

TA Session 4: Duration Models

Microeconometrics with Joan Llull
IDEA, Fall 2024

TA: Conghan Zheng

October 16, 2024

Overview

- 1 Duration Data
- 2 Continuous Duration
- 3 Discrete Duration
- 4 Appendix

Duration Data

Duration Data

- *Duration data*: data on a variable that measures the length of time spent in a state before transition to another state
- TA4.dta: college dropouts data (single-record data, one obs. per individual)

	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
- Empirical concern: the spell length may be incompletely observed (censored, individuals leave the study before the spell ends).

Duration Data

- Set the duration data structure based on variable duration. Commands begin with st (survival-time).
 - . stset duration, failure(event=1) 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: analysis time when record begins (the calendar time could be different for different individuals)
 - _t: analysis time when record ends
 - _d: 1 if failure, 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 key difference: grouping

- Continuously distributed durations
 - Time index is still "discrete", you have natural numbers $t = 1, 2, \dots$, 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 measurements 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 the 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.

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

Estimation of survival functions:

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

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	$h(t)$	$\lim_{h \rightarrow 0} \frac{P(t \leq T \leq t+h T \geq t)}{h}$	$h(t) = \frac{f(t)}{S(t)}$
Cumulative hazard	$H(t)$	$H(t) = \int_0^t h(s)ds$	$H(t) = -\ln S(t)$

- For each t , $h(t)$ is the instantaneous rate of leaving per unit of time.

$$h(t) = \lim_{\Delta \rightarrow 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta | T \geq t)}{\Delta}$$

and for “small” Δ ,

$$\mathbb{P}(t \leq T \leq t + \Delta | T \geq t) \approx h(t) \cdot \Delta$$

Thus, the hazard function can be used to approximate a conditional probability in much the same way that the height of the density of T can be used to approximate an unconditional probability.

The Kaplan-Meier Estimator

- Kaplan-Meier estimator or product limit estimator of the survivor function

$$\hat{S}(t) = \prod_{j|t_j \leq t} \frac{\text{# Spells at risk}(t_j) - \text{# Spells ending}(t_j)}{\text{# Spells at risk}(t_j)}$$

Kaplan-Meier survivor function

Time	At risk	Fail	Net lost	Survivor function	Std. 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

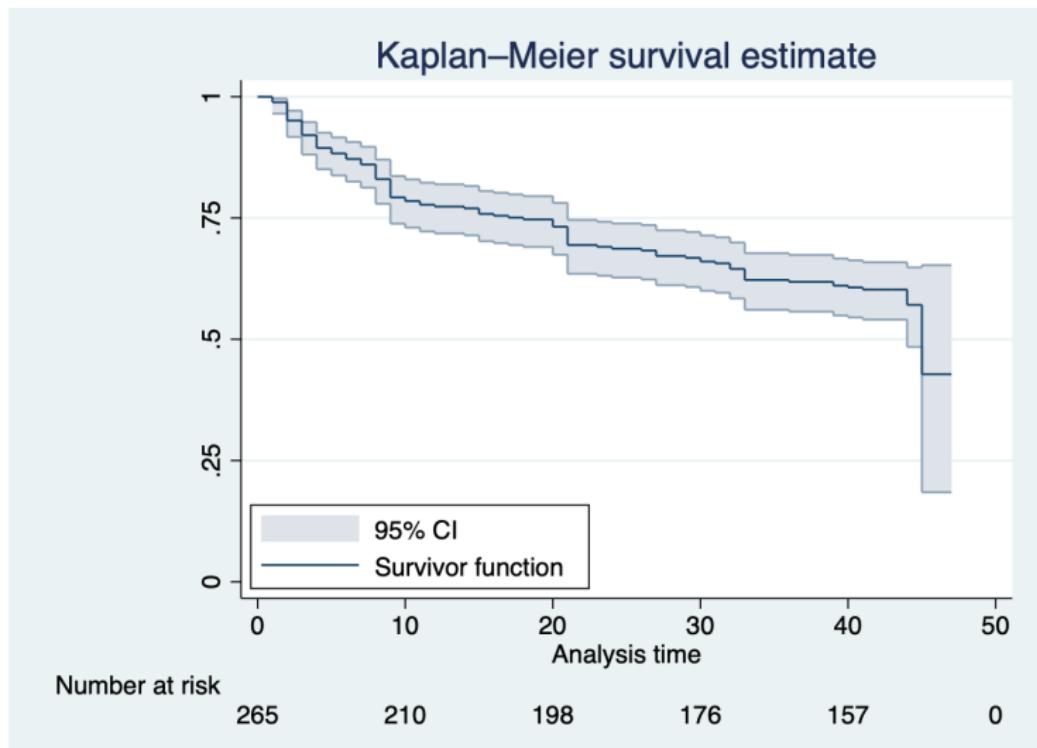
- At risk: at school; Fail: dropped out; Net Lost: censored
- Example:**

