Duration analysis

CONTENTS:

Hazard functions

Some response variables in economics come in the form of a **duration,** which is the time elapsed until a certain event occurs. A few examples include the number of weeks unemployed, days until arrest after incarceration, and quarters until a firm files for bankruptcy.

Traditional duration analysis begins by specifying a population distribution for the duration, usually conditional on some explanatory variables (covariates) observed at the beginning of the duration. For example, for the population of people who became unemployed during a particular period, we might observe education levels, experience, marital status, all measured when the person becomes unemployed, plus wage on prior job, and a measure of unemployment benefits. Then, we specify a distribution function for the unemployment duration conditional on the covariates. Any reasonable distribution reflects the fact that unemployment duration is nonnegative. Once the conditional distribution is specified we apply maximum likelihood methods. We are usually interested in estimating the effects of the covariates on the expected duration.

Recent treatments in duration analysis tend to focus on the **hazard function**. The hazard function allow us to approximate the probability of exiting the initial state within a short interval, conditional on having survived up to the starting time of the interval.

Hazard functions

Formally, let $T \geq 0$ denote the duration, which has some distribution in the population; t denotes a particular value of T. T is the time at which a person (family, firm) leaves the initial state. For example, if the initial state is unemployment, T would be the time, measured in, for example, weeks until a person becomes employed. The **cumulative distribution function** (cdf) of T is denoted by $F(t)$. Then the probability that the duration is less than t:

$$F(t) = P(T \leq t), t \geq 0 \quad (1)$$

A complementary concept to the cdf is the probability that the duration equals or exceeds t ("surviving" past time t), called the **survival function**:

$$S(t) \equiv 1 - F(t) = P(T > t) \quad (2)$$

$P(t \leq T < t + h | T \geq t)$ is the probability of leaving the initial state in the interval $[t, t + h)$, given survival up until time t. The **hazard function** of T is defined as

$$\lambda(t) = \lim_{h \downarrow 0} \frac{P(t \leq T < t+h | T \geq t)}{h} \quad (3)$$

For each t, $\lambda(t)$ is the instantaneous probability of leaving state conditional on survival to time t.

Note that the hazard equals the change in the log-survival function: $\lambda(t) = -\frac{dln(S(t))}{dt}$.

Another related function is the **cumulative hazard function** or integrated hazard function

$$\Lambda(t) = \int_0^t \lambda(s)ds \tag{4}$$

$\Lambda(t)$ can be thought as the sum of the risk you face going from duration 0 to t.

An important feature about **survival data** is that it is usually **censored,** as some spells are incompletely observed. [1] That is, the lifetime is only known to lie in certain intervals. For example, instead of observing the length of the completed spell of unemployment, data is usually captured at a particular point in time such that only the length of an incomplete spell of unemployment is observed (i.e. some people will still be unemployed the following months after the data is collected). For **right-censoring** we observe spells from time 0 until a censoring time c. **Left-censoring** occurs when spells are known to end at some time in the interval (0,c) but the exact time is unknown. Survival analysis has focused on right-censoring.

EXAMPLE 1 (Unemployment Duration)

If T is the length of time employed, measured in weeks, then $\lambda(20)$ is (approximately) the probability of becoming employed between weeks 20 and 21. The sentence "becoming employed" reflects the fact that the person was unemployed up to, including week 20. That is, $\lambda(20)$ is roughly the probability of becoming employed between weeks 20 and 21, conditional on having been employed through week 20.

EXAMPLE 2 (Live Duration)

The us2006s.dta dataset contains US data for 2006 on two variables: age and survival function (the probability that a person's life equal or exceed a certain age t) by single years of age, for ages 0 to 110. From this, we can plot the survival function, find the hazard function (from the survival function) and plot the hazard function.

