Applied Econometrics

Lecture 10: Analysis of Duration Data

Måns Söderbom*

11 April 2008

*University of Gothenburg. mans.soderbom@economics.gu.se

1. Introduction

We are sometimes interested in modelling **duration**, which is the time that elapses between the 'beginning' and the 'end' of some specified state. The most common example is unemployment duration, where the 'beginning' is the day the individual becomes unemployed and the 'end' is when the individual exits from the state of unemployment - for example because she gets a new job. Other examples are the duration of wars, duration of marriages, time between first and second child, the lifetimes of firms, the length of stay in graduate school, time to adoption of new technologies, length of financial crises etc etc.

Econometric analysis of duration data is a little 'different' - compared to what we have done so far in this course - primarily in two ways:

1. We are often interested in characterising the **distribution** of the duration variable, because this can often shed light on economic questions (more on this below).

2. Data on durations are often **censored**, either to the right (common & easy to deal with) or to the left (not so common, not so easy to deal with) or both (even less common & less easy to deal with). Right censoring means that we don't know from the data when a certain duration ended; left censoring means that we don't know when it began. Of course we talked about censoring last week when studying the censored regression model (estimated by tobit).

For now, let's concentrate on the first of these points, namely why we should be concerned with the distribution of durations, and how we can do it. We thus abstract from censoring for the moment.

Useful references on duration data analysis

- Wooldridge (2002), Chapter 20.

- A more comprehensive - and, at least in parts, more intuitive - exposition is the review by Nicholas Kiefer (1988), Economic Duration Data and Hazard Functions, Journal of Economic Literature XXVI: 646-679.

2. Distributions of durations

- Let $T \geq 0$ denote duration, i.e. this is the variable that we are modelling. Define the cumulative distribution function of T as

$$F(t) = \Pr(T \leq t).$$

This simply measures the likelihood that a randomly drawn duration from the population of individuals in the relevant state is shorter than or equal to length t.

- A closely related function is the **survivor function**, defined as

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

This measures the probability that a randomly drawn duration from the population is longer than t. We can thus interpret the survivor function as the **probability of surviving** (in the state) past time t.

- Another important function in duration data analysis, which is related to both $F(t)$ and $S(t)$, is the **hazard function**, which measures the instantaneous rate at which individuals **exit** (i.e. no longer survive, 'die') from the state at time t, **given that they have not exited before** time t. The formal definition of the hazard function is

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

So if T is, say, the length of unemployment in weeks, then $\lambda(20)$ can be interpreted (approximately - cf. continuous vs. discrete time) as the probability of getting a job between weeks 20 and 21:

$$\lambda(20) = \frac{\Pr(20 \leq T < 21 | T \geq 20)}{1}.$$

Notice that

$$\begin{aligned}
\Pr(t \le T < t+h|T \ge t) &= \frac{\Pr(t \le T < t+h)}{\Pr(T \ge t)} \\
\Pr(t \le T < t+h|T \ge t) &= \frac{F(t+h) - F(t)}{1 - F(t)},
\end{aligned}$$

and so, for a small h, we get the instantaneous rate of exiting per unit of time:

$$\begin{aligned}
\lambda(t) &= \lim_{h \downarrow 0} \frac{F(t+h) - F(t)}{h} \frac{1}{1 - F(t)} \\
\lambda(t) &= f(t) \frac{1}{1 - F(t)} \\
\lambda(t) &= \frac{f(t)}{S(t)}, \qquad (2.1)
\end{aligned}$$

i.e. the hazard function is the density of exits at time t divided by the survivor function at time t. This shows that, as asserted above, the hazard function measures the rate at which individuals exit from the state at time t, given that they have survived (not exited) until time t. Notice that, approximately, this is the **probability** that an individual exits from the state at time t, given that she has not exited before time t

- When we analyse duration data we are typically interested in two things:

  1. How does the hazard rate vary with time?

  2. Do variables other than time itself impact on the hazard rate? If so, what are those variables and how do they impact on the hazard rate?

Let's discuss these issues in turn.

3. Duration Dependence

Duration dependence means that the hazard $\lambda(t)$ varies with t. there is...

- **positive duration dependence** if $\frac{d\lambda(t)}{dt} > 0$;

- **negative duration dependence** if $\frac{d\lambda(t)}{dt} < 0$; and

- **no duration dependence** if $\frac{d\lambda(t)}{dt} = 0$.

- Now consider the hazard of marriage/partnership dissolution - i.e. the divorce rate at time t, given that the marriage has lasted up until time t. The relevant population is the sub-set of the total population consisting of married couples. Do we expect this hazard to exhibit duration dependence?

- Suppose the nature of the typical partnership is such that it goes 'from strength to strength', so that the two indivduals in the relationship become ever more closely linked to each other over time (or slightly less romantically: you have fewer outside options the longer you remain in a partnership). This would be an example of negative duration dependence, i.e. the hazard would slope downwards as t increases. Get over the first few difficult (high-risk-of-dissolution) years, and chances are the marriage will last for a long time.

