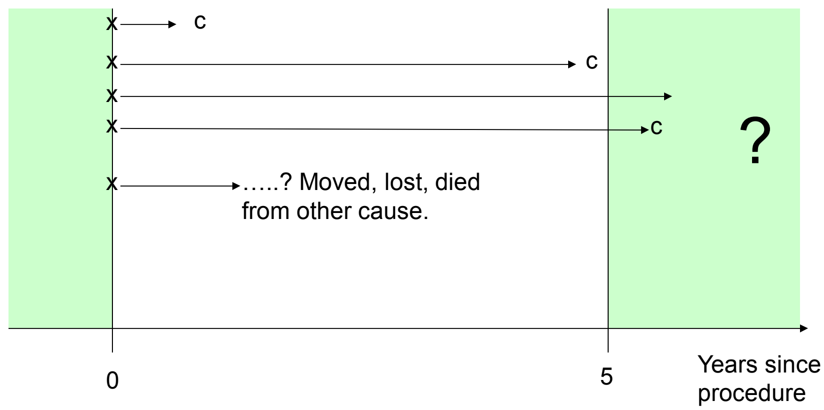
Better than OLS, but still has problems:
- Some had cancer again right away, others after 4 years 11 months (not using all the information we have)
- Some may get cancer 1 day after 5 years
- SES might have changed over time (“time varying covariates”), no way to handle in either OLS or simple logit model

Calendar Time



Calendar time is not particularly relevant. Person 2 and 4 got cancer on same day, but survived different lengths of time. You need the time at risk.

Exposure Time



People who get cancer may do so at different times.
Some people are right censored, some people are truncated.

Marriage Data

```
Contains data from marriage.dta
  obs:         5,000                          1930s cohort: Height and Marriage
 vars:             3                          17 Nov 2009 17:46
 size:       115,000                          (_dta has notes)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
birth           str11  %11s                   Birth Date
mar1date        str11  %11s                   Date of First Marriage
height          byte   %8.0g                  Height (inches)

. list in 1/7

+------------------------------------+
         birth      mar1date   height
 ------------------------------------
1.    7-Feb-1937                  69
2.   30-Oct-1934    4-Jul-1972    71
3.   11-Dec-1939    4-May-1958    73
4.   15-May-1939    3-Feb-1958    70
5.   13-Feb-1936   10-Jan-1955    70
 ------------------------------------
6.    6-Nov-1932                  68
7.    5-Nov-1932    4-Dec-1954    69
```

This is “span” data.
One observation per
person.

Discrete Version of Marriage Data

```
. des

Contains data from mar_discrete.dta
  obs:        91,761                          1930s cohort: Height and Marriage
 vars:             5                          21 Nov 2013 11:37
 size:     1,559,937                          (_dta has notes)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
id              float  %9.0g
year            float  %9.0g
age             float  %9.0g
height          byte   %8.0g                  Height (inches)
firstmar        float  %9.0g
-------------------------------------------------------------------------------
Sorted by:  id
     Note:  dataset has changed since last saved
```

This is person-year data.
Multiple observations per person.
Continues until the event occurs or the study ends.

Discrete Time Method

Panel Data:

ID	Month	Y
Bob	1	0
Bob	2	0
…		
Bob	60	0
Joe	1	0
Joe	2	0
Joe	3	1
Dave	1	0
Dave	2	0
Etc………		

h(t) = **probability** individual has event at time *t*, given still at risk. (When using discrete time, the hazard = probability of death this period *given* survival to this period, and is therefore bounded by 0 and 1.)

Alternative specifications:

$$h_{ti} = \beta_1 + \beta_2 X_i + \beta_3 Z_{ti} + \ldots + u_{ti}$$

$$\ln\left(\frac{h_{ti}}{1-h_{ti}}\right) = \beta_1 + \beta_2 X_i + \beta_3 Z_{ti} + \ldots + u_{ti}$$

$$\ln\left(\frac{h_{ti}}{1-h_{ti}}\right) = \beta_{1t} + \beta_2 X_i + \beta_3 Z_{ti} + \ldots + u_{ti}$$

Interpretation of Coefficients

$$\ln\left(\frac{h_{ti}}{1-h_{ti}}\right) = \beta_1 + \beta_2 X_{2i} + \beta_3 Z_{ti} \ldots + u_i \longrightarrow \frac{h_{ti}}{1-h_{ti}} = e^{\mathbf{x}_{ti}\boldsymbol{\beta}}$$