[1] Spell length or duration refers to the time spent in a given state.

```
. use us2006s.dta, clear

. help twoway

. use us2006s.dta, clear

. twoway line survival_function age, title("Survival Function, U.S. 2006")  name(a,replace) nodraw

. gen H = - log(survival_function)

. gen h = H[_n] - H[_n-1]
(1 missing value generated)

. list in 1/5
```

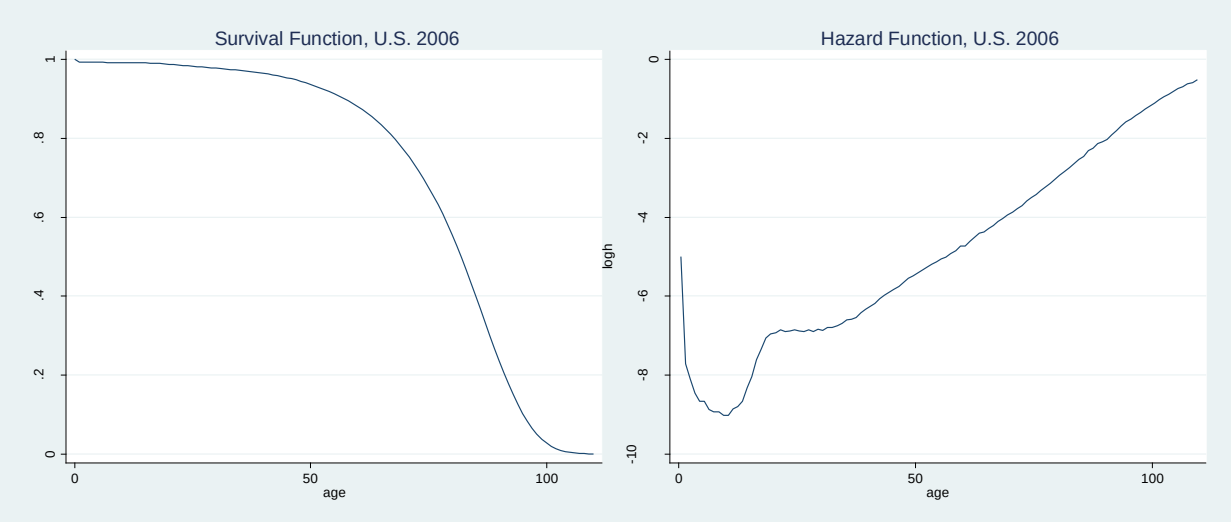
	age	surviv~n	H	h
1.	0	1	0	.
2.	1	.99327	.0067527	.0067527
3.	2	.99283	.0071958	.0004431
4.	3	.99253	.007498	.0003022
5.	4	.99232	.0077096	.0002116

```
. gen logh = log(h)
(1 missing value generated)

. gen agem = age - 0.5 if h < .
(1 missing value generated)

. twoway line logh agem,  xtitle("age") title("Hazard Function, U.S. 2006") name(b,replace) nodraw

. graph combine a b, xsize(7) ysize(3)
```



From the survival function graph we can see that for a person in the US in year 2006, the probability of being alive past 20 years from birth is around 99% while the probability of being alive past 80 years from birth is around 56%.[2] As expected, this probability declines over time being very low past 105 years from birth.

The hazard function indicates the probability of dying between age t and $t+1$. Our hazard function indicates that it is much more probable to die at birth than when you are 10 years old. Then, the probability increases until the mid-twenties where it remains more or less constant until the mid-thirties, and then increases again.

Fitting duration data

The shape of the hazard function is of primary interest in many empirical applications. There are several ways to fit this shape such as using non-parametric, parametric or semi-parametric models.

In order to keep it simple we will expose here two parametric specifications and an example. Note that different model specifications will lead to different hazard and survival functions.

In the simplest case, the hazard function is constant:

$$\lambda(t) = \lambda, \text{ all } t \geq 0 \tag{5}$$

This function means that the process driving T is *memory less*: the probability of exiting the next interval does not depend on how much time has been spent in the initial state.

Another popular parametric specification is that T has a Weibull distribution and its hazard function is given by

$$\lambda(t) = \gamma\alpha t^{\alpha-1} \tag{6}$$

If $\alpha > 1$ the hazard is monotonically increasing meaning that the hazard exhibits positive duration dependence. If $\alpha < 1$ the hazard is monotonically decreasing. When $\alpha = 1$, the Weibull distribution reduces to $\lambda = \gamma$.

