Event History/Survival/Duration/Hazards Models

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

April 3, 2020

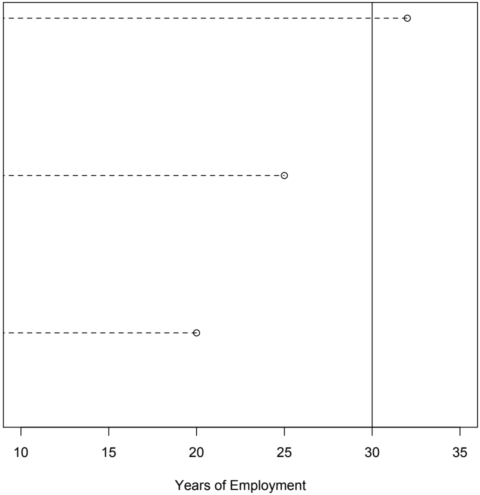
Overview

1. Intro
2. Basics
3. Cox Model
4. Parametric Models

Survival/Duration/Event History Data

- Observations represent the occurrence of a particular event over a period of time
- Fundamental goal of analysis is to determine survival time or'how long' it takes for some event to occur
- Initial analysis of duration data involved fitting OLS regression lines to data
  - Underlying theory is that time is continuous
  - Problem is that some events have not occurred at end of observation (i.e. censored)
  - How does one model censoring?

Example of Duration Data: When to Retire



Solutions for Censored Data

- Treat censored observation as equivalent to last observed data point
- Eliminate censored observation(s)
  - This solution only works if the factors which contribute to the censoring (i.e. extended life beyond the sample) are unrelated to the factors promoting an event's occurrence
  - If factors are related, than elimination of censored observations leads to biased estimates
- Create a binary indicator variable (coded '1' if event occurs and '0' otherwise)
  - Problem is that the dummy variable cannot capture the variation in duration time, which is precisely what we try to model
  - New indicator variable does not bias estimates, but leads to inefficiency in the model

Logic of Survival/Duration/Event History Models

- Underlying premise is that the survival/duration/ time-until-event of some process is modeled
- Technique originated from biostatistics to predict how long an individual will live after given specific medical treatments
- Overall approach involves modeling three related concepts
  1. Survivor function
  2. Occurrence of an event
  3. Hazard rate

Survival Data Basics

- Suppose we're modeling the life-span of patients in a hospital. Each patient, i, is observed across periods of time, t, where the length of their survival is denoted T_i.
- We might want to model the probability that a patient expires on or before a given period of time, t.
- Denote this cumulative probability as:

$$Pr(T_i \leq t) \equiv F(t) = \int_0^t f(t)dt.$$

- $F(t)$ is the probability of death on or before t.
- Conversely, we can get the probability of survival to t as:

$$Pr(T_i \geq t) \equiv S(t) = 1 - F(t).$$

Hazard Rates

- We'd like to know the probability of observing an event at t, provided that we haven't observed it already: $Pr(T_i = t \mid T_i \geq t)$.
- This figure is known as the "hazard" and is denoted as $h(t)$.
- The hazard can be expressed as a proportion of $f(t)$ (the probability density of f at some t) and $S(t)$ (the cumulative probability of survival to some t):

$$h(t) = \frac{f(t)}{S(t)}.$$

Assumptions about the Hazard Rate

- Assumptions most often based on the rate's dependency, or relationship, to time
  - Is the rate constant?
  - Does it increase or decrease?
- If rate is constant (i.e. time invariant)
  - We can estimate it using an exponential distribution
  - The hazard rate at any given time point is equal to the hazard rate at any other point in time: $h(t) = h$
  - Graphical depiction produces a flat line

Assumptions about the Hazard Rate

- If rate is time dependent
  - Need to determine whether event is affected by discrete time (i.e. finite categories) or continuous time
- Discrete Time
  - Goal of these models is to use the statistical model to derive estimates of the underlying hazard probability of a unit experiencing an event
  - Whether or not event is experienced is determined by the observed dependent variable
  - Since an event can occur only at discrete time intervals, we can assume that the probability of event T occurring at time t is also observable

