Technology, skills, and job duration

Hugo Castro Silva

Francisco Lima

CEG-IST, Universidade de Lisboa

Objective

Understand how technology and skills relate to job duration

As the title suggests, the objective of this work is to study the relationship between technology and job duration.

But because this is a Stata meeting…

Discrete-time Duration Models

My main aim today is to show you how we used discrete-time duration models with Stata in our analysis.

Duration Models

Time to an event of interest, like death or job separation

Continuous-time models

Discrete-time models

Duration models are used to model the time to some event of interest, like death, failure or job separation.

We have to types of models:

Those that treat time as continuous variable,

nd those that treat time as if it was discrete.

Continuous

Stata has *st* suite

streg, stcox …

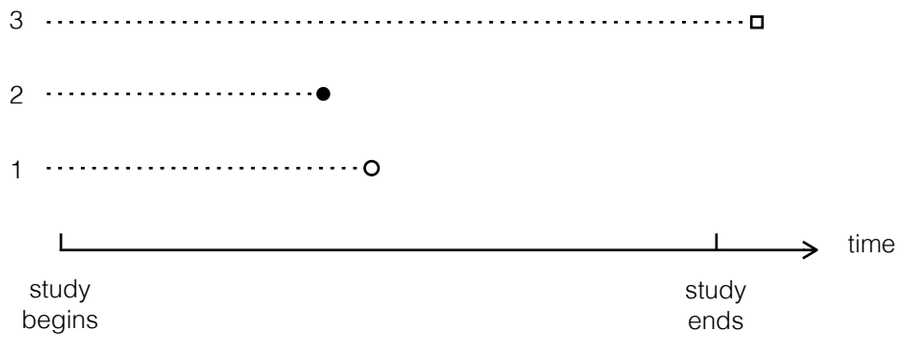
Very simple to use and popular

Distributions of time: exponential, Weibull, “Cox”, etc.

Stata has several commands for continuous-time models in the st suite which are very handy and easy to use.

Maybe because of that simplicity, nowadays we see them often in the economics literature.

These commands offer many choices of continuous distributions to represent time, including semi-parametrical Cox model.



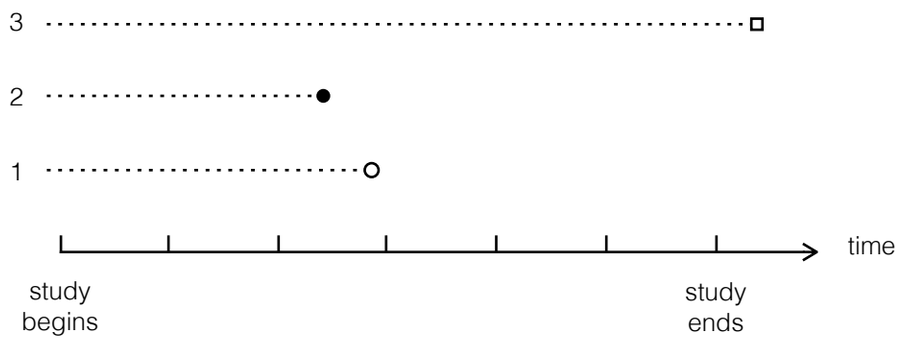
To better understand why the choice between continuous and discrete is important, let’s consider this diagram.

In a continuous model, we assume that time is measured as a continuous variable and with infinite precision.

We know that subject 1 fails precisely at that instant.

Subject 2 fails a little earlier.

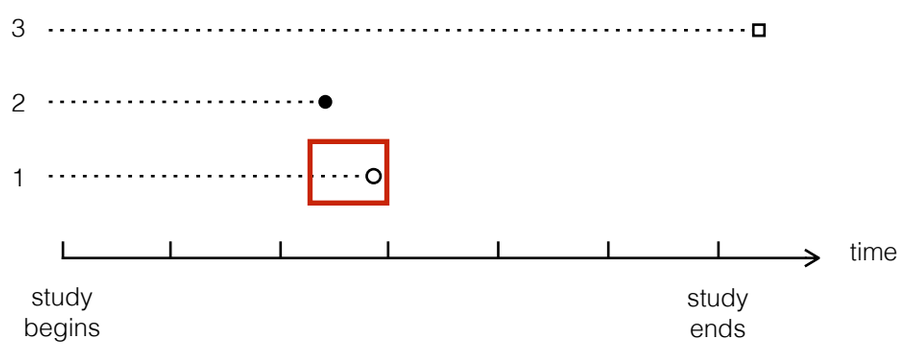
And subject 3 fails some time after the study ends.



