TA Session 5: Duration Models

Microeconometrics with Hanna Wang
IDEA, Fall 2022

TA: Conghan Zheng

Overview

1. Duration Data

2. Continuous Duration
   - Nonparametric Approach
   - Semiparametric Approach
   - Parametric Approach

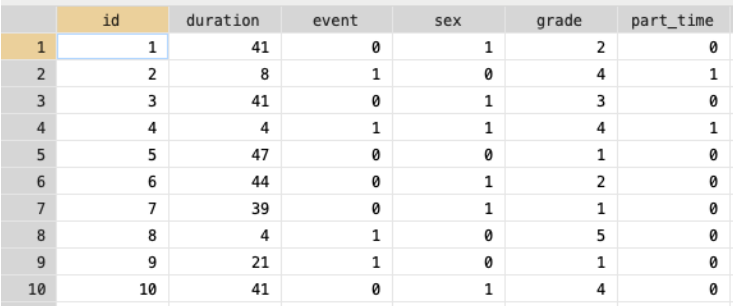
3. Discrete Duration

4. Appendix

Duration Data

Duration Data

- `TA5_1.dta`: college dropouts data



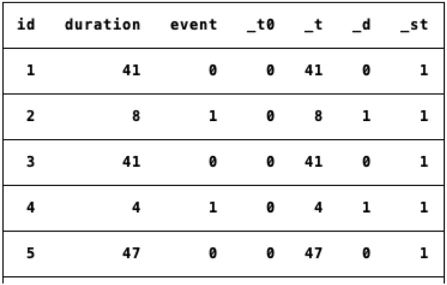
	id	duration	event	sex	grade	part_time
1	1	41	0	1	2	0
2	2	8	1	0	4	1
3	3	41	0	1	3	0
4	4	4	1	1	4	1
5	5	47	0	0	1	0
6	6	44	0	1	2	0
7	7	39	0	1	1	0
8	8	4	1	0	5	0
9	9	21	1	0	1	0
10	10	41	0	1	4	0

- `event`: the event of interest, 1 = dropout, 0 = censored

Duration Data

- Set the duration data structure based on variable duration.

```
.  stset duration, failure(event) id(id)
```



id	duration	event	_t0	_t	_d	_st
1	41	0	0	41	0	1
2	8	1	0	8	1	1
3	41	0	0	41	0	1
4	4	1	0	4	1	1
5	47	0	0	47	0	1

- Variables newly generated by the command:

  _t0: onset of risk (the calendar time could be different for different individuals)

  _t: analysis time when record ends

  _d: 1 if we observe the complete spell, 0 if the spell is censored

  _st: 1 if the record is to be included in analysis; 0 otherwise

Continuous Duration vs. Discrete Duration

The core: grouping.

- Continuously distributed durations
  - Time index is sill 'discrete', you have natural numbers $t = 1, 2, ...$, not something like $t = 1.4142$. Continuous means time is in its fairly precise unit, consecutively observed, not grouped.
- Discretely distributed durations: grouped data
  - When the meansurements are in aggregated time intervals, it can be important to account for the discreteness in the estimation.
  - In grouped duration data, each duration is only known to fall into a certain time interval, such as a week, a month, or even a year.
  - Why we can't address this discreteness using the continuous duration model: explained later in section Discrete Duration.

Continuous Duration

Estimation approaches

1. **non-parametric**: letting the data speak for itself and making no assumption about the functional form of the survivor function, the effect of covariates are not modeled either.
2. **semi-parametric**: no parametric form of the survivor function is specified, yet the effect of the covariates is still assumed to take a certain form (to alter the baseline survivor function that for which all covariates are equal to zero). The Cox(1972) model is the most popular semiparametric model.
3. **fully parametric**: analogous to a Tobit model with right-censoring, has the limitation of heavy reliance on distributional assumptions (in order for parameter estimates to be consistent).

Censoring

- One important problem of survival data is that they are usually censored, as some spells are incompletely observed. In practice, data may be
    - **right-censoring/censoring from above**: we observe spells from time 0 until a censoring time c, the unknown end lies in (c, ∞).
    - **left-censoring/censoring from below**: the spells are incomplete with an unknown end lies in $(0, c)$. For example when we talk about unemployment spell, this individual ends unemployment before her entering the study.
    - **interval censoring**: the censored spell ends between two known time points $[t_1^*, t_2^*)$.
- The survival analysis literature has focused on right-censoring.

Nonparametric Approach

Assumption

- Each individual in the sample has a completed duration T_i^* and censoring time C_i^*. What we observe for each spell is the minimum of T_i^* and C_i^*.
- For standard survival analysis methods to be valid, the censoring mechanism needs to be one with **independent (noninformative) censoring**.
- This means that parameters of the distribution of C^* are not informative about the parameters of the distribution of the duration T^*.

