Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Models With Time-To-Event

Now that we have introduced a majority of the key definitions and implications appropriate to time-to-event studies, we can proceed to discuss the many different types of models used in such studies.

The models first divide into those for discrete outcomes and those for continuous outcomes. The most familiar methods for continuous outcomes have been available in software since the 1980's. Methods for discrete outcomes have been implemented in most software much more recently. Most references begin with the continuous case first. We will take this option.

Modelling Continuous Time-To-Event Outcomes

In health research, the modelling of the log of the hazard dominates the literature. [Hazard models] In certain specific situations, one sees modelling of the log of time itself. [Accelerated Failure Time models]

Models for the hazard will have regression coefficients as usual but then separate into two classes:
1) models that assume that the log of a 'baseline' hazard function assumes forms determined by parameters.
2) models that do not require the baseline hazard function assumes any parametric form.

The second type of model that does not need a determination the baseline hazard now dominates in the health literature. The development of these methods began with the remarkable work of DR Cox in the 1970's. Such models are now almost always called Cox models.

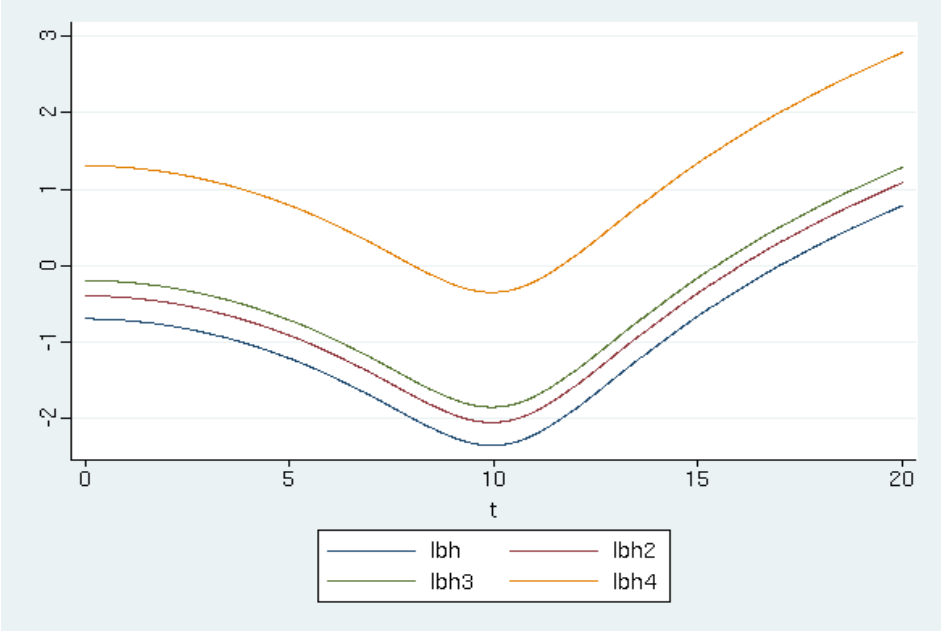
The next major separation is based on the assumption of proportional hazards. Again, proportional hazards models are widely seen along with techniques to assess the proportional hazards assumption. There are a very wide range of models that relax the proportional hazards assumption in various ways. The two [most often seen] use either stratification or time varying explanatory variables [or both stratification and time varying]

Lets start with proportional hazards models.
We could consider a model like:

$$\log h(t) = \log h_0(t) + \sum_{j=1}^{k} \beta_j x_j$$

We can think of a 'baseline' hazard $h_0(t)$ and, then, additive contributions on the logarithmic scale are like relative contributions on the original scale. In other words, the model above is then a fairly general example of a proportional hazards model. Depending on the software, an estimate of an apparent β_0 may be listed. This number is a part of the log of the baseline hazard function. The examples below will clarify this.

