"Further Modeling Issues in Event History Analysis

by

Robert E. Wright

University of Strathclyde, CEPR-London, IZA-Bonn and Scotecon

Topics

1. Duration dependence
2. Competing risks and Repeatable Events
3. Time-varying covariates
4. Unobserved heterogeneity

Duration Methods

- Event history analysis
- Duration econometrics
- Failure time analysis
- Hazard modeling
- Survival analysis

Transition is the key concept

t_1 ---→ t_2

e.g. $t_2 - t_1 = 12$ *months*

Two possible outcomes at t_1 and t_2:

Poor at t_1*?: Yes or No*

Poor at t_2*?: Yes or No*

Transition Matrix

		t_2	
	<u>Poor?</u>	**Yes**	**No**
t_1	**Yes**	*Poor*	*Exits poverty*
	No	*Enters poverty*	*Not poor*

Censoring

- Left censoring
- Right censoring

Basic Concepts

1) Hazard Rate

$$h(t) = \lim_{\Delta t \to 0} \frac{Prob\ (t,\ t + \Delta t)}{\Delta t}$$

2) Hazard Model

$$h(t) = f(\boldsymbol{X}, \theta, t)$$

3) Standard Proportional Hazard Model

$$\ln h(t) = \boldsymbol{\beta X} + \alpha(t)$$

Proportionality assumption

- Rarely tested for formally
- Key assumption of the model
- Only the case for fixed covariates
- Needs to be taken seriously

Estimation

- Maximum likelihood
- Partial likelihood ("Cox's model")

1. Duration Dependence

- Simply how the underlying hazard rate varies with duration (time)
- Ignored by sociologists!
- Fixated on by economists!
- Cox's model "assumes away" duration dependence.

$$\ln h(t) = \boldsymbol{\beta X} + \alpha(t)$$

What is the nature or shape of *α(t)*?

- Does it increase or decrease with time?
- Does it remain constant over time?
- Does it increase and then decrease over time?
- Does it decrease and then increase over time?

Lots of duration dependence specifications

- Exponential
- Linear
- Weibull
- Polynomial
- Log-normal
- Box-Cox
- Piecewise-constant

Exponential: $\alpha(t) = \alpha_0$

Log linear: $\alpha(t) = \alpha_0 + \alpha_1 t$

Weibull: $\alpha(t) = \alpha_0 + \alpha_1 \ln t$

Polynomials:

Quadratic: $\alpha(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$

Cubic: $\alpha(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3$

Order k: $\alpha(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \ldots + \alpha_k t^k$

Log-normal:

$$\alpha(t) = \alpha_0 + (2\pi)^{-½}\alpha_1{}^{t-1} exp\{[-\alpha_1(ln\alpha_2 t)^2/2]\}$$

Box-Cox (Heckman and Singer variant):

$$\alpha(t) = \alpha_0 + \alpha_1 exp(\alpha_2 t^{\alpha_3} + \alpha_4 t^{\alpha_5})$$

Various other shapes are nested within this specification:

e.g. $\alpha_1 = \alpha_3 = 1$ *and* $\alpha_5 = 2$ *then:*

$$\alpha(t) = \alpha_0 + \exp(\alpha_2 t + \alpha_4 t^2) \qquad \textit{Quadratic}$$

e.g. $\alpha_1 = \alpha_3 = 1$ *and* $\alpha_4 = 0$ *then:*

$$\alpha(t) = \alpha_0 + \exp(\alpha_2 t) \qquad \textit{Monotonic}$$

Piece-wise constant

$$\alpha(t) = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_4 D_4$$

	D_1 = *0-1 months*
Duration	D_2 = *2-4 months*
Dummies	D_3 = *5 to 7 months*
	D_4 = *8 to 12 months*

Competing Risks and Repeatable Events

- Agents are at risk to more than one event occurring at any point in time
- Many events are repeatable
- Complicates estimation of models from both theoretical and statistical points of view

Competing Risks

Example of competing risks when modeling divorce

- Population exposed to risk of divorce are married individuals
- A married individual can divorce and therefore experience the event of interest
- However a married individual can die and therefore is no longer at risk to divorce
- Both divorce and death are "competing" events since any individual is at risk to both events happening

At t_1 **At t_2**

Married

1) Married
2) Divorced
3) Dead

Time ----------------------------------→

Example of competing risks when modeling unemployment

When individuals "exit" unemployment they can enter full-time employment; part-time employment; or leave the labour force altogether

At t_1 **At t_2**

Unemployed	*1) Unemployed*
	2) Employed full-time
	3) Employed part-time
	4) Out of the labour force

Time --→

- To a certain extent modeling competing risks are not a problem
- In the case of the divorce example you essentially fit two models by assuming the events are independent.