Nonparametric approach

Nonparametric estimation of survival functions:

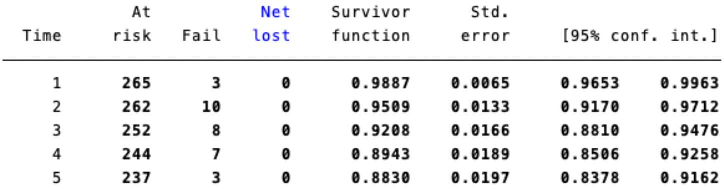
1. Estimate the survivor or hazard function in the presence of independent censoring.
2. No regressors are included.

Table 1: Key concepts of survival analysis

Function	Symbol	Definition	Relationship
Density	$f(t)$		$f(t) = \frac{dF(t)}{dt}$
Distribution	$F(t)$	$P(T \leq t)$	$F(t) = \int_0^t f(s)ds$
Survivor	$S(t)$	$P(T > t)$	$S(t) = 1 - F(t)$
Hazard	$h(t)$	$\lim_{h \to 0} \frac{P(t \leq T \leq t+h \mid T \geq t)}{h}$	$h(t) = \frac{f(t)}{S(t)}$
Cumulative hazard	$H(t)$	$H(t) = \int_0^t h(s)ds$	$H(t) = - \ln S(t)$

The Kaplan-Meier estimator

- `TA5_1.dta: college dropouts data`



```
                At             Net    Survivor      Std.
  Time        risk    Fail    lost    function     error     [95% conf. int.]
─────────────────────────────────────────────────────────────────────────────
     1         265       3       0      0.9887    0.0065     0.9653    0.9963
     2         262      10       0      0.9509    0.0133     0.9170    0.9712
     3         252       8       0      0.9208    0.0166     0.8810    0.9476
     4         244       7       0      0.8943    0.0189     0.8506    0.9258
     5         237       3       0      0.8830    0.0197     0.8378    0.9162
```

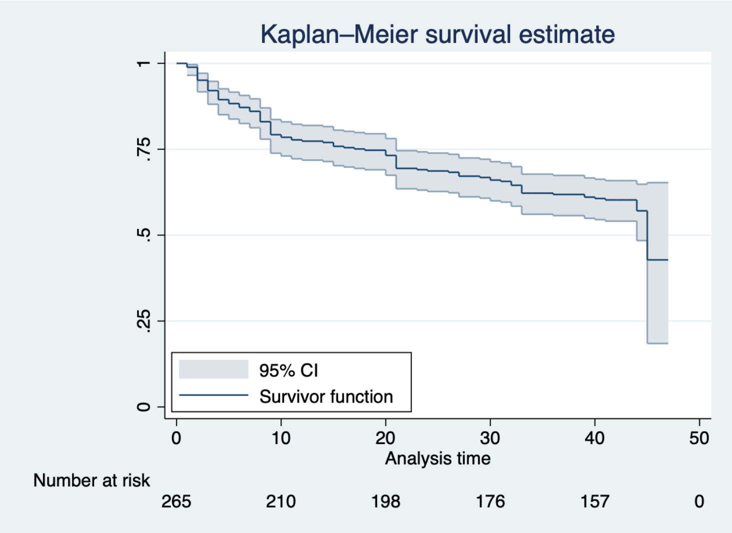
- `At risk`: # at school at t
- `Fail`: # dropped out at t
- `Net Lost`: # censored
- **Example**:

  The probability of survival beyond $t = 1$ is $\frac{262}{265} \approx 0.9887$.

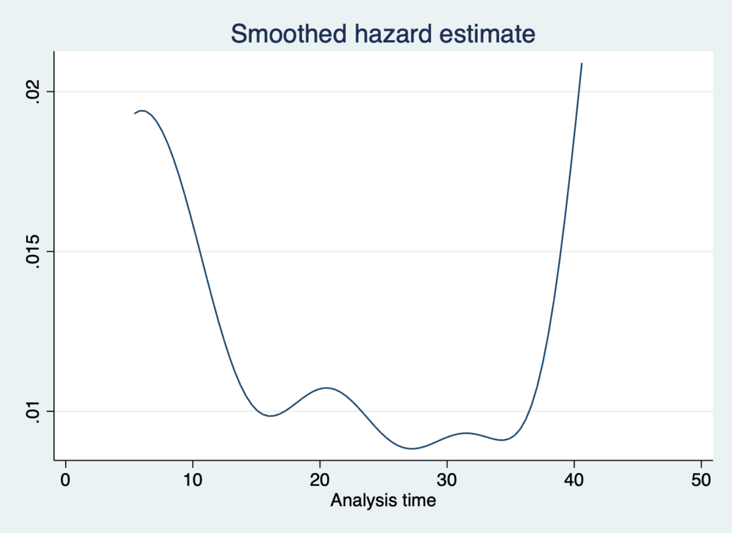
  The probability of survival beyond $t = 2$ is $\frac{262}{265} \times \frac{252}{262} = \frac{252}{265} \approx 0.9509$.

  ...

