Duration Models

Nathaniel Beck
Department of Political Science
University of California, San Diego
La Jolla, CA 92093
beck@ucsd.edu
http://weber.ucsd.edu/∼nbeck

Summer, 1999

Duration data

Duration models (event history analysis) have a dependent variable which measures how long it takes for something to happen. Note this dependent variable must be positive.

Duration data comes in lots of forms. May be a complicated event history, that is, at time 1 this happened, at time 2 that happened, time 1 had different values for independent variables than time at time 2. Or simple, a single record per observation, dependent variable is time until failure, ind vars characterize the observation and do not change over the course of time. For now assume latter.

Duration models are related to binary dependent variable models (which ask whether something happens) and event count models (which ask how many things happen). Thus a duration model treats how long it takes before the first event occurs, an event count model treats the number of events that occur and a binary model is appropriate for whether a single event occurs in some time interval (or for a binary variable which is one if one or more events occur). See the Alt, King and Signorino paper.

Start with characterizatation of duration data.

Applications

- Recidivism in criminal justice
- Length of time to complete Ph.D.
- How long an individual is unemployed
- How long an individual is married
- Lengths of coalitions
- Length of tenure on a Congressional committee
- Time until announce support of a bill
- How long a leader is in power

- How long a war lasts
- How long a peace lasts (the democratic peace)

Survivor function, hazard rates, etc.

Let $F(t)$ be the distribution of coalition 'failure' times, and f(t) the associated density. (Note that here we use t for time, rather than y. This is standard, and I hope causes no confusion.) While the distribution can be discrete or continuous, for now think of continuous distributions; we return to discrete time models on Friday. Also assume that F is differentiable so that f exists. For now leave F unspecified.

$$S(t) = 1 - F(t) \tag{1}$$

is called the SURVIVOR function, that is the proportion of coalitions (or whatever) surviving past t. The hazard rate is then

$$h(t) = \frac{f(t)}{S(t)}. \tag{2}$$

Since $f(t) = F'(t)$, the numerator is simply $\lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t}$ which is the rate of failure per unit time in the infinitesimal interval starting at t. The denominator, $S(t)$ makes for a conditional rates of failure, conditional on surviving to t.

Think about human mortality. Hazard is initially high, then declines, then rises. This is easier than thinking about the distribution of lifetimes. Same thing for coalitions and probably most other durations.

(Beware: Some authors use $\lambda(t)$ to be the hazard rate, but that can be confusing given the different uses of λ in duration modeling.

Note: Since $f = F'$ and $S = 1 - F$ and $h = f/S$, there is only one free component here. Once you have defined F, the survivor and hazard functions are also defined. Conversely, once you define h, the failure and survivor functions are also defined.

To see this, we need to look at the integrated hazard, $H(t)$.

$$H(t) = \int_0^t h(\tau) d\tau. \tag{3}$$

It is then easy to show that

$$H(t) = -\log S(t). \tag{4}$$

To see this, just remember that $\int \frac{1}{t} dt = \log(t)$, $f = F' = -S'$

$$\int h(\tau) d\tau = \int \frac{f(\tau)}{S(\tau)} d\tau \tag{5}$$

$$= -\log S(\tau) \tag{6}$$

and since $S(0) = 1$, Equation 4 follows.

Since we can also produce H if we know h, Equation 4 allows us to produce S if we know h. (With finite data or discrete distributions we will replace the integral with a sum. Some authors call H the cumulative hazard.)

Kaplan-Meier plots

If we are just interested in observing $S(t)$ or $h(t)$ for some duration variable (say coalition lengths) *without any covariates* we can use the Kaplan-Meier estimates.These are non-parametric estimates of the survivor and hazard functions.

Let exits occur at $T_1, \ldots, T_K$ and let n_i be the number "at risk" (that is, the number still "alive" and not censored at T_{i-1}) in the interval from T_{i-1} to T_i ($T_0 = 0$). Let d_i be the number that "die" in that interval (remember we only observe deaths at discrete time points).

The the probability of surviving beyond T_i is the the product of the conditional probabilities of surviving through each interval, conditional on having made it to that interval (with the probability of surviving to 0 being 1). So the KM ("product-limit") estimate of the survivor and hazard functions are

$$\hat{S}_{t_i} = \prod_{i=1}^{k} \frac{n_i - d_i}{n_i} \tag{7}$$

$$\hat{h}(t_i) = \frac{\frac{d_i}{n_i}}{t_i - t_{i-1}}. \tag{8}$$

Note: Standard errors can be calculated for these estimates (they are, after all, maximum likelihood). There are a number of different standard errors floating around, all of them more or less equally good. (There are also a number of estimators of S floating around, again all more or less equally good, differing in how they handle simultaneous (tied) exits.) STATA does a good job with all this. (Use option gwood in STATA.)

The survivor estimate is much cleaner than the hazard estimate, which jumps around a lot. So in general we only look at the K-M survivor plot estimate.

CENSORING - Right censoring

Sometimes the data is (right) censored. Coalitions might last forever, but law demands a new coalition form every three or four or five years. Thus cannot observe a coalition longer than this length. The data are CENSORED.

(If we are studying unemployment duration, someone may still be unemployed when we do the study. That observation is censored.)

Need to assume that the censoring mechanism is independent of what we are studying. In studying heart transplants, people may die in an automobile accident. We don't know how long they would have lived. This is easy right censoring, because dying in an automobile accident is not related to the transplant.

But suppose they had died of kidney damage, which could be related to the heart transplant. Is this simple right censoring?

Suppose we study the lengths of strikes, but some end by government injunction? Are the enjoined lengths censored????

Left censoring

Suppose we have a panel and for each period (month) we ask whether someone is on welfare. Some people will start and stop a welfare spell during the analysis, so we have complete duration data. Other people will begin off welfare, but end on welfare - they are right censored and can be handled as above.

Others begin on welfare but end off welfare. They are left censored. They can be handled exactly like the right censored, since all that we know is their duration is at least how long we observed them on welfare.

