Duration Analysis & Count Data[1]

Tyler Ransom

Duke University

August 6, 2012

[1]Based on Wooldridge (2003) Chapters 19-20

Introduction

- Today we'll discuss how to estimate a model where the dependent variable takes on two irregular types: a time interval, and a positive integer
- When the dependent variable is a time interval, we use duration analysis to recover the parameters of interest
- When the dependent variable is a positive integer, we use count data methods to estimate the parameters

When y is a time interval

When our dependent variable y is a time interval, we use duration analysis to estimate the model, which helps us understand the cause(s) of the dependent variable.

- Examples include:
- weeks unemployed until finding a job
- months married until divorce
- days until arrest after incarceration
- survey rounds until attrition
- weeks pregnant until delivery

Background

- For now, we assume that time is continuous
- The dependent variable is known as the "**initial state**" (e.g. marriage is the "initial state" in the divorce example. Someone exits the initial state when they become divorced)
- We also assume that there is no re-entry into the initial state
  - Therefore, we only consider spells until the first exit
  - Thus, y would be comprised of first marriages (for the divorce example) or first-time offenders (in the arrest example)
- For now, we also assume there are no covariates, i.e. we only look at unconditional time to exit

Hazard function

In order to better understand the econometrics of duration models, let's introduce some notation:

- Let T be a random variable with outcomes t and continuous distribution F with support $(0, \infty)$
- $F(t) = \Pr(T \leq t), \ t \geq 0$
- The **survivor function** is the probability of surviving past t and is denoted $1 - F(t)$
- The **hazard function** $\lambda(t)$ is the instantaneous rate of exiting the initial state per unit of time and is defined by

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

Hazard function

Some examples of the hazard function:

- If T is unemployment length (in weeks) then $\lambda(20)$ is approximately the probability of accepting a job offer between weeks 20 and 21 (conditional on not having accepted a job before then)
- If T is length of first marriage (in months) then $\lambda(84)$ is the approximate probability of becoming divorced between months 84 and 85

Hazard function

- You can use math to crank out the following using the definition of the hazard function given previously:

$$\lambda(t) = \frac{f(t)}{1 - F(t)}.$$

- The simplest form of a hazard function is $\lambda(t) = \gamma$ (some constant)
- Thus we have

$$\begin{aligned} \gamma &= \frac{f(t)}{1 - F(t)} \\ f(t) &= \gamma(1 - F(t)) \end{aligned}$$

which is an ordinary differential equation. The solution is

$$F(t) = 1 - e^{-\gamma t}$$

which is the CDF of the exponential distribution.

Duration dependence

- The exponential distribution has the property of **memorylessness**
- This means there is no **duration dependence** (i.e. the probability of exiting the initial state is the same regardless of how long you've been in the initial state)
- In practice, this is usually not a realistic assumption
- We usually want distributions that exhibit duration dependence, meaning $\lambda(t)$ is actually a function of t (not a constant)
- Common distributions that exhibit duration dependence but aren't too hairy to work with:
  - Weibull distribution
  - Log-logistic distribution

Introducing covariates

If we want to study the hazard rate conditional on a matrix of covariates X, for X's that do not vary over time (i.e. are measured at the start of the spell):

$$\lambda(t; X) = \frac{f(t \mid X)}{1 - F(t \mid X)}$$

- This leads us to the **proportional hazards model**:

$$\lambda(t; X) = \kappa(X)\,\lambda_0(t)$$

where we assume that the hazard function is separable between X and t. In the case, we call λ_0 the **baseline hazard**

Functional form

- In a proportional hazards model, it is common to assume

$$\kappa(\cdot) = e^{X\beta}$$

- Plugging this back into the conditional hazard function gives us

$$\lambda(t; X) = \lambda_0(t) \exp(X\beta) \tag{1}$$

- Simplifying, we have

$$\ln(\lambda(t; X)) = X\beta + \ln(\lambda_0(t))$$

- Remember, X cannot be time-varying!