The Kaplan-Meier estimator



The Kaplan-Meier estimator



Semiparametric Approach

The Cox proportional hazards model

- To estimate the role of individual observed heterogeneity while controlling for duration dependence, we consider the Cox proportional hazards regression model (Cox, 1972):

  $$h(t|x) = h_0(t)e^{x\beta}$$

  The Cox model is semiparametric in the sense that $h_0(t)$ is estimated non-parametrically, and the scale up part $e^{x\beta}$ is assumed to be depending on regressors. The only parameter β to estimate in this model comes from the second part.
- The Cox model has no intercept since

  $$h_0(t)e^{\beta_0 + x\beta} = \underbrace{h_0(t)e^{\beta_0}}_{\text{new baseline hazard}} \quad e^{x\beta}$$

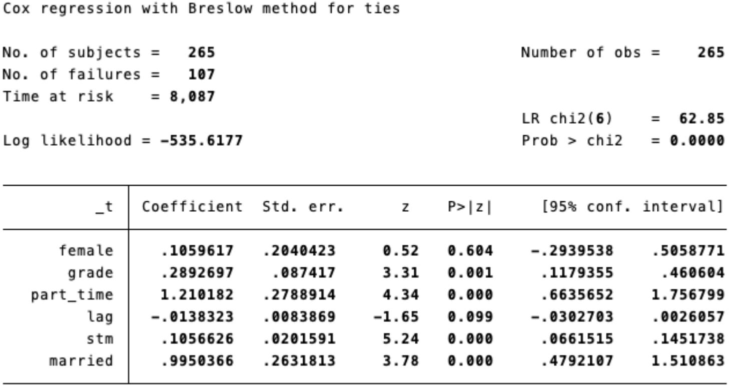
  Any intercept along with the regressors is not identified, since any value works as well as any other.

The Cox proportional hazards model

- Effects of regressors on the time until college dropout: the βs from $e^{x\beta}$

  `TA5_1.dta`: college dropouts data

  `.   stcox $x, nohr`



```
Cox regression with Breslow method for ties

No. of subjects =   265                              Number of obs =    265
No. of failures =   107
Time at risk    = 8,087
                                                     LR chi2(6)    =  62.85
Log likelihood = -535.6177                           Prob > chi2   = 0.0000

------------------------------------------------------------------------------
          _t | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   .1059617   .2040423     0.52   0.604    -.2939538    .5058771
       grade |   .2892697    .087417     3.31   0.001     .1179355     .460604
   part_time |   1.210182   .2788914     4.34   0.000     .6635652    1.756799
         lag |  -.0138323   .0083869    -1.65   0.099    -.0302703    .0026057
         stm |   .1056626   .0201591     5.24   0.000     .0661515    .1451738
     married |   .9950366   .2631813     3.78   0.000     .4792107    1.510863
------------------------------------------------------------------------------
```

The Cox proportional hazards model

- If the jth regressor in $x = (x_1, x_2, ..., x_k)$ is increased by 1 unit,

$$h(t|x+\Delta) = h_0(t)e^{\beta_\mathbf{1} x_\mathbf{1} + \ldots + \beta_j(x_j+1) + \ldots + \beta_k x_k} = h_0(t)e^{x\beta+\beta_j} = e^{\beta_j} h(t|x)$$

- Therefore, changes in regressors can be interpreted as having a multiplicative effect on the original hazard (semi-elasticity), as

$$\frac{\partial h(t|x)}{\partial x_j} = h_0(t)\frac{\partial e^{x\beta}}{\partial x_j} = h_0(t)e^{x\beta}\beta_j = h(t|x)\beta_j$$

- This is consistent with the noncalculus result as $e^x \simeq 1 + x$ (recall taylor series and equivalent infinitesimal).
- $\beta_{female} \approx 0.1060$, *harzard rate* is higher for female students. But the magnitude of the effect is not immediately clear.
- The effect size: *hazard ratio* for time-invariant variable sex is $e^{0.1059617} \approx 1.111779$,

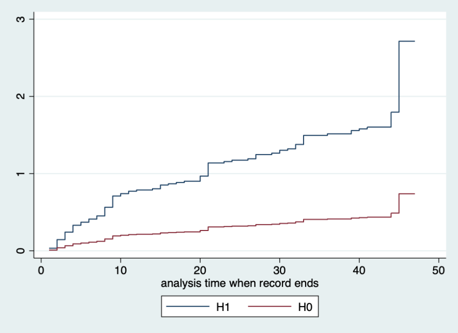
The Cox proportional hazards model

- The cumulative hazard:

$$H(t|x) = \int_0^t h(s|x)ds = e^{x\beta} \int_0^t h_0(s)ds = e^{x\beta} H_0(t)$$

- After including one binary regressor (part-time student) whose estimate is $\beta_1 \approx 1.3011$, we have

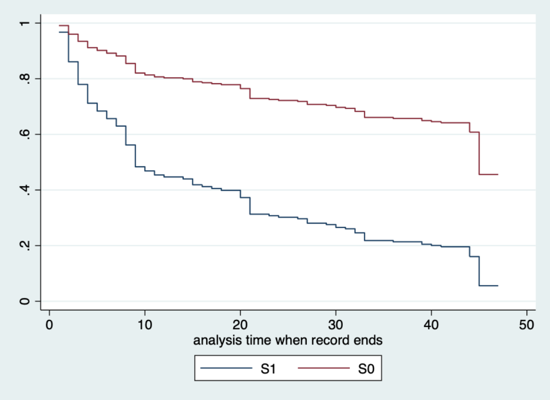
$$H(t|x=1) = e^{1.3011} H_0(t) \approx 3.6734 H_0(t)$$