- Or it could be that the hazard increases with time, perhaps because newly-weds are happier than couples that have been in a relationship for a long time.

- Or it could be that the hazard of divorce is constant over time - in which case the process exhibits no duration dependence.

- See Figure 5.1 in appendix for an estimate of the hazard function in Denmark (taken from: Svarer, Michael (2002) "Determinants of Divorce in Denmark" Working Paper No. 2002-19, Aarhus University).

- Thinking about **unemployment** now, do you think there might be duration dependence? If so, is it likely to be positive or negative? What about **civil wars**? (Figure 2 in appendix.)

We saw above that the hazard function is intimately linked to the distribution function of T in the population. We saw that

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

Hence, for a given distribution function $F(t)$ it is straightforward to obtain the hazard function, and once we know the hazard function we can easily investigate the nature of the duration dependence.

3.1. Nonparametric analysis

The obvious starting point in duration analysis is to use a nonparametric estimator of the hazard function, which is based entirely on the distribution of durations in the sample. The Kaplan-Meier estimator is the most common estimator of this kind. The sample survivor function for a sample of N observations (with no censoring) is simply

$$S(t) = \frac{\# \text{ of sample points } \geq t}{N},$$

and the sample hazard function is

$$\lambda(t) = \frac{(\# \text{ of sample points } \leq t+1) - (\# \text{ of sample points } \leq t)}{\# \text{ of sample points } \geq t}$$

- Illustration in appendix, Table 1.

Nonparametric hazards are easy to compute but can be misleading if there is heterogeneity in the hazard rates across, say, groups of individuals. EXAMPLE: Appendix Tables 1-2 and Fig 3. It is therefore usually desirable to control for (observed) heterogeneity - indeed, sometimes determining the effect of an x-variable on the hazard is what we are primarily interested in. Doing so nonparametrically is very difficult (for essentially the same reasons that estimating any multivariate regression model nonparametrically is difficult), which is why typically parametric methods are used.

Before looking into models allowing for x-variables, let's consider two widely used parametric distributions for T in this literature, namely the exponential distribution and the Weibull distribution.

3.2. Parametric analysis

3.2.1. The exponential distribution

The simplest case is when T follows an **exponential distribution:**

$$F(t) = 1 - \exp(-\gamma t), \gamma > 0.$$

- The density function is obtained by taking the derivative of F with regard to t, thus

$$f(t) = \gamma \exp(-\gamma t).$$

- The survivor function is $1 - F(t)$, thus simply

$$S(t) = \exp(-\gamma t).$$

- The hazard function is

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)}, \\ \lambda(t) &= \gamma. \end{aligned}$$

Thus, if T follows an exponential distribution, then the hazard function exhibits **no duration dependence**. This is why the exponential distribution sometimes is termed **memoryless**: the exit rate is independent of how long you have survived.

Of course, the exponential distribution is quite a special case.

3.2.2. The Weibull distribution

A more general distribution is the Weibull distribution, for which

$$F(t) = 1 - \exp(-\gamma t^{\alpha}), \gamma > 0, \alpha > 0.$$

- We see straight away that a special case of the Weibull distribution is given by $\alpha = 1$. Further, it should be clear that

  $$f(t) = \gamma \alpha t^{\alpha - 1} \exp(-\gamma t^{\alpha}),$$

  and

  $$S(t) = \exp(-\gamma t^{\alpha}),$$

  thus

  $$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)}, \\ \lambda(t) &= \gamma \alpha t^{\alpha - 1}. \end{aligned}$$

  Notice how the duration dependence is determined by the parameter α: if $\alpha > 1$, then there is positive duration dependence, while if $\alpha < 1$ there is negative duration dependence. Estimating α is thus of interest if we want to investigate the nature of the duration dependence.

While more flexible than the Exponential distribution, the Weibull distribution does constrain the hazard function to be monotonic in time - i.e. processes in which the $d\lambda(t)/dt$ changes sign will not be approximated well by the Weibull distribution. Another alternative approach, which is quite flexible, is to use dummy variables for a suitable number of intervals. One such framework is the piecewise exponential model.

3.2.3. The piecewise exponential hazard

Recall that for the exponential hazard we have

$$\lambda(t) = \gamma,$$

i.e. the hazard rate does not vary with time. A generalisation of this model is to divide the time axis into W different segments, and only restrict the hazard to be constant **within** each segment. That is, the hazard rate may 'jump' at certain pre-specified points in time. More precisely, define W 'duration dummies' as follows:

$$\begin{aligned}
d_1(t_1) &= 1 \text{ if } T \leq t_1, \text{ zero otherwise,} \\
d_2(t_2) &= 1 \text{ if } t_1 < T \leq t_2, \text{ zero otherwise,} \\
d_3(t_3) &= 1 \text{ if } t_2 < T \leq t_3, \text{ zero otherwise,} \\
& (...) \\
d_W(t_W) &= 1 \text{ if } T > t_W, \text{ zero otherwise,}
\end{aligned}$$

where $t_1, t_2, ..., t_W$ are pre-specified points in time (by the researcher). Then define the hazard rate as

$$\lambda(t) = \sum_{w=1}^{W} \gamma_w d_w(t),$$

where $\gamma_1, \gamma_2, ..., \gamma_W$, are non-negative constants. This is the **piecewise exponential** hazard function. Notice that if $\gamma_1 = \gamma_2 = ... = \gamma_W$, this gives the exponential hazard discussed above. Figure 3 shows what the piecewise exponential hazard for civil wars 1960-2000 looks like.[1]