This assumes censoring is independent of everything else.

Duration models with covariates

We are interested in the effect of covariates. For simple factors, such as majority/minority, we could just produce K-M plots at each level of the factor (and then test for differences). This is very common in biomedical work, where the iv's tend to be few and quite often just factors (treatment or not). In political science, we have lots of continuous covariates, and hence would like something analogous to multiple regression.

Actually, early work just used multiple regression. Could start with what goes wrong in regression, but since duration data do not look like they are generated by standard linear regression, why bother. Maybe regression is okay, but why not just do things right?

But if you care to start off with what goes wrong with regression ....

What goes wrong with linear regression

$$y_i = \vec{x_i}\beta + \epsilon_i \tag{9}$$

1. Predicted Value Could be Negative

2. Prediction Duration increases linearly with X, doubling any x doubles predicted duration

3. Like probit, effect of a change in x depends on the current expected value of y. If this is large, x has little effect. If x is small, may also have little impact, or may have huge impact.

4. How handle censoring?

5. What to do with more complicated data, where covariates change over time? And lots of other complications that are okay with duration/event history methods but impossible to handle in regression setup.

The simple (exponential) parametric model

Let y_i be the duration of, say, cabinet i with characteristics $\vec{x_i}$. In regression the x's are called independent variables, in statistics (and most duration models) they are called covariates. Assume the covariates are stable over the cabinet (no time-varying covariates.) We deal with the issue of time varying covariates later.

Assume that y_i has a conditional (on its covariates) exponential distribution

$$y_i \sim f_e(y_i|\lambda_i) = \lambda_i e^{-\lambda_i y_i}; \quad \lambda_i > 0, y_i > 0 \tag{10}$$

This is just a particular density which is admissible for failure times (that is, it only has support on positive reals, and can take on any positive real

$\frac{1}{\lambda}$ is the mean of the distribution, and hence must be positive. (Variance of distribution is square of mean.) In English, this means that coalition i will, on average last $\frac{1}{\lambda}$ months.

One way of parameterizing λ is

$$\lambda = e^{-\vec{x_i}\beta} \tag{11}$$

where β is a vector of parameters to be estimated. This is same as parameterizing expected duration

$$E(y_i) = e^{\vec{x_i}\beta} \tag{12}$$

Why this way?

Any function returning $\lambda > 0$ would do. Why this parameterization?

Think about how changes in the covariates affect expected duration. What is the effect of the same change in x when expected duration is high and low? It should be greater in the former case. (It is "harder" to move durations when they are short. For cabinets, a change from 6 months to 1 year is in some sense a bigger move than from 30 months to 3 years.

Note that for some given individual

$$E(y_i) = \frac{1}{\lambda_i} = e^{\vec{x_i}\beta} \tag{13}$$

so

$$\frac{\partial E(y_i)}{\partial x_k} = \beta_k e^{\vec{x_i}\beta} \tag{14}$$

$$= \beta_k E(y_i) \tag{15}$$

which is what is wanted.

Note memoryless property:

$$E(y > t + t0 | y > t0) = E(y > t) \tag{16}$$

This is easy to derive directly, but also follows from the fact that the exponential duration distribution is the duration analogue of the Poisson event count distributions. If an event count dist is Poisson, the time until the first event is exponential. The memoryless property follows then from the analogous property of Poisson event counts.

We can see this directly by noting that for an exponential the survivor function has the form $S(t) = e^{-\lambda t}$. Hence

$$P(S > t_2 | S | t_1) = e^{-\lambda(t_2 - t_1)} \tag{17}$$

which depends only on $t_2 - t_1$, not the specific t_1.

Likelihood

We can estimate an exponential duration model via maximum likelihood.

For an uncensored exit, that is, if $y_i = t_i$, the contribution of this observation to the likelihood is just $f_e(t_i)$ which is a function of $\lambda_i = e^{-\vec{x_i}\beta}$.

What about censored obs, that is, though that are censored before failure.

Suppose for observation i, the censoring point is C_i which is fixed and known to the investigator. Suppose $t_i = C_i$. What is the probability of this occurring (for an exponential)?

$$\text{Prob}(y_i \geq C_i) = \int_{t=C_i}^{\infty} \lambda_i e^{-\lambda_i t} dt \tag{18}$$

$$= e^{-\lambda_i C_i} \tag{19}$$

Thus the likelihood of the sample $y_1, y_2, \ldots\ y_n$ with known censoring points C_i is

$$\left[\prod_{y_i<C_i} f(y_i|\lambda_i)\right]\left[\prod_{y_i\geq C_i} P(y_i \geq C_i)\right] \tag{20}$$

which, for the exponential case is

$$\left[\prod_{y_i<C_i} \lambda_i e^{-\lambda_i y_i}\right]\left[\prod_{y_i\geq C_i} e^{-\lambda_i C_i}\right] \tag{21}$$

which can be rewritten

$$\prod \lambda_i^{d_i} e^{-\lambda_i y_i} \tag{22}$$

where d_i is one if the observation is not censored and zero otherwise. We can then take the log likelihood and maximize it by our usual methods, putting in

$$\lambda_i = e^{-x_i \beta}. \tag{23}$$

Interpreting results

STATA (or whatever) will produce estimates of β and standard errors. We can use these to test simple hypotheses, since the estimates are asymptotically normal. We can test more complicated hypotheses with likelihood ratio tests.

A nice way to display results is to calculate expected durations and derivatives at interesting combinations of the independent variables (or perhaps for the countries in the coalition example). Thus we need to show $\hat{\lambda}_i = \lambda = e^{-\vec{x_i} b\hat{e}ta}$.

NOTE THE HAT IN THE ABOVE INDICATES WE DON'T KNOW THE TRUE β BUT ARE USInG AN ESTIMATED VALUES. Thus we should use the methods that King et al., describe in their CLARIFY procedures, though these are not implemented by King et al., for duration models. But would be easy to do so.)