Now, think of a baseline hazard function and 3 other groups with proportional hazards. On the log hazard scale, we might have curves like:



The log of the baseline hazard might be the blue curve (the lowest curve) and, then, with each other groups, there is a fixed additive difference that does not depend on time. This is analogous to the notion that 'analysis time' does not modify the group comparisons on the log hazard scale. So then $h_0(t)$ is the blue curve and $\sum_{j=1}^{3} \beta_i \delta_j$ determines the spacings between the curves. For this picture, k=3 and we could think of 3 indicator variables δ_1 δ_2 δ_3 for groups 1, 2 and 3 respectively. Now, for say group 0, there is a baseline curve (the log of the baseline hazard). For this model, each regression coefficient is an assumed common difference between the curves.

β_1 : between red and blue, β_2 : between green and blue and β_3 : between orange and blue

Each of these coefficients would be positive indicating increased log hazard for each group compared with the baseline group.

Let us now suppose that in:

group 0, participants had exposure to neither E_1 nor E_2

group 1, participants had exposure to E_1 only

group 2, participants had exposure to E_2 only

group 3, participants had exposure to both E_1 and E_2

The model: $\log h(t) = \log h_0(t) + \beta_1 E_1 + \beta_2 E_2 + \beta_3 E_1 E_2$ would have a 'large' β_3 since:

When E_2 is present, the comparison between those with and without E_1 is $\beta_1 + \beta_3$

When E_2 is absent, the comparison between those with and without E_1 is β_1

So we see that β_3 reflects the interaction of the two exposures:

[orange minus green] minus [red minus blue].

So, while there is an assumption of additivity among the log hazard curves, one can still explore notions of modification, interaction and so on using the regression coefficients as always.

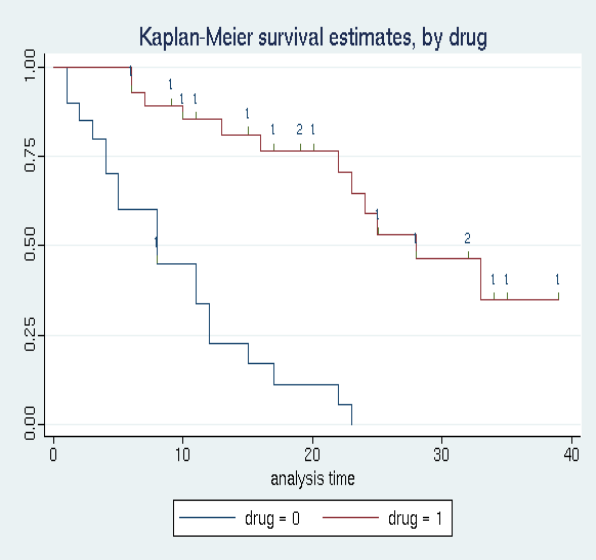
Indeed, we can reexpress the model as:

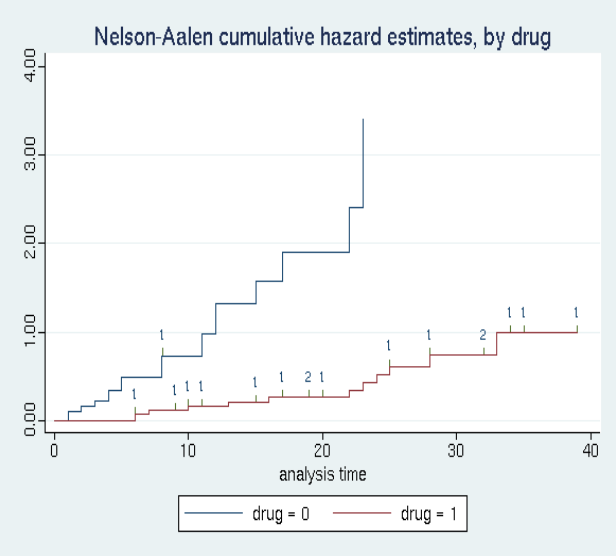
$$\log h(t) - \log h_0(t) = \log \frac{h(t)}{h_0(t)} = \sum_{j=1}^{k} \beta_j x_j$$

So the regression coefficients can be interpreted as in logistic regression, for example, except that we speak of the log of hazard ratios rather than the log of odds.

