Survival Analysis, Lecture 7

Part 1: Time-dependent variables

Part 2: Frailty models

Aurélien Belot

Inequalities in Cancer Outcomes Network
London School of Hygiene and Tropical Medicine

Intended Learning Outcomes - Time-dependent variables

At the end of this lecture, you should be able to

- Identify Time-Dependent(TD) explanatory variables
- Distinguish between external and internal TD variables
- Write down equations showing how TD variables are incorporated into the Cox model and in the partial likelihood
- Fit survival models with TD variables, and interpret the results

What is a time-dependent variable?

Time-dependent variable

Definition

A time-dependent variable is an explanatory variable which values may change over time. It's also called time-varying or time-updated variable

Examples of time-dependent variable:

- Individuals with cystic fibrosis: some patients may have a lung transplant. The transplant status is a time-updated (binary) variable
- Association between the level of air pollution and asthma attacks: the level of air pollution is a time-dependent (continuous) variable
- Association between blood pressure and heart disease in an observational cohort: blood pressure is a time-updated (continuous) variable

Time-dependent variable

Definition

A time-dependent variable is an explanatory variable which values may change over time. It's also called time-varying or time-updated variable

Examples of time-dependent variable:

- Individuals with cystic fibrosis: some patients may have a lung transplant. The transplant status is a time-updated (binary) variable
- Association between the level of air pollution and asthma attacks: the level of air pollution is a time-dependent (continuous) variable
- Association between blood pressure and heart disease in an observational cohort: blood pressure is a time-updated (continuous) variable

Analysis of a Time-Dependent variable: why is it important?

Illustrative example: the Stanford Heart Transplant programme

- Group of patients eligible for heart transplant
- When a donor becomes available, physicians choose transplant recipients according to various medical criteria
- We want to evaluate whether receiving a heart transplant is beneficial for patients

Naive analysis: Consider those who received a heart transplant as treated (others being not treated), and compare between these 2 groups the survival times from the date they entered the study. What's wrong?

Analysis of a Time-Dependent variable: why is it important?

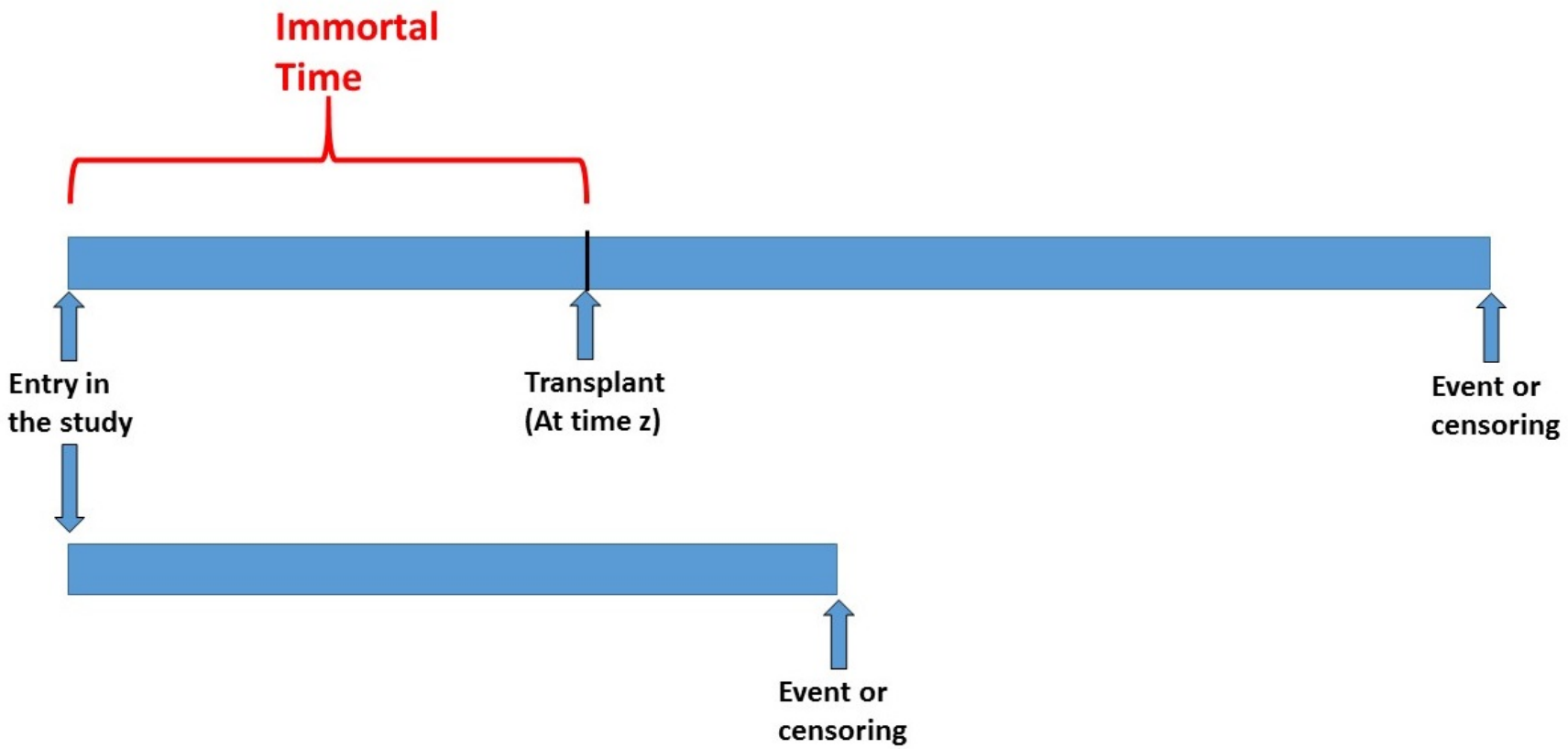
Immortal Time Bias

Using Time-fixed variable (Ever-treated *vs.* not-treated) to model time-varying treatment induces systematic bias toward a "Protective" effect

This "Immortal Time Bias" is due to misclassification of the true exposure (before transplantation) when using the time-fixed variable:

- An ever-treated subject has (by definition) to "survive" until his treatment initiation time z, i.e. is effectively "immortal" until time z
- Using time-fixed variable incorrectly "credits" this survival time to the treatment while the subject was not treated during that time (for $0 < t < z$)

Analysis of a Time-Dependent variable: why is it important?



