Fitting
Cox Regression Models

(Chapters 14 and 15, *ALDA*)

Judy Singer & John Willett

Harvard University Graduate School of Education
May, 2003

What we will cover

Data Example: Recidivism among former inmates

(ALDA, Section 14.1, p. 504)

- *Research Question*: Whether, and if so, when former inmates released from a medium security prison are re-arrested.
- *Citation*: Henning and Freuh (1996).
- *Design*: 194 inmates tracked for up to 3 years from release. 54.6% (106) were re-arrested (recorded to the nearest *day*)

Person-level data set (note, we do not use a person-period data set)

ID	*PERSONAL*	*PROPERTY*	*CENTERED AGE*	*DAY*	*MONTHS*	*CENSOR*
22	0	0	0.258	52	1.7084	1
8	1	1	22.451	19	0.6242	1
187	1	0	−7.200	1065	36.0000	1
26	0	1	−7.302	72	2.3655	0
5	1	1	−7.165	9	0.2957	0
130	0	1	22.391	486	15.9671	1
106	0	0	16.203	356	11.6961	0
33	1	0	27.061	85	2.7926	1

Identifies the 31.4% who have ever been convicted of a person-related crime (assault or kidnapping)

Identifies the 81.4% who have ever been convicted of a property-related crime

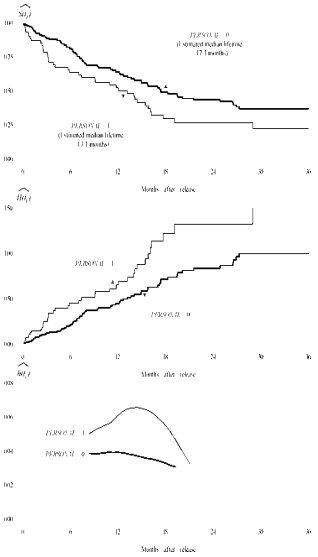
Age upon release from prison (centered around the sample mean, 30.7)

Event time measured in both days and months

Indicates whether the event time indicates re-arrest (CENSOR=0) or time out at end of data collection (CENSOR=1)

Towards a statistical model for continuous-time hazard
Sample functions by levels of PERSONAL

(ALDA, Section 14.1.1, p. 504, Fig 14.1, p 505)



Survivor functions:
Recidivism is high in both groups, although those with a history of person-related crimes are clearly at greater risk (ML of 17.3 *vs.* 13.1)

Cumulative hazard functions:
Approximately linear immediately after release and soon accelerates (but at slightly different times); eventually both decelerate.
Suggests that each underlying hazard function is initially steady, then rises, then falls.

Kernel-smoothed hazard functions: Don't describe risk immediately after release, but by month 8, we can see that the hazard for PERSONAL=1 is consistently higher than the hazard for PERSONAL=0

Intuitively, a continuous-time hazard model should look like a DT hazard model, where a transformation of hazard is expressed as the sum of two components:

- A *baseline function*, the value of transformed hazard when all predictors are 0
- A *weighted linear combination of predictors*

But, because we lack a complete picture of hazard, we instead develop the model in terms of **cumulative hazard.** After doing so, we use transformation to specify an equivalent model in terms of hazard.

Developing a statistical model for cumulative hazard: What is the impact of the predictor PERSONAL?

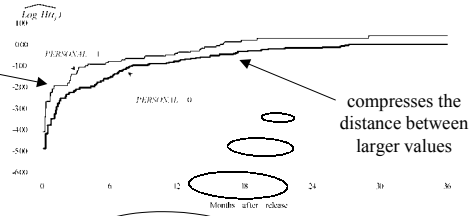
(ALDA, Section 14.1.2, p. 507, Fig 14.2, p 508)

Problem:
Cumulative hazard is semi-bounded from below by 0
Can't use logits (which are undefined for values >1)

→

Solution:
Model *log* cumulative hazard
Defined for any positive value
(*log negative log survivor function* or the *log-log survivor function)*

Expands the distance between small values



compresses the distance between larger values

What kind of statistical model should we use?
What would provide a reasonable representation of the population relationship between log cumulative hazard and predictors?