The Cox proportional hazards model

- The survival function:

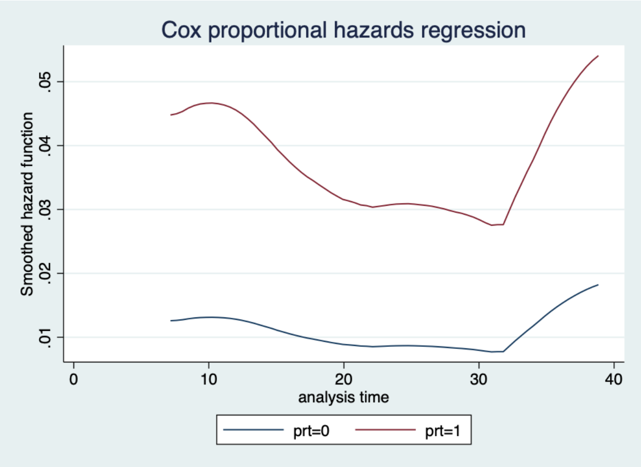
$$S(t|x) = e^{-H(t|x)} = e^{-e^{x\beta}H_{\mathbf{0}}(t)} = S_0(t)^{e^{x\beta}}$$



- Part-time students (S_1) survive much worse.

The Cox proportional hazards model

- Hazards:



- The hazards are indeed proportional, and if graphed on a log scale they would be parallell.

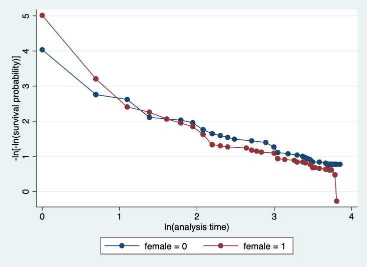
The Cox proportional hazards model

Model diagnostics

1. PH implies proportional integrated hazards:

$$H(t|x) = \int_0^t h(s|x)\mathrm{d}s = e^{x\beta} \int_0^t h_0(s)\mathrm{d}s = H_0(t)e^{x\beta}$$
$$\Rightarrow \ln H(t|x) = \ln H_0(t) + x\beta$$

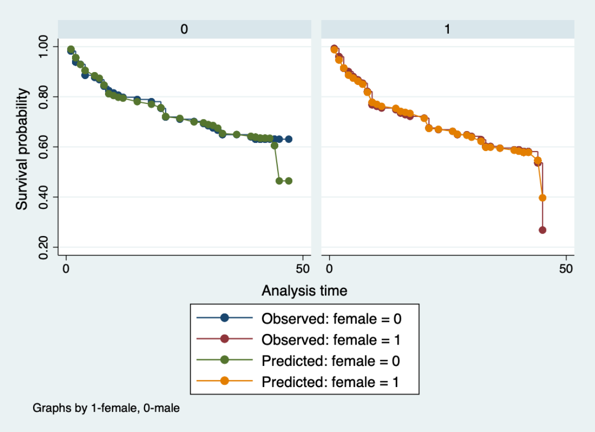
Graphical test: therefore under PH, the log-integrated hazard curves $\ln H(t|x)$, also called the log-log survivor curves, should be parallel at different values of the regressors x ($\Leftrightarrow$ no t in $x\beta$).



The Cox proportional hazards model

Model diagnostics

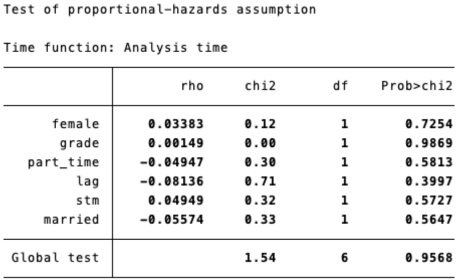
2. Also a graphical test: the fitted survivor function from Cox regression and the (nonregression) Kaplan-Meier estimate of the survivor function should be similar if PH holds.