The Weibull model

Exponential has only one parameter. (It is a lot like the Poisson.) Its variance is just the square of its mean - cannot separately estimate a variance once we have estimated a mean.

A more flexible model is that y has a WEIBULL distribution.

$$y \sim f_w(y|\lambda, p) = \lambda^p p y^{p-1} e^{-\lambda y^p} \quad (y, \lambda, p > 0) \tag{24}$$

where

$$\lambda = e^{-\vec{x}\beta} \tag{25}$$

$$E(y) = \frac{\Gamma(1 + \frac{1}{p})}{\lambda} \tag{26}$$

The Weibull is not symmetric, it easier to understand and compute percentiles

$$t(p) = \left[\frac{1}{\lambda} \log\left(\frac{100}{100 - p}\right)\right]^{\frac{1}{p}} \tag{27}$$

for the p'th percentile, so median is

$$= \left[\frac{1}{\lambda} \log 2\right]^{\frac{1}{p}} \tag{28}$$

At this point can do maximum likelihood just as with exponential, substituting Weibull density for exponential density for non-censored observations and the integral of the Weibull for the integral of the exponential for censored observation.

Note that the exponential is special case of ('nested in') the Weibull. When $p = 1$ the Weibull is same as exponential. Can test $H_0 : p = 1$ by usual methods ($\frac{p-1}{se.(p)} \sim N(0, 1)$).

Can gauge substantive impact of Weibull by noting how expected durations (as a function of the covariates, perhaps at their mean or median) change in moving from the exponential to the Weibull. The formulae above are most useful.

The effect of a small change in any covariate is similar to what we saw with the exponential:

$$\frac{\partial E(y_i)}{\partial x_k} = \beta_k e^{\vec{x_i}\beta} \Gamma(1 + \frac{1}{p}) \tag{29}$$

$$= \beta_k E(y_i) \tag{30}$$

Note: Be careful, some programs report p, some report $\frac{1}{p} = \sigma$, some report β, some $-\beta$, some $\frac{\beta}{\sigma}$ and so forth. So be careful! (In STATA, use the option 'time, nohr' to make consistent with slide.

Hazard rate modeling

Note that the hazard rate for the exponential is

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{(-\lambda t)}}{e^{-\lambda t}} = \lambda \tag{31}$$

so the exponential model implies constant hazard rate.

Weibull has monotonically increasing or decreasing hazard, depending on p, given by

$$h(t) = \lambda p(\lambda t)^{p-1}. \tag{32}$$

Monotonicity is better than constancy, but still troublesome.

Alternative Parametric Models

A simple non-monotonic hazards function is the log-logistic

$$S(t) = \frac{1}{1 + (\lambda t)^p} \tag{33}$$

$$h(t) = \frac{\lambda p(\lambda t)^{p-1}}{1 + (\lambda t)^p}. \tag{34}$$

This distribution first rises and then falls if $p > 1$ and monotonically falls if $p \leq 1$ (it can not monotonically rise!). While it looks more flexible than the monotonic Weibull, it is not, since both are two parameter distributions. Once you have estimated the mean and variance of the log-logistic, its shape is fixed. (It is a bit like estimated $y = a + bx^2$ without the linear term, so where the bend occurs is fixed, not estimated.)

There are many other distributions that are used: Pareto, Gompertz, log normal, gamma. In general you can cover most possibilities with the Weibull and the log logistic. Also remember that each book seems to have its own parameterization.

Very (too?) flexible distributions

Finally, note that there are some very general distributions, like the generalized gamma, of which gamma, Weibull and exponential are special cases. The generalized gamma has two free parameters (note that the gamma, even though it has a bend, only has one, and density must be concave) with

$$f(t) = \frac{(\lambda p)(\lambda t)^{p\theta - 1} e^{(-(\lambda t)^p)}}{\Gamma(\theta)} \tag{35}$$

where λ is parameterized as usual.

If $\theta = 1$ this reduces to a a Weibull, and if $p = 1$ it reduces to a simple gamma (which in practice looks like a Weibull), and if both parameters are one, it is an exponential. Thus the generalized gamma nests the other survival distributions of interest, and can, *in principal*, be used to discriminate among them. Thus we could start with the the generalized gamma and the test $H_0 : \theta = 1$ to see if the common Weibull is acceptable.

Cox Partial Likelihood

Another possibility is the Cox Partial Likelihood for proportional hazard models

$$h_i(t) = h_0(t) e^{-\vec{x}_i \beta} \tag{36}$$

where $h_0(t)$ is an unknown 'baseline' hazard which varies with time but not across individuals. This is a member of the 'proportional hazards' family since,

$$\frac{h_i(t)}{h_j(t)} = e^{(\vec{x}_j - \vec{x}_i)\beta} \tag{37}$$

and so is independent of t (always remains in same proportion for all t.

How do we estimate such a model? Cox came up with a non-parametric method called partial likelihood which turns out to be easy to estimate and is almost fully efficient. The idea is that suppose we have n individuals at risk at time t, and that i exits at time t. Given the parameters and covariates, then, conditional on either someone exiting at time t, the probability that i is the exiter is

$$PL(i) = \frac{e^{-\vec{x}_i \beta}}{\sum_{k=1}^{n} e^{-\vec{x}_k \beta}} \tag{38}$$

which then can be maximized as a function of β.

It turns out that this estimate yields asymptotically normal estimates with a covariance matrix which yields, as in maximum likelihood, asymptotic standard errors. Maximizing the partial likelihood is a lot lot maximizing the full likelihood.

Why is this so? Because the only information that Cox ignores is the timing of the exits; at the exits the method uses all the information about which unit is most likely to exit. It turns out that this is most of the information in a duration model.

Note that the Weibull model is a proportional hazards model, since

$$\frac{h_i(t)}{h_j(t)} = \left[\frac{\lambda_i}{\lambda_j}\right]^p \tag{39}$$

which depends only on the covariates but not time. Thus we note that the Cox estimates of the β are usually similar to the Weibull estimates. Thus while proportional hazards is a strong assumption, it must be weaker than the assumptions which underlie the Weibull.