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

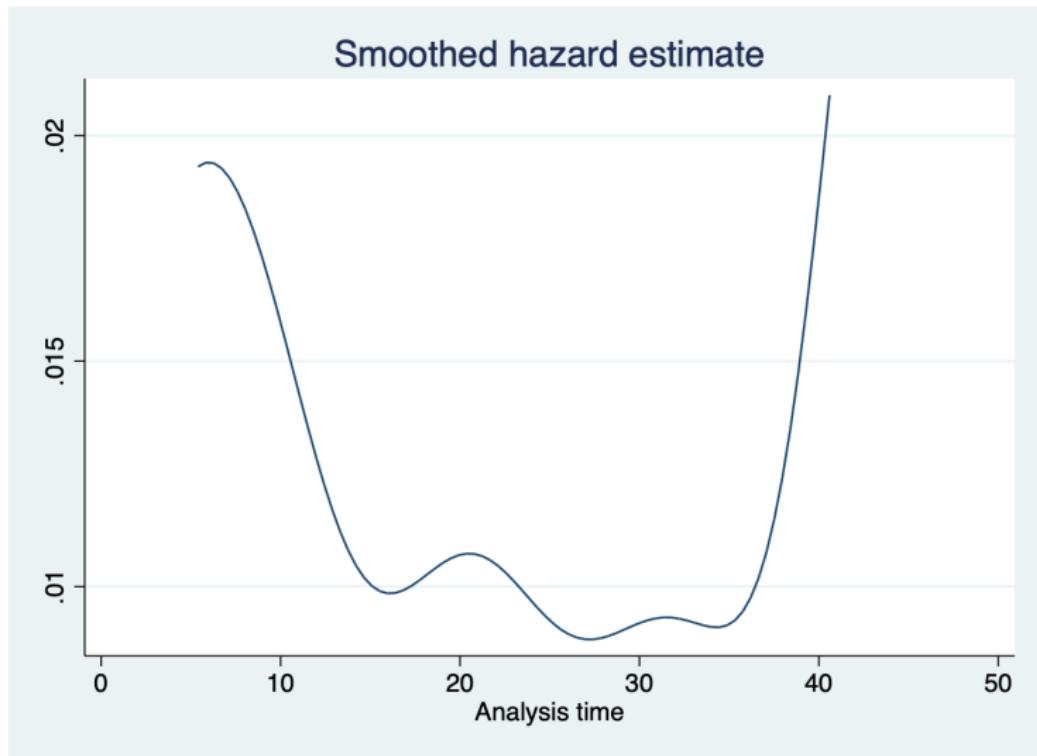
The probability of survival beyond $t = 2$ is $\frac{262}{265} \times \frac{252}{262} = \frac{252}{265} \approx 0.9509$.

...