Again, a dual partition makes sense, where log cumulative hazard is expressed as the sum of two parts:

- A *baseline function*, now the value of log cumulative hazard when all predictors are 0
- A *weighted linear combination of predictors*

But, how do we specify the baseline?
As in DT, use a completely general unconstrained shape:

- Let's call the general baseline $\log H_0(t_j)$
- Might think this vagueness creates problems for estimation, but it doesn't

Specifying the Cox model in terms of log cumulative hazard

(ALDA, p 509, Fig 14.2, p. 508)

$$\log H(t_{ij}) = \log H_0(t_j) + \beta_1 PERSONAL_i$$

when PERSONAL = 0
$\log H(t_{ij}) = \log H_0(t_j)$

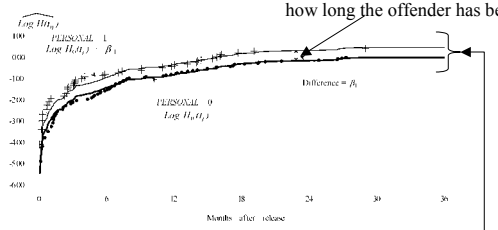
when PERSONAL = 1
$\log H(t_{ij}) = \log H_0(t_j) + \beta_1$

When PERSONAL=1, the *Baseline Function* shifts "vertically" by β_1

Mapping the model onto sample log cumulative hazard functions

(using +'s and !'s to denote estimated subsample values)

Vertical distance between functions assesses the magnitude of the predictor's effect.
(We assume that the effect is constant regardless of how long the offender has been out of prison)



Curves are hypothesized population log cumulative hazard functions
(they should go through sample data but we don't expect them to fit perfectly)

Antilogging yields a Cox model in *cumulative hazard* form

(ALDA, p 510, Fig 14.2, p. 508)

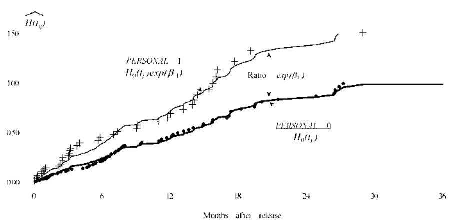
$$H(t_{ij}) = H_0(t_j)e^{\beta_1 PERSONAL_i}$$

when PERSONAL = 0	*when PERSONAL = 1*
$H(t_{ij}) = H_0(t_j)$	$H(t_{ij}) = H_0(t_j)e^{\beta_1}$

When PERSONAL=1, the *Baseline Function* is no longer simply shifted "vertically." Instead it is multiplied by e^{β_1}

Mapping the model onto sample cumulative hazard functions

(using +'s and !'s to denote estimated subsample values)



Ratio of cumulative hazard functions

$$\frac{H_0(t_j)\exp[\beta_1]}{H_0(t_j)} = \exp[\beta_1]$$

When the outcome is raw cumulative hazard, the functions are magnifications and dimunitions of each other—**they are *proportional***

Yet we still say the effect is constant over time, but instead of their vertical distance being constant, their ***ratio* is constant**

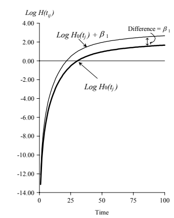
Hazard function representation of the Cox regression model

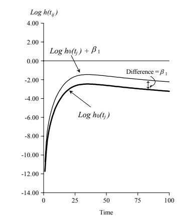
(ALDA, Section 14.1.3, p. 512, Fig 14.3, p. 513)

Through calculus, we can show that the Cox models just developed in terms of cumulative hazard are *identical* to those expressed in terms of raw hazard

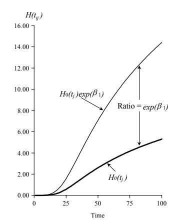
Cumulative hazard form $\quad log\ H(t_{ij}) = log\ H_0(t_j) + \beta_1 X_i$

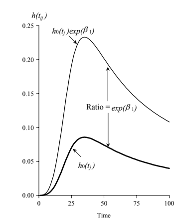
Hazard form $\quad log\ h(t_{ij}) = log\ h_0(t_j) + \beta_1 X_i$





Log scale
β_1 Constant vertical distance





Raw scale
e^{β_1} Proportional vertical distance

$$H(t_{ij}) = H_0(t_j)e^{\beta_1 X_i} \quad \rightarrow \quad h(t_{ij}) = h_0(t_j)e^{\beta_1 X_i}$$

Practical consequences of this equivalence

1. We can do exploratory data analysis using cumulative hazard
2. We can interpret parameter estimates in terms of predictors' effects on hazard
3. Because effects are *proportional* for raw hazard, the Cox model is often called a *proportional hazards model*

Fitting the Cox regression model to data

(ALDA, Section 14.2, p. 516)

General representation of the Cox model

$$h(t_{ij}) = h_0(t_j)\exp\left[\beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_P X_{Pij}\right]$$

$$log\ h(t_{ij}) = log\ h_0(t_j) + \left[\beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_P X_{Pij}\right]$$

In addition to specifying a particular model for hazard, Cox developed an ingenious method for fitting the model to data: *partial maximum likelihood* estimation (available in every major statistical package (See Section 14.2)).