More on Ties

In continuous time there can be no tied exits. But we frequently record data discretely, so ties are very common in practice. There is some conventional wisdom that a non-trivial proportion of ties causes problems for Cox P-H, but that is not obviously the case.

What goes wrong?

Say units 1 and 2 both exit at time t. If we knew that 2 exited after 1 (say with 3 left), the contribution to partial likelihood would be

$$PL(1) = \frac{e^{-\vec{x}_1\beta}}{e^{-\vec{x}_1\beta} + e^{-\vec{x}_2\beta} + e^{-\vec{x}_3\beta}} \tag{40}$$

$$PL(2) = \frac{e^{-\vec{x}_2\beta}}{e^{-\vec{x}_2\beta} + e^{-\vec{x}_3\beta}} \tag{41}$$

$$\tag{42}$$

whereas if 2 had really exited first we would have

$$PL(1) = \frac{e^{-\vec{x}_1\beta}}{e^{-\vec{x}_1\beta} + e^{-\vec{x}_3\beta}} \tag{43}$$

$$PL(2) = \frac{e^{-\vec{x}_2\beta}}{e^{-\vec{x}_1\beta} + e^{-\vec{x}_2\beta} + e^{-\vec{x}_3\beta}} \tag{44}$$

$$\tag{45}$$

As you see, the only problem is caused by the "second" denominator. The quick and dirty method, due to Breslow, is to use the same denominator (the sum of all the $e^{-\vec{x}_i\beta}$) for all the tied exits. This is quick, and with few ties it isn't bad.

A better method is due to Efron. He argued that either 1 or 2 could have equally left first. Thus he takes the denominator for the second exit (say 2) as $.5e^{-\vec{x}1\beta} + .5e^{-\vec{x}2\beta} + e^{-\vec{x}3\beta}$ with similar weightings for more than two ties.

It used to be too hard to compute the exact partial likelihood, but with modern computers and good software (like S-PLUS), this can now be done. Here we would just compute the probability of 1 and 2 both exiting at time t, given the risk set.

Testing the proportionality assumption

We can test the assumption of proportionality by seeing whether the impact of the covariates changes over time. The simplest way is to estimate a model which has

$$h_i(t) = h_0(t)e^{(\vec{x_i}\beta + \vec{x}_i t\gamma)} \tag{46}$$

which can be estimated using time varying methods to be discussed later.

If γ is not zero by a standard t-test, then we know the effect of some covariates changes with time. Here we impose a strong linear form, but we could test to see if a period dummy changes things, or any other function of time.

We can also examine the estimated baseline survivor function at different values of some discrete covariate and see if they are parallel, which is what proportional hazards implies.

There are lots of other tests, many implemented in STATA/

Estimating the baseline hazard

Once we have estimate the β we can estimate the components of h_0 (or we could do jointly, but never is done that way). Given the hazard function, and the general result on survivor functions. we have

$$S(t) = S_0(t)^{e^{-\vec{x}\beta}} \tag{47}$$

since

$$S(t) = e^{-H(t)} \tag{48}$$

but

$$H(t) = \int_0^t h(\tau)d\tau \tag{49}$$

$$= e^{\vec{x}\beta} \int_0^t h_0(\tau)d\tau \tag{50}$$

$$= e^{\vec{x}\beta} H_0(t) \tag{51}$$

where $H_0(t)$ is the "baseline integrated hazard" and $S_0(t)$ is the "baseline survivor function."

To estimate the baseline function, since exits occur only at $t_1, \ldots, t_K$, we note that maximum likelihood estimates of the baseline hazard must be zero at other than these exit points. (We thus really have a discrete failure time model, which will be discussed on Friday.)

Letting α_k be the baseline probability of surviving at t_k, given survival through t_{k-1}, $1 - \alpha_k$ is the baseline hazard at t_k. The baselines survivor is then

$$F_0(t_i) = \prod_{j=0}^{i-1} \alpha^j \tag{52}$$

where α_0=1.

Thus, given the β, the likelihood of the data as a function of the α_k is

$$\prod_{i=1}^{K} \left[\prod_{j \in \mathcal{D}_j} (1 - \alpha_i^{e^{-\vec{x}\beta}}) \prod_{j \in \mathcal{R}_j - \mathcal{D}_i} \alpha_i^{e^{-\vec{x}\beta}} \right] \tag{53}$$

where $\mathcal{D}_i$ denotes the units dying at t_i and $\mathcal{R}_i$ denoting the units still at risk just before t_i.

This can then be maximized as a function of the α to allow for estimation of the baseline hazard function. The baseline survival function is then obtained by computing the baseline integrated hazard and exponentiating. Since the hazards are only non-zero for discrete times, the integrated hazard is obtained by simple addition (and is a step function).

In the absence of covariates, the estimated baseline survival is identical to the Kaplan-Meier estimate.

Time Varying Covariates

In the King model, duration is a function of characteristics of coalition which don't change. What if you think that a coalition fails partly as a function of the state of the economy. Then we have time varying covariates (TVC's). Usually easiest to handle in terms of Cox model, though

can do with full likelihood. (May be quite tedious to setup in dataset and handle in software, but this is getting easier.)

STATA is quite good for these types of data sets. It uses a "counting process" setup, based on Flemming and Harrington, *Counting Processes and Survival Analysis*, Wiley, 1991.

In this setup each record (line of data) gives the value of covariates that are constant between a beginning and ending time point, and whether an event (failure) has occurred by the ending time point. Thus we can keep "count" of the number of failures for multiple failure data, and otherwise handle complicated data sets.

Note for relatively continuously varying measures, such as the economy, we might wish to take them as constant over a year or so in order to simplify data handling (or not, the counting process notation would have no trouble with monthly varying covariates). But covariates do have to be constant over some discrete time interval. This is not a serious limitation since data comes to us in discrete from in any event. (You can have variables which are a continuous function of time, such as time itself, but such variables are handled differently and are not TVC's.)