$\Rightarrow$ Time-varying variables are necessary to avoid immortal time bias

Time-dependent variable

External (exogeneous) or internal (endogeneous) TD variable

- A time-dependent variable is **external** if its knowledge does not require the patient to be alive.
- **Internal** TD variables can only be measured when an individual is alive and still in the study and they cannot be determined without contact with the patient.

Caution in interpretation for Internal TD variable (More on this at the end of part 1)

Time-dependent variable

External (exogeneous) or internal (endogeneous) TD variable

- A time-dependent variable is **external** if its knowledge does not require the patient to be alive.
- **Internal** TD variables can only be measured when an individual is alive and still in the study and they cannot be determined without contact with the patient.

Caution in interpretation for Internal TD variable (More on this at the end of part 1)

Time-dependent variable

External (exogeneous) or internal (endogeneous) TD variable

- A time-dependent variable is **external** if its knowledge does not require the patient to be alive.
- **Internal** TD variables can only be measured when an individual is alive and still in the study and they cannot be determined without contact with the patient.

Caution in interpretation for Internal TD variable (More on this at the end of part 1)

Analysis of a Time-Dependent variable

Extension of the Cox model

Let's define $x(t)$ as the time-dependent variable.
The Cox model could be extended to accommodate time-dependent variables:

$$h(t; x(t)) = h_0(t) \exp(\beta x(t))$$

Note: In this formulation, we assume that only the current value of the TD variable at time t affects the hazard

Interpretation

The following Cox model was fitted, where $x(t)$ is the time-dependent variable denoting transplant status (yes/no) :

$$h(t; x(t)) = h_0(t) \exp(\beta x(t))$$

Exercise 7.1: Suppose patient r is transplanted at time 5, and patient s is never transplanted. What is the ratio of hazards for these two individuals

(a) at time 4?

(b) at time 6?

(we assume both individuals are at risk at times 4 and 6)

Interpretation

(a) at time 4:
For patients r and s the hazard is the same: $h(t; x(t) = 0) = h_0(t)$.
So the hazard ratio is 1 at time 4.

(b) at time 6:
For patients r the hazard is $h(t; x(t) = 1) = h_0(t) \exp(\beta)$. For patient s, the hazard is $h(t; x(t) = 0) = h_0(t)$.
So the hazard ratio is $\exp(\beta)$ at time 6.

Interpretation

More generally, the hazard ratio between two individuals r and s is given by the following formula:

$$\frac{h(t; x_r(t))}{h(t; x_s(t))} = \frac{h_0(t)\exp(\beta x_r(t))}{h_0(t)\exp(\beta x_s(t))} = \exp(\beta(x_r(t) - x_s(t)))$$

So β is the log hazard ratio for 2 individuals r and s whose variable differs by 1 unit at time t.

Note: This hazard ratio depends on time t, such that we can no longer call this model a "Proportional Hazard model".

Data management for the analysis of a Time-Dependent variable

Before fitting a Cox model with a TD variable, a step of data-management is necessary.

- This step simply requires a good understanding of what a TD variable means
- Then we can build an appropriate dataset for the analysis

.

Key to understanding: when the TD variable changes, the hazard changes.
So the analyst needs to create a new row each time the TD variable changes.

Data management for the analysis of a Time-Dependent variable

Illustration with the data from the Stanford Heart Transplant programme

Individual	Date of study entry	Date of transplant	Date of event or censoring	Indicator of event or censoring
6	13jun1968	.	15jun1968	1
7	12jul1968	31aug1968	17may1970	1
8	01aug1968	.	09sep1968	1
.	.	.	.	
.	.	.	.	

The data-management step involves creating a new row for each patient receiving an heart transplant.

Data management for the analysis of a Time-Dependent variable

Illustration with data from the Stanford Heart Transplant programme (`Stata users`/`R users`)

Individual	`_t0`/`tstart`	`_t`/`tstop`	`post`/`trt`	`_d`/`death`
6	0	0.00547	0	1
7	0	0.13689	0	0
7	0.13689	1.84531	1	1
8	0	0.10677	0	1
.	.	.	.	
.	.	.	.	

Exercise 7.2: What is contained within `_t0`/`tstart`, `_t`/`tstop`, `post`/`trt`, `_d`/`death`?

Data management for the analysis of a Time-Dependent variable

Exercise 7.2: What is contained within `_t0/tstart`, `_t/tstop`, `post/trt`, `_d/death`?

`_t0/tstart` records the time origin and `_t/tstop` records the survival or censoring time. The time origin is updated on each new row of data for a given individual.
So in the pre-transplant period `_t0/tstart` records the time of entry to the study (which is always 0 because the time scale is time-in-study), and in the post-transplant period, `_t0/tstart` records (just after) the time of transplant
Similarly, in the pre-transplant period, `_t/tstop` records the time of transplant if the person had a transplant, or the time of death or censoring if this occurred before they had the chance to receive a transplant.

Data management for the analysis of a Time-Dependent variable

Exercise 7.2: What is contained within `_t0/tstart`, `_t/tstop`, `post/trt`, `_d/death`?

`_t0/tstart` records the time origin and `_t/tstop` records the survival or censoring time. The time origin is updated on each new row of data for a given individual.
So in the pre-transplant period `_t0/tstart` records the time of entry to the study (which is always 0 because the time scale is time-in-study), and in the post-transplant period, `_t0/tstart` records (just after) the time of transplant
Similarly, in the pre-transplant period, `_t/tstop` records the time of transplant if the person had a transplant, or the time of death or censoring if this occurred before they had the chance to receive a transplant.

Data management for the analysis of a Time-Dependent variable

Exercise 7.2: What is contained within `_t0`/`tstart`, `_t`/`tstop`, `post`/`trt`, `_d`/`death`?

`post`/`trt` is the indicator of whether an individual is in the pre- or post- transplant period.
For example, person 7 entered the study on 12 of July 1968 and then had a transplant on 31 August 1968 (*ie* 0.13689 year latter). So between those 2 dates they are recorded as pre-transplant (`post`/`trt`=0). After their transplant they were observed up until 17 may 1970 and during this period were recorded as being post-transplant (`post`/`trt`=1).