EXAMPLE 3(Weibull Model for Reoffending Duration)

The variable of interest is the length of time, in months, until an inmate is arrested after being released from prison (*durat*). Although the duration is rounded to the nearest month, we treat *durat* as a continuous variable with a Weibull distribution. We are interested on how certain covariates affect the hazard function for recidivism (reoffending), and also whether there is positive or negative duration dependence, once we have conditioned on covariates. The binary indicator for participation in a prison work program (*workprg*) is of particular interest.

The data in RECID.RAW is a random sample of convicts released from prison during the period July 1, 1977, through June 30, 1978. The data are retrospective in that they were obtained by looking at records in April 1984, which served as the common censoring date. The variable indicating which observation is censored (*cens*) is an indicator coded 1 if the observation was censored. That is, the individual had not returned to prison. Because of the different starting times, the censoring times vary from 70 to 81 months.

[2] Note that S(0) =1 (since the event is sure not to have occurred by duration 0)

Before using any of Stata's survival commands, we have to *stset* the data. This will tell Stata that we have duration data and specify the time variable and the failure indicator. In this example the latter variable needs to first be calculated:

```
. use http://www.stata.com/data/jwooldridge/eacsap/recid, clear

. gen fail = 1 - cens

. stset durat, failure(fail)

     failure event:  fail != 0 & fail < .
obs. time interval:  (0, durat]
 exit on or before:  failure

------------------------------------------------------------------------------
     1445  total obs.
        0  exclusions
------------------------------------------------------------------------------
     1445  obs. remaining, representing
      552  failures in single record/single failure data
    80013  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =        81
```

We can now use a Weibull model using as predictors an indicator of participation in a work program (*workprg*), the number of previous convictions (*priors*), the time served rounded to months (*tsserved*), an indicator for felony sentences (*felon*), an indicator for alcohol problems (*alcohol*), an indicator for drug use history (*drugs*), an indicator for African Americans (*black*), an indicator if married when incarcerated (*married*), the number of years of schooling (*educ*) and the age in months (*age*).

Let's first fit a proportional hazard model:

```
. streg workprg priors tserved felon alcohol drugs black married educ age, distrib(weibull)

        failure _d:  fail
  analysis time _t:  durat

Fitting constant-only model:

Iteration 0:   log likelihood = -1739.8944
Iteration 1:   log likelihood = -1716.1367
Iteration 2:   log likelihood = -1715.7712
Iteration 3:   log likelihood = -1715.7711

Fitting full model:

Iteration 0:   log likelihood = -1715.7711
Iteration 1:   log likelihood = -1669.1785
Iteration 2:   log likelihood = -1634.3693
Iteration 3:   log likelihood = -1633.0405
Iteration 4:   log likelihood = -1633.0325
Iteration 5:   log likelihood = -1633.0325

Weibull regression -- log relative-hazard form

No. of subjects =         1445                 Number of obs   =      1445
No. of failures =          552
Time at risk    =        80013
                                               LR chi2(10)     =    165.48
Log likelihood  =   -1633.0325                 Prob > chi2     =    0.0000
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
workprg	1.095148	.0992728	1.00	0.316	.9168814	1.308074
priors	1.092848	.014683	6.61	0.000	1.064445	1.122008
tserved	1.013655	.0017037	8.07	0.000	1.010321	1.017
felon	.7412054	.0785485	-2.83	0.005	.6021898	.9123128
alcohol	1.564179	.165389	4.23	0.000	1.271406	1.92437
drugs	1.325064	.1296765	2.88	0.004	1.093791	1.605237
black	1.574149	.1390031	5.14	0.000	1.32398	1.871587
married	.8593436	.0938794	-1.39	0.165	.6937084	1.064527
educ	.9769709	.0189724	-1.20	0.230	.9404845	1.014873
age	.9962823	.000523	-7.09	0.000	.9952577	.997308
_cons	.0333035	.0100249	-11.30	0.000	.0184613	.0600781
/ln_p	-.2158398	.0389149	-5.55	0.000	-.2921115	-.1395681
p	.8058644	.0313601			.7466852	.8697338
1/p	1.240904	.0482896			1.149777	1.339252

.

Note that we do not specify the outcome, as this has been already done with *stset*; we just specify the explanatory variables. The Weibull parameter p (α in equation (6)) is 0.8, indicating that the risk of reoffending declines over time and the standard error leads to a strong rejection that the coefficient is statistically different from zero. This means that for a particular ex-convict the instantaneous rate of being arrested decreases with the length of time out of prison.

By default Stata exponentiates the coefficients to show the hazard ratios. Use the option *nohr* to obtain the coefficients.