Time-varying covariates

- We can introduce time-varying covariates $X(t)$, but we need to be a lot more careful
- In particular, the X's must be strictly exogenous (meaning $X(t)$ needs to have a well defined path once an agent leaves the initial state)
- If these stricter assumptions are met, we can write the proportional hazard as

$$\lambda(t; X(t)) = \kappa(X(t))\,\lambda_0(t)$$

with

$$\kappa(\cdot) = \exp(X(t)\,\beta)$$

- If we assume $T \sim$log-logistic then

$$\lambda(t; X(t)) = \frac{\alpha t^{\alpha-1} \exp(X(t)\,\beta)}{1 + t^{\alpha} \exp(X(t)\,\beta)}$$

and we can estimate α, β by MLE

Estimation

Estimation of duration data depends on the sampling scheme. Let's suppose T is the duration of unemployment for a male in the year 1998. To get data, we can sample in one of two ways:

1. flow sampling: Randomly sample men who became unemployed in 1998
2. stock sampling: Randomly sample men who were unemployed in the last week of 1998

- In both cases, we need to correct for right censoring, because there will be people who are still in the initial state at the time of sampling
- In the case of stock sampling, we also need to correct for left censoring
- Unobserved heterogeneity is much more difficult to correct for in duration analysis, so we won't cover it today

Estimation of expected duration

Suppose we weren't interested in estimating the hazard function, just the effect of X on the expected duration $\mathbb{E}[t]$.

- If there were no censoring, we could just run OLS

$$\ln(t^*) = X\delta + e$$

  where t^* is a vector of actual duration times for staying in the initial state

- If we assume $T \sim$Weibull, we can still estimate the expected duration using OLS, but

$$\delta_j = \frac{-\beta_j}{\alpha}$$

  where δ_j is the effect of X_j (the jth column of X) on expected duration and α is a parameter of the Weibull distribution

- If there is right censoring, need to do standard tobit to estimate δ

Estimation of the hazard function

When estimating parameters of the hazard function (e.g. equation 1) with flow sampling, need to do maximum likelihood and pay attention to truncation. The likelihood function looks like

$$\begin{aligned}\mathcal{L}\left(t_i; X_i, \theta\right) &= \prod_i f\left(t_i \mid X_i; \theta\right)^{d_i} \left[1 - F\left(t_i \mid X_i; \theta\right)\right]^{1-d_i}, \\ & \\ & d_i = 1 \left[\text{uncensored}\right]\end{aligned}$$

Likelihood for stock sampling

- With stock sampling, we need to introduce some more notation:
- a_i is the starting time for person i
- b is the maximum length of the spell (e.g. 52 weeks for the "unemployed in year 1998" example)
- The likelihood function is then

$$\mathcal{L}\left(t_i; X_i, \theta\right) = \prod_i \frac{f\left(t_i \mid X_i; \theta\right)^{d_i} \left[1 - F\left(t_i \mid X_i; \theta\right)\right]^{1-d_i}}{1 - F\left(b - a_i \mid X_i; \theta\right)}$$

Proportional hazard models in Matlab

- The Matlab command for estimating a Cox proportional hazard model (the most common type of duration analysis) is `coxphfit`
- The full syntax is
- `[b,logl,H,stats] = coxphfit(X,y,'name',value)`
- `H` is a two-column matrix that contains y values in the first column and cumulative hazard values in the second
- `'name'` is how you invoke optimization options

Grouped data

The assumption that time is continuous can be a hard pill to swallow, especially when the data is measured on a weekly or monthly basis. For example:

- We only have weekly data on arrests
- but arrests could have happened any time in that week
- We need to account for this discreteness in our estimation

Estimation of grouped data

We observe y_m, c_m, X where m indicates the time range (e.g. $y_m = 1$ [exited initial state in week m of 1998], $c_m = 1$ [censored in week m], and X is a matrix of time-invariant covariates