But often in the social sciences, time is registered as a discrete quantity.

because the data are usually collected as cross sections or panel data with some time interval.

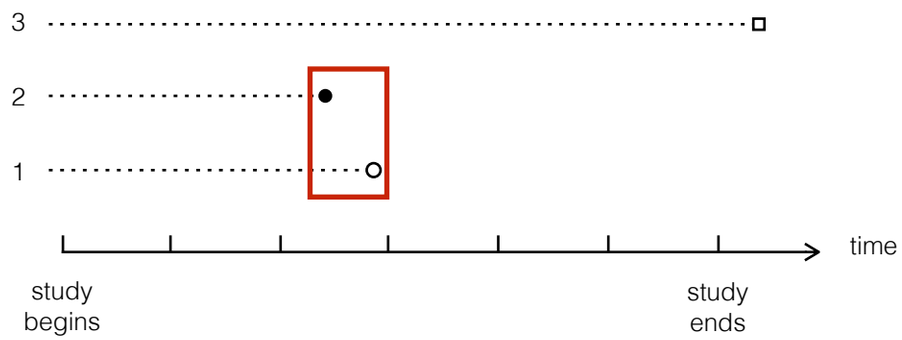
When this happens, we have interval-censoring.



Interval-censoring leads to a loss of precision.

For example, we know that subject 1 had a job in period 2, but by the end of period 3 she was no longer working. We don’t know exactly when the separation happened, just that it happened sometime between periods 2 and 3.

Interval-censoring



Loss of precision

Tied failure times

Another issue is that, because subject 2 also lost his job sometime between periods 2 and 3, we can no longer say who failed first. We have tied failure times.

And this is where the discrete models come in.

Discrete

Rule of thumb: if *interval/(typical spell)* is large

Models can be estimated with maximum likelihood

h(t) = Pr(event in t | time > t-1)

Naturally, all data are collected in a discrete way because we don’t have infinite precision.

To decide between using continuous or discrete, there is a sort of rule of thumb. The larger the ratio of length of the interval to the typical duration, the more appropriate it will be to use discrete models.

With some reorganization to the data, these models can be easily estimated as binary dependent variable models using maximum likelihood estimation

Typically these models represent the hazard of an event as a function of time, which is the probability of the event happening in time t, given that no event happened before

Discrete

Some commands in Stata:

cloglog - no unobserved heterogeneity (frailty)

xtcloglog - Normal

pgmhaz8 - Gamma (Jenkins 2004; Meyer 1990)

hshaz - Non-parametric (Jenkins 2005; Heckman and Singer 1984)

Stata has several commands for this

Examples of commands for proportional hazards models:

Complementary log-log model which does not account for unobserved heterogeneity

Complementary log-log model with Normally distributed unobserved heterogeneity.

cloglog with Gamma distributed frailty.

cloglog with non-parametric frailty.

This is another advantage of discrete models versus continuous.
In Stata's continuous commands, frailty for each subject can be used but, at least in version 13, is restricted to a relatively small number of subjects.

Another family of duration models are the proportional odds models. These can be estimated using logit or xtlogit with random effects.

Proportional hazards are popular in the literature but there is nothing in the literature suggesting that the PH assumption is more reasonable than other models in economics applications than

Discrete

Steps for easy estimation (Allison 1984, Jenkins 1995)

1. **Data organization**: one observation per person-period
2. **Create variables**: interval identifier; censoring; duration dependence
3. **Choose model and estimate**

Jenkins proposed three steps for easy estimation of discrete duration models

The first step is to organize the data set so that, for each subject, we have one observation for every period she is at risk.
This may result in a very large data set.

In the second step we have to create an interval identifier variable, which is a sequence of integers indicating: this is the first period, this is the second period, etc…

Then we need a variable that will be 1 in the period when the event of interest takes place, and zero in the remaining periods. This will be our dependent variable.

And then we must decide on the functional form of the duration dependence. That is to say: how are we going to represent time in our model?

Finally we will choose which model is appropriate to our question, based on the literature and other considerations

Application

I think an example makes everything easier to understand so

Let’s see how these steps take place in our work

Application

Understand how technology and skills relate to job duration

Using discrete-time duration models in Stata

Remember, we want to find out how technology and skills relate to job duration and job separation

Using discrete-time duration models

Step 1: Data Organization

‘Quadros de Pessoal’ - yearly matched worker-firm panel, covering 1995-2007

Rule of thumb: *interval/(typical spell)* = 1/3

Naturally organized: one observation per worker-year

Just one spell per worker

We use Quadros de Pessoal, which is a very detailed Portuguese data set that matches workers and firms.