Data management for the analysis of a Time-Dependent variable

Exercise 7.2: What is contained within `_t0`/`tstart`, `_t`/`tstop`, `post`/`trt`, `_d`/`death`?

`_d`/`death` is the event indicator. Each individual can have more than one row of data, to account for the time before and after transplant. `_d`/`death` is 1 for an individual in the time interval which ends with the event of interest (death), but it is 0 in earlier intervals.
Each individual only has one row of data which has `_d`/`death` equals to 1.

Analysis of a Time-Dependent variable

Partial Likelihood

$$L_P(\beta) = \prod_j \frac{\exp(\beta^\top x_{i_j}(t_j))}{\sum_{k \in R_j} \exp(\beta^\top x_k(t_j))}$$

Note: we need to know the value of the TD variable at each event-time for patients at-risk:

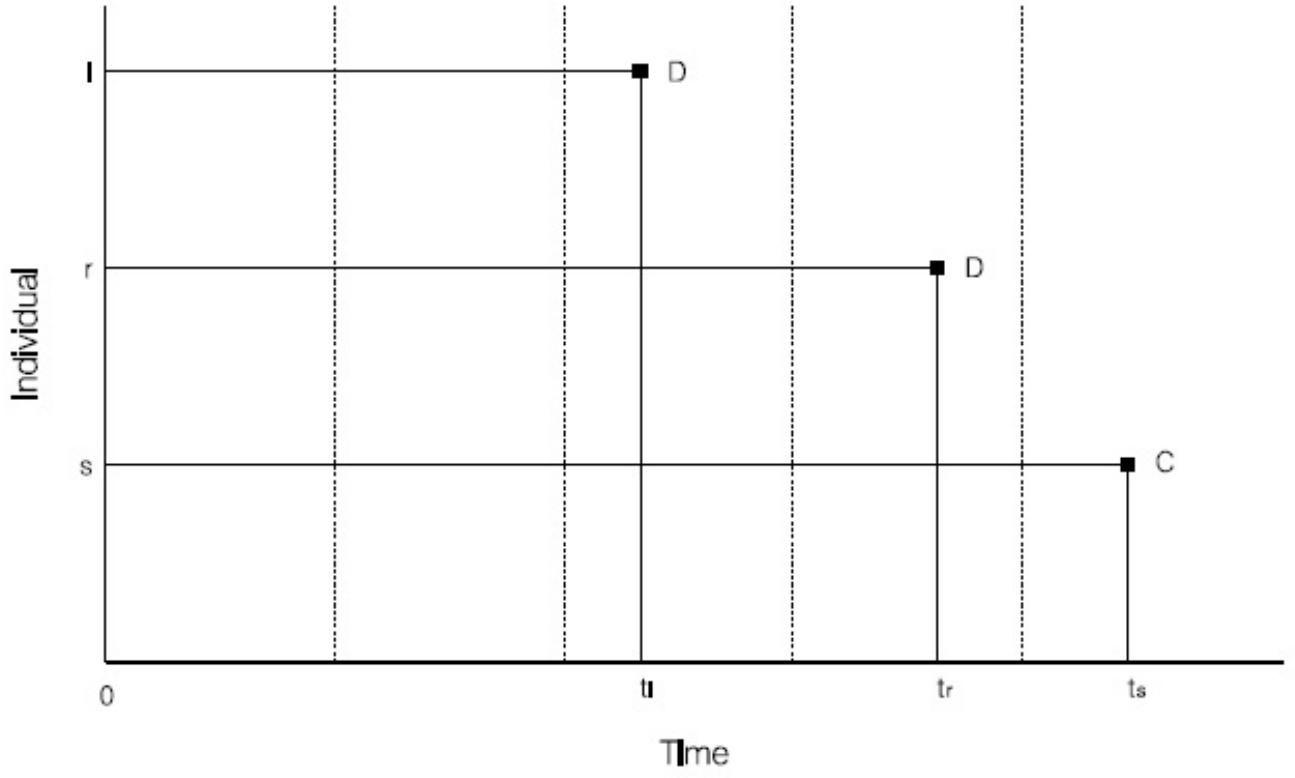
- Not a problem for a binary TD variable that can change only once (e.g. surgery yes/no)
- May be more problematic for a continuous TD variable (e.g. a biomarker), or for a binary variable that can change many times in successive period (under treatment/ not treated/ under treatment . . .)

Illustration of the partial likelihood

- Let's consider a study aiming at quantifying the association between a continuous biomarker and the mortality hazard.
- The biomarker is measured at regular intervals, and the analysis should use the time-updated value of the biomarker (which is more likely to influence the mortality hazard at time t than the measure made at entry) .

Illustration of the partial likelihood

Let's say that individual i died at time t_i, and that only two other individuals are still in the risk-set (individuals r and s) .



What is the contribution of the i^{th} individual to the partial likelihood?

Illustration of the partial likelihood

Contribution of the i^{th} individual to the partial likelihood

$$\frac{\exp(\beta^{\mathsf{T}} x_i(t_i))}{\exp(\beta^{\mathsf{T}} x_i(t_i)) + \exp(\beta^{\mathsf{T}} x_r(t_i)) + \exp(\beta^{\mathsf{T}} x_s(t_i))}$$

- It shows that the values of the TD variable is needed for each individual (i, r and s) at the death time t_i.
- By default, the partial likelihood is maximised using the last known value of the TD variable.