Three important practical consequences of Cox's method:

- **The shape of the baseline hazard function is irrelevant.** Unlike parametric methods—and there are many—we need not make *any assumptions* about the shape of the baseline hazard function—therefore, no assumptions about event time distributions are violated.
- **The precise event times are irrelevant; only their rank order matters.** Cox regression is semi-parametric. The very data you took pains to collect precisely is effectively converted into ranks!
- **Ties can create analytic difficulties.** Even though the specific values are irrelevant their ranking does matter. In theory, there should be no ties; in reality, there always are. All packages have one or more approximations (we use Efron's method).

Interpreting parameter estimates in a fitted Cox model

(ALDA, Section 14.3.1, p. 524, Table 14.1, p. 525)

Simple uncontrolled models

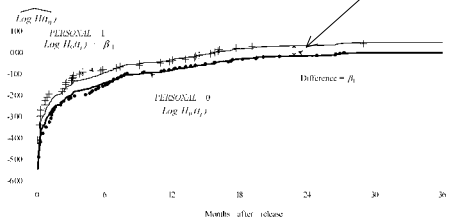
Overall model

	Model A	Model B	Model C	Model D
Parameter Estimates and Asymptotic Standard Errors				
PERSONAL	0.4790*			0.5691**
	(0.2025)			(0.2052)
PROPERTY		1.1946***		0.9358**
		(0.3493)		(0.3509)
AGE			−0.0681***	−0.0667***
			(0.0156)	(0.0168)

Interpreting parameter estimates: Each assesses the effect of a 1-unit difference in the associated predictor on log hazard (controlling for all other predictors in the model)

Example: Log hazard function for someone with a history of personal offenses is 0.479 units higher than for someone without this history

Is there any intuitive way of understanding this?
Returning to the sample log cumulative hazard functions by PERSONAL, we estimate that in the population, the average distance between them is 0.479



Interpreting hazard ratios in a fitted Cox model

(ALDA, Section 14.3.1, p. 524, Table 14.1, p. 525)

Antilogged parameter estimates are fitted hazard ratios associated with a 1-unit difference in the predictor

The estimated hazard of recidivism among offenders with a history of property offenses is three times that of those with no such history

	Model A	Model B	Model C	Model D
Parameter Estimates and Asymptotic Standard Errors				
PERSONAL	0.4790*			0.5691**
	(0.2025)			(0.2052)
PROPERTY		1.1946***		0.9358**
		(0.3493)		(0.3509)
AGE			−0.0681***	−0.0667***
		$e^{1.1946}$	(0.0156)	(0.0168)
Hazard Ratios and Their Asymptotic Standard Errors				
PERSONAL	1.6144***			1.7659***
	(0.3268)			(0.3624)
PROPERTY		3.3022***		2.5482**
		(1.1535)		(0.8941)
AGE			0.9342***	0.9355***
			(0.0146)	(0.0157)

For continuous predictors
Compute the %age difference in hazard associated with a 1-unit difference in the predictor: 100*(hazard ratio-1)

100*(0.9342-1)=-6.58%
The hazard of recidivism is 6.6% lower for each additional year of age upon release

Careful: Only make *comparative* statements about hazard
You can say that the hazard for one group is three times higher than that of another, but you cannot say how high, or low, either function is
This is the compromise associated with Cox regression

Evaluating the goodness of fit of the Cox model

(ALDA, Section 14.3.2, p. 528, Table 14.1, p. 525)

Log Likelihood statistics (LL & -2LL)

- LL statistics increase across models suggesting that each fits better than the previous one.
- Similarly –2LL statistics decrease (note, this is not a deviance statistic as there is no saturated model that can reproduce the sample data)

