ECONOMETRICS II, Fall 2024.

Bent E. Sørensen

1 Duration Models

In economics, we often want to model how long it takes for something to happen. I am currently working with a paper on mortgage refinancing and it is not so interesting to see whether Jones or Smith refinance without knowing when. A mortgage typically lasts 30 years from origination and for the lender it makes a huge difference if someone refinances (pay it off and gets another one, although the lender only care about the paying off as the new one typically is with someone else) after 1 year or after 20 years. Similarly for default. Another area is labor: how long does it take workers to find a job—that may be crucial for the business cycle.

If we just care about how many people get a job in a year, sometimes we will estimate a Poisson model. Under some assumptions, modelling how long you wait for a cab in a certain street or modeling how many cabs are likely to pass in an hour are two sides of the same coin. I assume you have a little familiarity with the Poisson model and I will not talk further about it here.

Define a continuous random variable T_i are the time until an event happens, although we usually use the model for discrete intervals like days or months. Events can be getting a job, defaulting, dying (whence the name)...whatever is your focus. "Time" here is usually time from some event started (purchase of new car, onset of unemployment, origination of mortgage or other loan,...) so if you have a sample of events "started" at different times, you have to make sure to keep track of what

you mean by time. We define some simple concepts namely, the density

$$f(t)$$

(for $t \geq 0$) and the survival function

$$S(t) = P(T_i > t)\,,$$

which of course is one minus the distribution function: $S(t) = 1 - F(t)$. It therefore contains the same amount of information, but in survival analysis we focus on the ones that are still "alive." We also make use of the hazard function

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

The hazard function is the density in the conditional distribution that conditions on still being alive or, a little imprecisely, the probability of dying conditional on being alive. Very often, it the hazard that we are interested in because this is more natural to model in economics than the density. For example, does the hazard of leaving unemployment go up or down with the length of an unemployment spell, or maybe the hazard go up when unemployment benefits are about to run out.

The most commonly used distribution is the exponential and I expect that you know that by heart as it can be seen as the benchmark distribution. The exponential distribution is characterized by a **constant hazard** h. The density function is (no surprise here) exponential:

$$f(t) = \theta \exp(-\theta t)\,,$$

and distribution (cumulative density) function

$$F(t) = 1 - \exp(-\theta t)\,.$$

So the survival function is

$$S(t) = \exp(-\theta t)\,,$$

and you can verify that the constant hazard is the parameter θ. Notice, that θ needs to be positive for this to make sense.

In most applications that I can think of, you would have "regressors;" that is, you would want the distributions to depend on co-variates such as gender, age, income etc. A common suggestion is that, in order to make sure the hazard is positive, one parameterizes

$$\theta_i = \exp X_i\beta\,,$$

where β is the vector of parameters to be estimated. For example, you might reasonably want to know if duration of unemployment is longer on average for, say, older or wealthier individuals. In your Matlab program, you would have a code for the hazard and the hazard goes into the likelihood function, but some textbooks like to substitute this in and you get, for example, the survivor function

$$S(t_i) = \exp[-\exp(X_i\beta)\,t\,]\,,$$

which I think looks confusing/scary with all those exponentials. Do not look at it! In most of what we do, think modularly: you know how the hazard rate goes into the likelihood and you can have a formula for the hazard. I have not personally worked with this specification and I wonder if it makes sense, if, say, the coefficient to income is positive and someone is having very high income then the exponential in the hazard makes it absolutely huge, so I would not use that specification without seeing if some observations (outliers) drive all the results. There is no generally agreed on alternative, I might just do $X_i\beta$ and see if any of the hazards go negative, or you can do $X_i\alpha^2$ which is an old trick which forces $\beta = \alpha^2$ to be posi-

tive. (If you want to test hypotheses about β you would need to use the Delta rule.)

You can have the hazard depend on calendar time s, but not duration t! The duration dependence is captured by the shape of the survival function.

For specific purposes, you can use other distribution functions, in principle, any distribution function defined for positive numbers. One is the log-normal. There is also a Weibull, which allows for increasing, the decreasing hazards. If you are interested, look at their shapes online, but in order to now use too much time on this subject, I will only ask about duration models with exponential densities.

Consider a sample of, say, unemployed individuals. You would follow the sample for some interval of time $[0, T]$. Period 0 is the time unemployment starts. (This could be a different calendar time for different individuals, in which case the end time would vary by individual if you stop at a given calendar time, but that does not create issues). If someone gets a job before period T, we talk about a *completed spell.* (I like that word, makes me think of magic spells in Harry Potter.) If we have to stop following people because we have to write up the research report, some people will still be unemployed and we talk about incomplete spells.