The piecewise exponential hazard model is flexible in that the hazard function can move up and down relatively freely. There are several other models that do this as well, but I do not go into further details

[1] From Collier, Paul, Anke Hoeffler and Måns Söderbom, 2004, "On the Duration of Civil War," 2004, Journal of Peace Research 41:3, pp. 253-273.

here (Stata has a big family of potential distributions that one can use). Instead I move on to discuss determinants of the hazard rate other than time itself.

4. Explanatory Variables and the Hazard Rate

In many economic applications it seems quite likely that the duration variable T depends on a set of explanatory variables. Clearly, if there exist a vector of explanatory variables x that impact on T, then these variables x also impact on the hazard rate. Suppose that the variables in the x vector are all time invariant - i.e. they do not change. To take into account the possibility that time invariant explanatory variables affect the hazard rate, we can write down a **proportional hazard model** as follows:

$$\lambda\left(t; x, \beta\right) = \kappa\left(x, \beta\right) \lambda_0\left(t\right),$$

where $\kappa\left(x, \beta\right)$ is a non-negative function and $\lambda_0\left(t\right)$ (also non-negative) is called the **baseline hazard**. By writing it like this, we are assuming that the that the baseline hazard is common to all individuals while the hazard rate at a given point in time t differs proportionally across individuals depending on the x variables. We typically assume that

$$\kappa\left(x, \beta\right) = \exp\left(\beta_1 + \beta_2 x_2 + \beta_2 x_2 + ... + \beta_K x_K\right) \equiv \exp\left(\boldsymbol{x\beta}\right).$$

Now things start to look very familiar. Clearly, if we can estimate the $\boldsymbol{\beta}$-parameters and the baseline hazard, we will be able to answer two important questions:

1. How does the hazard rate vary with time? The estimated baseline hazard will shed light on this.
2. What are the other determinants of the hazard rate and what is their impact? The estimates of the $\boldsymbol{\beta}$-parameters will shed light on this. Notice that β_1, for instance, is interpretable as the

semi-elasticity of the hazard rate with respect to x_1:

$$\frac{d \ln \lambda\left(t; x, \beta\right)}{dx_1} = \beta_1.$$

Thus, a positive β_1 would imply that...

- ...an **increase** in x_1 is associated with an **increase** in the hazard rate and thus an **decrease** in the expected duration; and
- ...an **decrease** in x_1 is associated with an **decrease** in the hazard rate and thus an **increase** in the expected duration.

To proceed towards an estimable equation, we need to specify a functional form for the baseline hazard $\lambda_0\left(t\right)$. Several options are open to us:

- If the baseline hazard is exponential, then

$$\lambda\left(t; x, \beta\right) = \exp\left(x\beta\right)\gamma.$$

- If the baseline hazard is Weibull, then

$$\lambda\left(t; x, \beta\right) = \exp\left(x\beta\right)\gamma\alpha t^{\alpha-1}.$$

- If the baseline hazard is piecewise exponential, then

$$\lambda\left(t; x, \beta\right) = \exp\left(x\beta\right)\sum_{w=1}^{W}\gamma_w d_w\left(t\right).$$

Naturally, one can use different functional forms for both $\kappa\left(x, \beta\right)$ and the baseline hazard. The principle is the same, however, so I do not dwell further on this issue here.

4.1. Analysis of Single-Spell Data with Time-Invariant Explanatory Variables

The simplest case in duration data analysis is when our data consists of single durations (or 'spells') - i.e. each individual/country/whatever is only observed once - and the variables in the x vector do not change over time.

4.1.1. Flow or stock sampling?

The first issue we should worry about is how the data have been sampled from the population. The most common way of sampling individuals is to sample from the flow of individuals entering the state at some point in time during the interval $[0, b]$. This is known as **flow sampling**.

- If we are studying the duration between the first and the second birth, for instance, flow sampling would involve drawing a random sample of women from the population of women that became first-time mothers during some suitable time interval, say the two-year window 1 January 2006 - 31 December 2007. The x vector would consist of data on the relevant explanatory time-invariant variables, e.g. education, location, health, income etc. We would follow this group of women over time and, for each individual, measure the time it takes until the second child is born. This information gives us the duration data.

