1st R Summer School @ AUEB
Mixed Effects Models & Survival Analysis

Dimitris Rizopoulos
Department of Biostatistics, Erasmus University Medical Center

d.rizopoulos@erasmusmc.nl

Center of Professional Education, A.U.E.B.

June 25-26, 2014, Athens

Contents

III Relative Risk Models 63

IV Practical 106

What is this Part About

- Often we are faced with data collected in follow-up studies

- Longitudinal outcomes
  - ▷ biomarkers, patient parameters, . . .

- Survival outcomes
  - ▷ death, relapse of disease, . . .

What is this Part About (cont'd)

- We will introduce two popular modeling paradigms for analyzing such data:

Mixed Effects Models & Relative Risk Models

Learning Objectives

- **Goals:** After this course participants will be able to
  - ▷ identify settings in which mixed models are required,
  - ▷ construct and fit an appropriate mixed model to the data, and
  - ▷ correctly interpret the obtained results

- The course will be explanatory rather than mathematically rigorous
  - ▷ emphasis is given on sufficient detail in order for participants to obtain a clear view on the different mixed modeling approaches, and how they should be used in practice

Agenda

- **Part I:** Introduction
  - ▷ Data sets that we will use throughout the course

- **Part II:** Review of Linear Mixed Models
  - ▷ Features of repeated measurements data
  - ▷ Naive approaches
  - ▷ Linear mixed models

Agenda (cont'd)

- **Part III:** Review of Survival Analysis
  - ▷ Features of survival data
  - ▷ Basic functions in survival analysis
  - ▷ Relative risk models

Structure of the Course & Material

- Lectures & short software practicals using R

- Material:
  - ▷ Course Notes
  - ▷ R code in soft format

- Within the course notes there are several examples of R code which are denoted by the symbol '`R> `'

Software Requirements

- The recent version of R and Rstudio; downloadable from
  - ▷ http://cran.r-project.org/
  - ▷ http://www.rstudio.com/

- No additional packages will be required
  - ▷ we will use the recommended packages **nlme**, **survival** and **lattice**

References

- Standard texts in longitudinal data analysis
  - ▷ Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.
  - ▷ Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.
  - ▷ Fitzmaurice, G., Laird, N., and Ware, J. (2011). *Applied Longitudinal Analysis*, 2nd Ed. Hoboken: Wiley.
  - ▷ Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd edition. New York: Oxford University Press.

References (cont'd)

- Standard texts in survival analysis
  - ▷ Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data, 2nd Ed.*. New York: Wiley.
  - ▷ Therneau, T. and Grambsch, P. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.
  - ▷ Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.
  - ▷ Andersen, P., Borgan, O., Gill, R. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.
  - ▷ Klein, J. and Moeschberger, M. (2003). *Survival Analysis - Techniques for Censored and Truncated Data*. New York: Springer-Verlag.

Part I
Motivating Data Sets

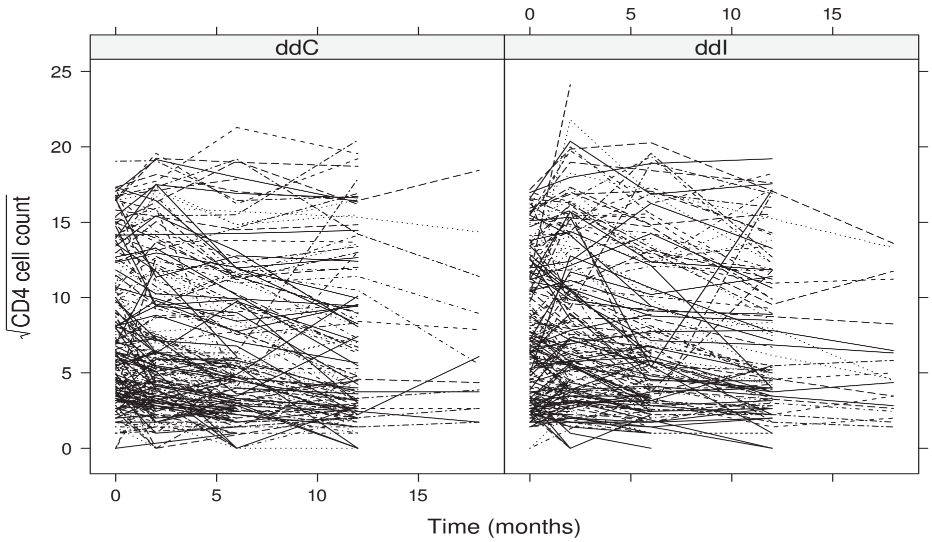
1.1 Motivating Longitudinal Studies

- AIDS: 467 HIV infected patients who had failed or were intolerant to zidovudine therapy (AZT) (Abrams et al., NEJM, 1994)