For the exponential with hazard θ_i for agent i (which you can specify as just discussed), in the case of all spells completed (at t_i for agent i), we get (assuming θ_i is a function of "regressors" and parameters β the log-likelihood

$$l(\beta) = \Sigma_{i=1}^{N} \log \theta_i - \theta_i t_i \, .$$

The typical situation, however, is one where say the first N_1 spells are complete and the other N_2 spells are incomplete at duration T. We often say the incomplete

spells are *censored* at period T. (You should yourself be able to generalize to the situation where agent is spell end at T_i, that is the kind of small generalizations that appear in midterms or exams.) For this situation, we just know that the censored agent would have a duration in the set $t > T$. But, as we saw for censored regression models, there is no problem in mixing probabilities and densities in the likelihood function and the probability that agent i is censored at t is simply the survival function $S(T; \theta_i)$. We get

$$l(\beta) = \Sigma_{i=1}^{N_1}(\log \theta_i - \theta_i t_i) - \Sigma_{i=N_1+1}^{N_2} \theta_i T \, .$$

Sometimes econometricians like to write the likelihood in terms of the hazard and the survival function by using the definition of the hazard to get

$$f(t) = h(t) \, S(t) \, .$$

It is a simple exercise to see how the likelihood changes if you substitute this expression for the density.

Instead of assuming an exponential form for the hazard in the exponential distribution with covariates ("regressors"), it is probably better in many applications to assume a logistic form

$$\theta_i = \frac{\exp X_i\beta}{1 + \exp X_i\beta} \, ,$$

as the logit will stay limited rather than taking very large value for large X values. For the logistic specification

$$l(\beta) = \Sigma_{i=1}^{N_1} \, \log(\frac{\exp X_i\beta}{1 + \exp X_i\beta}) - t_i * \frac{\exp X_i\beta}{1 + \exp X_i\beta} \, .$$

Many people estimate the hazard in an explicitly discrete time model. Consider the time series likelihood

$$\pounds(\theta) = f(x_T|X^{T-1}) * ... * f(x_2|x_1) * f(x_1) \, ,$$

where, for example. $f(x_2|x_1)$ is the discrete probability of survival in period 2 conditional on x_1, where x_1 is survival in period 1. If "death" happens, the sampling stops. So in the survival application, the first term is the probability of "death" at period T multiplied by the conditional probabilities of not dying in the first $T-1$ periods.

For a constant hazard (for simplicity of notation) model in discrete time, we have

$$\pounds(\theta) = \theta * (1-\theta) * ... * (1-\theta) = \theta * (1-\theta)^T .$$

Assume we observe the data at higher frequencies, say periods of length δ and also observe the outcomes at the same time range, so T/δ observations. The hazard over interval δ is $\theta\delta$ so we have

$$\pounds(\theta) = \delta\theta(1-\delta\theta)^{\frac{T}{\delta}-\delta} \approx \delta\,\theta e^{-\theta T} .$$

because $(1-\delta\theta)^{\frac{T}{\delta}-\delta} \to e^{-\theta T}$ for $\delta \to 0$. (This convergence you might know from continuous compounding of interest rates.) The probability of being in a small δ interval is δ time the density, so what we see is that for small time intervals, the probability from the discrete model approximates the continuous exponential duration density. So estimating in discrete time give the hazard suitably normalized to reflect the time period.

People often estimate the duration model in discrete time using logit estimation for the hazard (implicity, I think, having the above approximation in mind). It is perfectly valid to assume the hazard in discrete time takes the form $\theta_i = \theta(X_i) = \exp(X_i\beta)/(1+\exp(X_i\beta))$ (you could assume probit). So for "death" at period T, the likelihood would be $(\exp(X_i\beta)/(1+\exp(X_i\beta)) * (1/(1+\exp(X_i\beta))^{T-1}$ or for a sample of individuals you use the standard logit procedure and get

$$l(\beta) = \Sigma_i[(T_i-1) * \log\frac{1}{1+\exp(X_i\beta)} + \log\frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}] ,$$

where, if you use for example Stata, you can use the standard logit algorithm, but you code the data as a panel with (for each i), $T_i - 1$ 0's to indicate "non-death" and a 1 at the end to indicate the event happened. For censored observations, you have all 0's for unit i.