- An alternative way of sampling the data is to draw from the population who are in the state of interest at the point in time b. In our example, this would involve drawing from the population who on 31 December 2007 had exactly one child. That is we are sampling from the **stock** of women with exactly one child at this point in time. Of course, in several cases the age of the child may be more than two years, implying that the child was born before 1 January 2006. Such observations of long durations would not have been included if we were sampling from the flow but they would if we are sampling from the stock. Further, spells that started **and ended** between 1 January 2003 and 31 December 2004 will not be included if we are sampling from the stock - simply because such women had left the relevant 'state' at 31 December.

Thus, the difference in the sampling schemes can be summed up as follows:

- Flow sampling: we draw from the population of individuals that **entered** the state between time 0 and b;
- Stock sampling: we draw from the population of individuals that **were** in the state at a given point in time, say b.

This is an important difference, because the two sampling schemes will generally result in different distributions of durations in our sample. More precisely, if we are sampling from the stock our sample will consist of...

- more long durations, and
- fewer short durations

than if we are sampling from the flow. And because the distribution of durations determines the hazard rates and how these vary with time, it should be clear that how the data have been sampled has a direct implication for how our results (e.g. regarding duration dependence) should be interpreted. Notice that his is a form of sample selection problem: short durations are less likely to appear in your sample than long durations, a phenomenon often referred to as **length-biased sampling**.

In most cases flow sampling provides the best basis for empirical analysis, simply because it does not give rise to the sample selection problem discussed above. Econometric methods designed to correct of length-biased sampling exist, but I will not discuss these here. In what follows I will focus only on flow data.

4.1.2. Right censoring

Flow data are typically subject to **right censoring,** i.e. for some observations in the data set we do not know when the duration ended. It is easy to see why this may arise in practice: after having drawn a random sample from the population that entered (note) the state during the interval $[0, b]$, we follow this sample of individuals over time in order to get data on how long they remained in the state.

At some point, however, we must stop and begin the analysis of the data. Observations of those individuals who, at that time, have not yet completed their spells will be recorded in our data sets as right censored durations. We don't know how long these durations will turn out to be, all we know is that they will be **at least as long** as the tracking period.

In the time to second birth example, for instance, our sample consisted of women who gave birth to their first child some time during 1 January 2006 and 31 December 2007. We follow this group of women over time to measure the time until the second child is born, but for practical reasons we do this for a limited period only, perhaps five years. Thus, those women in the sample that by 31 December 2007 have not yet given birth to a second child will have right censored durations.

Right censoring is a feature of our sample that is not (typically) shared by the population. For this reason, we need to devise an econometric estimator that takes this form of censoring into account.

4.1.3. Maximum Likelihood Estimation

Now consider estimation of the duration model. Initially, suppose our sample consists of completed spells only. As discussed in the previous paragraph, this may not be very realistic but it is a useful starting point. Definitions:

- a_i = the time at which individual i enters the state of interest (the 'beginning');
- t_i = the actual duration,
- x_i = vector of explanatory variables.

We will use the method of maximum likelihood to estimate the parameters of the model. The density function of the duration variable is denoted

$$f\left(t|x_i;\theta\right),$$

i.e. the density is written as conditional on the explanatory variables in x_i and the parameter vector θ. Because there is no censoring, the log likelihood function is simply the sample sum of individual likelihood contributions:

$$\ln L = \sum_{i=1}^{N} \ln f\left(t_i|x_i;\theta\right),$$

which, once we have specified $f\left(t_i|x_i;\theta\right)$, is to be maximised with respect to the parameters θ. Suppose we use a proportional hazard model of the form:

$$\lambda\left(t^*;x,\theta\right) = \kappa\left(x,\beta\right)\lambda_0\left(t\right),$$

where

$$\kappa\left(x,\beta\right) = \exp\left(x\beta\right),$$

and suppose the baseline hazard function is Weibull:

$$\lambda_0\left(t\right) = \gamma\alpha t^{\alpha-1}.$$

By definition (see eq. (2.1)):

$$\lambda\left(t|x_i;\theta\right) = \frac{f\left(t|x_i;\theta\right)}{S\left(t|x_i;\theta\right)},$$

thus

$$\begin{aligned} f\left(t|x_i;\theta\right) &= \lambda\left(t|x_i;\theta\right)S\left(t|x_i;\theta\right) \\ &= \exp\left(x_i\beta\right)\alpha t^{\alpha-1}[\exp[-\left(x_i\beta\right)t^{\alpha}]], \end{aligned}$$

and this is the expression that goes into the log likelihood function. The intuition is reasonably clear: the likelihood observing a duration of length t_i, conditional on the explanatory variables x_i can be written as the product of the likelihood of surviving until time t_i and exiting from the state at time t_i.

If the duration of individual i is right censored, all we have is the survival function:

$$f\left(t|x_i;\theta\right) = [\exp[-\left(x_i\beta\right)t^{\alpha}]].$$

Hence, for a sample in which some observations are right-censored, you'd modify the likelihood slightly and write:

$$
\begin{aligned}
f\left(t|x_i;\theta\right) &= \left[\lambda\left(t|x_i;\theta\right)\right]^{\delta} S\left(t|x_i;\theta\right) \\
f\left(t|x_i;\theta\right) &= \left[\exp\left(x_i\beta\right)\alpha t^{\alpha-1}\right]^{\delta}\left[\exp[-\left(x_i\beta\right)t^{\alpha}]\right],
\end{aligned}
$$

where $\delta = 1$ for completed spells and zero for censored (incomplete) spells.