The Cox proportional hazards model

Model diagnostics

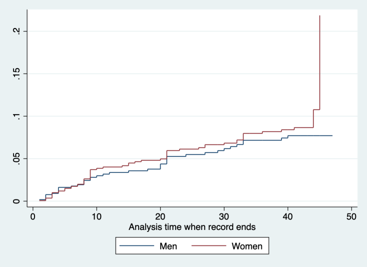
3. A formal statistical test on the key assumption of the Cox model: separable components, duration $h_0(t)$ and regressors $e^{x\beta}$.
   - Under the PH assumption, there should be no time (duration/spell) trend in the regressors part.
   - The test retrieves the scaled Schoenfeld residuals computed from the regressors, fit a smooth function of time to them, and then test whether there is a relationship.
   - Rejection of the null (no time trend / zero slope) indicates a deviation from the proportional-hazards assumption.



```
Test of proportional-hazards assumption

Time function: Analysis time
────────────┬────────────────────────────────────────────────
            │       rho            chi2       df       Prob>chi2
────────────┼────────────────────────────────────────────────
     female │   0.03383          0.12        1          0.7254
      grade │   0.00149          0.00        1          0.9869
  part_time │  -0.04947          0.30        1          0.5813
        lag │  -0.08136          0.71        1          0.3997
        stm │   0.04949          0.32        1          0.5727
    married │  -0.05574          0.33        1          0.5647
────────────┼────────────────────────────────────────────────
 Global test│                    1.54        6          0.9568
────────────┴────────────────────────────────────────────────
```

Stratified Cox Model

- If some variable (grade in last slide) doesn't fullfil the PH assumption, we can use it as a strata (group) variable.
- In the stratified Cox model, we relax the assumption that everyone faces the same baseline hazard.
- The baseline hazards are allowed to differ by group, while the coefficients β are constrained to be the same across groups (testable): $h_g(t|x) = h_{0g}(t)e^{x\beta}$, g indicates the gender groups[1].



- The cost of this model is that the effect of female is not identified.

[1] For simplicity, instead of using grade, we take female as the strata in this example.

Time-varying covariates

Extended Cox Model

- There are cases that require time-varying covariates: when one is repeatedly unemployed, the macroeconomic conditions change. Extended Cox model:

  $$h(t|x) = h_0(t)e^{x_t\beta}$$

  For two individuals i and j,

  $$\frac{h(t|x_{it})}{h(t|x_{jt})} = \frac{h_0(t)e^{x_{it}\beta}}{h_0(t)e^{x_{jt}\beta}} = e^{(x_{it}-x_{jt})\beta}$$

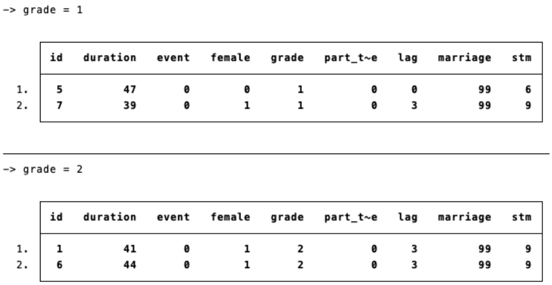
  This hazard ratio between two individuals is a function of t, the PH assumption no longer holds.
- One solution to time-varying covariates is to separate the record for one individual to several records where for each record the covariates are constant[2]. And then estimate it using the standard Cox model.

[2] We can do this because in practice, the data is always discrete.

Unobserved Heterogeneity

Group Effects

- `TA5_1.dta: college dropouts data`



```
-> grade = 1

     +-----------------------------------------------------------------------------+
     | id   duration   event   female   grade   part_t~e   lag   marriage   stm |
     |-----------------------------------------------------------------------------|
  1. |  5         47       0        0       1          0     0         99     6 |
  2. |  7         39       0        1       1          0     3         99     9 |
     +-----------------------------------------------------------------------------+

---------------------------------------------------------------------------------------
-> grade = 2

     +-----------------------------------------------------------------------------+
     | id   duration   event   female   grade   part_t~e   lag   marriage   stm |
     |-----------------------------------------------------------------------------|
  1. |  1         41       0        1       2          0     3         99     9 |
  2. |  6         44       0        1       2          0     3         99     9 |
     +-----------------------------------------------------------------------------+
```

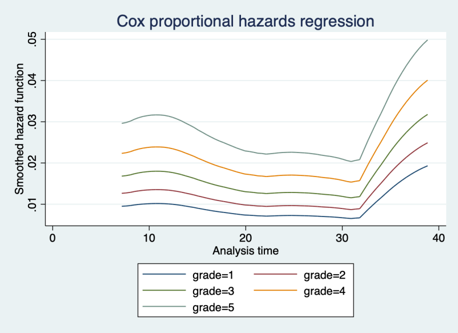
- `grade`: high school grades before college, $[1, 5]$, the lower the better
- Unobserved heterogeneity: skill levels. [3]

[3] Here we take group effects for an example, we can also consider individual effects.

Unobserved Heterogeneity

Group Effects

- `grade` does have an effect (Cox hazards for subsamples):



- Ways to account for group effects:
  1. Clustered errors: `. stcox age gender, nohr vce(cluster patient)`
  2. Random effects: `. stcox age gender, nohr shared(patient)`
  3. Fixed effects: `. stcox age gender i.patient, nohr`
- Let's compare these ways.

Unobserved Heterogeneity

Group Effects

1. If there indeed exists within-group correlation, the standard Cox model is misspecified. With **clustered errors**, although we can have valid measure of the variability of the coefficients, we don't have consistent measure for the point estimates.