Illustration of the partial likelihood

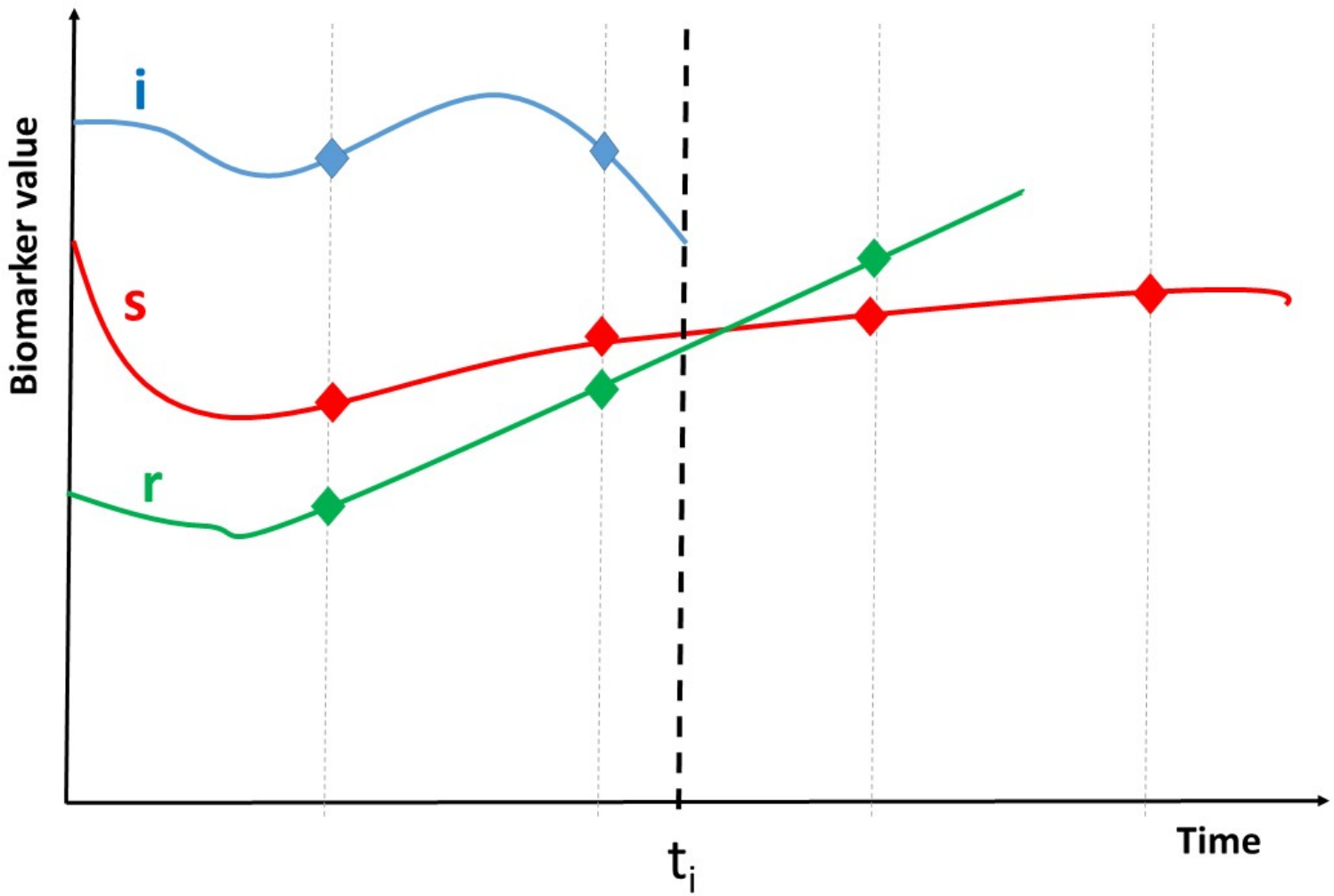


Illustration of the partial likelihood

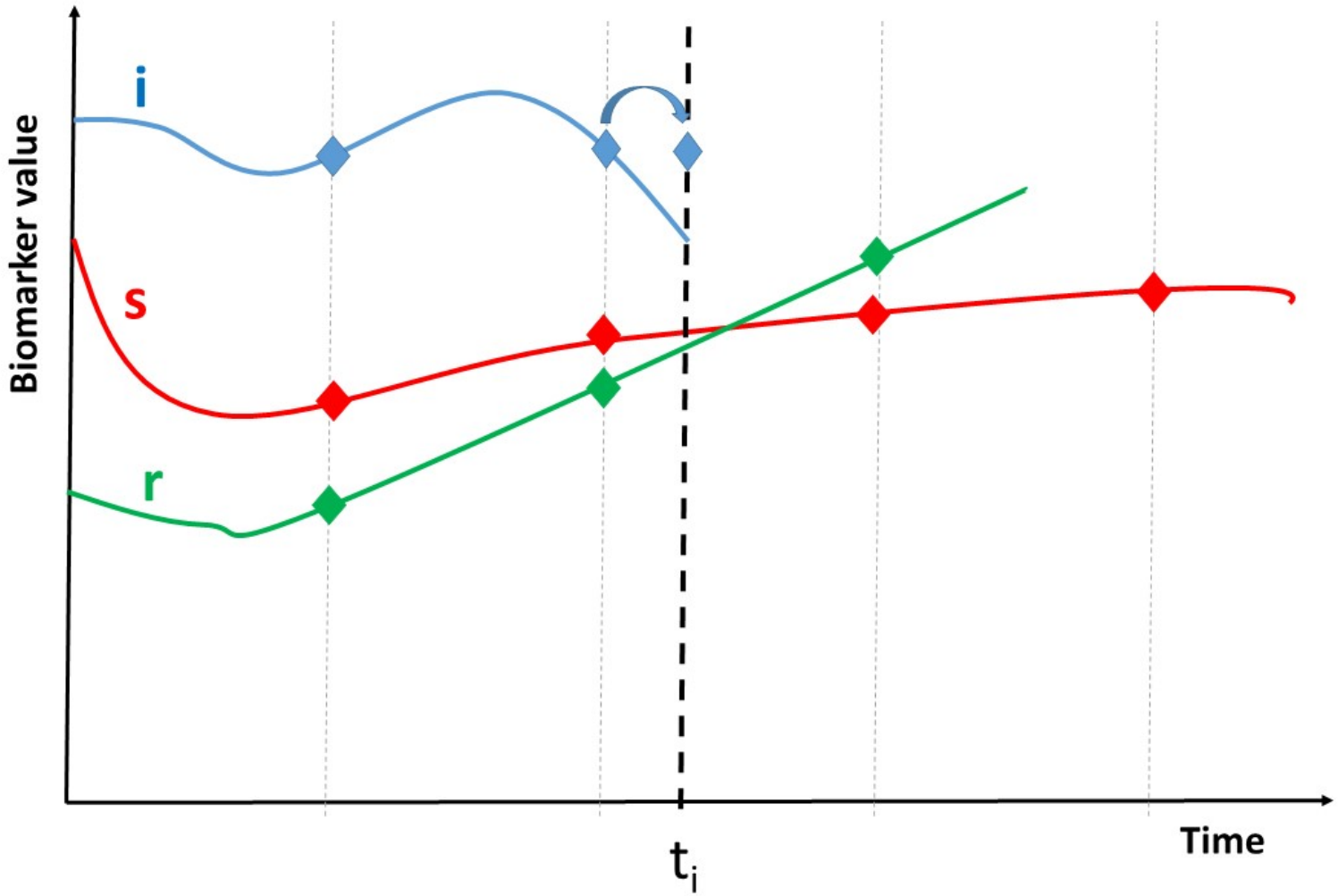


Illustration of the partial likelihood

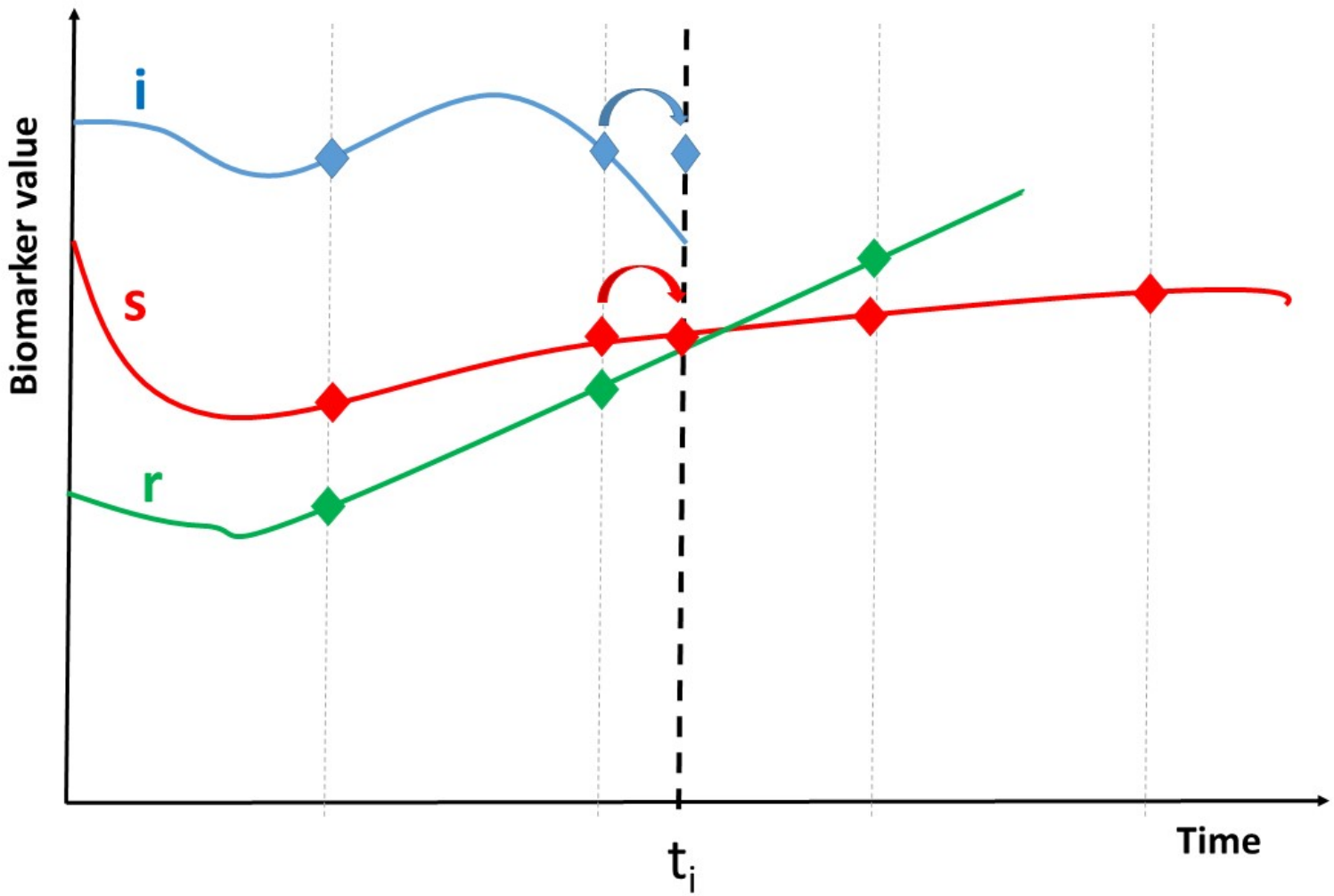