Discuss: **Unobserved** heterogeneity.

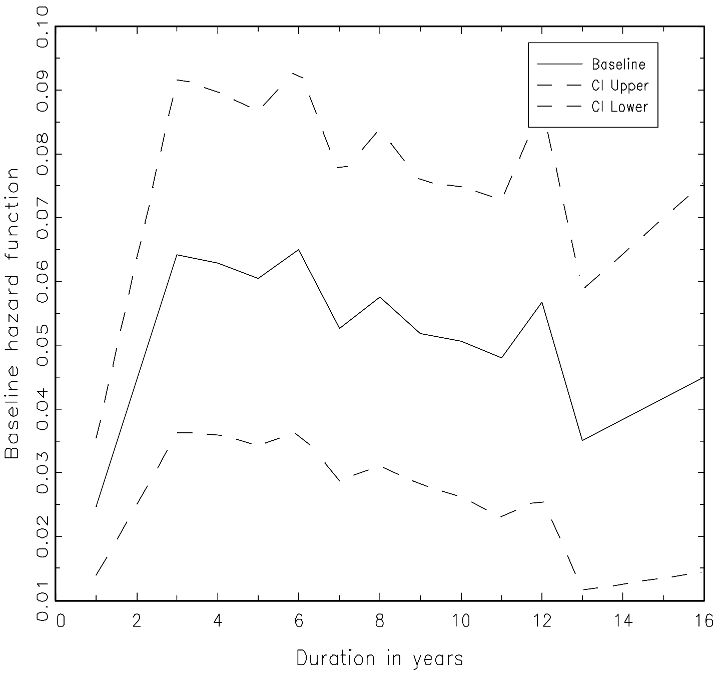
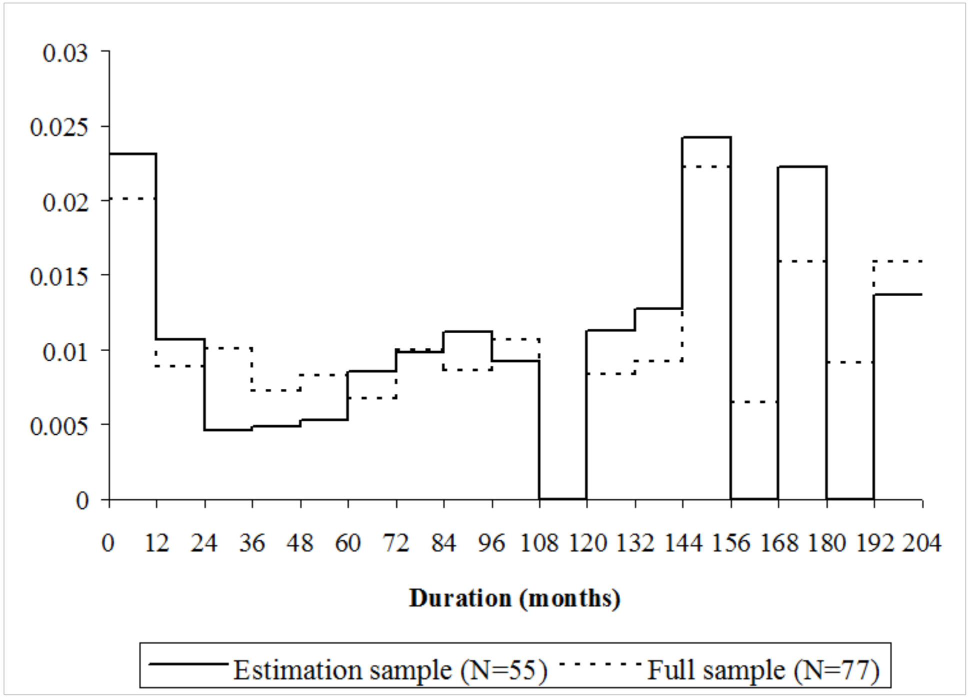


Figure 5.1: Baseline hazard

Figure 2: Nonparametric Hazard of Peace while at War



Source: Collier, Hoeffler and Söderbom (2004).

Table 1:
Illustration of Kaplan-Meier estimation of the hazard and survivor rates

Number of observations in ‘data set’: N = 100.

Duration	(a) Number of individuals 'at risk'	(b) Number of exits	(c) Number of survivors	Hazard rate: (b)/(a)	Survivor rate (c)/N
1	100	30	70	0.30	0.7
2	70	15	55	0.21	0.55
3	55	10	45	0.18	0.45
4	45	5	40	0.11	0.4
5	40	5	35	0.13	0.35
(…)					

Table 2: Illustration of the problem posed by heterogeneity in the hazard

Pooling of heterogeneous sub-samples (X = 0 or 1).