	Model A	Model B	Model C	Model D
Goodness-of-fit				
LL	−492.04	−486.60	−483.22	−475.22
−2LL	*984.08*	*973.20*	*966.43*	*950.44*
LR statistic	5.32	16.20	22.97	38.96
n parameters	1	1	1	3
p	0.0210	<0.0001	<0.0001	<0.0001
AIC	986.08	975.20	968.44	956.44
BIC	988.74	977.86	971.09	964.43
Likelihood-ratio Hypothesis Tests				
H_0: $\beta_{PERSONAL} = 0$	5.32* (1)			7.28(1)**
H_0: $\beta_{PROPERTY} = 0$		16.20(1)***		9.15(1)***
H_0: $\beta_{AGE} = 0$			22.97*** (1)	18.32(1)***

Evaluating goodness-of-fit in comparison to a null model

Every Cox model has a null model with no predictors
(in DT we fit it explicitly;
here, we fit it only implicitly as we never estimate the baseline hazard function).
The –2LL for the null model for these data is 989.402.
All tests reject: Each model fits better than the null (*big deal!).*

Likelihood ratio hypothesis tests

Used to compare nested models; here, only Model D provides unique tests
All tests in D reject, indicating that each predictor is statically significant, even on control for all other predictors in the model

As usual AIC and BIC are useful for comparing non-nested models

Summarizing findings using risk scores

(ALDA, Section 14.3.4, p. 532, Table 14.2, p. 533)

How might you compare each person's risk of event to that of the "baseline individual"
(the person who really has all predictor values = 0)?

$$\frac{h_0(t_j)\exp\left[\beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_P X_{Pij}\right]}{h_0(t_j)} = \exp\left[\beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_P X_{Pij}\right]$$

risk score

Because AGE is centered, the baseline individual is someone of average age on release (30.7) who has no history of PROPERTY or PERSONAL crime

ID	*PERSONAL*	*PROPERTY*	*CENTERED AGE*	Risk score	*DAY*	*MONTHS*	*CENSOR*
22	0	0	0.258	0.98	52	1.7084	1
8	1	1	22.451	1.01	19	0.6242	1
187	1	0	−7.200	2.86	1065	36.0000	1
26	0	1	−7.302	4.15	72	2.3655	0
5	1	1	−7.165	7.26	9	0.2957	0
130	0	1	22.391	0.57	486	15.9671	1
106	0	0	16.203	0.34	356	11.6961	0
33	1	0	27.061	0.29	85	2.7926	1

<u>At average comparative risk, but obtained that value in different ways</u>

- ID 22 was average age with no history of these crimes
- ID 8 had a history of both crimes but was 22 years older than the average inmate upon release

<u>At high comparative risk</u>

- All are younger than average on release
- ID 5 is over 7 times more likely than a baseline individual to re-offend

<u>At low comparative risk</u>

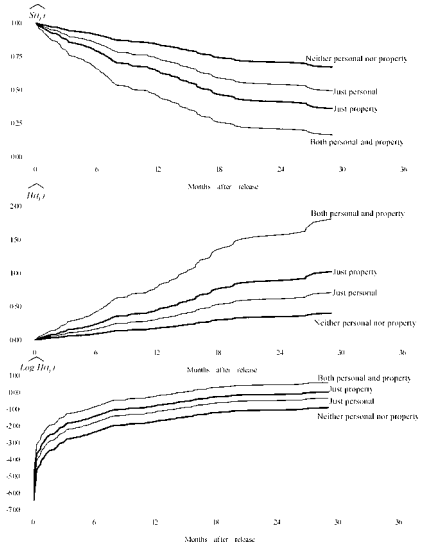
- All are much older than average on release
- None has history of both crimes

Careful: <u>Changing the baseline by centering predictors changes risk scores</u>

Recovered survivor and cumulative hazard functions

(ALDA, Section 14.4.2, p. 540, Fig 14.5, p. 541)

Yes, Virginia, there is a Santa Claus
Even though we have repeatedly stated that Cox regression provides *no* information about the baseline hazard function, it is possible to recover baseline functions from a model fit with time-invariant predictors (however, these are not predicted values)



Useful for documenting the combined effects of predictors
Here, we use Model D to control for AGE and show the combined effect of PERSONAL and PROPERTY, documenting the large differences in survival associated with variation in these predictors

See Section 14.1, *p. 535* for details

Including time varying predictors in a Cox model

(ALDA, Section 15.1, p. 544)