Info on human capital and date of admission, as well as firm characteristics

The interval duration is 1 year, whereas the typical job duration in our sample is around 3 years. That means the ratio is relatively large and so using discrete models is recommended.

Luckily for us, this is a yearly panel and so it is already in the required form of one observation for each person-period

And we restrict ourselves to only one spell per worker as some of the commands are restricted to single spell data

Worker	Year	Firm	Admission year
1	2001	1000	2001
1	2002	1000	2001
1	2003	1000	2001
2	2006	3526	2006
2	2007	3526	2006

In this table, we have an example of how our data initially looks like.

We have one observation for every year a worker is in a firm and we know when each spell starts.

Step 2: Create variables

Interval id: tenure = current year - admission year + 1

Censoring: Job separation

0 if same firm next year, 1 otherwise

Duration dependence: which form?

- tenure + tenure^2 ?
- log(tenure) ?
- dummy variables for each value of tenure

In the second step, we need to create some variables

The interval identification variable is simple: it's the same as the worker's tenure

Our censoring variable is 1 if there is a job separation next year

Which form should the duration dependence take?

If there is no indication in the literature or the theory, some of the most common are:
a polynomial of tenure
log of tenure

but both of these impose a restrictive, somewhat monotonic form to duration dependence

a more flexible alternative is to use dummy variables for each period. It allows the function to move freely.

This is the functional form that we will use.

Worker	Year	Firm	Admission year	Tenure	Job separation	Tenure 1	Tenure 2	Tenure …
1	2001	1000	2001	1	0	**1**	0	0
1	2002	1000	2001	2	0	0	**1**	0
1	2003	1000	2001	3	**1**	0	0	**1**
2	2006	3526	2006	1	0	**1**	0	0
2	2007	3526	2006	2	0	0	**1**	0

Starting with worker 1: because he is not in the same firm in 2004, we say there was a job separation, and so the variable will have the value 1.

Worker 2, however, did not go through a job separation during the study period.

Also look at the dummy variables for tenure

Step 3: Choose model

Proportional hazards model with frailty

- Normal?
- Heckman and Singer?
- Gamma (Abbring and van den Berg 2007)

Flexible duration dependence mitigates bias of wrong frailty (Dolton and van der Klaauw 1995)

Our choice falls on a proportional hazards model, with frailty

But which form should the frailty take?

Normal is always a nice distribution, but this model requires numerical methods for integration and so might take a long long time

Hickman and Singer makes no assumptions made about the form of the distribution. Powerful, but often harder to get results because it might not converge

The Gamma is easier/faster to estimate because there is a closed form for the integration

The literature has shown that often the distribution of the frailty converges to a Gamma distribution

Additionally, because we are using the dummy variables for time dependence, we mitigate any bias arising from a wrong choice of frailty.

So, if you can afford it, you could try using dummy variables

And truly, we estimated all three types of models for robustness sake, and the results are very similar

Model

Analyze role of technology through interactions:

1. i.education##(i.high-tech i.medium-tech)

2. i.skill_level##(i.high-tech i.medium-tech)

Finally, some specificities of our own analysis:

To understand the role of technology we have two different models:

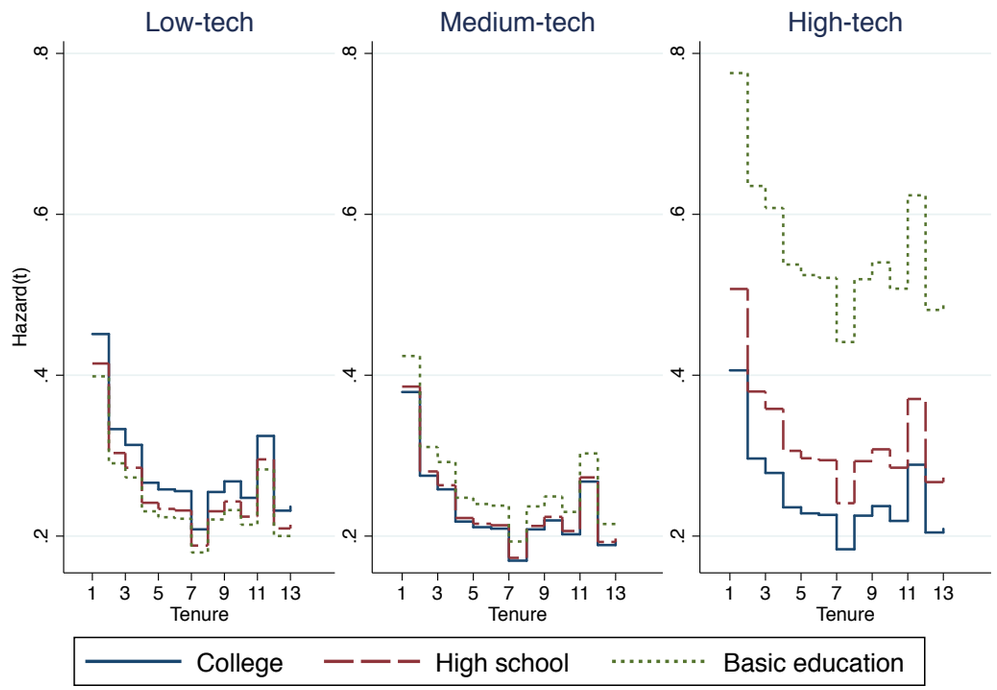
in the first one we interact technology dummy variables with education dummies

in the second model we have interactions between technology and dummies of skills

Results

Let’s see some results then, in the form of graphs

Education

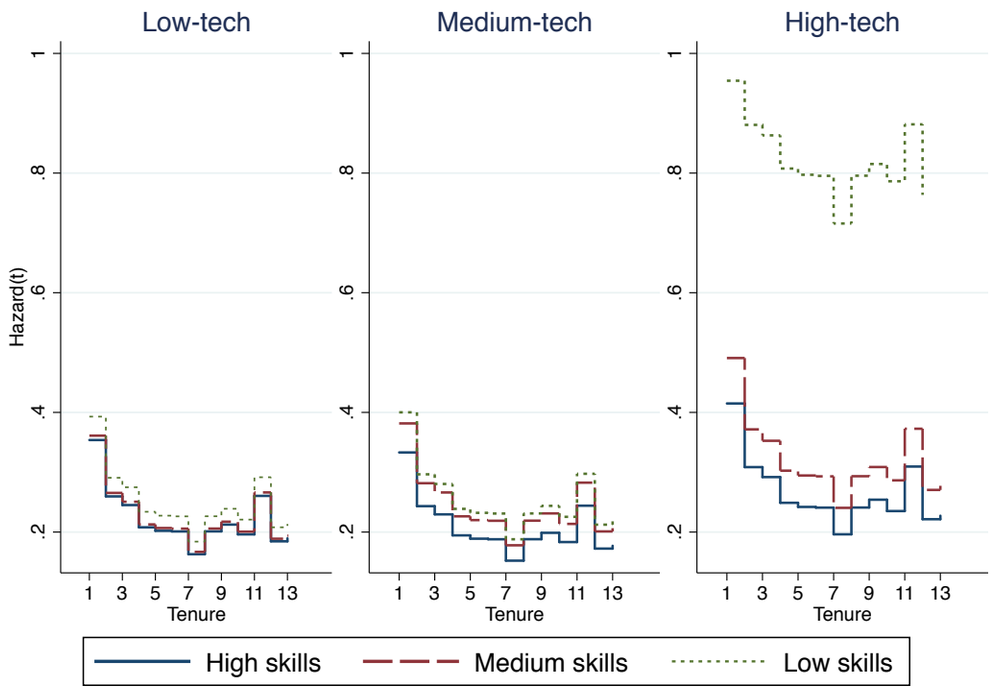


First we see how the hazard of job separation decreases with tenure.

Workers with college education face more or less the same hazards regardless of the level of technology

However, as technology intensity increases, less-educated workers face higher hazards.

Skills



as skills increase, the hazard decreases in every group

But again, we see that while the high skilled workers experience the same level of hazards, low-skilled workers have higher hazards in sectors with more technology intensity

Conclusions

In conclusion

Conclusions

Choice between continuous and discrete

pgmhaz8 and *hshaz* do not support multiple spells, factor variables or *margins*

Stata 14 has new and improved duration models

The choice between continuous and discrete models can be very important

The commands written by Jenkins, while practical, do not support multiple spells and are outdated: do not support factor variables, margins

However, xtcloglog which is part of Stata supports all of these

Stata 14 introduced multilevel and panel-data survival duration models which I'm very curious to try.

Thank you for your attention.