   ```
   stcox $x, nohr vce(cluster grade) nolog
   ```

Unobserved Heterogeneity

Shared frailty

- In duration analysis, the unobserved heterogeneity will lead to inconsistent estimates even if it's not correlated with the explanatory variables[4].
- Consider the example that there are groups of unemployed people differ by the unobserved skill level, which will affect their hazard function.
- For the *j*th subject in the *i*th group, the hazard is

$$\begin{aligned} h_{ij}(t) &= h_0(t)\alpha_i e^{x_{ij}\beta}, \ \ \alpha_i > 0 \\ &= h_0(t) e^{x_{ij}\beta + \nu_i}, \ \ \nu_i = \ln \alpha_i \end{aligned}$$

The unobserved heterogeneity enters multiplicatively on the hazard function as a group-level *frailty*: α_i. The log frailty ν_i is analogous to random effects[5] in panel data.

[4] Unlike in linear models, where the estimates will be consistent if the unobserved heterogeneity is not correlated with the regressors.

[5] In this shared frailty model, the group effects α_i is assumed to be random and is governed by a Gamma distribution with mean 1 and variance θ. We will also talk about the fixed effects case later.

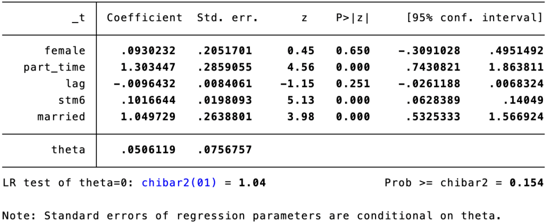
Unobserved Heterogeneity

Group Effects

2. **Random effects** (shared frailty):

```
.   stcox $x, nohr shared(grade)
```

Reliable estimates require sufficient number of observations for each group.



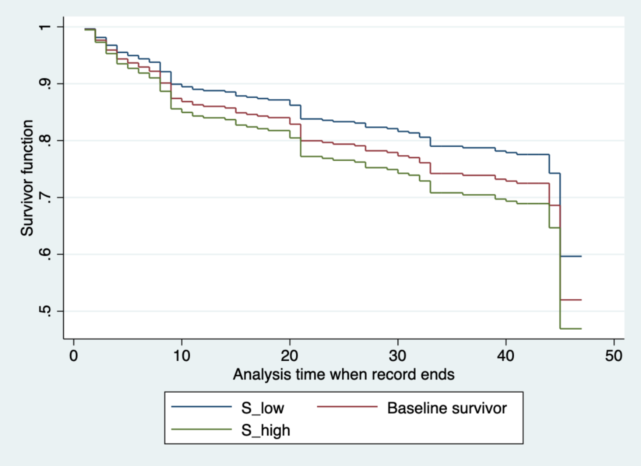
```
          _t   Coefficient  Std. err.      z    P>|z|     [95% conf. interval]

      female     .0930232   .2051701     0.45   0.650    -.3091028    .4951492
   part_time     1.303447   .2859055     4.56   0.000     .7430821    1.863811
         lag    -.0096432   .0084061    -1.15   0.251    -.0261188    .0068324
        stm6     .1016644   .0198093     5.13   0.000     .0628389      .14049
     married     1.049729   .2638801     3.98   0.000     .5325333    1.566924

       theta     .0506119   .0756757

LR test of theta=0: chibar2(01) = 1.04                 Prob >= chibar2 = 0.154

Note: Standard errors of regression parameters are conditional on theta.
```

- The LR test (on the null that the variance of α_i is zero) tells us that there is somehow significant (0.15) within group correlation.
- Once we account for intragroup correlation via the shared frailty model, for a given level of group effects, the hazard for females is about 9 percent higher ($e^{0.093} \approx 1.09$) than that for the males.

Unobserved Heterogeneity

Shared frailty

- The least frail group (lower ν_i from better grade) have survival experiences that are far superior.



Unobserved Heterogeneity

Group Effects

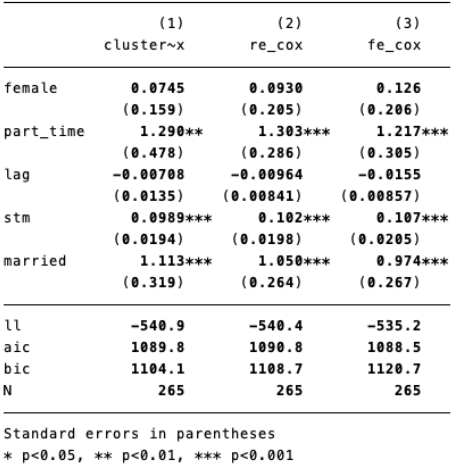
3. **Fixed effects**:

```
.  stcox $x i.grade, nohr
```

- Using random effects model (shared frailty) can help us to make inferences about the group effects based on the observed random sample.
- If we don't want to make this inference and are only interested in the observed sample, we can treat the group effects as fixed and account for them by including indicator variables indentifying the groups.
- In the fixed effects case, groups still have a direct multiplicative effect on the hazard function.
- That is, all individual records share the same baseline hazard, the effect of a group multiplies this baseline hazard up or down depending on the estimated coeffecients for the group indicator variables.