It is time for a small but illustrative example from a cancer drug trial.

```
. use cancer.dta
. stset studytime died
. sts graph, by(drug) cen(number)
. sts graph, na by(drug) cen(number)
```





The little numbers above the curves show us the number censored at those times. Only one censored value among the placebo group and 16 censored in the active group. Also, we can see that there are many 'ties'.

```
. table studytime drug

----------------------
Months to |
death or  | Drug type
end of    |(0=placebo)
exp.      |    0     1
----------+-----------
        1 |    2
        2 |    1
        3 |    1
        4 |    2
        5 |    2
        6 |          3
        7 |          1
        8 |    4
        9 |          1
       10 |          2
       11 |    2     1
       12 |    2
       13 |          1
       15 |    1     1
       16 |          1
       17 |    1     1
```

```
       19 |          2
       20 |          1
       22 |    1     1
       23 |    1     1
       24 |          1
       25 |          2
       28 |          2
       32 |          2
       33 |          1
       34 |          1
       35 |          1
       39 |          1
----------------------

. gen da=drug*age
```

Let us consider a Weibull proportional hazards model first:

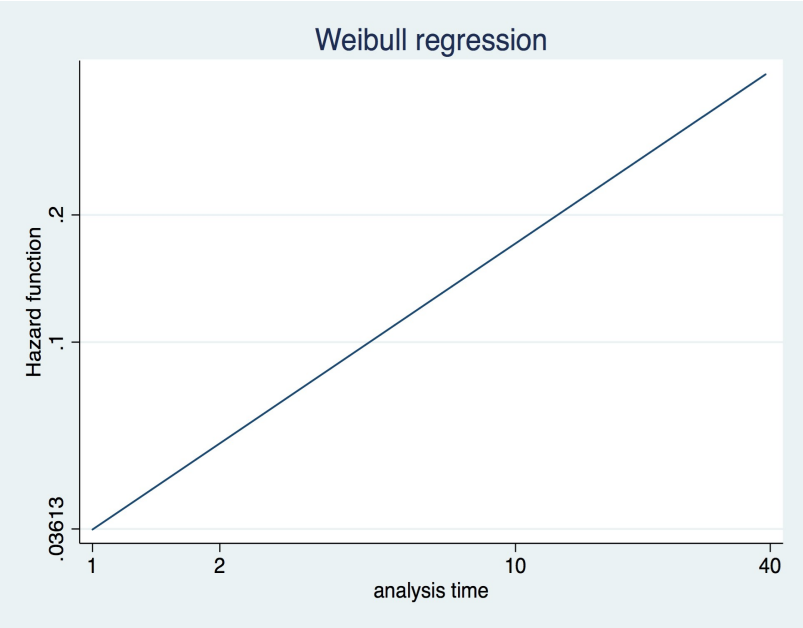
$$\log h(t)=\log h_{W0}(t)+\sum_{j=1}^{k}\beta_j x_j=\log(\lambda)+\log(p)+(p-1)\log(t)+\sum_{j=1}^{k}\beta_j x_j$$

```
. streg drug age da,d(w) nohr

Weibull regression -- log relative-hazard form

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(3)      =     35.47
Log likelihood  =   -42.887722                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |   -3.42893   4.204572    -0.82   0.415    -11.66974     4.81188
         age |   .1108742   .0489165     2.27   0.023     .0149996    .2067487
          da |   .0216165   .0733299     0.29   0.768    -.1221075    .1653405
       _cons |  -10.04656   2.949979    -3.41   0.001    -15.82841   -4.264704
-------------+----------------------------------------------------------------
       /ln_p |   .5170442   .1395839     3.70   0.000     .2434648    .7906235
-------------+----------------------------------------------------------------
           p |   1.677063    .234091                      1.275661    2.204771
         1/p |   .5962805   .0832311                      .4535619    .7839071
------------------------------------------------------------------------------
. gen ac=age-56
. gen dac=drug*ac
. streg drug ac dac,d(w) nohr

Weibull regression -- log relative-hazard form

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |  -2.218406   .4196609    -5.29   0.000    -3.040926   -1.395886
          ac |   .1108742   .0489165     2.27   0.023     .0149996    .2067487
         dac |   .0216165   .0733299     0.29   0.768    -.1221075    .1653405
       _cons |  -3.837604   .6400264    -6.00   0.000    -5.092032   -2.583175
-------------+----------------------------------------------------------------
       /ln_p |   .5170442   .1395839     3.70   0.000     .2434648    .7906235
-------------+----------------------------------------------------------------
           p |   1.677063    .234091                      1.275661    2.204771
         1/p |   .5962805   .0832311                      .4535619    .7839071
------------------------------------------------------------------------------
```

The centring of age provides a 'meaningful' interpretation for the baseline log hazard.