Modeling Discrete Time

- $\lambda(t) = Pr(T = t | T \geq t)$
- Where $\lambda(t) =$ the discrete time hazard function
- $\lambda(t)$ can be interpreted as the probability that a unit experiences an event at time t, given the event has yet to be experienced

Modeling Discrete Time

- Most analysts want to know how specific independent variables affect the hazard rate
- $\lambda(t) = \Pr(T = t | t \geq t; \alpha, \boldsymbol{X\beta})$
  - where α represents a baseline probability (when covariates equal zero) and $\boldsymbol{X\beta}$ represents matrix of independent variables and their parameters
- Cox (1972) demonstrates that the λ probabilities can be parameterized through the logistic distribution

$$\lambda(t) = \frac{1}{1 + \exp^{-[\alpha + \boldsymbol{X\beta}]}}$$

Modeling Discrete Time

- Estimating this equation requires a logistic transformation

$$\ln \frac{\lambda(t)}{1-\lambda(t)} = \alpha + \boldsymbol{X\beta}$$

- This model can be estimated with a variation of the logit model, called the proportional hazards model

Cox Proportional Hazards Model

- Logic behind the proportional hazards model

$$\lambda(t) = \frac{\text{probability of failing between times } t \text{ and } t + \Delta t}{(\Delta t)(\text{probability of failing after time } t)}$$

- Note: the data MUST be `stset` in Stata (using the `stset` command) to designate that observations are based on 'survival time'
- Syntax for the command is: `stset [timevar [if] [, id(idvar), failure(failvar[==numlist])]`

stset Example: Judicial Retirements

```
stset years_on_court, id(jcode) failure(voluntary==1)

                id:  jcode
     failure event:  voluntary == 1
obs. time interval:  (years_on_court[_n-1], years_on_court]
 exit on or before:  failure

------------------------------------------------------------------------------
       3601  total observations
         72  observations end on or before enter()
------------------------------------------------------------------------------
       3529  observations remaining, representing
        388  subjects
        144  failures in single-failure-per-subject data
       7250  total analysis time at risk and under observation
                                                 at risk from t =         0
                                      earliest observed entry t =         0
                                           last observed exit t =        60
```

Cox Proportional Hazards Model

- Stata syntax for estimating Cox Model:
- `stcox [varlist] [if] [in] [, options]`

```
. stcox vested ideoagree minority sex if appointed==1

Cox regression -- Breslow method for ties

No. of subjects =          146                 Number of obs    =      1,532
No. of failures =           63
Time at risk    =         3058
                                               LR chi2(4)       =      31.49
Log likelihood  =   -222.84145                 Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      vested |   11.32882   6.900349     3.99   0.000     3.433368    37.38083
   ideoagree |   1.195356   .3236574     0.66   0.510     .7031108    2.032219
    minority |   .8834718   .6404115    -0.17   0.864     .2133896    3.657734
         sex |   .5658492   .4108675    -0.78   0.433     .1363442     2.34836
------------------------------------------------------------------------------
```

