Labor Economics with STATA

Liyousew G. Borga

March 9, 2016

Duration Analysis

Outline

Duration Models

"models of the length of time spent in a given state before transition to another state"

- event history analysis
- survival analysis
- failure-time models
- hazard modeling

What is Duration Analysis?

- Methods for the analysis of length of time until the occurrence of some event
- The dependent variable is the duration until event occurrence

Examples

- Health: Age at death; duration of hospital stay
- Demography: Time to first birth; time to first marriage; time to divorce
- Economics: Duration of an episode of employment or unemployment
- Education: Time to leaving full-time education; time to exit from teaching profession
- Politics: Time before a political party losing election; regime change
- Others: Recidivism; conflict

Questions we are interested in:

- What are the chances I get this job?
- What are the chances I'll graduate with PhD?
- How likely is it my business will fail?
- Will the government survive the election?
- Will the criminal return to crime?

Essential Elements of Duration Analysis

- The Event: Something can happen; there is a “chance” this event may occur
- Timing: Given that something hasn’t happened, what are the chances it will happen subsequently?
- The Risk: a relationship between the chances that something can happen relative to the chances that it hasn’t happened yet

Example:

- Event: Divorce
- Timing: Years Married
- The Risk: “Given a couple has remained married 10 years, what is the likelihood they will divorce next year?”

Definition of some terms

- Failure: The unconditional probability that an event will occur
- Survival: The probability that "up until now" the event has not yet occurred
- Risk: The conditional failure rate - given that the event has not yet occurred, what are the chances it will occur?

$$\text{Risk} = \frac{\text{“Chance that Something Happens”}}{\text{“Chance that it hasn't Happened Yet”}}$$

$$\text{Hazard} = \frac{Pr(\text{“}failure\text{''})}{Pr(\text{“}survival\text{''})}$$

Types of duration data

- Dates of start of exposure period and events, e.g. dates of start and end of an employment spell
    - Usually collected retrospectively
    - Sources include panel and cohort studies (partnership, birth, employment and housing histories)
- Current status data from panel study, e.g. current employment status at each year (Collected prospectively)
- Durations are always positive and their distribution is often positively skewed

Special features of duration data

Censoring

- There are usually people who have not yet experienced the event when we observe them, but may do so at an unknown time in the future
- In general, censoring occurs whenever an observation's full event history is unobserved

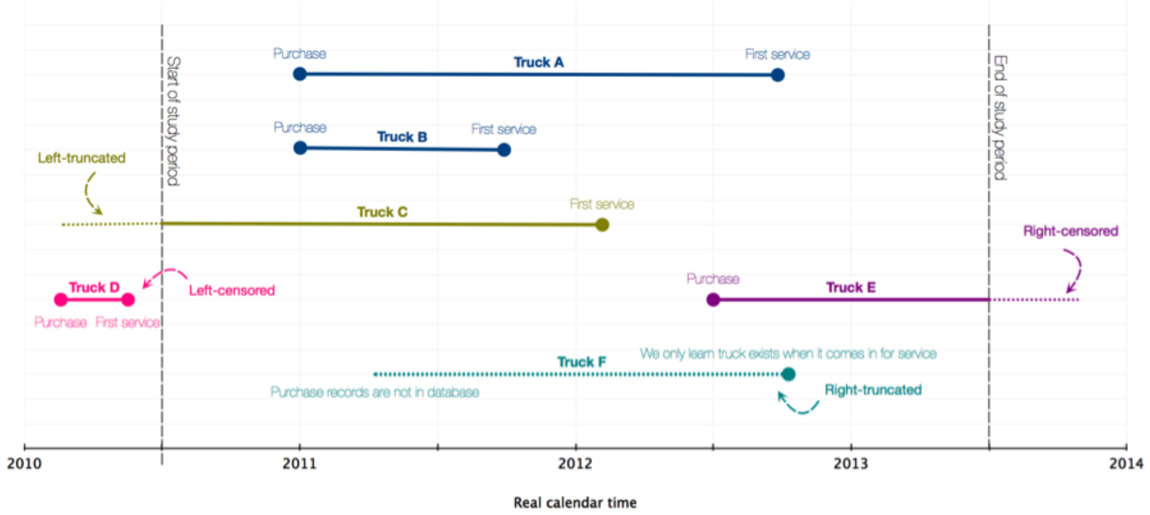
Time-varying covariates

- The values of some covariates may change over time

Types of Censoring

- start and end time known
- Right censoring: end time outside observation period
- Left-truncated: start time outside observation period, i.e.
- start and end time outside observation period

Types of Censoring



Types of Censoring

- Right-censoring is commonly observed in event history data sets
- Durations are right-censored if the event has not occurred by the end of the observation period
- Excluding right-censored observations (e.g. still married) leads to bias and may drastically reduce sample size
- Usually assume censoring is non-informative (i.e. assume that the censored history is missing at random)

Event times and censoring times

- Denote the event time (duration, failure or survival time) by the random variable T
- t_i event time for individual i
- δ_i censoring/event indicator

$$\begin{cases} = 1 & \text{if uncensored (i.e. observed to have event)} \\ = 0 & \text{if censored} \end{cases}$$

- But for a right-censored case, we do not observe t_i
- We observe only the time at which they were censored, c_i
- Our outcome variable is $y_i = min(t_i, c_i)$;
- Our observed data are (y_i, δ_i)