```
. streg, nohr

Weibull regression -- log relative-hazard form

No. of subjects =         1445                 Number of obs   =      1445
No. of failures =          552
Time at risk    =        80013
                                                LR chi2(10)     =    165.48
Log likelihood  =   -1633.0325                  Prob > chi2     =    0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
workprg	.0908893	.0906478	1.00	0.316	-.0867772	.2685558
priors	.0887867	.0134355	6.61	0.000	.0624535	.1151198
tserved	.0135625	.0016808	8.07	0.000	.0102682	.0168567
felon	-.2994775	.105974	-2.83	0.005	-.5071826	-.0917723
alcohol	.4473611	.1057353	4.23	0.000	.2401236	.6545985
drugs	.2814605	.0978644	2.88	0.004	.0896499	.4732711
black	.4537147	.0883037	5.14	0.000	.2806426	.6267867
married	-.1515864	.1092454	-1.39	0.165	-.3657035	.0625307
educ	-.0232984	.0194196	-1.20	0.230	-.0613601	.0147633
age	-.0037246	.000525	-7.09	0.000	-.0047536	-.0026956
_cons	-3.402094	.3010177	-11.30	0.000	-3.992077	-2.81211
/ln_p	-.2158398	.0389149	-5.55	0.000	-.2921115	-.1395681
p	.8058644	.0313601			.7466852	.8697338
1/p	1.240904	.0482896			1.149777	1.339252

For small β_j estimations, we can multiply the coefficient by 100 to obtain the semi elasticity of the hazard with respect to x_j. For example, if *tserved* increases by one month, the hazard shifts up by about 1.4 percent, and the effect is statistically significant. Another year of education reduces the hazard by about 2.3 percent, but the effect is insignificant at even the 10% level.

The sign of the *workprg* coefficient is unexpected; at least if we expect the work program to have positive benefits after the inmates are released from prison. The results are not statistically different from zero. The reason could be that the program is ineffective.

For large β_j estimations, we should exponentiate and subtract unity to obtain the proportionate change. For example, at any point in time, the hazard is about 100[exp(0.447)-1]=56.3 percent grater for someone with an alcohol problem than for someone without.

EXERCISE 1 (Practical exercise)

Use the data in RECID.RAW ("use http://www.stata.com/data/jwooldridge/eacsap/recid ") for this problem.

a) Using the covariates in EXAMPLE 3, estimate the log-normal duration model. Verify that the log-likelihood value is -1,597.06. Note that the obtained estimated coefficients are semi elasticities-or elasticities if the covariates are in logarithmic form-of the covariates on the expected duration.
b) Plug in the mean values for *priors*, *tserved*, *educ* and *age* and the values *workprg*=0, *felon*=1, *alcohol*=1, *drugs*=1, and *married*=0, and plot the estimated hazard for the lognormal distribution. Describe what you find.
c) Using only the uncensored observations, perform an OLS regression of log(*durat*) on the covariates in EXERCISE 3. Compare the estimates on *alcohol* and *drugs* with those from part a). What would you conclude?
d) Now compute an OLS regression using all the data- that is, treat the censored observations as if they are uncensored. Compare the estimates on *alcohol* and *drugs* with those from part a) and c).

EXERCISE 2 (Practical exercise)

Use the data in RECID.RAW ("use http://www.stata.com/data/jwooldridge/eacsap/recid ") to answer these questions:

a) To the Weibull model, add the variables *super* (=1 if release from prison was supervised) and *rules* (number of rules violations while in prison). Does the coefficient estimates on these new variables have the expected signs? Are they statistically significant?
b) Add *super* and *rules* to the lognormal model, and answer the same questions as in part a)
c) Compare the estimated effects of the *rules* variable on the expected duration for the Weibull and lognormal models. Are they practically different? (NOTE: The lognormal models directly estimates the proportional effect of rules violations on the duration but to obtain comparable Weibull estimate we need to find $-\hat{\beta}_{rules}/\hat{\alpha}$)