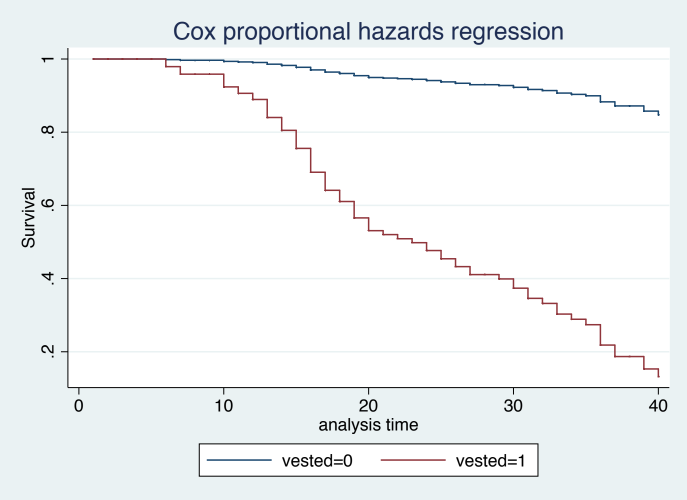
Interpreting Hazards Ratios

- Hazard ratios equal to one indicate that in the presence of a covariate, the hazard of a failure is no more or less than in the absence of a covariate.
- Hazard ratios greater than one indicate an increasing hazard of failure, while rates less than one indicate a decreasing hazard of failure.
- For example, with a hazard ratio of 11.33 on the variable, "vested," we learn that the hazard of a voluntary retirement for a judge who has vested in her pension is over 11 times greater than a similarly situated judge who has not vested, all else equal.

Cox Proportional Hazards Model: Postestimation

- Post-estimation graphing commands
- Basic syntax: `stcurve, hazard` or `stcurve, survival`
- Alternatively: `stcurve, hazard at1(varname=value) at2(varname=value)`
  or `stcurve, survival at1(varname=value) at2(varname=value)`

stcurve Example



Cox Model Assumptions

1. Non-Informative Censoring — mechanisms responsible for censoring observations unrelated to the likelihood of an event occurring
2. Proportional Hazards Assumption — if an explanatory variable is altered the new hazard rate will be proportional to the old one
   - This is easy to test for in Stata. Use the command `estat phtest` post estimation.

Exponential and Weibull Models

- Limitation of the Cox regression
  - Estimates baseline survival function without a theoretical justification for the statistical distribution
  - Offers no assumptions about the relation of the hazard rate to time
- Exponential Models
  - Assumes that the hazard rate remains constant
  - Therefore, ‘failures’ assumed to occur randomly
- Weibull Regressions
  - Assumes that the hazard rate either increases or decreases over time

Exponential and Weibull Models

- How do we know which model to use?
  - Need to examine and identify trends in the baseline hazard
- Kaplan-Meier survival estimate graph
  - Based on following equation

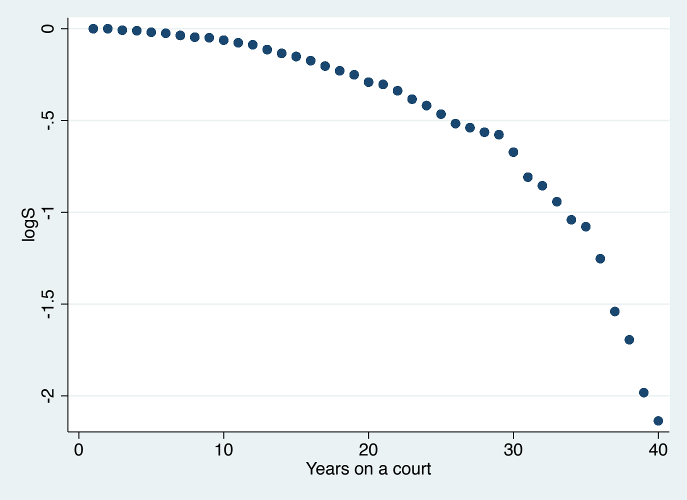
$$S(t) = \prod_{j=t_0}^{t} \frac{(n_j - d_j)}{n_j}$$

- Where $n_j = \#$ of observations that have not failed and are not censored, and $d_j = \#$ failures occurring at time t

Exponential and Weibull Models

- Limitations of Kaplan-Meier graphs
  - Unadjusted graphs are somewhat misleading because the hazard rate will always fluctuate over time
  - To correct for this, we graph the natural log of survival time $\ln S(t)$ versus time
  - If line appears relatively straight, then the Exponential Model is more appropriate
- Stata syntax for Kaplan-Meier log versus time graph:
  - `sts gen S = S`
  - `gen logS = ln(S)`
  - `graph twoway scatter logS timevar`
    - Note: `timevar` above is the variable that you `stset` your data by