Unobserved Heterogeneity

Group Effects



	(1) cluster~x	(2) re_cox	(3) fe_cox
female	0.0745	0.0930	0.126
	(0.159)	(0.205)	(0.206)
part_time	1.290**	1.303***	1.217***
	(0.478)	(0.286)	(0.305)
lag	-0.00708	-0.00964	-0.0155
	(0.0135)	(0.00841)	(0.00857)
stm	0.0989***	0.102***	0.107***
	(0.0194)	(0.0198)	(0.0205)
married	1.113***	1.050***	0.974***
	(0.319)	(0.264)	(0.267)
ll	-540.9	-540.4	-535.2
aic	1089.8	1090.8	1088.5
bic	1104.1	1108.7	1120.7
N	265	265	265

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- The random effects (shared frailty) model is slightly better.

Parametric Approach

Parametric Models

- Proportional hazard specification: $h(t|x) = h_0(t)e^{x\beta} \Rightarrow$ flexible hazard functions

Semi-parametric model:

$$\text{Cox PH:} \quad h(t|x) = \underbrace{h_0(t)}_{\substack{\text{left}\\\text{unparameterized}}} \quad e^{x\beta}$$

Parametric models:

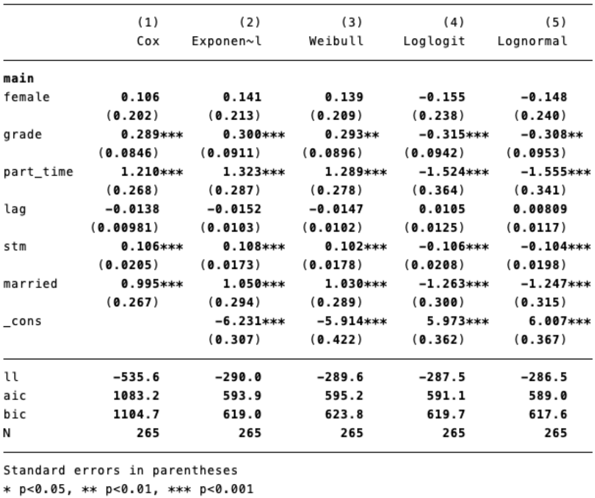
$$\text{Weibull:} \quad h(t|x) = h_0(t,\alpha,\gamma)e^{x\beta} = \alpha t^{\alpha-1}e^{\gamma}e^{x\beta} \rightarrow (\alpha,\gamma,\beta)$$

$$\text{Exponential:} \quad h(t|x) = h_0(t,\alpha)e^{x\beta} = e^{\alpha}e^{x\beta} \rightarrow (\alpha,\beta)$$

Notice that there is no constant term in vector x.

- The estimates from the parametric PH model should be roughly similar to that from the Cox model. Otherwise there is evidence of a misparameterized underlying baseline hazard.

Parametric Models



	(1) Cox	(2) Exponen~l	(3) Weibull	(4) Loglogit	(5) Lognormal
main					
female	0.106	0.141	0.139	-0.155	-0.148
	(0.202)	(0.213)	(0.209)	(0.238)	(0.240)
grade	0.289***	0.300***	0.293**	-0.315***	-0.308**
	(0.0846)	(0.0911)	(0.0896)	(0.0942)	(0.0953)
part_time	1.210***	1.323***	1.289***	-1.524***	-1.555***
	(0.268)	(0.287)	(0.278)	(0.364)	(0.341)
lag	-0.0138	-0.0152	-0.0147	0.0105	0.00809
	(0.00981)	(0.0103)	(0.0102)	(0.0125)	(0.0117)
stm	0.106***	0.108***	0.102***	-0.106***	-0.104***
	(0.0205)	(0.0173)	(0.0178)	(0.0208)	(0.0198)
married	0.995***	1.050***	1.030***	-1.263***	-1.247***
	(0.267)	(0.294)	(0.289)	(0.300)	(0.315)
_cons		-6.231***	-5.914***	5.973***	6.007***
		(0.307)	(0.422)	(0.362)	(0.367)
ll	-535.6	-290.0	-289.6	-287.5	-286.5
aic	1083.2	593.9	595.2	591.1	589.0
bic	1104.7	619.0	623.8	619.7	617.6
N	265	265	265	265	265

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- The lognormal model provides the best fit. The loglogistic model is the next-best model.