The Kaplan-Meier Estimator



The Kaplan-Meier Estimator



- Kernel smoothing: the weighted (kernel) average of neighboring observations

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):

$$h(t|x) = \underbrace{h_0(t)}_{\text{baseline hazard}} \cdot \underbrace{e^{x\beta}}_{\text{relative hazard}}$$

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 Cox model has no intercept since

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

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

The Cox Proportional Hazards Model

Partial Likelihood Estimation

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

- Partial likelihood estimation (Cox, 1972, 1975)
 - For now, we consider only time-invariant regressors, but later we will relax this assumption.
 - “Partial”: we estimate β without estimating $h_0(t)$.
 - Partial likelihood minimization $\rightarrow \hat{\beta}$
 - Nonparametric KM estimation $\rightarrow \hat{h}_0(t)$

The Cox Proportional Hazards Model

- Effects of regressors on the time until college dropout: the β s from $e^{x\beta}$
 - . 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		
Log likelihood =	-535.6177	LR chi2(6) =	62.85
		Prob > chi2 =	0.0000

<u>_t</u>	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 magnitude of these effects is not immediately clear. Why?

The Cox Proportional Hazards Model

Effect Size

- If the j th regressor in $x = (x_1, x_2, \dots, x_k)$ is increased by 1 unit,
- $$h(t|x + \Delta) = h_0(t)e^{\beta_1 x_1 + \dots + \beta_j(x_j + 1) + \dots + \beta_k x_k} = h_0(t)e^{x\beta + \beta_j} = e^{\beta_j} h(t|x)$$
- Therefore, changes in regressors can be interpreted as having a multiplicative effect on the original hazard (semi-elasticity), as

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

- The coefficients:
 - $\beta_{\text{female}} \approx 0.106 > 0$, hazard rate is higher for female students;
 - $\beta_{\text{grade}} \approx 0.289 > 0$, hazard rate is higher for college students with worse high school performance (high grade).
- The effect size:
 - hazard ratio for time-invariant variable `female` is $e^{0.106} \approx 1.112$;
 - A one unit increase in grade (high school grades before college, the lower the better) leads to the hazard rate being $e^{0.289} \approx 1.335$ times higher.

The Cox Proportional Hazards Model

Baseline

- Concern: the baseline

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

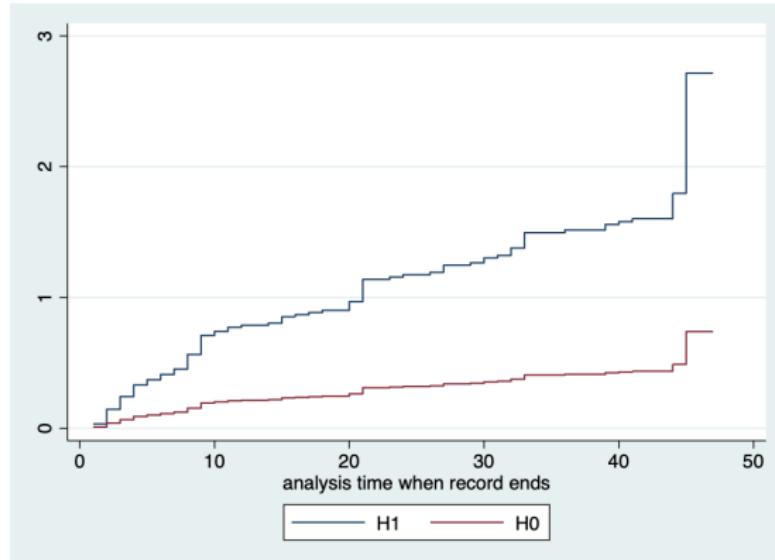
- Problem: our $x = (\text{female}, \text{grade}, \text{part_time}, \text{lag}, \text{stm}, \text{married})$, variable stm never goes to zero in our sample, $\min(\text{stm}) = 6$
- Solution: recenter the variable
 - . generate $\text{stm6} = \text{stm} - 6$
 - . `stcox $x stm6, shared(grade)`
- Now the baseline survivor estimate (S_0) corresponds to a male full-time student, not married and $\text{stm} = 6$.