Sample with X=1

Duration	(a) Number of individuals 'at risk'	(b) Number of exits	(c) Number of survivors	Hazard rate: (b)/(a)	Survivor rate (c)/N
1	1024	256	768	0.25	0.75
2	768	192	576	0.25	0.56
3	576	144	432	0.25	0.42
4	432	108	324	0.25	0.32
5	324	81	243	0.25	0.24
6	243	243	0		0.00

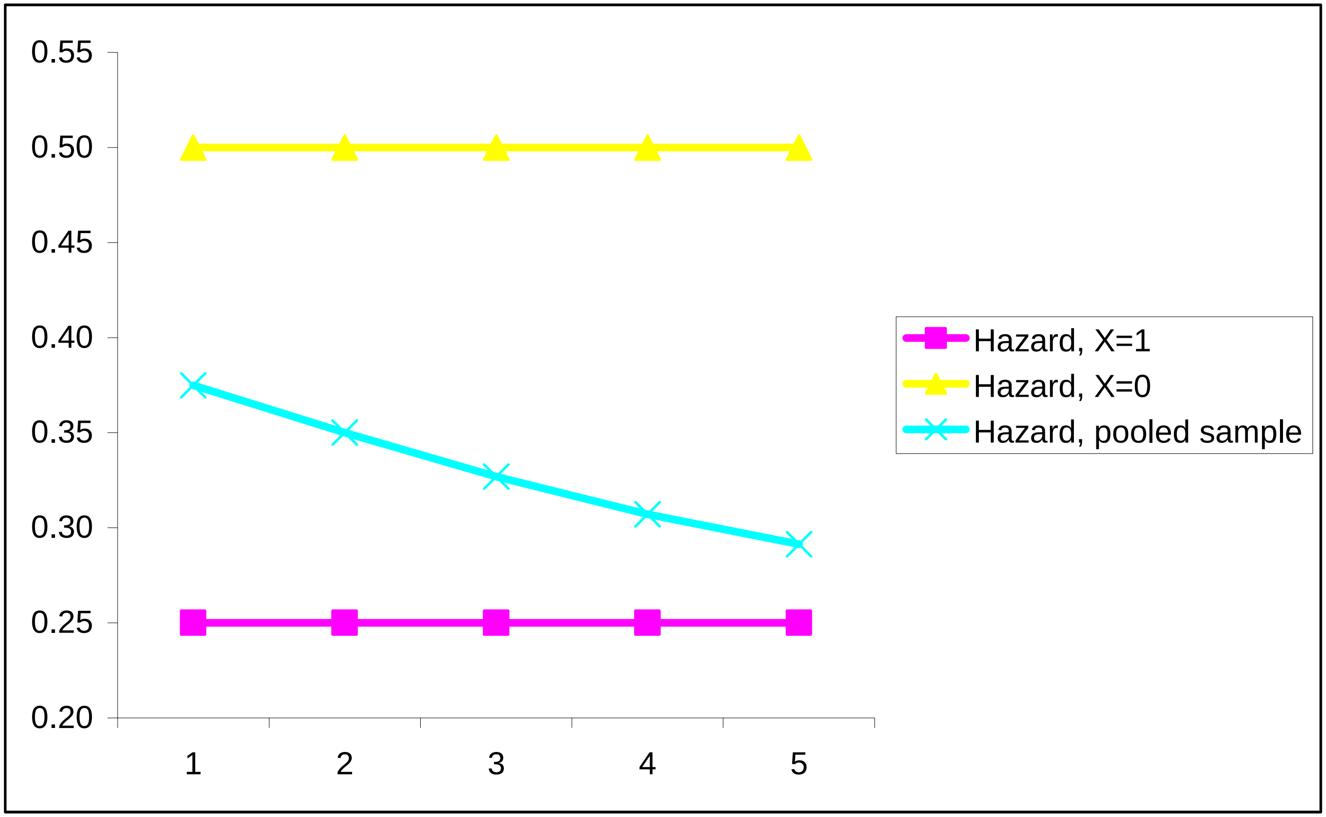
Sample with X=0

Duration	(a) Number of individuals 'at risk'	(b) Number of exits	(c) Number of survivors	Hazard rate: (b)/(a)	Survivor rate (c)/N
1	1024	512	512	0.50	0.50
2	512	256	256	0.50	0.25
3	256	128	128	0.50	0.13
4	128	64	64	0.50	0.06
5	64	32	32	0.50	0.03
6	32	32	0		0.00

Pooled sample

Duration	(a) Number of individuals 'at risk'	(b) Number of exits	(c) Number of survivors	Hazard rate: (b)/(a)	Survivor rate (c)/N
1	2048	768	1280	0.38	0.63
2	1280	448	832	0.35	0.41
3	832	272	560	0.33	0.27
4	560	172	388	0.31	0.19
5	388	113	275	0.29	0.13
(…)					0.00