The estimates of $\log(\lambda)$ and log(p) provide the 'intercept' when log(t)=0 or t=1. Notice that : $\beta_0 = \log(\lambda)$ here.

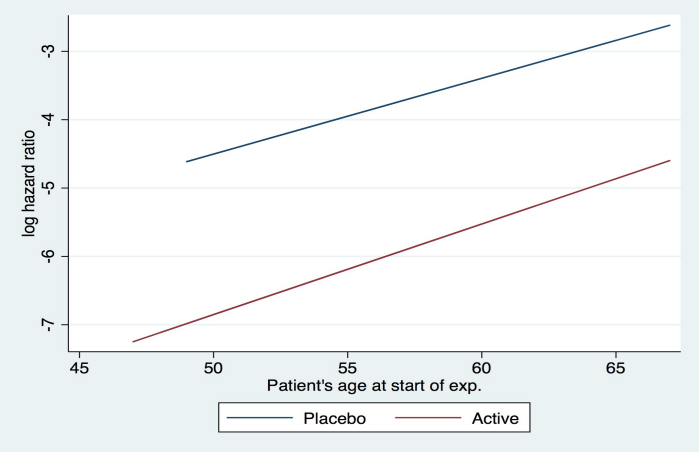
```
 disp exp(-3.8376 +0.5170)
 0.03613115
 stcurve, hazard at1(drug=0 ac=0) yscale(log) xscale(log) xlabel(1 2 10 40) ylabel(0.03613
0.1 0.2)
```



This model suggests non-constant hazard since the estimate of p appears to be greater than one. We can then see that the log hazard is linear in log time.

Then we can display the estimates of the regression coefficients [as usual]

```
. predict lhr,xb
. twoway (line lhr age if drug==0,legend(label(1 "Placebo")) ytitle("log hazard ratio"))(l
ine lhr age if drug==1,legend(label(2 "Active")))
```



Let us now consider a Cox proportional hazards model.

$\log h(t) = \log h_0(t) + \sum_{j=1}^{k} \beta_j x_j$ The method provides for estimates of the regression coefficients using a 'Partial Likelihood' approach. [Cox 1971]. An 'estimate' of the function $\log h_0(t)$ is not needed to

give us regression coefficient estimates using the method of Cox. Nevertheless, other methods can be used to provide an estimate of the log hazard function using smoothing techniques.

```
. stcox drug age da, nohr

Cox regression -- Breslow method for ties

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(3)      =     33.33
Log likelihood  =   -83.245435                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |  -3.934271   4.297017    -0.92   0.360    -12.35627    4.487727
         age |   .1013904   .0486087     2.09   0.037      .006119    .1966618
          da |   .0293675   .0745665     0.39   0.694    -.1167802    .1755152
------------------------------------------------------------------------------

. stcox drug ac dac, nohr

Cox regression -- Breslow method for ties

No. of subjects =           48                     Number of obs   =        48
No. of failures =           31
Time at risk    =          744
                                                   LR chi2(3)      =     33.33
Log likelihood  =   -83.245435                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |  -2.289689   .4695969    -4.88   0.000    -3.210082   -1.369296
          ac |   .1013904   .0486087     2.09   0.037      .006119    .1966618
         dac |   .0293675   .0745665     0.39   0.694    -.1167802    .1755152
------------------------------------------------------------------------------

stcurve, hazard at1(drug=0) at2(drug=1) yscale(log) xscale(log)
gen plac=1-drug
```

Reverse coding the drug indicator indicates that the 'exposed' group is those without the active form of treatment.

```
stcox plac ac
```