The Cox Proportional Hazards Model

- The cumulative hazard:

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

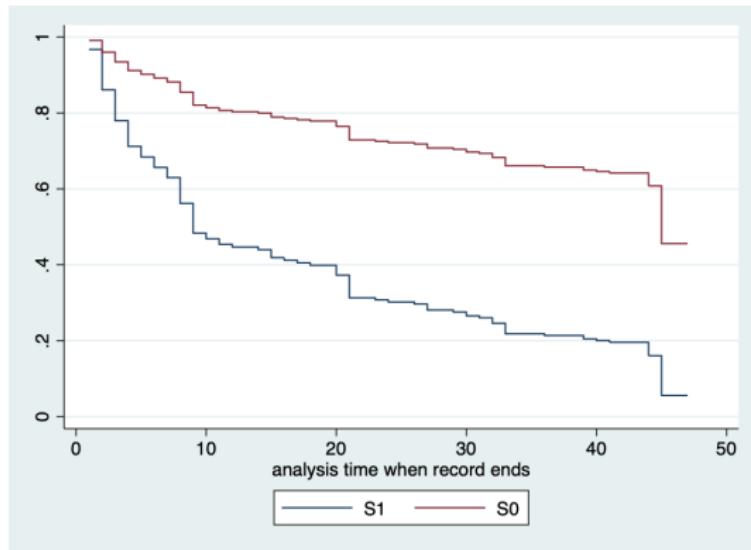
- After including one binary regressor (part-time student) whose estimate is $\beta_1 \approx 1.210$, we have $H(t|x=1) \approx e^{1.210} H_0(t) \approx 3.353 H_0(t)$.



The Cox Proportional Hazards Model

- The survival function:

$$S(t|x) = e^{-H(t|x)} = e^{-e^{x\beta} H_0(t)} = [e^{-H_0(t)}]^{e^{x\beta}} = S_0(t)^{e^{x\beta}}$$

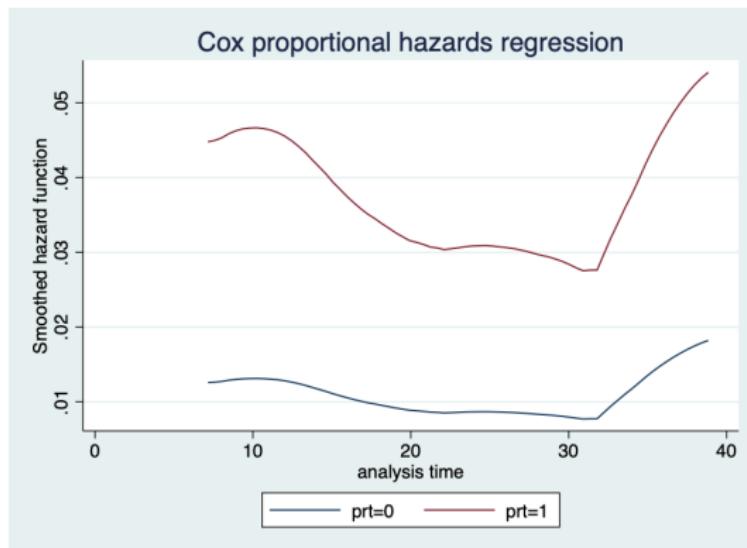


- Part-time students (S_1) survive much worse:
 $S(t|x=1) \approx S_0(t)^{e^{1.210}} \approx S_0(t)^{3.353}$, higher power $e^{x\beta}$ makes $S(t|x)$ more convex.

The Cox Proportional Hazards Model

- Hazards:

$$h(t|x=1) = h_0(t) \cdot e^{1.210} = h_0(t) \cdot 3.353$$



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

The Cox Proportional Hazards Model

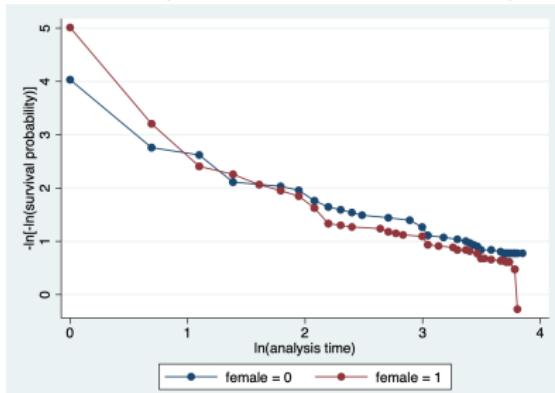
Model Diagnostics

- ① PH implies proportional integrated hazards:

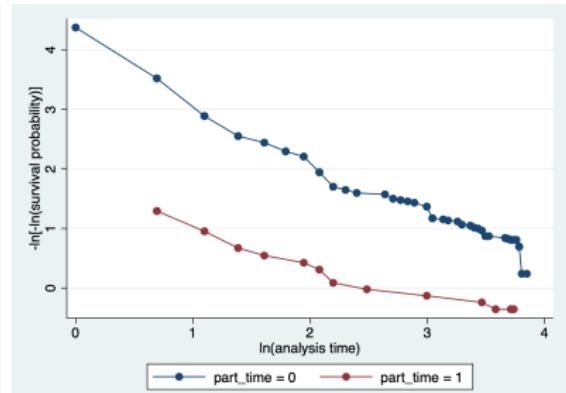
$$H(t|x) = \int_0^t h(s|x)ds = e^{x\beta} \int_0^t h_0(s)ds = H_0(t)e^{x\beta}$$

$$\Rightarrow \ln H(t|x) = \ln H_0(t) + x\beta$$

Therefore under PH, the log-integrated hazard curves $\ln H(t|x)$ (the *log-log survivor curves*), should be parallel at different values of the time invariant regressors x (as there is no t in $x\beta$) → a graphical test on PH



$$e^{\beta_{\text{female}}} \approx 1.112$$

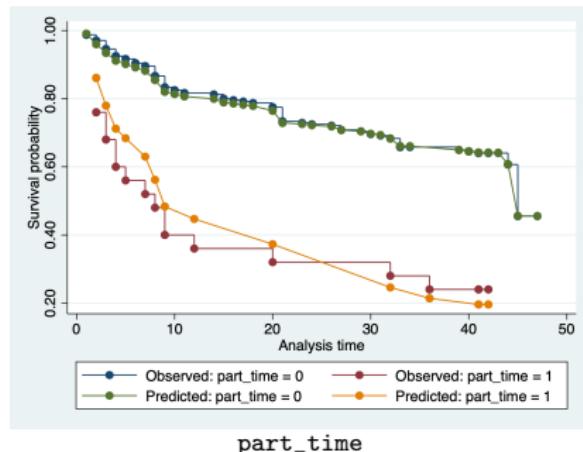
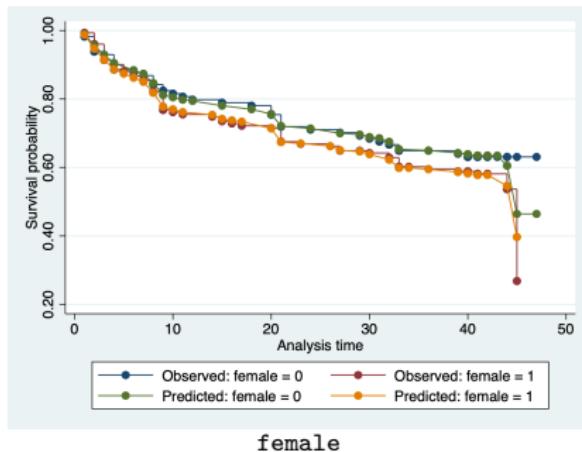


$$e^{\beta_{\text{part_time}}} \approx 3.353$$

The Cox Proportional Hazards Model

Model Diagnostics

- ② The predicted survivor function from Cox regression and the (nonregression) Kaplan-Meier estimate (observed) of the survivor function should be similar if PH is appropriate. → another graphical test on PH



The PH model is reasonable for female but does not do so well for part_time.

The Cox Proportional Hazards Model

Model Diagnostics

- ③ A formal residual-based statistical test on the key assumption of the Cox model: separable components, duration part $h_0(t)$ and regressors part $e^{x\beta}$. Under the standard PH assumption, there should be no time (duration/spell length) trend in the regressors part. Rejection of the null (no time trend / zero slope) indicates a deviation from the proportional-hazards assumption.