Parametric Models

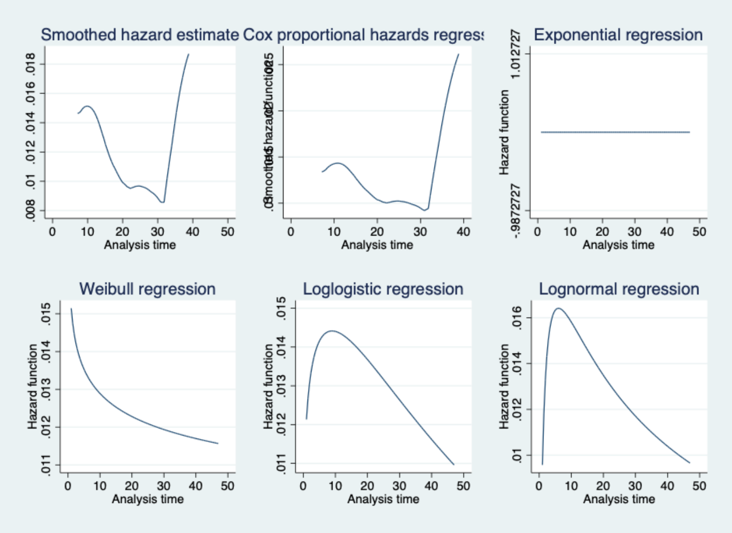


Figure 1: Hazard rates from various models, evaluated at the mean of the regressors

Discrete Duration

Discrete-time hazards

- The T periods indexed by $t = 1, \dots, T$ are grouped into A intervals indexed by $a = 1, \dots, A$, unequally spaced intervals are allowed.

$$h(t_a|x) = \mathbb{P}(t_{a-1} \leq T < t_a | T \geq t_{a-1}, x(t_{a-1}))$$

- Why discrete durations is a problem: we need to consider three indexes i, t, a in the derivation.
  - PH model of continuous durations:

$$h(t|x) = h_0(t)e^{x\beta}$$

  - PH model of discrete durations associated with the continuous model:

$$h(t|x) = h_0(t)e^{x(t_{a-1})\beta}$$

    The regressors are constant within the interval (a) but can vary across intervals, and $h_0(t)$ can vary within the interval (a).

Discrete-time hazards

- Two solutions:
  1. Use index a, group $h_0(t)$: logit.
     - Consider a binary choice model for transitions:

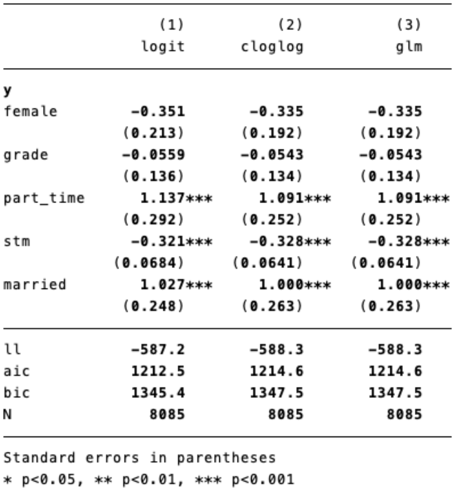
       $$d = \begin{cases} 1, & \text{if the spell ends} \\ 0, & \text{otherwise} \end{cases}$$

     - And we fit a simple (stacked) logit model on it:

       $$\mathbb{P}(t_{a-1} \leq T < t_a | T \geq t_{a-1}, x) = F(\lambda_a + x(t_{a-1})\beta)$$

       where β is restriced to be constant over time, and the intercept λ_a is allowed to vary across intervals.
  2. Use index t, add group indicators for each a: complementary log-log, GLM.

Discrete-time hazards



	(1) logit	(2) cloglog	(3) glm
y			
female	-0.351	-0.335	-0.335
	(0.213)	(0.192)	(0.192)
grade	-0.0559	-0.0543	-0.0543
	(0.136)	(0.134)	(0.134)
part_time	1.137***	1.091***	1.091***
	(0.292)	(0.252)	(0.252)
stm	-0.321***	-0.328***	-0.328***
	(0.0684)	(0.0641)	(0.0641)
married	1.027***	1.000***	1.000***
	(0.248)	(0.263)	(0.263)
ll	-587.2	-588.3	-588.3
aic	1212.5	1214.6	1214.6
bic	1345.4	1347.5	1347.5
N	8085	8085	8085

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- Three models yield similar results.

Appendix

Baseline

$$h_{ij}(t|x,\nu) = h_0(t)e^{x_{ij}\beta+\nu_i}$$

- Another concern: baseline estimates

$$\begin{aligned}\Rightarrow h_{ij}\Big(t|x=0,\nu_i=0\Big) &= h_0(t)e^{\beta_1\cdot 0+\cdots+\beta_k\cdot 0+0} \\ &= h_0(t)\cdot e^0 = h_0(t)\end{aligned}$$

- Problem: our $x =$ (`female,grade,part_time,lag,stm,married`), variable `stm` never goes to zero in our sample, $min($`stm`$) = 6$
- Solution: change the baseline

```
.   generate stm6 = stm - 6
.   stcox $x stm6, shared(grade)
```

- Now the baseline survivor estimate (S_0) corresponds to a male full-time student, not married and `stm` $= 6$, with mean frailty ($\nu_i = \ln\alpha_i = 0$).

References

- Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: methods and applications. Cambridge university press. Chapter 17.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press. Chapter 22.
- Cameron, A. C., & Trivedi, P. K. (2022). Microeconometrics using stata (Second Edition). Stata press. Chapter 21.
- Cleves, M., Gould, W., Gould, W. W., Gutierrez, R., & Marchenko, Y. (2010). An introduction to survival analysis using Stata. Stata press.