Problems of OLS in Duration analysis

Typically researchers do the following to analyze an event:

- Have some theory or hypothesis relating timing and other factors (i.e. independent variables or covariates) to some event
- Observe some "sample" over time
- Record whether or not some event of interest occurs over time
- Collect data on important covariates
- Model the "event" or the "time until the event" as a function of covariates, and perhaps, time itself

Problems with OLS:

- OLS may return negative predicted values -an impossibility: "survival times" must be positive
- Duration data are often right-skewed, often times, heavily so
- OLS does not easily distinguish "censored" from "uncensored" cases
- OLS cannot easily accommodate covariates that change value over time (TVCs).
- Assumed linearity in the survival times may be unrealistic

"FIX":

- Treat $log(t)$ as the response variable: mitigates the skewness problem to some degree
- Parametric and Nonparametric modeling strategies

Notation:

- Let X denote the random variable time-to-event.
- $f(x)$ is a probability density function
- F(x) denotes cumulative distribution function

The distribution of X can be described by several equivalent functions:

- Survival function,

$$S(x) = Pr(X > x)$$
$$= 1 - F(X)$$

- Hazard function,

$$h(x) = \lim_{\Delta x \to 0} \frac{Pr[x \leq X \leq x + \Delta x | X \geq x]}{\Delta x} = \frac{f(x)}{S(x)}$$

- Cumulative hazard function

$$H(x) = \int_0^x h(x) dx$$

Discrete Data

- It is very common for a duration to be measured as an interval
- E.g. data may indicate that a transition occurred in a particular week, but the exact time in the week is not given
- In such cases the transition times are said to be grouped and it is assumed that the hazard within the interval is constant
- Discrete-time hazard models deal with such data

Nonparametric Estimation

- If the data were not censored, the empirical estimate of the survival function, $\hat{S}(t)$, is the proportion of individuals with event times greater than t
- If there are censored observations, then $\hat{S}(t)$ is not a good estimate of the true $S(t)$, so other non-parametric methods must be used to account for censoring
  - The Kaplan-Meier estimator
  - The Nelson-Aalen estimator

Examples of parametric distribution families:

- Exponential distribution:

  $f(x) = \lambda e^{-\lambda x}$

  $S(x) = e^{-\lambda x}$

  $h(x) = \lambda$

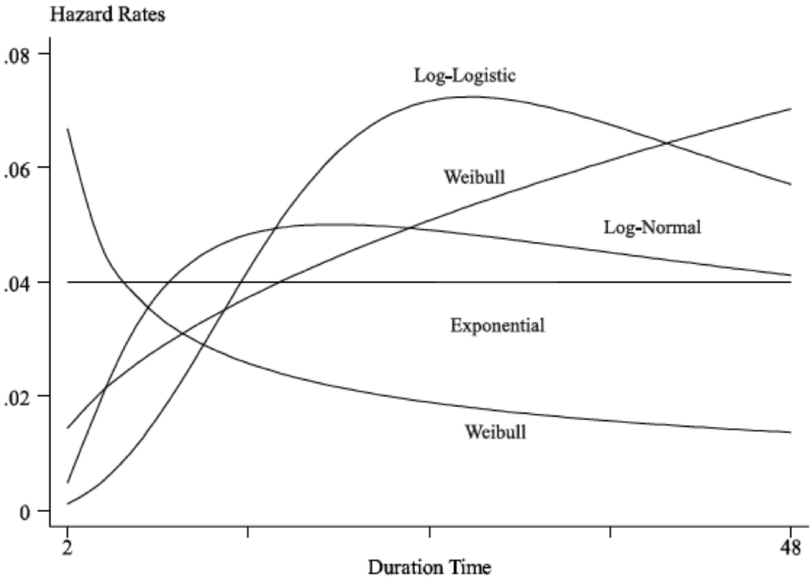
- Weibull distribution:

  $f(x) = \lambda \alpha x^{\alpha-1} e^{-\lambda x^{\alpha}}$

  $S(x) = e^{\alpha} x^{\lambda x^{\alpha}}$

  $h(x) = \lambda \alpha x^{\alpha-1}$

Functional forms for common parametric distributions



Parametric Models With Covariates:

Approaches to modeling survival data with covariates:

- The Proportional Hazard Form:

  assumes that the effect of the covariates is to increase or decrease the hazard by a proportionate amount at all durations
- The Accelerated Life Form:

  assumes that the covariates rescale time directly

- Stata's `st (survival time)` suite of commands provide sophisticated tools for these tasks

```
stset       Declare data to be survival-time data
stdes       Describe survival-time data
stsum       Summarize survival-time data
sts         Generate, graph, list, and test
stcox       Fit Cox proportional hazards model
streg       Fit parametric survival models
```

Syntax of the `stset` command

```
stset timevar [if] [weight] , failure(failvar[==numlist]) [options]
```

For example,

```
stset survtime, failure(dead==1)
```

The `stset` command creates 4 variables

- _t0 - analysis time when record begins (time at which individual becomes at risk)
- _t - analysis time when record ends (time at which individual stops being at risk)
- _d - failure indicator: 1 if failure, 0 if censored
- _st - 1 if the record is included in st analyses, 0 if excluded

All the survival analysis (st) commands use these variables