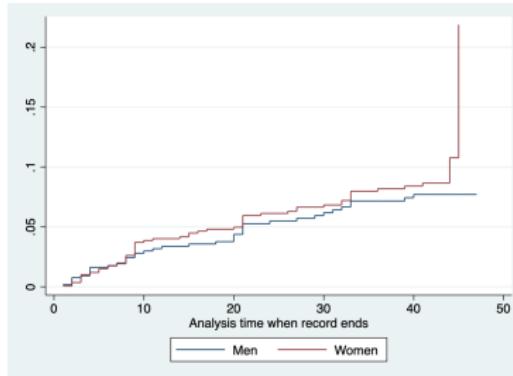
Test of proportional-hazards assumption

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 does not fulfill 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. Ex: $h_g(t|x) = h_{0g}(t)e^{x\beta}$, where g indicates the gender groups.



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

Time-Varying Covariates

Extended Cox Model

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

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

- 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.

- Estimation: in the likelihood, x_i is replaced by $x_i(t_j)$...

Parametric Models

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

Semi-parametric model:

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

Parametric models:

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

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

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 Comparison

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

- Better model fit but counterintuitive signs of coef. for some models?
- Can be more precise on coef.; Low robustness to distribution misspecification.

Hazards from Various Models

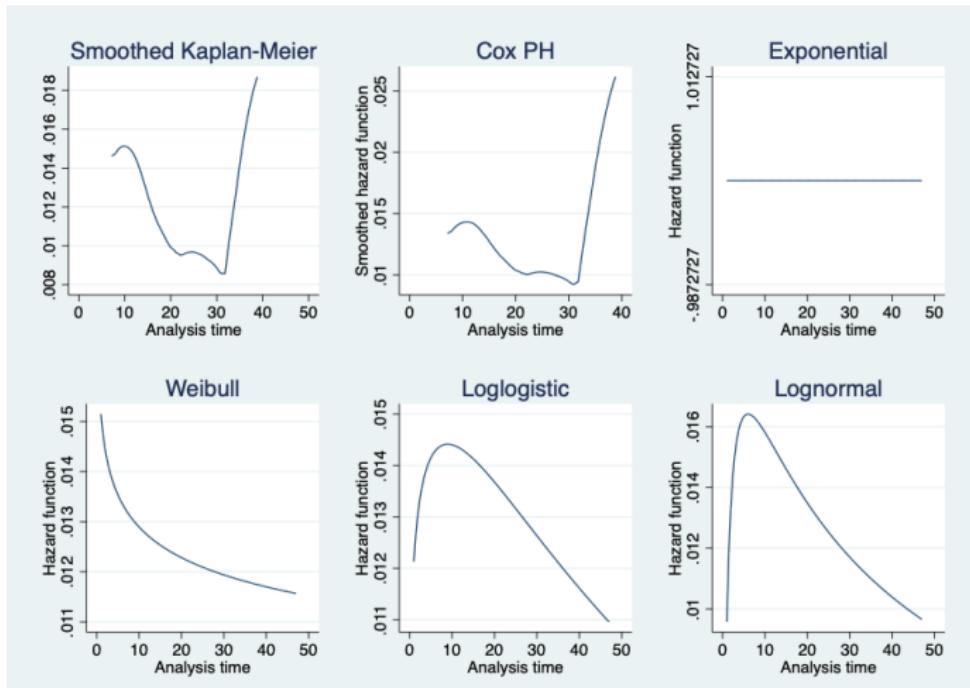


Figure 1: Hazard rates from various models, evaluated at the mean of the regressors

- Exponential: constant hazard; Weibull: monotonic hazard; Loglogistic and Lognormal: inverted U-shaped hazard; Cox PH: flexible hazard.