It is important to note when one does not use the nohr option, one is receiving estimates of the exponent of the corresponding regression coefficients. Stata labels these estimates 'Haz. Ratio'. Depending on the model one fits, only some of these estimates will be estimates of hazard ratios. The number beside 'ac' is the exponent of the estimate provided with the nohr option.The unexponentiated number is an estimated rate of change of a log hazard ratio per year of age assumed common to both drug and placebo groups. The exponent has a rather specialized interpretation and is not as easy to explain.

```
stcurve, surv at1(plac=0) at2(plac=1)
stcurve, cumhaz at1(plac=0) at2(plac=1)
predict lhr,xb
lab var age "baseline age"
twoway (line lhr age if drug==0,legend(label(1 "Placebo")) ytitle("log hazard ratio"))(l
ine lhr age if drug==1,legend(label(2 "Active")))
```

The Cox model provides much the same message as the Weibull model and the Cox model does not require an assumption as to the form of the baseline hazard function. There is considerable empirical and theoretical support for the Cox model. The methods of estimation with the Cox model are typically not impaired by the absence of the hazard function form assumption. The proportional hazard assumption, though, is so critical here and is the key issue in many health research studies.

Methods are now available which enable the use of restricted cubic splines to provide a baseline hazard. For example, for the Weibull hazard, we have that the log cumulative hazard is linear in log time. Replacing the line with restricted cubic splines enables a wide range of hazard forms including non-monotone forms. Implementations are available with Stata [stpm2] and with R [flexsurv]

```
. stpm2 drug ac dac,df(1) scale(hazard)

Log likelihood = -42.887722                     Number of obs     =         48

------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
        drug |  -2.218406   .4196609    -5.29   0.000    -3.040926   -1.395885
          ac |   .1108742   .0489165     2.27   0.023     .0149996    .2067487
         dac |   .0216165   .0733299     0.29   0.768    -.1221075    .1653405
       _rcs1 |   1.456171    .203258     7.16   0.000     1.057793    1.854549
       _cons |   .2711763   .2306783     1.18   0.240    -.1809448    .7232974
------------------------------------------------------------------------------

. stpm2 drug ac dac,df(4) scale(hazard)

Log likelihood = -42.299792                     Number of obs     =         48

------------------------------------------------------------------------------
             |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
        drug |  -2.349561   .4587812    -5.12   0.000    -3.248755   -1.450366
          ac |   .1059472   .0490748     2.16   0.031     .0097623     .202132
         dac |   .0309635   .0748127     0.41   0.679    -.1156667    .1775938
       _rcs1 |   1.435409   .2085075     6.88   0.000     1.026741    1.844076
       _rcs2 |  -.0610769   .1580855    -0.39   0.699    -.3709188    .2487651
       _rcs3 |  -.0553187   .1019401    -0.54   0.587    -.2551177    .1444803
       _rcs4 |  -.0691038   .0798791    -0.87   0.387    -.2256639    .0874563
       _cons |   .3321558   .2397368     1.39   0.166    -.1377197    .8020312
------------------------------------------------------------------------------
```

When df=1, one gets the Weibull analysis. The single restricted cubic spline is just a location and scale shift of log(t). rcs1(t) =a + b*log(t) and so _rcs1 and _cons can be obtained as :
$\beta_0+\beta_1\log(t)$ or $\alpha_0+\alpha_1(a+b\log(t))$ and so $\alpha_1=\beta_1/b$ $\alpha_0=\beta_0-\beta_1 a/b$
Many authors suggest that using cubic splines in this way may be preferred to the Cox approach. Perhaps there are arguments based on prediction and extrapolation matters, in particular.