Beware of endogeneity or spurious causation

Once we allow time varying covariates, we must be very careful of using endogenous or jointly causal covariates. For example, in a model of marriage duration, we might use the time varying covariate, whether there are children of the marriage. But we might expect (hope!) that people decide to have children if they think their marriage is stable, so finding that children increases marriage duration might well be spurious.

Similarly, in the cabinet data, we might have a variable as to whether a key minister has resigned. But such resignations might come because the minister recognizes an imminent breakup, and so in spite of a large coefficient, the effect of this variable is spurious, not causal.

One simply must be careful here. Is the variable under the control of the unit? The economy is a pretty good exogenous measure (usually), but any strategic variable must be suspect. There is little difference between thinking about this in a duration and a simple multiple regression context. There are no statistical means to assess the suitability of time varying covariates.

We also must be careful about exactly what we mean by time varying covariates. Consider a study of length of unemployment spells, and assume that in some states people have 26 weeks of benefits, in others 39 weeks. We might thing that weeks left on benefits is a good time varying covariate. But it is not. People know the rules throughout their spell. Thus there is nothing that is changing from week to week in terms of the rules. Better would be to have a dummy variable which marks which rule a person is under. This would be a non-time varying covariate.

Estimation

Cox proportional hazards

In the Cox setup there is no reason that P(exit by i at time t_i) should not be a function of $\vec{x}_{t_i}$ rather than just a function of $\vec{x}_i$. One needs to be a bit careful here, since maybe the economy at time t has an effect with a lag, or maybe it is not the value of the covariate at time t but rather a discounted sum of past values that should be entered, but these are standard modeling issues whenever we have temporal data. So once one gets the data setup AND THE MODEL right, TVC's present no special problem. Note that with TVC's hazards are no longer proportional (that is $\frac{h_i(t)}{h_j(t)}$ will vary over time, as the covariates change over time. This is not a problem. STATA handles this setup quite well.

Parametric Methods

The TVC setup is much more complicated in the fully parametric models. This is because we don't get to just make probability calculations at exit times.

Remembering that once we know the survivor function everything we need for likelihood can be obtained, all we need do is handle the survivor function. Assume that the time varying covariates are measured discretely at intervals $t_0, t_1, \ldots, t_k$ (which might vary by individual, but let us ignore that - it is only a complication of the data setup, and the counting process data setup handles this easily enough).

By the laws of conditional probability

$$S(t_k) = \prod_{j=1}^{k} P(T > t_j | T > t_{j-1}). \tag{54}$$

Letting the hazard in the interval (t_{j-1}, t_j) depend on the value of the covariates in that interval, x_j (they are constant during that interval), each of the terms in the product in Equation 54 has the form (from the basic integrated hazard equation)

$$P(T > t_j | T > t_{j-1}) = e^{-\int_{t_{j-1}}^{t_j} h(s|x_j)ds} \tag{55}$$

where $h(s|x_j)$ is the hazard function (which depends on the covariates for the relevant subinterval).

For those units that exit at t_k, the contribution to the likelihood is $f(t_k) = S(t_k)h(t_k)$ while for those that are censored at t_k, the contribution is just $S(t_k)$.

Letting δ_i be the usual (non)censoring indicator, we thus have an individual who exits or is censored at t_k contributing

$$\log L_i = \delta_i \log h(t_k) - \sum_{j=1}^{k} \int_{t_{j-1}}^{t_j} h(s)ds \tag{56}$$

to the log likelihood. (For censored observations, with $\delta_i = 0$, this reduces to the log of the survivor function; for non-censored observations it is $\log h(t_k) + \log S(t_k)$.)

Depending on the the distribution, these integrals are more or less easy to evaluate. They are particularly easy for the Weibull and log-logistic (and Weibull with gamma heterogeneity).

Generalized Residuals

We can see how our model fits by looking at the difference between observed and predicted durations (using the formulae for $E(y_i|\vec{x}_i)$. These can be very helpful, but these differences do not have the nice mathematical properties that ordinary OLS residuals have. Cox (with Snell) has defined a "generalized residual" as some function of the data and estimated coefficients such that if the model is correct these generalized residuals should have a known distribution.

For duration models the generalized residuals are the estimated integrated hazards (that is, the log of the estimated survival function). If the model is correct, these generalized residuals should look like draws from a unit exponential distribution

But the most common method for checking specification is graphical. Note that if $S(t)$ follows a Weibull,

$$S(t) = e^{-\lambda t^p} \tag{57}$$

$$\log(-\log(S(t)) = \log \lambda + p \log(t) \tag{58}$$

. For a unit exponential, both λ and p are one. Remembering that the integrated hazard is the negative log of the estimated survivor function, we thus can test for whether S follows a unit exponential by plotting the log of the integrated hazard against the log of time, and seeing whether the intercept is zero and the slope is one.

We could also plot these residuals against possible independent variables to see if they relate to survival - could also pick up non-linear effects this way.

There are lots of other residuals - modified Cox residuals which account for censoring, martingale residuals, deviance residuals and score residuals. They all are transformations of the Cox residuals. See Collett or Box-Steffensmeier and Jones (forthcoming).

Weibull with Gamma heterogeneity

Note that if we have heterogeneous exponential durations, then it will appear that probability of exit is declining over time. Why? Because those with highest λ will exit earliest. As time goes on, our risk set contains only those with smaller and smaller λ's.

Suppose we think that on average the Weibull model is, on average, correct, but that the survival rate randomly varies from individual to individual in an unobserved manner (that is, we

won't model as a function of covariates). We can write

$$S(t|\nu) = \nu e^{-(\lambda t)p} \tag{59}$$

where ν measures random variation. ν must be non-negative and should have mean one. A distribution which leads to nice math is that ν is Gamma with mean one and variance θ. Since we only know the distribution of ν, all we can do is take expected values (that is, "integrate it out"). (THIS IS A NICE EXAMPLE OF IDEAS ABOUT MIXING DISTRIBUTIONS.) Letting someone else do the work, we get

$$S(t) = [1 + \theta(\lambda t)^p]^{-\frac{1}{\theta}}) \tag{60}$$

which yields a hazard rate

$$\lambda p(\lambda t)^{p-1}[S(t)]^{\theta} \tag{61}$$

which nicely reduces to the Weibull when there is no variance in ν, that is, $\theta = 0$.