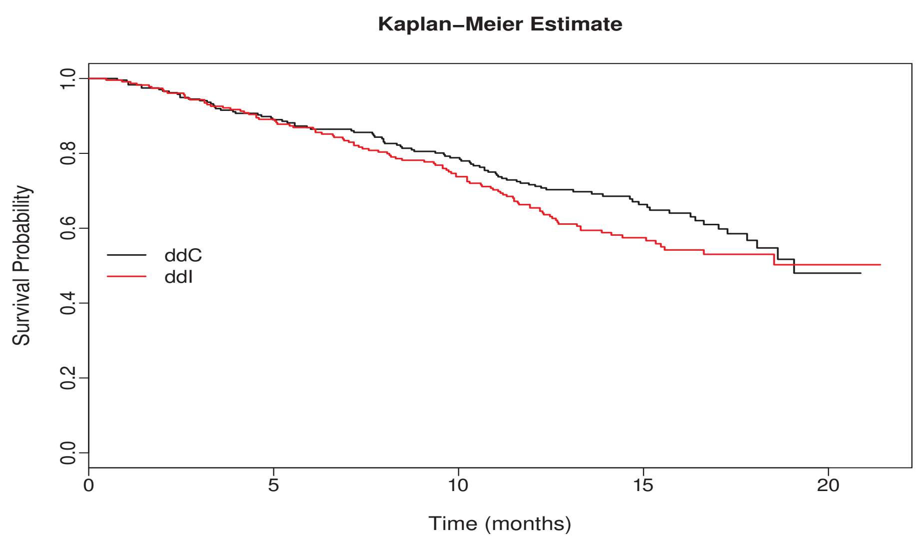
- The aim of this study was to compare the efficacy and safety of two alternative antiretroviral drugs, didanosine (ddI) and zalcitabine (ddC)

- Outcomes of interest:
  - ▷ time to death
  - ▷ randomized treatment: 230 patients ddI and 237 ddC
  - ▷ CD4 cell count measurements at baseline, 2, 6, 12 and 18 months
  - ▷ prevOI: previous opportunistic infections

1.1 Motivating Longitudinal Studies (cont'd)



1.1 Motivating Longitudinal Studies (cont'd)



1.1 Motivating Longitudinal Studies (cont'd)

- Research Questions:
  - ▷ How CD4 cell count evolves in time for this cohort of patients?
  - ▷ Does treatment improve average longitudinal evolutions?
  - ▷ How strong is the association between CD4 cell count and the risk for death?

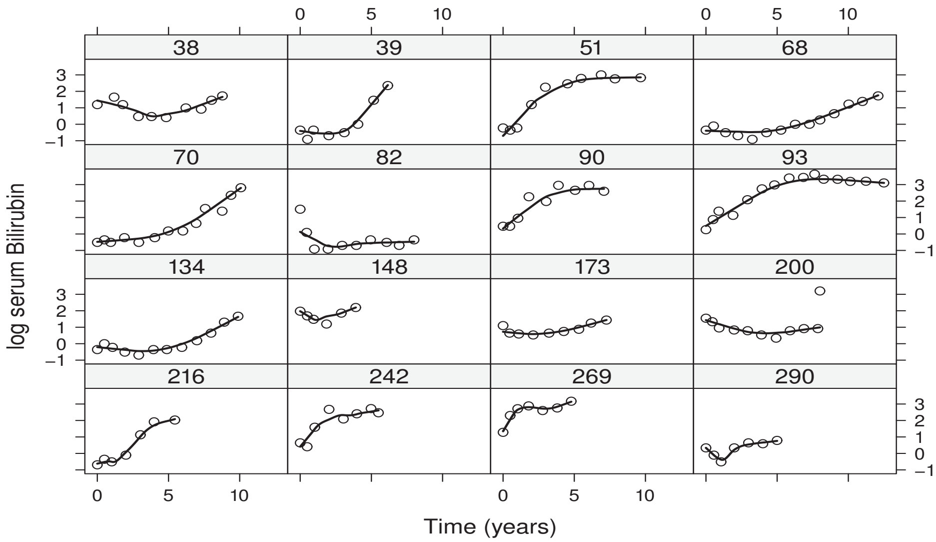
1.1 Motivating Longitudinal Studies (cont'd)