Models for Subhazard
Now let us consider models in the competing events world. In particular, we will consider the Fine & Gray Proportional Subhazards models [Fine & Gray 1999]. Now we will have:

$$\log \bar{h}_1(t)=\log \bar{h}_{10}(t)+\sum_{j=1}^{k}\beta_j x_j$$

So the log of the subhazard function is expressed in terms of the log of the baseline subhazard plus the usual linear combination of regression coefficients.
Interpretation of the subhazard is quite elaborate but it is instructive to consider the implied graphs of the Failure functions and graphs showing log subhazard ratios versus our explanatory variables, as usual.

```
. use byar.dta
. gen tstage=stage-3
. summ age,d

                        Age: years
-------------------------------------------------------------
      Percentiles      Smallest
 1%           51             48
 5%           56             49
10%           60             49       Obs                 505
25%           70             50       Sum of Wgt.         505

50%           73                      Mean           71.44158
                        Largest       Std. Dev.      7.081516
75%           76             87
90%           78             87       Variance       50.14787
95%           80             88       Skewness      -1.047976
99%           84             89       Kurtosis       4.080304

. gen ac=age-73
. gen tac=treatment*ac
. stcrreg treatment ac tac tstage,compete(status == 2 3) noshr

Competing-risks regression                    No. of obs       =       505
                                              No. of subjects  =       505
Failure event   : status == 1                 No. failed       =       155
Competing events: status == 2 3               No. competing    =       201
                                              No. censored     =       149

                                              Wald chi2(4)     =     65.36
Log pseudolikelihood = -897.10587             Prob > chi2      =    0.0000

------------------------------------------------------------------------------
             |               Robust
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   treatment |   -.363127   .1736585    -2.09   0.037    -.7034913   -.0227626
          ac |  -.0337263   .0138147    -2.44   0.015    -.0608027   -.0066499
         tac |   .0327927    .020657     1.59   0.112    -.0076943    .0732797
      tstage |   1.122175   .1664584     6.74   0.000     .7959226    1.448428
------------------------------------------------------------------------------

. stcurve,cif at1(tstage=0 treatment=0) at2(tstage=1 treatment=0) at3(tstage=0 treatment=1)
at4(tstage=1 treatment=1)
. predict lshr,xb
twoway (line lshr age if treat==0 & tstage==0)(line lshr age if treat==1 & tstage==0)(line
lshr age if treat==0 & tstage==1)(line lshr age if treat==1 & tstage==1), legend(off)
```

The estimates of the regression coefficients are available in the same way as Cox proportional hazards models except that we have the log of the subhazard rather than the log of the hazard in the descriptions. The graphs are then estimates of the Failure function using the proportional subhazards assumption.

Discrete Time Models

Now we return to the study of discrete time-to-event. We will see that the software needed to construct, fit and assess models in discrete time is not new to us. However, we will also see that there are a

number of new steps needed to prepare the data for the analysis. Most times, we will see the need for two datasets. These two datasets are often called:
1) the 'person-level' dataset
2) the 'person-period' dataset
Typically, an investigator has the 'person-level' dataset first. Then one must construct the 'person-period' dataset.

A collection of Stata commands [Dinno] can be very helpful with this construction. You can download this family of commands from within Stata by typing findit dthaz

We will consider a study [by Capaldi, Crosby, and Stoolmiller's (1996)] of the grade when a sample of at-risk adolescents males had heterosexual intercourse for the first time. Among 180 boys tracked from seventh grade, 54 (30.0%) were still virgins (were censored) when data collection ended in 12th grade. The outcome is time to first sex. We will start our example by considering two explanatory variables pt [parental transition before seventh grade] and pas [an index of the parent's antisocial behavior]. pt is dichotomous and we will assume that the relationship between the outcome and pas is linear.

The person-level data is in capaldi_pl.dta and the person-period data is in capaldi_pp.dta

We will now be considering models of the form:
$g(h(t))=\sum_{j=1}^{l}\alpha_j d_j+\sum_{i=1}^{k}\beta_i x_i$ where g is a link function. [We will see two types of links: logit and complementary log log]
The first piece of the 'right hand side': $\sum_{j=1}^{l}\alpha_j d_j$ will take on the role of baseline hazard and is the part that is a function of time. The d_j are the indicators for the time intervals. The second piece $\sum_{i=1}^{k}\beta_i x_i$ will give us the usual regression coefficients and the explanatory variables.