Note that using the arguments about pseudo-maximum likelihood (GMT), if you ignore the heterogeneity your parameter estimates will be right but your standard errors can be wrong. These errors can be fixed using the usual robust maximum likelihood standard errors.

Truncation

Suppose we only observe durations of a certain length, e.g. suppose unemployment spells are only recorded if they exceed four weeks. Then we know that $f(t) = 0, t <= 4$. We can use standard truncation methods to fix up the likelihood, that is, if say we think that unemployment durations are Weibull (f_w), the truncated density,

$$f_{wtrunc(t)} = \frac{f_w(t)}{1 - \int_0^4 f_w(s)ds}, t > 4. \tag{62}$$

Split Population

An interesting model has been studied by Schmidt and Witte. Modeling recidivism, they assume that of those people who have not returned to jail, some will (that is are censored) but some never will. Note that those who are already back in jail are clearly of the former type! It is easy to set this model up (and easy to estimate in LIMDEP).

Let R_i be the probability that individual i is the type who might eventually return to prison. Thus an individual will not have returned either if they are censored or of the non-returning type. The probability of this (which goes in the likelihood) is

$$(1 - R_i) + R_i \; P(\text{return after censoring point}) \tag{63}$$

and the density for those whose return is observed is just R_i times the original duration density.

It then remains to estimate the R equation as a function of covariates; this is just a logit (or probit). The setup in LIMDEP is straightforward.

Multiple Destination (Competing Risks)

So far have assumed that only possibility is that something ends (coalition breaks up, divorce, death). Only other possibility so far is right censoring.

But there may be a variety of "destinations." A member of the House and retire, be defeated for re-election, die, run for the Senate, etc. A coalition could break-up, add new members, change at election time, etc. A marriage could end in divorce, annulment or death. These are all multiple destination problems (sometimes called "competing risks" though Lancaster prefers to reserve that term for a simple subset of multiple destination problems.

How to model? Say the destinations are labelled $1, 2, \ldots, K$.

Let $h_k(t)$ be the "transition intensity" for exiting to state K at time t. (It could also be called a state specific hazard, which is why we use λ.)

These can be defined just like hazard functions, except in terms of exit to destination k in a small interval after time t. All of the other definitions in terms of survivor functions goes through analogously, and all the destination specific functions can be composed to generate unconditional hazard and survivor functions. (See Lancaster.)

The Medical Model

The simplest multiple destination model is the medical "competing risk" model. A patient could die of lots of things. Let T_k represent a patient's time until death from cause k. Then we observe the minimum of all the T_k, and we can treat the various T_k of independent (how long would it be until the patient died of prostate cancer if he hadn't died of lung cancer first seems meaningful; can we model how long a Member of the House would have served before being defeated if said Member hadn't been elected to the Senate?).

For competing risks

$$P(\text{exit to k in interval t, t+dt}) \tag{64}$$

$$= h_k(t) e^{\left[-\sum_{j=1}^{K} \int_0^t h_k(u)du\right]} dt \tag{65}$$

which is a pretty simple model.

Models with Binary Dependent Variables

Things are easy with a continuous dependent variable (and normal errors). What if the dependent variable is binary (even worse, what if is multichotomous or tobit, but let us start with

easy case!). Suppose the dependent variable is whether a dyad is at war in a given year, with iv's being characteristics of the dyad. The usual way to estimate such a model is via logit or probit, but this ignores any temporal dependencies in the data, and so will probably yield inferior parameter estimates, wrong standard errors and will not even estimate the dynamics.

In Beck and Katz (1997) we run some simulations of the following form:

$$y^*_{i,t} = \vec{x}_{i,t}\beta + \epsilon_{i,t} \tag{66a}$$

$$y_{i,t} = 1 \text{ if } y_{i,t} > 0 \tag{66b}$$

where

$$x_{i,t} = \rho_x x_{i,t-1} + \nu_{i,t} \tag{67}$$

$$\epsilon_{i,t} = \rho_\epsilon \epsilon_{i,t-1} + \mu_{i,t} \tag{68}$$

varying the two ρ's.

We found that with very high ρ (both) the probit standard errors underestimated variability by almost 50% in some cases, and that with small T's the mean squared error of the estimated $\hat{\beta}$ could be quite high.

There are a couple of fixes to the standard errors that might work.

Huber grouped standard errors

While this does not get the time series structure right, it does correct the standard errors for interrelationships of the errors within each unit.

Huber showed that if one has a maximum likelihood estimator (such as logit or probit) with grouped data that the appropriate variance-covariance matrix of the parameter estimates is

$$\text{VCV}(\hat{\beta}) = H^{-1}\left[\sum_{i=1}^{N}\left(\sum_{j=1}^{T}\vec{u}_{i,j}\right)'\left(\sum_{j=1}^{T}\vec{u}_{i,j}\right)\right]H^{-1} \tag{69}$$

where H is the usual matrix of second derivatives (Hessian) and $\vec{u}$ is the vector of scores (first derivatives of the likelihood with respect to the parameters for each observation). Note that H^{-1} is the usual estimate of the covariance matrix of the estimator.

Huber is clearly not optimal in this situation, but it must be better than doing nothing.

Note it is easy to do this in STATA with a "robust" subcommand to xtprobit (it could also be used for normal models).

Effects - back to TSCS data!

Since most work is on large N panels, the standard wisdom is that you cannot used fixed effects with binary dependent variables. But this is an incidental parameters issue, not a binary dependent variable issue. Thus with bit T, fixed N TSCS data, you can use simple fixed effects. Alternatively, one could do random effects probit - the algorithm that drives this model appears to work decently for T's as large as 100. Note that random effects probit is easy; random effects logit is not. It is not obvious that the Chamberlain conditional approach for logit will work well with large T.