Therefore: $$\frac{\left(\frac{h_2}{1-h_2}\right)}{\left(\frac{h_1}{1-h_1}\right)} = \frac{e^{\mathbf{x}_i\boldsymbol{\beta}+\delta\beta_k}}{e^{\mathbf{x}_i\boldsymbol{\beta}}} = e^{\delta\beta_k}$$

*So, given different values of X, the **odds** are proportional. If greater than 1, hazard is increasing, etc.*

Proportional Odds Model

- In discrete time hazard model, if all variables are time independent (race, gender, etc.), the model implies a proportional odds structure.

$$\mathbf{x}_i\boldsymbol{\beta} = \beta_1 + \beta_2 male_i + \beta_3 white_i + \ldots + u_i$$

$$OR_{M|F} = \frac{\left(\frac{P_{male}}{1-P_{male}}\right)}{\left(\frac{P_{Female}}{1-P_{Female}}\right)} = e^{\beta_2}$$

So regardless of the values of the other variables, the odds of males and females have a constant proportionality.

The Base Hazard

- If all covariates are zero:

$$\left(\frac{h_0}{1-h_0}\right) = e^{\beta_1}$$

$$\left(\frac{h_i}{1-h_i}\right) = e^{\beta_1+\beta_2 X_{2i}+\beta_3 X_{3i}+\ldots}$$

$$= e^{\beta_1} e^{\beta_2 X_{2i}} e^{\beta_3 X_{3i}} \ldots$$

$$= \left(\frac{h_0}{1-h_0}\right) e^{\beta_2 X_{2i}} e^{\beta_3 X_{3i}} \ldots$$

So all hazards are proportional to the base hazard.

Continuous Time Models

- Focus is on **hazard rates**, not odds
- Smaller data files, 1 obs per person
- t = time to failure/event, or dates
- Finer measure of time, more flexible
- Model the hazard rate directly
- *Many different models*
- Main difference: assumption about the form of *duration dependence of the hazard*
- A little harder (not impossible) to model time-varying covariates

Parametric Regression Models

A: $\ln(h_t) = \beta_1 + \beta_2 X_i$

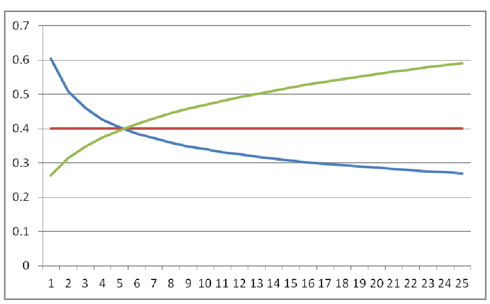
- The hazard is constant (after controlling for X)
- Implies an exponential distribution for the time until an event occurs.
- Thus, it is called the "exponential regression model"

B: $\ln(h_t) = \beta_1 + \beta_2 X_i + \beta_3 t$

- Gompertz distribution for t
- Known as "Gompertz regression model"

Models continued

C: $\ln(h_t) = \beta_1 + \beta_2 X_i + \beta_3 \ln(t)$, restriction: $\beta_3 > -1$



- Known as Weibull regression model

For A-C: no disturbance term. Variance/randomness comes in by whether you die, given your hazard rate.
B-C: hazard can increase or decrease, but not both.

Parametric Estimation

For uncensored $P_i = f(t) = \lambda(t) S(t)$

For censored $P_i = S(t)$

$$\mathcal{L} = \prod_{i=1}^{n} [P_i] = \left(\prod_{event} [f(t)] \right) \left(\prod_{censored} [S(t)] \right)$$
$$= \left(\prod_{event} [\lambda(t) S(t)] \right) \left(\prod_{censored} [S(t)] \right)$$
$$= \left(\prod_{event} [\lambda(t)] \right) \left(\prod_{all} [S(t)] \right)$$

Plug in exponential, Gompertz, Weibull, etc.

Cox Proportional Hazards Model

$$\frac{h(t)}{h_0(t)} = e^{\sum \beta_k X_{kt}} \qquad h(t) = h_0(t) e^{\sum \beta_k X_{kt}}$$

- "Semi-parametric" – baseline relationship with t not specified, t only orders the observations
- Can assume time dependence without assuming a specific form, good for weird/unknown forms of duration dependence
- Baseline hazard function varies by time in an *unspecified* way (model will determine how)
- Not as efficient as parametric model, if correct parameterization is known.