We have noted that, in the discrete time setting, h(t) is a probability. We can consider [first] the log of the odds of this probability and use logistic regression.

```
. use capaldi_pp.dta
. logit event d7 d8 d9 d10 d11 d12 pt pas, nocons

Logistic regression                                     Number of obs   =        822
                                                        Wald chi2(8)    =     269.81
Log likelihood = -314.57348                             Prob > chi2     =     0.0000

------------------------------------------------------------------------------
       event |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d7 |  -2.893237   .3206302    -9.02   0.000     -3.52166   -2.264813
          d8 |  -3.584759   .4231479    -8.47   0.000    -4.414114   -2.755404
          d9 |  -2.150233    .277458    -7.75   0.000    -2.694041   -1.606426
         d10 |   -1.69318   .2646518    -6.40   0.000    -2.211888   -1.174472
         d11 |  -1.517695   .2757453    -5.50   0.000    -2.058146   -.9772446
         d12 |  -1.009884   .2811314    -3.59   0.000    -1.560891   -.4588762
          pt |   .6605301   .2367273     2.79   0.005     .1965532    1.124507
         pas |   .2963606   .1253784     2.36   0.018     .0506235    .5420976
------------------------------------------------------------------------------

. gen pas0 = -2.893237*d7 - 3.584759*d8 - 2.150233*d9 - 1.69318*d10 - 1.517695*d11 -
1.009884*d12 + .6605301*pt

. gen pas1 = pas0 + .2963606
```

```
. gen pasneg1 = pas0 - .2963606

. collapse (mean) pas0 pas1 pasneg1, by(period pt)

. twoway (line pas1 period if pt==0)(line pas1 period if pt==1)(line pas0 period if pt==0)
(line pas0  period if pt==1)(line pasneg1 period if pt==0)(line pasneg1 period if
pt==1),xtitle("Grade") ytitle("Fitted Log Odds of Hazard") legend(ring(0) pos(10) col(1)
lab(1 "PAS=1, PT = 0") lab(2 "PAS=1, PT = 1") lab(3 "PAS=0, PT = 0") lab(4 "PAS=0, PT = 1")
lab(5 "PAS=-1, PT = 0") lab(6 "PAS=-1, PT = 1"))
```

It has been shown [Prentice & Gloeckler (1978)] that the likelihood from a [continuous time] proportional hazards model is the same as a [discrete time] model with the cloglog link. The regression coefficients from these two models have identical interpretations.

```
. cloglog event d7 d8 d9 d10 d11 d12 pt pas, nocons
Complementary log-log regression                Number of obs     =        822
                                                Zero outcomes     =        696
                                                Nonzero outcomes  =        126

                                                Wald chi2(8)      =     344.98
Log likelihood = -314.55927                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       event |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          d7 |  -2.876523   .3001068    -9.58   0.000    -3.464722   -2.288325
          d8 |  -3.551336   .4071691    -8.72   0.000    -4.349373   -2.753299
          d9 |  -2.207147   .2535697    -8.70   0.000    -2.704135    -1.71016
         d10 |  -1.792702   .2363678    -7.58   0.000    -2.255975    -1.32943
         d11 |  -1.638841   .2452343    -6.68   0.000    -2.119491   -1.158191
         d12 |  -1.194946    .238931    -5.00   0.000    -1.663242   -.7266496
          pt |   .5953676   .2138192     2.78   0.005     .1762897    1.014446
         pas |   .2572451   .1088811     2.36   0.018      .043842    .4706481
------------------------------------------------------------------------------
```

Comparing the logit link with the cloglog link:

```
clear
set obs 1001
range p 0 1
gen lgp=log(p/(1-p))
gen cllp=log(-log(1-p))
line lgp cllp p
line lgp cllp p if p<0.2
gen diff=lgp-cllp
line diff p
line diff p if p<0.2
line diff p,yline(0.2)
line lgp cllp p
line lgp cllp p if p>0.2
line lgp cllp p if p>0.8
```

For values of $p < 0.2$, the differences are 'small' but for $p>0.8$ the differences are 'large'.

So for 'small' h(t), the logit analysis and the cloglog analysis aught to be 'close'. The analogy between the discrete cloglog and the continuous log, may give an edge to a cloglog choice, especially if one can conceptualize time as continuous but one can only observe time in grouped form. Both are seen quite widely though, so one may wish to choose based on your content area literature.