- If we assume a **piecewise-constant proportional hazard**, then

$$\lambda(t; X, \theta) = \kappa(X\beta)\,\lambda_m,\ a_{m-1} \leq t < a_m$$

and we get the likelihood function

$$\mathcal{L}(X; \theta) = \prod_{i=1}^{N} \left[\prod_{h=1}^{m_i - 1} \alpha_h(X_i; \theta) \right] [1 - \alpha_{m_i}(X_i; \theta)]^{d_i}$$

where

$$\alpha_m = \lambda_m (a_m - a_{m-1}) \exp(-\exp(X\beta))$$

and a_m is a time period (e.g. one week); d_i is the same as before

Kaplan-Meier estimator

One of the most famous estimators in duration analysis (used frequently in the finance literature) is called the **Kaplan-Meier estimator of the survivor function**.

- No covariates
- Want to estimate λ_m with grouped data

$$\hat{s}\left(a_m\right) = \prod_{r=1}^{m} \frac{N_r - E_r}{N_r}$$

- where N_r is the number of people who have neither left the initial state nor been censored at a_{r-1}
  E_r is the number of people leaving the initial state between a_{r-1} and a_r

Time-varying covariates in grouped data

If we want to estimate grouped data with time-varying covariates, then the covariates X_m must not change within each a_m

- If this is the case, then we can rewrite the likelihood function from before, but add an m subscript to the X matrix:

$$\mathcal{L}(X;\theta) = \mathcal{L}(X;\theta) = \prod_{i=1}^{N} \left[\prod_{h=1}^{m_i-1} \alpha_h (X_{im};\theta) \right] [1 - \alpha_{m_i} (X_{im};\theta)]^{d_i}$$

where

$$\alpha_m = \lambda_m (a_m - a_{m-1}) \exp(-\exp(X_m \beta))$$

Competing risks models

If we want to generalize our duration analysis to having more than one exit state, this is called a **competing risks model**

- In our employment example, we considered employment as the "exit state"
- We could consider a model in which the individual can move either from unemployment to employment, or from unemployment to out of the labor force
- In a competing risks model, each of the "risks" is assumed to be independent of one another
- The econometrics get a lot more hairy, so we won't cover it today

Models with count data

Count data is a set of data where the dependent variable takes on values that are nonnegative integers. In this case, y is called a **count variable**. Examples include

- Number of cigarettes smoked in a day
- Number of plane crashes experienced by an airline over a certain period of time
- Number of days spent in a hospital last year
- Number of times visited a recreational site

Count data

Just like with binary models (where y only takes on values in $\{0, 1\}$), we need to estimate the model differently

- The main reason is that if we performed OLS on $y = X\beta + \varepsilon$, we would find that $\hat{y}$ would be negative for some values of X
  - This is the same reason we do a logit or probit model for binary data
- In order to ensure that $\hat{y}$ is always positive, we make our favorite transformation: $\exp(X\beta)$
- We also need to assume a better distribution for ε

Poisson regression

Let's assume $y|X \sim$Poisson with mean parameter $\mu = \exp(X\beta)$

- The pdf of this distribution is

$$f(y|X) = \frac{\exp(-\mu(X\beta))[\mu(X\beta)]^y}{y!}$$

- Our quasi MLE (QMLE) likelihood function is then

$$\ell = \sum_i y_i X_i \beta - \exp(X_i\beta) - \ln(y!)$$

Interpretation

Interpretation of Poisson regression coefficients is unlike typical regression models:

$$\beta_j = \frac{\partial \ln\left(\mathbb{E}\left[y|X\right]\right)}{\partial X_j}$$

Turning this into a differential and rearranging terms, we get

$$\Delta\mathbb{E}\left[y|X\right] \approx 100\beta_j \Delta X_j$$

Thus β_j is a semielasticity

Other count distributions

Besides the Poisson distribution, we can also use the following distributions to fit our count data:

- Negative binomial
- Binomial
- Exponential
- Fractional logit

See Wooldridge for more information regarding these. Also be aware that inference is slightly more complicated with count models.

Conclusion

- Today we discussed how to estimate models where the dependent variable takes a specific form that is different than what we're used to
- You will likely come across duration analysis in the future
- You may not come across count data, but if you do, you know where to look for more information