Graphical comparison of estimated hazard functions



Estimation of Duration Data Models:
The Case of Civil Wars

The examples below are based on data on the duration of civil wars 1960-2000. The data set can be downloaded at

http://users.ox.ac.uk/~ball0144

See "On the Duration of Civil War," by Paul Collier, Anke Hoeffler and Mans Soderbom, Journal of Peace Research, 41:3, 2004, pp. 253-73, for details on this research.

```
> use "d:\warduration\jpr_revised03\estsample.dta", clear;

/* First, I declare the data to be duration data. I have monthly duration data
(mo), each war is indexed by the variable indsp, and cens is a dummy variable
equal to 1 if the war had ended by 31/12/2000, and 0 if it hadn’t (in which
case it would have been right censored). */

. stset mo, id(indsp) f(cens);

                id:  indsp
     failure event:  cens != 0 & cens < .
obs. time interval:  (mo[_n-1], mo]
 exit on or before:  failure

------------------------------------------------------------------------------
     4625  total obs.
        0  exclusions
------------------------------------------------------------------------------
     4625  obs. remaining, representing
       55  subjects
       48  failures in single failure-per-subject data
     4625  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =       364
```

```
/* Some simply summary statistics for the duration variable */

. stdes ;

         failure _d:  cens
   analysis time _t:  mo
                 id:  indsp

                                   |-------------- per subject --------------|
Category                   total        mean         min     median        max
------------------------------------------------------------------------------
no. of subjects               55
no. of records              4625    84.09091           1         73        364

(first) entry time                         0           0          0          0
(final) exit time                   84.09091           1         73        364

subjects with gap              0
time on gap if gap             0           .           .          .          .
time at risk                4625    84.09091           1         73        364

failures                      48    .8727273           0          1          1
------------------------------------------------------------------------------
```

Next I will consider the following variables as determinants of the hazard rates:

rgdpch	Per capita income
elf	Ethnic fractionalization
gini_m	Income inequality
ginmis	Missing inequality
logpop	ln Population
y70stv	Dummy for 1970s
y80stv	Dummy for 1980s
y90stv	Dummy for 1990s

More general specifications are considered in the Collier et al. paper.

Model 1: Exponential

```
streg gini_m ginmis rgdpch elf elf2 logpop y70stv y80stv y90stv,
dist(exponential) nohr;
```

```
Exponential regression -- log relative-hazard form

No. of subjects =           55                     Number of obs   =      4625
No. of failures =           48
Time at risk    =         4625
                                                   LR chi2(9)      =     34.93
Log likelihood  =   -84.172747                     Prob > chi2     =    0.0001

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gini_m |  -.1040502   .0241029    -4.32   0.000    -.1512911   -.0568093
      ginmis |   -4.85638   1.058398    -4.59   0.000    -6.930801   -2.781959
      rgdpch |   .3342252   .1169375     2.86   0.004      .105032    .5634184
         elf |  -.0571514   .0253769    -2.25   0.024    -.1068893   -.0074135
        elf2 |   .0548429   .0267493     2.05   0.040     .0024152    .1072706
      logpop |  -.3174559   .1252571    -2.53   0.011    -.5629553   -.0719565
      y70stv |   .1912412   .4537672     0.42   0.673    -.6981263    1.080609
      y80stv |  -1.161242   .4812326    -2.41   0.016    -2.104441    -.218044
      y90stv |  -.6954615   .4734954    -1.47   0.142    -1.623495    .2325724
       _cons |   6.299516   2.624996     2.40   0.016     1.154617    11.44441
------------------------------------------------------------------------------
```

Note: The option **nohr** specifies that coefficients rather than exponentiated coefficients are reported.

Model 2: Weibull

```
streg gini_m ginmis rgdpch elf elf2 logpop y70stv y80stv y90stv, dist(weibull)
nohr;
```

```
Weibull regression -- log relative-hazard form

No. of subjects =           55                     Number of obs   =      4625
No. of failures =           48
Time at risk    =         4625
                                                   LR chi2(9)      =     30.17
Log likelihood  =   -84.082074                     Prob > chi2     =    0.0004

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gini_m |  -.1090997   .0270188    -4.04   0.000    -.1620555   -.0561439
      ginmis |  -5.100231   1.212371    -4.21   0.000    -7.476434   -2.724027
      rgdpch |   .3568975    .129353     2.76   0.006     .1033702    .6104247
         elf |  -.0598764   .0262609    -2.28   0.023    -.1113469   -.0084059
        elf2 |   .0570066   .0273157     2.09   0.037     .0034687    .1105445
      logpop |  -.3271305   .1272727    -2.57   0.010    -.5765804   -.0776806
      y70stv |   .1627748   .4584631     0.36   0.723    -.7357963    1.061346
      y80stv |  -1.236041   .5131423    -2.41   0.016    -2.241782   -.2303007
      y90stv |  -.8034153   .5397958    -1.49   0.137    -1.861395     .254565
       _cons |   6.513506   2.684383     2.43   0.015     1.252212     11.7748
-------------+----------------------------------------------------------------
       /ln_p |   .0549738   .1277022     0.43   0.667    -.1953178    .3052655
-------------+----------------------------------------------------------------
           p |   1.056513    .134919                      .8225732    1.356985
         1/p |   .9465099   .1208714                      .7369277    1.215697
------------------------------------------------------------------------------
```

No evidence of duration dependence.