Model specification is easy
(just add subscript *j* to time varying predictors)

$$h(t_{ij}) = h_0(t_j)\exp\left[\beta_1 X_{1i} + \beta_2 X_{2ij}\right]$$

Data demands can be high (sometimes insurmountable)

Need to know the time-varying predictor's value —for *everyone* still at risk— at every moment when *someone* experiences the event

- Requirement holds whether there are 10, 100, or 1000 unique event times
- True in DTSA, but less problematic because:
  - Number of unique event times was relatively small
  - Event occurrence and predictors are typically assessed on the same schedule
- Typically can't set the data collection schedule to coincide with event occurrence for everyone still at risk

Practical implications

- If you're interested in TV predictors, research design is crucial—Don't wait until the data are collected
- Non-reversible dichotomies—that themselves represent event occurrence—are easiest (eg, 1st marriage, HS graduation)
- Reversible dichotomies and continuous predictors usually require imputation (discussed in Section 15.1.2 and 15.1.3*)*

Data example: Time-varying predictors that are non-reversible dichotomies

(ALDA, Section 15.1.1, p. 545)

- *Research Question*: **Is use of marijuana and other drugs** (e.g, amphetamines, psychedelics) **a precursor to cocaine use?**
- *Citation*: Burton, Johnson, Ritter, & Clayton (1996).
- *Design*:
  - 1,658 men, drug histories take twice (in 1974 & 1985)
  - 382 (23.0%) started using cocaine between ages 17 and 41.
- *Three time-invariant predictors*:
  - **EARLYMJ** and **EARLYOD** indicate whether the respondent had initiated marijuana (7.2%) or other drugs (3.7%) so early that he could be characterized as a previous user at t_0 (age 17)
  - **BIRTHYR** (1961-1985), to account for societal changes (included as a control predictor in every model
- *Four time-varying predictors:* **USEDMJ$_j$, SOLDMJ$_j$ USEDOD$_j$, SOLDOD$_j$**
  - Identifies, at each age t_j , whether the respondent had previously used or sold marijuana or other drugs
  - Conceptually, think about a person-period data set in which these variables switch from 0 to 1 in the relevant year and stay at 1 thereafter.
  - In reality, we do not use a person-period data set but rather computer code in a person-level data set (Section 15.1, p. 547)
- Rather than using *contemporaneous values* of the TV predictors, we *lag* them by one year. Addresses issues of rate- and state-dependence (discussed in Section 12.3.3, *p. 440)*

Interpreting Cox models with time-varying predictors

(ALDA, Table 15.1, p. 548)

A: Only time-invariant predictors
All 3 stat sig.

B: Substitute TV *use* predictors:
- Effects much larger (still sig.)
- Fit much better (use AIC)

	Model A	Model B	Model C	Model D
Parameter Estimates, Asymptotic Standard Errors, and Deviance-Based Hypothesis Tests				
BIRTHYR	0.1551***	0.1074***	0.0849***	0.0835***
	(0.0199)	(0.0215)	(0.0218)	(0.0226)
Marijuana use				
EARLYMJ	1.2171***			0.0753
	(0.1640)			(0.1709)
$USEDMJ_j$		2.5518***	2.4592***	2.4525***
		(0.2810)	(0.2836)	(0.2843)
$SOLDMJ_j$			0.6899***	0.6789***
			(0.1226)	(0.1250)
Other drug use				
EARLYOD	0.7912***			−0.0803
	(0.1962)			(0.2033)
$USEDOD_j$		1.8539***	1.2511***	1.2543***
		(0.1292)	(0.1566)	(0.1572)
$MOREOD_j$			0.7604***	0.7638***
			(0.1307)	(0.1322)
Goodness-of-fit				
−2LL	5277.228	4669.096	4580.537	4580.311
AIC	5283.228	4675.096	4590.537	4594.311
Δ −2LL	247.830***	55.962***	88.559***	0.226 (*ns*)
(*df*)	(3)	(3)	(2)	(2)
Comparison	Null	Null	Model B	Model C

~ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$.