One can also do RCM models here; the setup for hierarchical models is about as easy in logit or probit as for the normal continuous model (since it is GLM based).

Back to Event History

A perhaps better way to proceed is to note that binary TSCS data looks like event history data, with each 1 marking a failure and the time between 1's, that is, the number of 0's, being the time until or between failures. This works just fine so long as the number of 1's is not large, so we have a good distribution of failure times.

We could thus simply convert the binary TSCS data into event history data and use the methods we have already learned. The Cox semi-parametric approach is most appropriate (unless we have some idea about the shape of the hazard function).

Thinking about binary TSCS data as event history data gives us certain insights.

- The simple probit/logit approach is equivalent to the assumption of no duration dependence in event history analysis. Thus if we are willing to assume no duration dependence and use an exponential, then the simple probit/logit model is fine. But normally we test for duration dependence, no such test for logit/probit. (Re-examine Berry and Berry.)
- While it looks like we have $N \times T$ binary TSCS observations, this is the same at N duration observations. Thus even if $N \times T$ is large, we may not have all that much information. This is of particular importance in evaluating some critiques of the democratic peace analyses.
- While we think of probit/logit as having troubles with rare events, such rare events are the lifeblood of event history analysis. Non-rare events yield very short durations, which are difficult; rare events yield a wide variety of durations, which is good.
- Binary TSCS data in event form may have more than one event per unit. In the probit/logit setup, we assume that second and subsequent events can be modeled just like first events; events history modelers realize that life is more complex. Thus we have both duration dependence for time until failure, and complexities involved in modelling subsequent failures for the same unit.

Discrete Time Duration Models

Sticking with one event per unit for now, we can model the time until an event in binary TSCS data by a discrete time duration model. These are in many ways simpler than continuous time models.

Let us assume we have time measured in equal discrete intervals, $0, 1, \ldots, t, \ldots$. We only observe whether someone dies in the interval $(t-1, t]$, (open on the left, closed on the right) and will assume this is a death at t. Then we need discrete time analogues of the survivor and hazard function. The survivor function will simply be a step function, with steps at $1, 2, \ldots, t, \ldots$.

Let y_i be the duration for the i'th unit; y_i is a discrete random variable, with support at the positive integers.

Using the same notation as for continuous time (note that we have to be careful here differentiating $y > t$ from $y \geq t$ since we are in discrete time)

$$\begin{aligned} S(t) &= P(y > t) \\ &= P(y > t | y > t-1) P(y > t-1) \\ &= \prod_{i=0}^{t-1} P(y > t-i | y > t-i-1) \end{aligned} \tag{70}$$

where $S(0) = P(y > 0) = 1$. We still have

$$F(t) = 1 - S(t). \tag{71}$$

Since F is discrete, we have an associated discrete density with support on the positive integers,

$$f(t) = F(t) - F(t-1). \tag{72}$$

Here the density is a probability; it is the unconditional probability of dying in the interval $(t-1, t]$.

We can define the discrete hazard analogously to the continuous time hazard, though we can now use probability interpretations. Letting $h(t)$ be the hazard of dying at time t (or in the interval $(t-1, t]$ given that one survived until $t-1$, we have

$$h(t) = \frac{f(t)}{S(t-1)}. \tag{73}$$

Since $1 - h(t)$ is the conditional probability of surviving at t given survival through $t-1$, substituting in Equation 70, we get

$$S(t) = \prod_{i=0}^{t-1} [1 - h(t-i)] \tag{74}$$

Note that we could have unequal interval lengths which just makes the notation more complicated (and is of little practical interest).

We also could start with a continuous time model, and then use integrated hazards to produce discrete quantities (so that the discrete hazard is the integrated hazard from $t-1$ to t and so forth). This is the way Sueyoshi (JAE, 1996) proceeds. But it is not obvious that there is a practical gain to this other than to show the obvious underlying unity of discrete and continuous time models. Note that a continuous time Weibull will induce a discrete time "Weibull" and so forth. Note also that the Cox proportional hazards model is really a discrete time model.

Estimation of discrete duration models via logit

If we estimate a binary dependent variable model with dependent variable $y_{i,t}$ being whether unit i failed in the interval $(t-1,t]$ (by probit, logit or any other binary model), as a function of covariates $\vec{x}_{i,t}$, we are estimating a model for $h(t)$. (If the dependent variable is scored as 1 for non-failure, then we have a model for $1-h(t)$. Thus if we start with a logit or probit model we induce a survival model (or, as Sueyoshi or Alt, King and Signorino do, we can start with a survival model which induces the binary model).

If we estimate a simple binary dependent variable model (for shorthand I say "logit"), with the $x_{i,t}$ as independent variables, we are assuming the hazard rate is time invariant (that is,

$$h_{i,t} = h(\vec{x}_{i,t}) \tag{75}$$

so that the relationship between the covariates and the conditional probability of failure is the same at all time points) or that the failure process is memoryless.

These are the same set of assumptions that led to the exponential. While the simple independent logit is not the exact discrete time analogue of the exponential, the two models are very similar.

To allow for duration dependency we would need to estimate a binary model (with $y_{i,t}$ being one for the failure of unit i in the interval $(t-1,t]$ which has

$$h_{i,t} = h_t(\vec{x}_{i,t}). \tag{76}$$

But this is too general. A possible compromise to allow for different intercepts at each time point, so the hazard changes with time. Thus we might have

$$h_{i,t} = a_t + h(\vec{x}_{i,t}) \tag{77}$$

which would be estimated by putting in a period dummy in the logit. But these are perhaps to jagged, so we might be better off using

$$h_{i,t} = s(t) + h(\vec{x}_{i,t}) \tag{78}$$

where s is some function of time. In my own work I model s as a "smoothing spline" but a natural spline or even a polynomial in time would probably do. Note that one just adds this term to the basic logit, so easy enough to do.