- PBC: Primary Biliary Cirrhosis:
  - ▷ a chronic, fatal but rare liver disease
  - ▷ characterized by inflammatory destruction of the small bile ducts within the liver

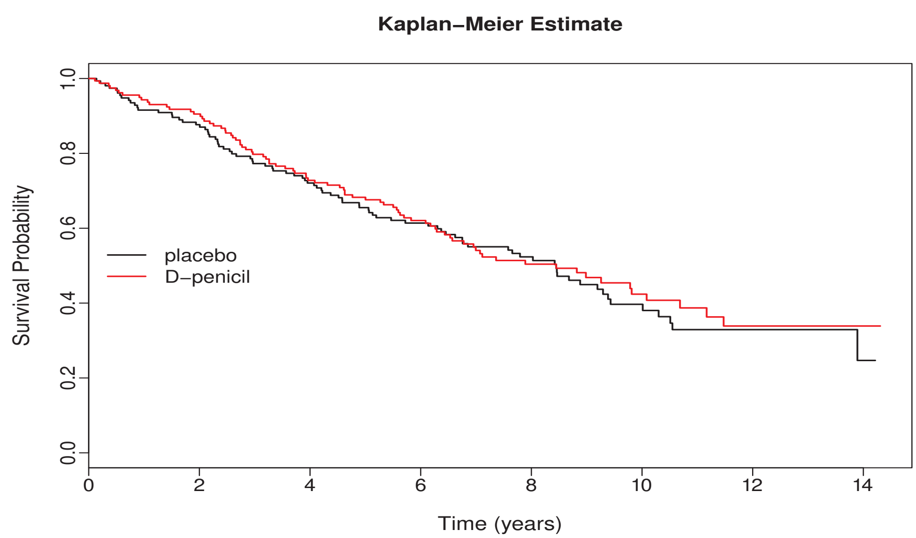
- Data collected by Mayo Clinic from 1974 to 1984 (Murtaugh et al., Hepatology, 1994)

- Outcomes of interest:
  - ▷ time to death and/or time to liver transplantation
  - ▷ randomized treatment: 158 patients received D-penicillamine and 154 placebo
  - ▷ longitudinal serum bilirubin levels

1.1 Motivating Longitudinal Studies (cont'd)



1.1 Motivating Longitudinal Studies (cont'd)



1.1 Motivating Longitudinal Studies (cont'd)

- Research Questions:
  - ▷ Do men have higher serum bilirubin during follow-up than women?
  - ▷ Is there a difference in the average longitudinal evolutions of serum bilirubin when we correct for age differences at baseline and gender differences during follow-up?
  - ▷ How strong is the association between bilirubin and the risk for death?
  - ▷ How the observed serum bilirubin levels could be utilized to provide predictions of survival probabilities?

Part II

Linear Mixed-Effects Models

2.1 Features of Longitudinal Data

- Repeated evaluations of the same outcome in each subject in time
  - ▷ CD4 cell count in HIV-infected patients
  - ▷ serum bilirubin in PBC patients

- Visiting process
  - ▷ some times fixed by design (e.g., in randomized trials) but often not everybody adheres to them
  - ▷ completely determined by the physicians and/or the patients

2.1 Features of Longitudinal Data (cont'd)

Measurements on the same subject are expected to be (positively) correlated

- This implies that standard statistical tools, such as the t-test and simple linear regression that assume independent observations, are not optimal for longitudinal data analysis

2.1 Features of Longitudinal Data (cont'd)

- Let's see why: The simplest case of longitudinal data are paired data

- Example: We consider the baseline and 6-month longitudinal measurements of square root CD4 cell count from the AIDS dataset

	n	mean	sd
$month = 0$	294	7.73	4.69
$month = 6$	294	6.71	4.96

2.1 Features of Longitudinal Data (cont'd)

- There is an average decrease of about 1 unit

- The classical analysis of paired data is based on comparisons within subjects:

$$\Delta_i = Y_i(t = 0) - Y_i(t = 6), \qquad i = 1, \ldots, n$$

- A positive Δ_i corresponds to a decrease of the square root CD4 cell count, while a negative Δ_i is equivalent to an increase