C: Add TV *sales* predictors
- Creates *ordinal variable* when paired with *use* predictors
- Both *use* and *sales* are sig.
- Hazards add up: Someone who both used and sold MJ and OD has a hazard ratio of exp(5.1606)=**164.27!**
- Best fitting model so far

D: Add back time-invariant predictors
- Estimates are not sig.
- D fits no better than C
- Prefer Model C

Note: Diminishing BIRTHYR effects
- Uncontrolled estimate=0.2026
- Drops from .1551 to 0.0849 from A to C
- Effects previously attributable to BIRTHYR get absorbed by TV drug use (known as *substitution effects*)

Data example: Fitting non-proportional hazards models by including interactions with TIME

(ALDA, Section 15.3, p. 562)

- *Research Question*: **Does provision of comprehensive mental health services reduce the length of adolescent in-patient hospital stays?**
- *Citation*: Foster (2000).
- *Design*:
  - 174 teens admitted to a psychiatric hospital
  - Half *(n=88)* had traditional services (**TREAT=0**); the other half *n=86*) were randomly assigned to the innovative program **TREAT=1**).
  - Tracked for up to 3 months to determine whether and, if so, when they were released

Fitted Cox model for TREAT

	Model A
TREAT	0.1457
	(0.1542)
Goodness-of-fit	
−2LL	1436.628
n parameters	1
AIC	1438.628

Does this non-significant effect mean the treatment has *no effect*?

Perhaps, but not necessarily. **It could be that the effect of the treatment varies over time**

Non-proportional hazards models via interactions with TIME

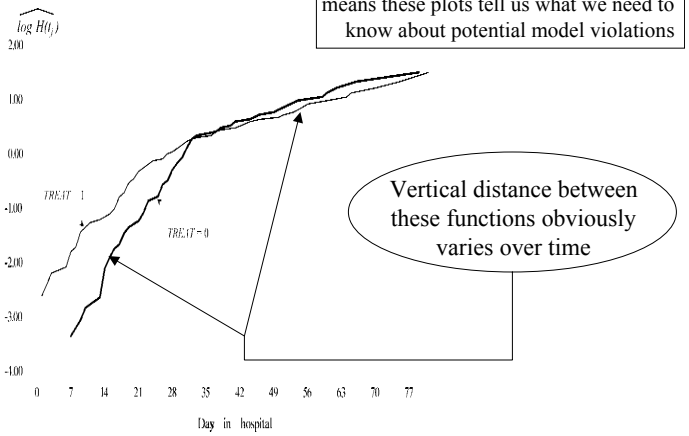
(ALDA, Section 15.3, p. 562, Fig 15.3, p.567)

In every Cox model so far, we've assumed that the *proportional hazards assumption* holds….

Proportional Hazards Assumption ⇒ *A constant difference in the elevation of the log-hazard profile* among groups defined by constant values of the predictor.

Problem: **We cannot plot sample hazard functions** to see if there's a violation

Solution: **Plot sample cumulative hazard functions**. Model equivalence means these plots tell us what we need to know about potential model violations



Vertical distance between these functions obviously varies over time

If the *proportionality assumption* is violated for a predictor, then there is an interaction between the predictor and *TIME*.

Three common specifications for interactions with TIME

(ALDA, Table 15.4, p.566)

Linear
- The effect of TREAT declines smoothly (linearly) over time
- By centering TIME on 1, .7064 is treatment effect on first day of hospitalization

Step-function
- Effect of TREAT differs across epochs (here weeks)
- AIC is superior to B
- Note decline in estimates in first few weeks

	Model B	Model C	Model D
TREAT	0.7064*** (0.0208)		2.5335*** (0.7613)
TREAT × (TIME-1)	−0.0208* (0.0092)		
TREAT1		1.5711* (0.6406)	
TREAT2		0.5678 (0.4929)	
TREAT3		0.8497 (0.3627)	
TREAT4		−0.3499 (0.3621)	
TREAT5		−0.7697~ (0.4160)	
TREAT6+		−0.0993 (0.3111)	
TREAT × L2(TIME)			−0.5301** (0.1619)
Goodness-of-fit			
−2LL	1431.374	1417.730	1423.062
n parameters	2	6	2
AIC	1435.374	1429.730	1427.062

exp(2.5335)=12.60 is the estimated hazard ratio on day 1

Estimated log hazard for TREAT declines by .5301 as length of stay doubles (1 to 2, 2 to 4, 4 to 8 etc.)
Day 1=12.60
Day 8 = 2.56
Day 32=0.89

Logarithmic
- Similar to linear but handles tails for TIME
- AIC is superior to C (because of parsimony)