Introduction to Event History Analysis
Objectives

- Introduce event history analysis
- Describe some common survival (hazard) distributions
- Introduce some useful Stata and SAS commands
- Discuss practical issues worth keeping in mind
Objectives

- Introduce event history analysis
- Describe some common survival (hazard) distributions
- Introduce some useful Stata and SAS commands
- Discuss practical issues worth keeping in mind
What is event history analysis?

- A set of statistical techniques used to analyze the time it takes an event to occur within a specified time interval

- Also called survival analysis (demography, biostatistics), reliability analysis (engineering), duration analysis (economics)

- The basic logic behind these methods is from the life table

- Types of “Events” – Mortality, Marriage, Fertility, Recidivism, Graduation, Retirement, etc.
Basic Concepts in Event History Analysis

- Events
  - Repeatable vs. Non-Repeatable
  - Single vs. Multiple

- Exposure to Risk (i.e., Measuring Time)
  - Risk Set
  - Discrete vs. Continuous Time
  - Censoring

- Hazard & Survival Functions
  - Non-Parametric vs. Semi-Parametric vs. Parametric
  - Distributional Assumptions (Proportionality)
Basic Concepts: Different Types of Events

- Event (the outcome) - A discrete transition between two “states”
  - Non-Repeatable Events
    - Transition can occur only once (absorbing state)
    - Examples: Alive → Dead, Nulliparous → First Birth
  - Repeatable Events
    - Transition can occur more than once (non-absorbing state)
    - Examples: Married ↔ Divorced, Healthy ↔ Disabled
  - Single Events – i.e., Alive → Dead
  - Multiple Events – i.e., Alive → Cancer Deaths vs. CVD Deaths
    - Methods for competing risks an extension of those for single events
Basic Concepts: Time (Exposure, Duration)

- Time is the core component of event history analysis
  - *Risk Set* – Individuals\(^1\) at risk of experiencing some event
    - Risk *exposure* occurs in an *observation interval* (study time)
    - The observation interval is when the “clock” begins and ends
  - One of two outcomes are possible in the observation interval
    - *Failure* – Event occurs in the interval (i.e., death)
    - *Censoring* – Event does not occur in the interval (i.e., survival)
  - Time usually is measured in *discrete* units (i.e., years, months)
  - Time theoretically can be measured in (quasi) *continuous* units (i.e., hours, minutes, seconds)

---

\(^1\) Note that the unit of analysis does not necessarily have to be individuals.
Basic Concepts: Time (Exposure, Duration)

Figure 2.1.1 Types of censoring in an observation window.

Objectives

- Introduce event history analysis
- Describe some common survival (hazard) distributions
- Introduce some useful Stata and SAS commands
- Discuss a few practical issues worth keeping in mind
**Basic Concepts: Hazard & Survival Functions**

- **Hazard Function** – Instantaneous probability\(^2\) that an event will occur at time \(t\), conditional that the event has not already occurred.

\[
\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t + \Delta t > T \geq t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}
\]

- Also called the *Hazard Rate* or the *Force of Mortality*

- With some additional math, you can get the *Survival Function*

\[
S(t) = \exp\{-\lambda t\}
\]

\(^2\) Note that strictly speaking the hazard rate is a probability only in discrete-time models.
Basic Concepts: Hazard & Survival Functions

- Models impose different distributional assumptions on the hazard
- Three basic types of hazard (survival) functions are common
  - Each one imposes different amounts of “structure” on the data
  - The ultimate decision to use one approach over another should be driven by:
    - Your specific research question
    - How well the model fits the actual data
    - Practical concerns – i.e., difficulty estimating with available software, interpretability, “typical” approach in previous research
Basic Concepts: Hazard & Survival Functions

- **Non-Parametric Models**
  - No assumptions about the baseline hazard distribution
  - Pros: Imposes the least structure, easy to estimate and interpret
  - Cons: Difficult to incorporate predictors (mostly descriptive)
  - Examples: Kaplan-Meier, Nelson-Alan, “Classic” Life Table

- **Parametric Models**
  - Baseline hazard assumed to vary in a specific manner with time
  - Pros: Easy to incorporate covariates, gives baseline hazard to calculate rates, smoothes “noisy” data
  - Cons: Imposes the most structure, need to be sure that estimated distribution matches the data
  - Examples: Weibull (decrease or increase), Gompertz (exponential increase), Exponential (constant)
Basic Concepts: Hazard & Survival Functions

- Semi-Parametric Models
  - Baseline hazard is not pre-determined, but it must be positive.
  - Pros: Covariates easily incorporated, less structure than parametric, smoothes “noisy” data
  - Cons: Does not provide the baseline hazard
    - Cannot calculate rates (absolute differences)
    - Can only interpret in terms of relative differentials
    - Any specification errors are “absorbed” into the coefficients
  - Examples: Cox Proportional Hazards (most popular model)

- Proportional Hazards Assumption
  - The hazard rate is equivalent over time across groups
  - Cox models must satisfy this assumption
  - Some parametric models - Weibull, Gompertz, Exponential, etc.
Objectives

- Introduce event history analysis
- Describe some common survival (hazard) distributions
- Introduce some useful Stata and SAS commands
- Discuss practical issues worth keeping in mind
How To Estimate Hazard Models

- SAS – “lifereg” (parametric models), “phreg” (Cox models), “lifetest” (Kaplan-Meier), other user-written macros available

- Stata3 - “streg” (parametric models), “stcox” (Cox models), “sts test” (“Kaplan-Meier”), other specialty packages as .ado files

- R – package “survival” (parametric and Cox models), “KMsurv” (Kaplan-Meier), other specialty packages “frailpack,” etc.