Model 1: "The divorce model"

Event indicator:	(0) Married or dead
	(1) Divorced

- If you die you are no longer exposed to the risk of divorce. If you die you are simply "right-censored" at the time of death

Model 2: “The death model”

Event indicator: (0) Married or divorced
(1) Dead

- If you divorce you are no longer exposed to the risk of death while married. If you divorce you are simply “right-censored” at the time of divorce

Three models for the unemployment example

Model 1: Entry to full-time employment

Model 2: Entry to part-time employment

Model 3: Entry to “out of the labour force”

What happens if the events are not independent?

Answer: You potentially have a big problem!

- There are no quick-fix solutions for this problem
- Essentially you have a problem of unobserved heterogeneity which I will discuss later
- There are models that attempt to deal with this problem statistically but these models are well beyond the scope of this current lecture

Example

- What happens if “sickly” people have both a higher risk of divorce and higher risk of death
- Therefore the events of divorce and death are not independent
- Control for health status (i.e. include measures of health status as covariates)

Repeatable Events

Many events are repeatable:

e.g. One can marry, divorce, remarry, and divorce again

e.g. One can exit and enter unemployment on a regular basis (“labour-market churning”)

Many events are not repeatable:

e.g. Death only occurs once!

e.g. Once you marry you cannot re-enter the “never-married” state

The basic issue here is that there is a likely a link between an event occurring in the past and the probability that it will occur in the future.

- e.g. If you marry once you might have a higher (lower) probability of remarrying in the future
- e.g. If you divorce once you might have a higher (lower) probably of divorcing in the future

This is the notion of “state dependence”

What to do:

1. Nothing—just pool all the information together and fit standard models treating each individual's experience as an independent observation
2. Include in the models variables that capture aspects of the individual's prior experience

   -The number of prior events experienced

   -The length of time spent in prior states
3. Statistical twists

Modeling divorce

- Include the number of times the individual has been previously divorced and length of time the individual was previously married

Modeling unemployment

- Include the number of times the individual has been previously unemployed and length of time the individual has been previously unemployed

Time-varying covariates

Essentially three types of covariates:

1. Fixed covariates
2. Time-dependent covariates
3. Time-varying covariates

People tend to confuse (2) and (3)—they are NOT the same

Consider the basic exponential proportional hazards model:

$$lnh(t) = \boldsymbol{\beta X} + \alpha_0$$

The **Xs** are all measured <u>before</u> the individual is exposed to the risk of the event of interest occurring. These are **FIXED COVARIATES**

$$lnh(t) = \boldsymbol{\beta X} + \boldsymbol{\gamma X_t} + \alpha_0$$

The $\mathbf{X_t}$**s** are all measured <u>after</u> the individual is exposed to the risk of the event of interest occurring. The values of these variables change over-time. This variables are **TIME-VARYING COVARIATES**

$$lnh(t) = \boldsymbol{\beta X} + \boldsymbol{\gamma X_t} + \boldsymbol{\delta f(Z,t)} + \alpha_0$$

The **Zs** are all measured <u>before</u> the individual is exposed to the risk of the event of interest occurring. However, the effect of these variables, unlike fixed covariates, are specified to have a differential impact as time proceeds. These are **TIME-DEPENDENT COVARIATES.** The researcher picks the function "*f*"

4. Unobserved heterogeneity

Perfect Specification Assumption

$$lnh(t) = \boldsymbol{\beta X} + \alpha(t)$$

- The included **X**s capture all the non-random variation in the hazard rate
- Seems unlikely! Specification bias.

If there is serious unobserved or residual heterogeneity then:

- Parameters estimates will be biased

  and/or

- Incorrect pattern of duration dependence will be observed (bias towards negative duration dependence)

- There are **NO** solutions to the problem of unobserved heterogeneity. It is a **DATA** problem
- However, there are things you can do that at their very best can help control for the degrading effects of unobserved heterogeneity

Nature of the problem

Standard regression model:

$Y_i = a + bX_i + e_i$

Standard panel regression model:

$Y_{it} = a + bX_{it} + e_{it}$

$Y_{it} = a + bX_{it} + e_{it}$

Decompose e_{it} into two components:

$e_{it} = \theta_i + \varepsilon_{it}$

Get the standard one-way fixed effects model:

$Y_{it} = a + bX_{it} + \theta_i + \varepsilon_{it}$

$$lnh(t)_i = \boldsymbol{\beta X}_i + \alpha(t) + \theta_i + \varepsilon_{it}$$

"Like" estimating a unbalanced fixed or random effects panel model

Need some assumptions:

- ε_{it} is random
- θ_i is uncorrelated with $\mathbf{X}_i$ and ε_{it}
- θ_i is normally distributed

Examples of theoretical relevant unobserved heterogeneity

- The notion of “frailty” in mortality research
- The idea of “ability bias” in labour economics research

SABRE can be used to estimate this model in discrete time

Most research into developing hazard models that include unobserved heterogeneity essentially try to relax these assumptions

“The **Identification Problem** in proportional hazards models”