One could also just use the time dummies, which is easier if a bit less elegant.

Remember, *the time variable is time since the last "event," not the particular period of the observation*.

Sometimes the time dummies or the spline indicate that we don't need to include time in the specification (using standard tests on the coefficients of all the time variables). At that point we can assume no duration dependence and use ordinary logit.

Or the spline may look linear, in which case we can just use time in the logit (again remembering that time is time since the last event).

A more formal derivation

So far the approach has been informal. We can formalize it by thinking of the binary TSCS data as having come from a *grouped time* duration model, where grouping is by year.

Start with a continuous time Cox proportional hazards model, so

$$h_i(t) = h_0(t)e^{0x_i,t\beta}. \tag{79}$$

Letting $S(t)$ be the probability of surviving beyond t, by the math of hazard rates we have basic identity that

$$S(t) = \exp(-\int_0^t h(\tau)d\tau). \tag{80}$$

We only observe whether or not an event occurred between time $t_k - 1$ and t_k (assuming annual data) and are interested in the probability of this event, $P(y_{i,t_k} = 1)$. This probability is one minus the probability of surviving beyond t_k given survival up to $t_k - 1$. Assuming no prior events (so $t0 = 0$), we then get

$$P(y_{i,t_k} = 1) = 1 - \exp(-\int_{t_k-1}^{t_k} h_i(\tau)d\tau) \tag{81a}$$

$$= 1 - \exp(-\int_{t_k-1}^{t_k} e^{0x_i,t_k\beta} h_0(\tau)d\tau) \tag{81b}$$

$$= 1 - \exp(-e^{0x_i,t_k\beta} \int_{t_k-1}^{t_k} h_0(\tau)d\tau) \tag{81c}$$

(Note that $0x$ is indexed by t_k not τ because we assume that the independent variables are only measured for an entire interval and not for every instant in the interval $t_k - 1$ to t_k.)

Since the baseline hazard is unspecified, we can just treat the integral of the baseline hazard as an unknown constant. Defining

$$\alpha_{t_k} = \int_{t_k-1}^{t_k} h_0(\tau)d\tau \text{ and} \tag{82}$$

$$\kappa_{t_k} = \log(\alpha_{t_k}) \tag{83}$$

we then have

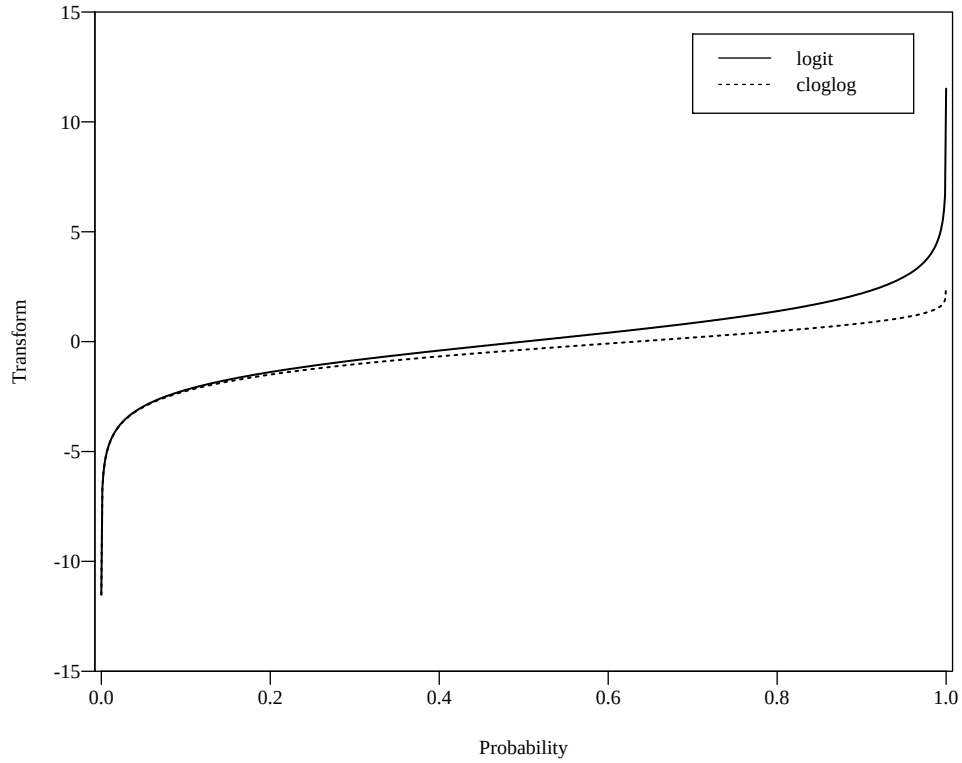
$$P(y_{i,t_k} = 1) = 1 - \exp(-e^{0x_{i,t_k}\beta}\alpha_{t_k}) \tag{84a}$$

$$= 1 - \exp(-e^{0x_{i,t_k}\beta+\kappa_{t_k}}) \tag{84b}$$

This is exactly a binary dependent variable model with a cloglog link and the κ (time dummy) terms added.

There is almost no difference in practice between estimated a logit model and a cloglog model (see Figure 1. The κ terms are the temporal dummies, which by a smoothness argument we can model as a spline in time.

Figure 1: Comparison of cloglog and logit transforms



Further Complications

Thinking about the war data as event history data leads to thinking about other issues. Dyads can fight a number of wars. Here time is coded so that the end of a war (which can last more than one year) starts the clock anew.

But no events history analyst would stop without thinking that perhaps the modelling of subsequent events is different from the first event: the hazard, conditional on a prior event, is different from the hazard conditional on no prior event (in general). Durations of second hospitalizations are longer than durations of first hospitalizations!

This is difficult to model. One solution is to add a variable to the hazard function which counts the number of previous failures. or perhaps to select only those cases up to the first failure and see if that model differs from the overall model.

Multiple spells is a very hot issue now in biostatistics, particularly in modelling hospitalizations for AIDS patients.