- SPSS – “coxreg” (Cox models), “km” (Kaplan-Meier), no parametric models available?

---

3 You must "stset" the data before estimating survival models in Stata. Type "help st" for details.
Stata Example: Exponential Model

```
. streg ib4.edcat4 ib0.female age, d(exponential)

Iteration 0: log likelihood = -134647.19
Iteration 1: log likelihood = -111943.39
Iteration 2: log likelihood = -104339.66
Iteration 3: log likelihood = -104000.24
Iteration 4: log likelihood = -103999.94
Iteration 5: log likelihood = -103999.94

Exponential regression -- log relative-hazard form

No. of subjects = 489830                         Number of obs  =  489830
No. of failures =  29199                         LR chi2(5)     =  61294.51
Time at risk    = 2882523.743                    Prob > chi2    =  0.0000
Log likelihood  = -103999.94

             _t | Haz. Ratio   Std. Err.     z     P>|z|    [95% Conf. Interval]
-------------|-------------|-----------------|-------|--------|-----------------------------|
       edcat4 |             |                 |       |        |                            |
        1     | 1.964928    |  .039325        |  33.75|  0.000  |  1.889345                 |
        2     | 1.583016    |  .0316355       |  22.98|  0.000  |  1.522211                 |
        3     | 1.421683    |  .031107        |  16.08|  0.000  |  1.362004                 |
       1.female| .6486875    |  .0076684       | -36.61|  0.000  |  .6338305                 |
       age    | 1.088775    |  .0004605       | 201.07|  0.000  |  1.087872                 |
       _cons  | .0000654    |  2.11e-06       | -299.33|  0.000  |  .0000614                 |
### Stata Example: Cox PH Models

```stata
. stcox ib4.edcat4 ib0.female age
   failure _d:  dead
   analysis time _t:  expoxy

Iteration 0:  log likelihood = -371246.33
Iteration 1:  log likelihood = -340946.22
Iteration 2:  log likelihood = -340166.37
Iteration 3:  log likelihood = -340165.32
Iteration 4:  log likelihood = -340165.32
Refining estimates:
Iteration 0:  log likelihood = -340165.32

Cox regression -- Breslow method for ties

<table>
<thead>
<tr>
<th>No. of subjects</th>
<th>Number of obs</th>
<th>489830</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of failures</td>
<td>29199</td>
<td></td>
</tr>
<tr>
<td>Time at risk</td>
<td>2882523.743</td>
<td></td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-340165.32</td>
<td></td>
</tr>
</tbody>
</table>

| t    | Haz. Ratio | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|------|------------|-----------|-------|------|---------------------|
| edcat4 |            |           |       |      |                     |
| 1    | 1.96006    | 0.0392274 | 33.63 | 0.000| 1.884664            |
| 2    | 1.583022   | 0.0316374 | 22.98 | 0.000| 1.522213            |
| 3    | 1.425253   | 0.0311864 | 16.19 | 0.000| 1.365421            |

1.female | 0.6435176 | 0.0076114 | -37.27 | 0.000 | 0.6287712           |

age      | 1.090236  | 0.0004665 | 201.90 | 0.000 | 1.089322            |
SAS Example: Cox Proportional Hazards Model

```
proc phreg data = nhis ;
   model expos*dead(0) = female age ;
run ;
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
<th>Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>lths</td>
<td>1</td>
<td>0.67266</td>
<td>0.020010</td>
<td>1129.662</td>
<td>&lt;.0001</td>
<td>1.959</td>
</tr>
<tr>
<td>hs</td>
<td>1</td>
<td>0.45942</td>
<td>0.019990</td>
<td>528.4468</td>
<td>&lt;.0001</td>
<td>1.583</td>
</tr>
<tr>
<td>scol</td>
<td>1</td>
<td>0.35431</td>
<td>0.021880</td>
<td>262.1947</td>
<td>&lt;.0001</td>
<td>1.425</td>
</tr>
<tr>
<td>female</td>
<td>1</td>
<td>-0.44060</td>
<td>0.011830</td>
<td>1387.689</td>
<td>&lt;.0001</td>
<td>0.644</td>
</tr>
<tr>
<td>age</td>
<td>1</td>
<td>0.08639</td>
<td>0.000428</td>
<td>40766.7</td>
<td>&lt;.0001</td>
<td>1.090</td>
</tr>
</tbody>
</table>
Objectives

- Introduce event history analysis
- Describe some common survival (hazard) distributions
- Introduce some useful Stata and SAS commands
- Discuss a few practical issues worth keeping in mind
Other Issues: Data Structure

- The data structure has important substantive implications
- The models shown here were estimated on individual-level data
- Models estimated on *person-period* data can be used to answer other substantive questions.
  - Easy to calculate various life table functions - central death rates ($m_x$), probabilities of death ($q_x$), etc.
  - Easy to incorporate *time-varying* covariates (age, etc.)
Other Issues: Alternative Models

- Always test model assumptions, evaluate model fit, etc.
  - Compare how well various models fit the data
    - Fit statistics - BIC, AIC, etc.
    - Fitted vs. Observed values – Do the distributions overlap?
  - Proportionality assumption (proportional hazards models)

- Other approaches often yield equivalent results
  - Count models - Poisson models, Negative Binomial models, etc.
  - Logistic Regression – Similar to Cox models, especially when few observations are censored