Illustration of the partial likelihood

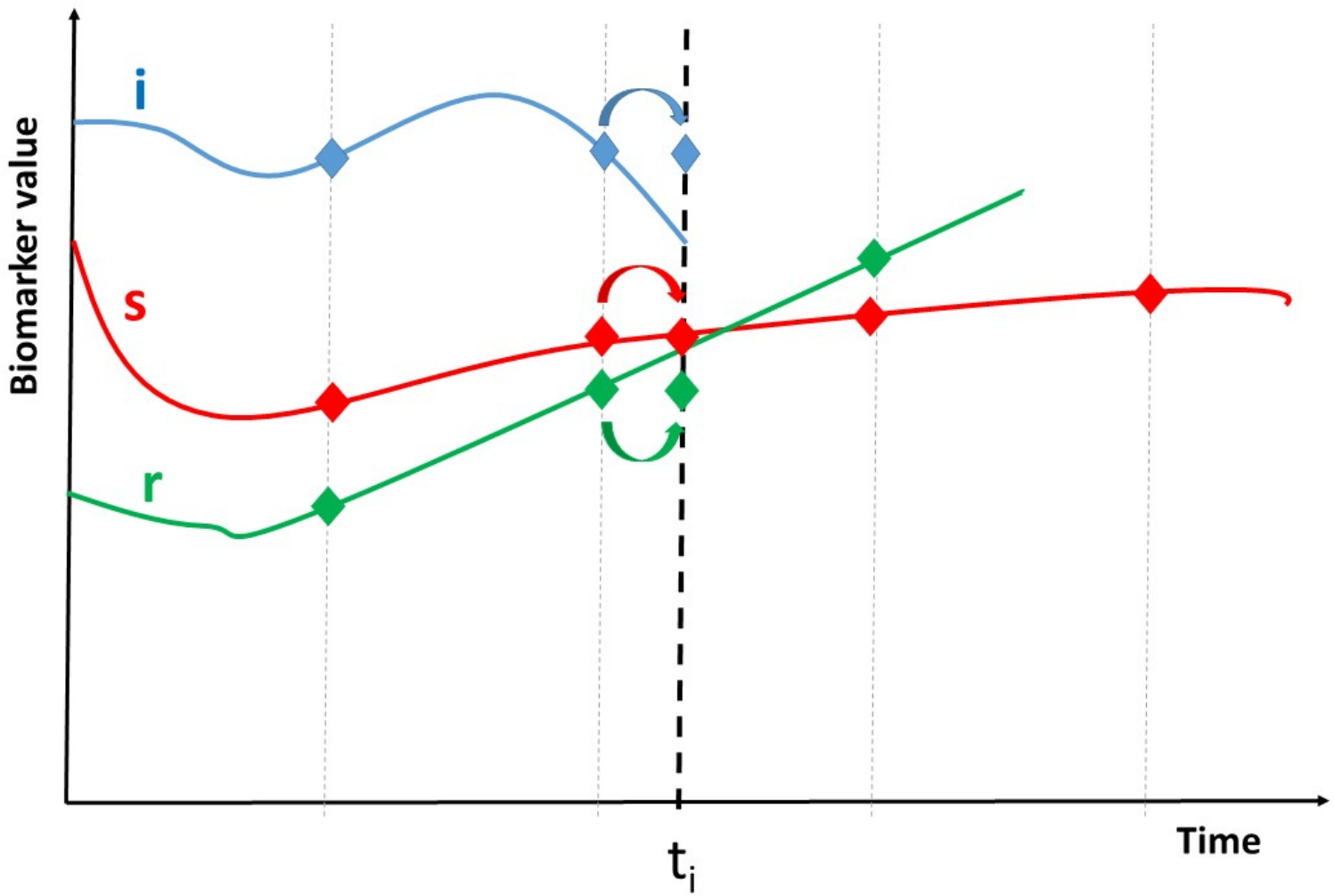
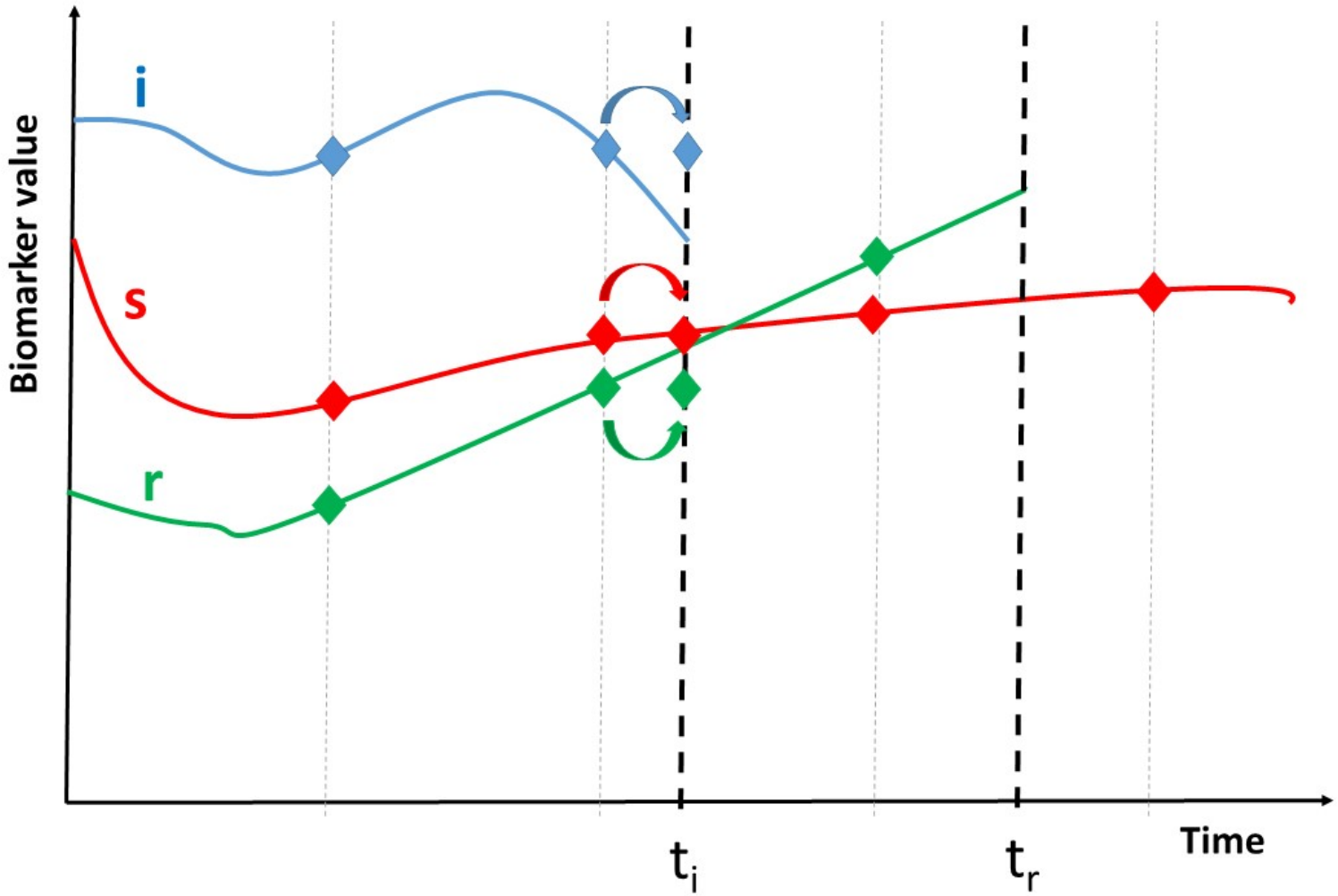


Illustration of the partial likelihood



Extension to other regression models

Parametric models

- Even if dealing with TD variables is an important feature of the Cox model, extending also parametric regression models (Weibull, Exponential) for TD variable is straightforward.
- However, extending the Accelerated Failure Time models is far more difficult (see Kalbfleisch & Prentice's book chapter 7.4.4.)