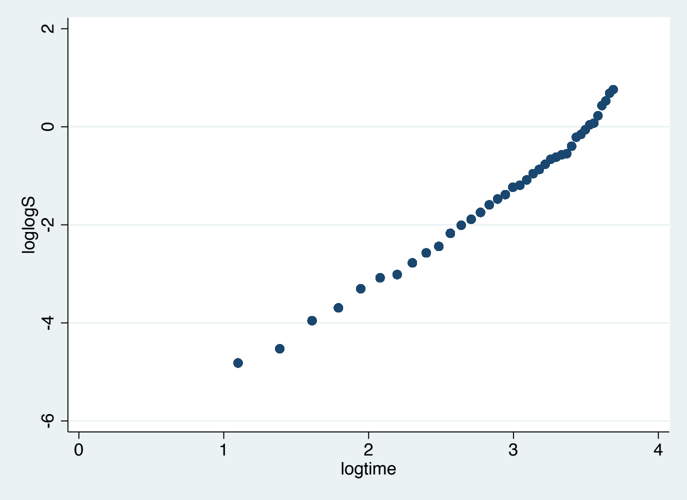
Log Versus Time Example



Exponential and Weibull Models

- One last adjustment needed to be confident that the Weibull model is not more appropriate
- Weibull distribution might appear curvilinear in the log versus time plot, but will be linear in a loglog plot $\ln[-\ln S(t)]$
- Exponential distribution will appear linear in both plots, and have a slope equal to 1 in the loglog plot
- Stata syntax for loglog plot:
  - `gen loglogS = ln(-ln(S))`
  - `gen logtime = ln(timevar)`
    - Note: `timevar` above is the variable that you `stset` your data by
  - `graph twoway scatter loglogS logtime`

Log-Log Example



Exponential and Weibull Models

- Estimation of Exponential or Weibull Models
- Stata syntax:
  - `streg [varlist] [if] [in] [, options]`
  - Key option:
    - `distribution(weibull)` when estimating Weibull model
    - `distribution(exponential)` when estimating Exponential model

Exponential Model Example

```
. streg vested ideoagree minority if appointed==1, distribution(weibull)

Weibull regression -- log relative-hazard form

No. of subjects =          145                  Number of obs     =      1,492
No. of failures =           59
Time at risk    =         2973
                                                LR chi2(3)        =      26.19
Log likelihood  =   -64.170637                  Prob > chi2       =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      vested |   10.18912   6.242528     3.79   0.000     3.066436    33.85626
   ideoagree |   1.384442   .3791402     1.19   0.235     .8094061    2.368006
    minority |   .8282316   .5984289    -0.26   0.794     .2009677    3.413323
       _cons |   .0000564   .0000575    -9.60   0.000     7.65e-06     .000416
-------------+----------------------------------------------------------------
       /ln_p |   .7615059   .1262756     6.03   0.000     .5140102    1.009002
-------------+----------------------------------------------------------------
           p |   2.141499   .2704191                      1.671983    2.742861
         1/p |   .4669627    .058966                      .3645828    .5980923
------------------------------------------------------------------------------
```

Weibull Model Example

- Note: the ρ parameter in the Weibull provides information about the hazard rate
  - If $\rho = 1$ then Weibull equals Exponential
  - If $\rho > 1$ then hazard increases over time if $\rho < 1$ then hazard decreases

Comparing Models

- Cox Proportional Hazards Model
  - Fewer parameters to estimate
  - Easier, more parsimonious model
  - If hazard rate is related to time, this model produces biased estimates
- Exponential or Weibull Model
  - More parameters to estimate
  - Models more susceptible to specification error
  - If hazard rate is not related to time, these models produce biased estimates
- Kaplan-Meier Graphs
  - Probably the best way to determine proper specification (unless there is a theoretical reason)