- Testing for a time effect is now equivalent to testing whether the average difference μ_Δ equals zero

2.1 Features of Longitudinal Data (cont'd)

- The paired t-test yields

```
        Paired t-test

data:  CD4 by obstime
t = 6.472, df = 293, p-value = 4.057e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.7105585 1.3315439
sample estimates:
mean of the differences
               1.021051
```


2.1 Features of Longitudinal Data (cont'd)

- What if we had ignored the paired nature of the data?

- We then could have used a two-sample (unpaired) t-test to compare the average CD cell count at the two time points

```
        Welch Two Sample t-test

data:  CD4 by obstime
t = 2.565, df = 584.229, p-value = 0.01056
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2392406 1.8028617
sample estimates:
mean in group 0 mean in group 6
       7.730128        6.709077
```

2.1 Features of Longitudinal Data (cont'd)

- We would still have found a significant difference (p = 0.0106), but the p-value would have been many times larger compared to the one obtained using the paired t-test

- The two-sample t-test does not take into account the fact that the measurements are not independent observations

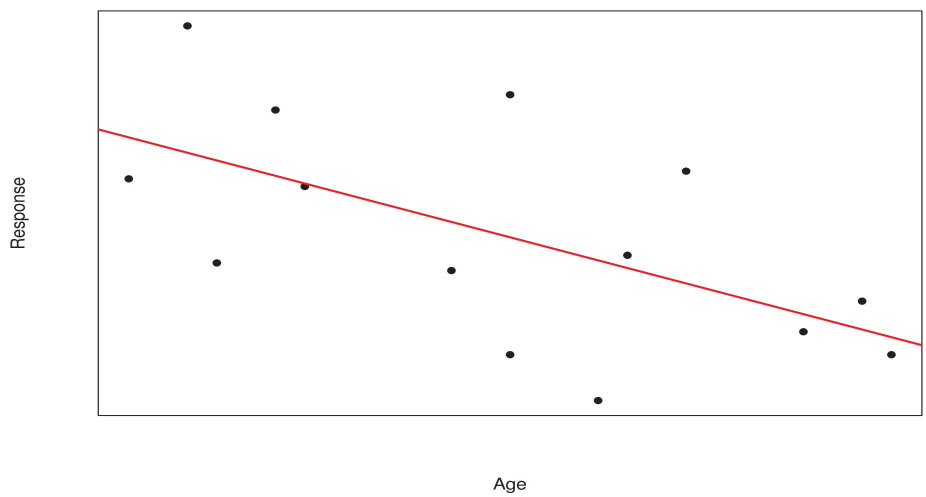
- This illustrates that classical statistical models which assume independent observations will not be valid for the analysis of longitudinal data

2.1 Features of Longitudinal Data (cont'd)

- Longitudinal studies allow to investigate
  1. how treatment means differ at specific time points, e.g., at the end of the study (*cross-sectional effect*)
  2. how treatment means or differences between means of treatments change over time (*longitudinal effect*)

- An example: Suppose it is of interest to study the relation between some response Y and age
  - ▷ a cross-sectional study yields the following data:

2.1 Features of Longitudinal Data (cont'd)

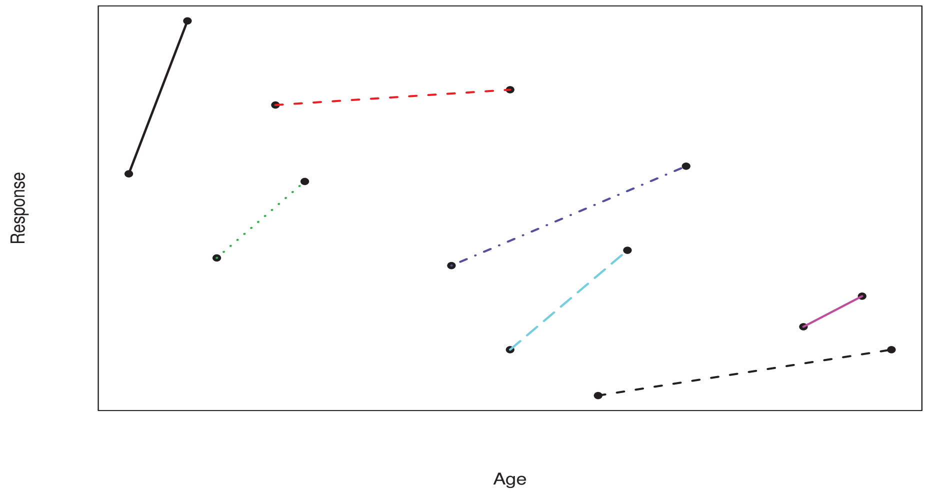


2.1 Features of Longitudinal Data (cont'd)

- The graph clearly suggests a negative relation between Y and age

- **Nevertheless**, exactly the same observations also could have been obtained in a longitudinal study, with 2 measurements per subject

2.1 Features of Longitudinal Data (cont'd)



2.1 Features of Longitudinal Data (cont'd)

Are we now still inclined to conclude that there is a negative relation between Y and age?

- <u>Conclusion:</u> Longitudinal data allow to distinguish differences between subjects from changes within subjects

2.2 Simple Methods

- The reason why classical statistical techniques fail in the context of longitudinal data is that observations within subjects are correlated
  - ▷ often the correlation between two repeated measurements decreases as the time span between those measurements increases

- The paired t-test accounts for this by considering subject-specific differences $\Delta_i = Y_{i1} - Y_{i2}$
  - ▷ this reduces the number of measurements to just one per subject, which implies that classical techniques can be applied again

2.2 Simple Methods (cont'd)

- In the case of more than 2 measurements per subject, similar simple techniques are often applied to reduce the number of measurements for the ith subject, from n_i to 1
  - ▷ Analysis at each time point separately
  - ▷ Analysis of Area Under the Curve (AUC)
  - ▷ Analysis of endpoints
  - ▷ Analysis of increments

2.2 Simple Methods (cont'd)

- **Analysis at each time point separately**
  - ▷ **General idea:** The data are analyzed at each occasion separately
  - ▷ **Advantages:**
    - * simple to interpret
    - * uses all available data

    Disadvantages:
    - * does not consider 'overall' differences
    - * does not allow to study the evolution of differences
    - * problem of multiple testing
    - * possible problems with missing data

2.2 Simple Methods (cont'd)

- **Analysis of area under the curve (AUC)**
  - ▷ **General idea:** For each subject, the area under her curve is calculated

$$\text{AUC}_i = (t_{i2} - t_{i1}) \times (y_{i2} + y_{i1})/2 + (t_{i3} - t_{i2}) \times (y_{i3} + y_{i2})/2 + \ldots$$

    Afterwards, these AUCs are analyzed

  - ▷ **Advantages:**
    - * no problems of multiple testing
    - * does not explicitly assume balanced data
    - * compares 'overall' differences

2.2 Simple Methods (cont'd)

- **Analysis of area under the curve (AUC)**
  - ▷ **Disadvantages:**
    - * uses only partial information
    - * possible problems with missing data

2.2 Simple Methods (cont'd)

- **Analysis of endpoints**
  - ▷ **General idea:** Assess differences only on the last time point
  - ▷ **Advantages:**
    - * no problems of multiple testing
    - * does not explicitly assume balanced data

    Disadvantages:
    - * applicable only in randomized trials
    - * does not consider 'overall' differences
    - * possible problems with missing data

2.2 Simple Methods (cont'd)

- **Analysis of increments**
  - ▷ **General idea:** A simple method to compare evolutions between subjects, correcting for differences at baseline, is to analyze the subject-specific changes $y_{in_i} - y_{i1}$
  - ▷ **Advantages:**
    - * no problems of multiple testing
    - * does not explicitly assume balanced data

    Disadvantages:
    - * uses partial information
    - * possible problems with missing data

2.2 Simple Methods (cont'd)

- The AUC, endpoints and increments are examples of summary statistics
  - ▷ such summary statistics summarize the vector of repeated measurements for each subject separately
- This leads to the following general procedure:
  - ▷ **Step 1:** Summarize the data of each subject into one statistic
  - ▷ **Step 2:** Analyze the summary statistics, e.g. analysis of covariance to compare groups after correction for important covariates
- This way, the analysis of longitudinal data is reduced to the analysis of independent observations, for which classical statistical procedures are available