Moreover, more refinements (than those presented today) are possible, depending on the research question (see the course manual).

Exercise 7.3

Question

The following Cox model was fitted on the Stanford data

$$h(t, trt(t), age) = h_0(t) \exp(\beta_1 trt(t) + \beta_2 age)$$

and $\exp(\hat{\beta}_1)$=0.91, 95% CI [0.49;1.67]
What is the interpretation?

The mortality hazard for individuals receiving transplant is equal to 0.91 the one for individuals not receiving transplant, after adjusting on age of acceptance.
No evidence that heart transplant would decrease the mortality hazard in this population.
Ref: Crowley, J., and Hu, M. (1977). “Covariance Analysis of Heart Transplant Survival Data.” Journal of the American Statistical Association 72:27–36

Exercise 7.3

Question

The following Cox model was fitted on the Stanford data

$$h(t, trt(t), age) = h_0(t) \exp(\beta_1 trt(t) + \beta_2 age)$$

and $\exp(\hat{\beta}_1)$=0.91, 95% CI [0.49;1.67]
What is the interpretation?

The mortality hazard for individuals receiving transplant is equal to 0.91 the one for individuals not receiving transplant, after adjusting on age of acceptance.
No evidence that heart transplant would decrease the mortality hazard in this population.

Ref: Crowley, J., and Hu, M. (1977). “Covariance Analysis of Heart Transplant Survival Data.” Journal of the American Statistical Association 72:27–36

Cautionary notes

For **external TD variables** the classical relationship between the conditional hazard and the conditional survival holds:

$$S(t|x(u), u \leq t) = \exp(-\int_0^t h_0(u) \exp(\beta^\mathsf{T} x(u)) du)$$

However, for **internal TD variables**, this classical relationship is broken: given that we measured the TD variable, the survival probability is 1.

Cautionary notes

Another important points to mention:

Models with TD variables carry a great risk of controlling for variables in the causal pathway.

For example, if the treatment was reducing the risk of Myocardial Infraction by reducing blood pressure, adjusting on the (Time-varying) blood pressure will erroneously mask/cancel the treatment effect.

Part 2 Frailty models

Intended Learning Outcomes - Frailty models

At the end of this lecture, you should be able to

- Distinguish between unshared and shared frailty models
- Write down survival models that include frailty terms, and outline in broad terms how such models are fitted
- Interpret the results of both unshared and shared frailty models
- Fit frailty models in Stata and R

Definition of frailty model

Definition

Frailty models are regression models to analyse time-to-event data which include a random effect

Frailty models are useful in two situations:

1. When you think that there is still heterogeneity between individuals (even after adjusting for covariates): individual (unshared) frailty models
2. When there is a clustering in you data which make unrealistic the assumption of independence between observed survival times (clustered randomised clinical trials, observational study with an hierarchical structure (patients nested in hospitals), repeated events per individual . . . ...): shared frailty models

Definition of frailty model

Definition

Frailty models are regression models to analyse time-to-event data which include a random effect

Frailty models are useful in two situations:

1. When you think that there is still heterogeneity between individuals (even after adjusting for covariates): individual (unshared) frailty models
2. When there is a clustering in you data which make unrealistic the assumption of independence between observed survival times (clustered randomised clinical trials, observational study with an hierarchical structure (patients nested in hospitals), repeated events per individual . . . ...): shared frailty models

Definition of frailty model

Definition

Frailty models are regression models to analyse time-to-event data which include a random effect

Frailty models are useful in two situations:

1. When you think that there is still heterogeneity between individuals (even after adjusting for covariates): individual (unshared) frailty models
2. When there is a clustering in you data which make unrealistic the assumption of independence between observed survival times (clustered randomised clinical trials, observational study with an hierarchical structure (patients nested in hospitals), repeated events per individual . . . ...): shared frailty models

Individual (unshared) frailty model

Motivation

Consider a population with 2 subgroups, with high and low constant mortality hazard

If we don't observe this grouping, we will model a common hazard for the whole population

What would be the shape of this common hazard?