Unobserved Heterogeneity

- In duration analysis, the unobserved heterogeneity will lead to inconsistent estimates even if it's not correlated with the explanatory variables¹. Consider for example that there are groups of unemployed people that differ by the unobserved skill level, which will affect their hazard function.

$$\begin{aligned} h_i(t) &= h_0(t)\alpha_i e^{x_i \beta}, \quad \alpha_i > 0 \\ &= h_0(t)e^{x_i \beta + \nu_i}, \quad \nu_i = \ln \alpha_i \end{aligned}$$

The unobserved heterogeneity enters the hazard function multiplicatively: α_i (which can also be extended to a group-level effect α_g). The log effect ν_i is analogous to random effects² in panel data.

¹Unlike in linear models, where the estimates will be consistent if the unobserved heterogeneity is not correlated with the regressors.

²The effects α_i are assumed to be random and follow a predefined distribution.

Unobserved Heterogeneity

```
. streg, dist(weibull) frailty(invga) vce(robust) nolog
nohr3
```

Weibull PH regression
Inverse-Gaussian frailty

No. of subjects = 265 Number of obs = 265
 No. of failures = 107
 Time at risk = 8,087
 Wald chi2(0) = .
 Log pseudolikelihood = -318.11508 Prob > chi2 = .

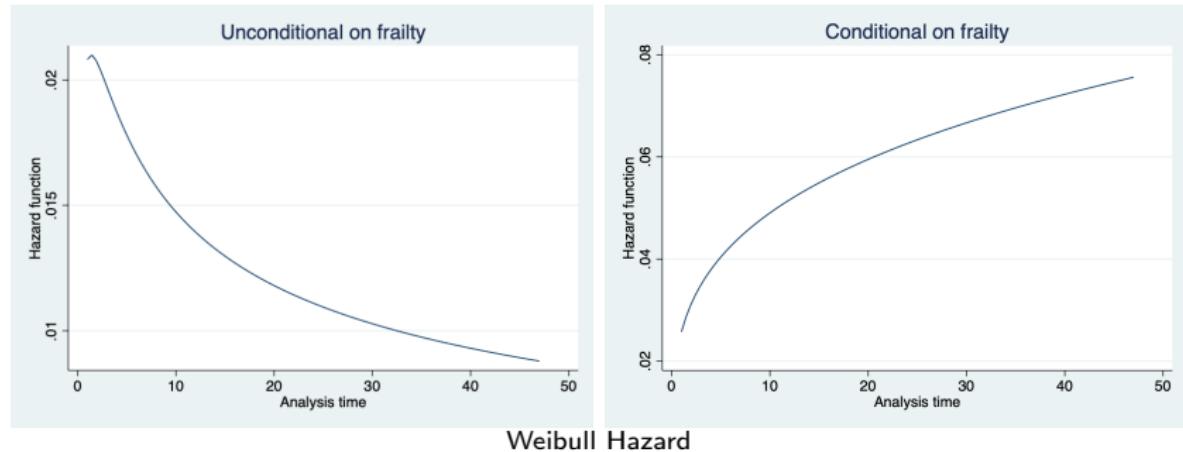
(Std. err. adjusted for 265 clusters in id)

<i>t</i>	Robust				
	Coefficient	std. err.	z	P> z	[95% conf. interval]
_cons	-3.906223	.2703001	-14.45	0.000	-4.436002 -3.376445
/ln_p	.2467046	.0768851	3.21	0.001	.0960125 .3973967
/lntheta	2.575637	.2272414	11.33	0.000	2.130253 3.021822
p	1.279801	.0983977			1.180773 1.487946
1/p	.7813715	.0600759			.6720673 .9084527
theta	13.13969	2.985881			8.416992 20.51225

- The log likelihood increases from -535.6177 (the Cox PH with 6 regressors) to -318.1151.

³You can check the code TA4.do for an example of Cox PH with Gamma-distributed random effects.

Unobserved Heterogeneity



Weibull Hazard

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 $h_0(t)$ (more common)

- 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 restricted 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 (dummies for each interval a are included as regressors)

- Complementary log-log: equivalent to a Cox PH, also called a grouped Cox PH.

Discrete-time hazards

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

Standard errors in parentheses

* p<0.05, ** p<0.01, *** p<0.001

Appendix

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.