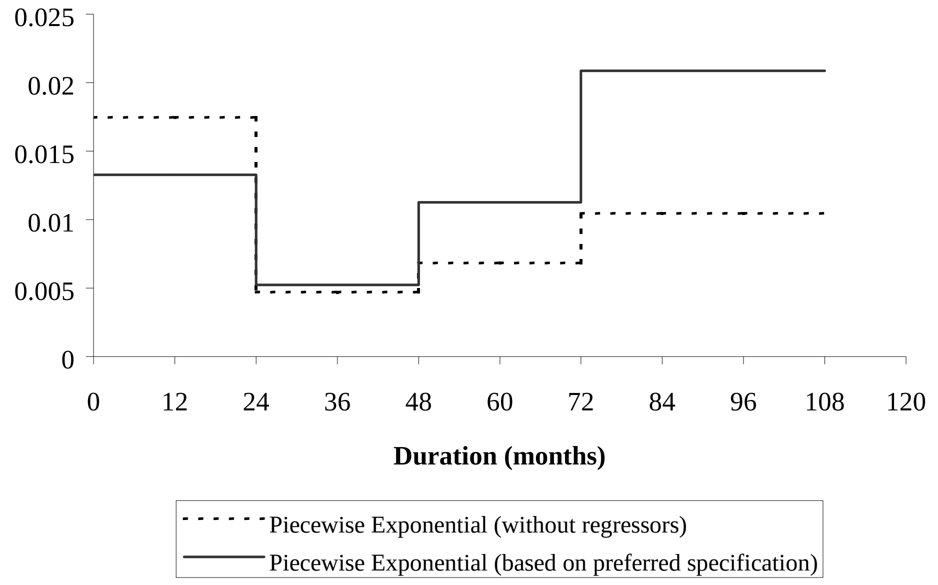
Model 3: Piecewise Exponential

```
d1          Duration dummy 1st and 2nd years of war
d2          Duration dummy 3rd and 4th years of war
d3          Duration dummy 5th and 6th years of war
d4          Duration dummy 7th year of war and beyond
```

```
streg gini_m ginmis rgdpch elf elf2 logpop y70stv y80stv y90stv d2-d4,
dist(exponential) nohr;

         failure _d:  cens
   analysis time _t:  mo
                 id:  indsp

Exponential regression -- log relative-hazard form

No. of subjects =           55                     Number of obs   =      4625
No. of failures =           48
Time at risk    =         4625
                                                   LR chi2(12)     =     42.41
Log likelihood  =   -80.429995                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gini_m |  -.1244463   .0284179    -4.38   0.000    -.1801444   -.0687482
      ginmis |  -5.867928   1.277403    -4.59   0.000    -8.371591   -3.364265
      rgdpch |   .3651031   .1322248     2.76   0.006     .1059472     .624259
         elf |  -.0628267   .0258742    -2.43   0.015    -.1135392   -.0121143
        elf2 |   .0581252   .0270411     2.15   0.032     .0051256    .1111247
      logpop |  -.3163905   .1230657    -2.57   0.010    -.5575948   -.0751863
      y70stv |   .0077905   .4625409     0.02   0.987    -.8987729    .9143539
      y80stv |  -1.420202   .5203341    -2.73   0.006    -2.440038   -.4003656
      y90stv |  -1.162059   .5416506    -2.15   0.032    -2.223675   -.1004433
          d2 |  -.8067415   .5742936    -1.40   0.160    -1.932336    .3188533
          d3 |  -.0010657   .5606172    -0.00   0.998    -1.099855    1.097724
          d4 |   .6098389   .4464024     1.37   0.172    -.2650937    1.484771
       _cons |   7.433105   2.707863     2.75   0.006     2.125791    12.74042
------------------------------------------------------------------------------

. exit;

end of do-file

. test d2=d4

 ( 1)  [_t]d2 - [_t]d4 = 0

           chi2(  1) =    5.86
         Prob > chi2 =    0.0155
```

A graph of the estimated hazard function is provided on the next page.

Figure 2: Piecewise Exponential Estimates of the Hazard Function



Note: The hazard function based on the preferred specification (i.e. the reference model) is calculated using the formula $[\exp[\bar{x}_{\tau}\hat{\beta}]\cdot\exp[\hat{\alpha}+\sum_{w}\hat{\lambda}_{w}d_{w}(t)]]$, where $\hat{\beta},\hat{\alpha},\hat{\lambda}_{2},...,\hat{\lambda}_{W}$ are the parameter estimates and $\bar{x}_{\tau}$ denotes a vector of sample means of the explanatory variables. The hazard function without regressors was calculated as explained in the notes to Figure 1. The underlying regression is not reported but is available on request from the authors.

Source: Collier et al. (2004).