Individual (unshared) frailty model

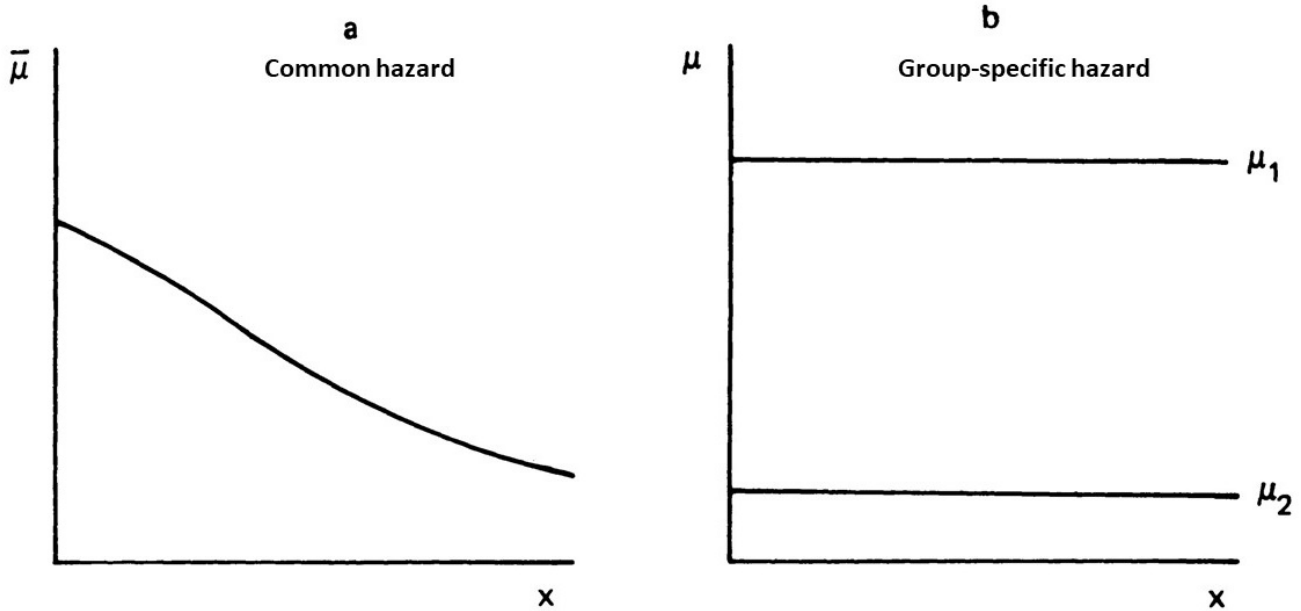
Motivation

Consider a population with 2 subgroups, with high and low constant mortality hazard

If we don't observe this grouping, we will model a common hazard for the whole population

What would be the shape of this common hazard?

(Vaupel and Yashin, The American statistician, Heterogeneity's ruses: some surprising effects of selection on population)



Individual (unshared) frailty model

The conditional hazard is defined as

$$h_i(t; x|\alpha) = \alpha_i h_0(t) \exp(\beta^T x)$$

where α_i is the individual frailty.

- α_i is a random variable (unobserved)
- It is assumed to follow from a specific distribution with positive support (usually gamma) with mean 1 and variance θ
- When $\alpha_i > 1$, individuals experience an increased hazard (they are more frail), and when $\alpha_i < 1$, individuals experience a lower hazard (less frail)

Individual (unshared) frailty model: parameter estimation

Derivation of the likelihood

- Assume a parametric distribution for the (conditional) hazard
- Assume a distribution for the frailty
- Derive the marginal hazard and survival from the conditional hazard and survival
- Maximise the full log-likelihood using those marginal quantities

Individual (unshared) frailty model: parametrisation

Distribution for the (conditional) hazard

- exponential
- Gompertz
- Weibull
- Log-normal
- ...

Distribution for the frailty

- Gamma
- Inverse Gaussian

Using one of those 2 distributions, we get a closed (analytical) form for 2 the marginal quantities (hazard and survival)

Individual (unshared) frailty model: Debates

- There are debates about the usefulness of individual (unshared) frailty models
- Some authors argued that they are of little practical use, mainly because one can not distinguish (from the data) if there is an unmeasured heterogeneity or if the model is misspecified (as it should include a time-dependent effect)...

See O'Quigley and Stare, Statistics in Medicine, 2002, and http://data.princeton.edu/pop509/UnobservedHeterogeneity.pdf for a discussion of this.

Shared frailty models

- All of the regression models considered until now assume independence between observed survival times
- In some cases, this assumption may be unrealistic, like in clustered randomised clinical trials, in observational study with an hierarchical structure (patients nested in hospitals), etc...

Shared frailty models

The conditional shared hazard is defined as

$$h_{ij}(t; x_{ij}|\alpha_j) = \alpha_j h_0(t) \exp(\beta^\mathsf{T} x_{ij}) = h_0(t) \exp(\beta^\mathsf{T} x_{ij} + w_j)$$

- where α_j is the frailty shared by individuals from cluster j
- Equivalently, w_j is the random effect shared by individuals from cluster j
- α_j (or w_j) is assumed to follow a specific distribution

Shared frailty models

Classical assumptions:

- Parametric distribution for T (Weibull, piecewise constant,...)
- Gamma distribution for the frailty α

Mainly due to practical reasons (analytical expression of the marginal likelihood)
But other possibilities are feasible, as for example assuming w_j to follow a normal distribution (mean 0 and variance σ^2)

Likelihood function: overview

Overview of the process to define the (log)-likelihood

1. Likelihood of one observation *i* in cluster *j*
2. Conditional Likelihood for cluster *j*
3. Marginal Log-Likelihood for cluster
4. Total Log-likelihood

See the notes for more details.

Example: Infection in kidney dialysis

Exercice 7.4

- The data contain times to recurrence of infection due to catheter insertion for kidney patients using portable dialysis equipment (38 patients)
- We may observe more than one infection time per patient.
- The aim is to investigate the association between age and gender with time to recurrence
- We fitted 2 models, a standard Cox model, and a shared frailty Cox model

Example: Infection in kidney dialysis

Results of the standard Cox model

Variable	HR	Std.Err	95 % CI
Age	1.00	0.01	[0.98, 1.02]
Female	0.45	0.13	[0.25, 0.81]

Results of the shared frailty Cox model

Variable	HR	Std.Err	95 % CI
Age	1.00	0.01	[0.98, 1.03]
Female	0.21	0.10	[0.08, 0.51]
θ	0.48	0.27	

LR Test of $\theta = 0$: χ^2=6.27, p=0.006

What is your interpretation of these results?

Example: Infection in kidney dialysis

The standard Cox model

This model is a standard Cox model with sex and age as explanatory variables

- Sex is strongly associated with reduced risk
- No evidence that the hazard ratio for age is different from 1

The shared frailty Cox model

This model takes account of the correlation between times to infection arising from the same individual

- θ is the variance of the frailty parameter and its estimate here is 0.48. The LR test shows strong evidence of intra-individual correlation (equivalently between-individuals heterogeneity): p-value = 0.006
- No evidence that the hazard ratio for age is different from 1
- The hazard ratio for sex moves further away from 1 and its confidence interval is narrower

Example: Infection in kidney dialysis

The standard Cox model

This model is a standard Cox model with sex and age as explanatory variables

- Sex is strongly associated with reduced risk
- No evidence that the hazard ratio for age is different from 1

The shared frailty Cox model

This model takes account of the correlation between times to infection arising from the same individual

- θ is the variance of the frailty parameter and its estimate here is 0.48. The LR test shows strong evidence of intra-individual correlation (equivalently between-individuals heterogeneity): p-value = 0.006
- No evidence that the hazard ratio for age is different from 1
- The hazard ratio for sex moves further away from 1 and its confidence interval is narrower