2.2 Simple Methods (cont'd)

- However, all these methods have the disadvantage that (lots of) information is lost

This has led to the development of statistical techniques that overcome these disadvantages

2.3 The Linear Mixed Model

- The direct approach to model longitudinal data ⇒ *multivariate regression*

$$y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, V_i),$$

where

  ▷ y_i the vector of responses for the ith subject

  ▷ X_i design matrix describing structural component

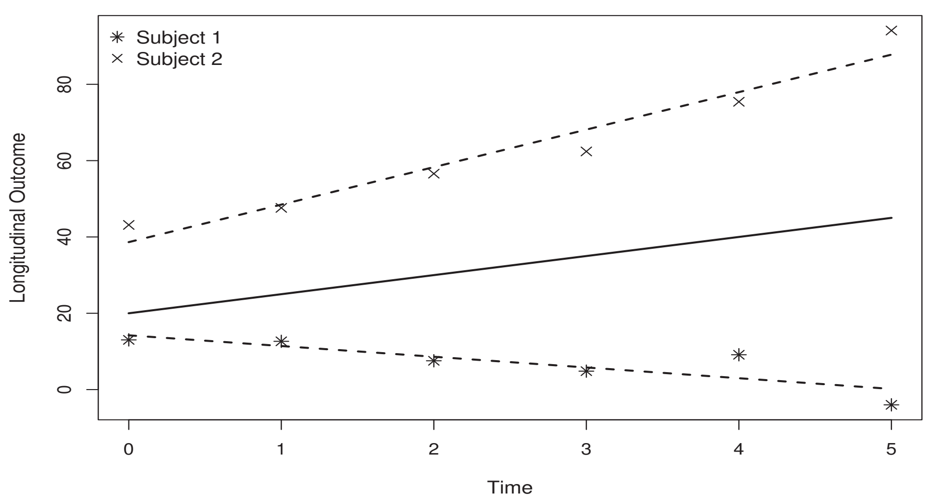
  ▷ V_i covariance matrix describing the correlation structure

- There are several options for modeling V_i, e.g., compound symmetry, autoregressive process, exponential spatial correlation, Gaussian spatial correlation, ...

2.3 The Linear Mixed Model (cont'd)

- **Alternative intuitive approach:** Each subject in the population has her own subject-specific mean response profile over time

2.3 The Linear Mixed Model (cont'd)



2.3 The Linear Mixed Model (cont'd)

- The evolution of each subject in time can be described by a linear model

$$y_{ij} \;=\; \tilde{\beta}_{i0} + \tilde{\beta}_{i1} t_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

  where

  ▷ y_{ij} the jth response of the ith subject

  ▷ $\tilde{\beta}_{i0}$ is the intercept and $\tilde{\beta}_{i1}$ the slope for subject i

- **Assumption:** Subjects are randomly sampled from a population ⇒ subject-specific regression coefficients are also sampled from a population of regression coefficients

$$\tilde{\beta}_i \sim \mathcal{N}(\beta, D)$$

2.3 The Linear Mixed Model (cont'd)

- We can reformulate the model as

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1}) t_{ij} + \varepsilon_{ij},$$

  where

  ▷ βs are known as the *fixed effects*

  ▷ b_is are known as the *random effects*

- In accordance for the random effects we assume

$$b_i = \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \sim \mathcal{N}(0, D)$$

2.3 The Linear Mixed Model (cont'd)

- Put in a general form

$$\begin{cases} y_i = X_i\beta + Z_i b_i + \varepsilon_i, \\ \\ b_i \sim \mathcal{N}(0, D), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathsf{I}_{n_i}), \end{cases}$$

with

▷ X design matrix for the fixed effects β

▷ Z design matrix for the random effects b_i

▷ $b_i \perp\!\!\!\perp \varepsilon_i$

2.3 The Linear Mixed Model (cont'd)

- Interpretation:

▷ β_j denotes the change in the average y_i when x_j is increased by one unit

▷ b_i are interpreted in terms of how a subset of the regression parameters for the ith subject deviates from those in the population

- Advantageous feature: population + subject-specific predictions

▷ β describes mean response changes in the population

▷ $\beta + b_i$ describes individual response trajectories

2.3 The Linear Mixed Model (cont'd)

- How do the random effects capture correlation:
  - ▷ Given the random effects, the measurements of each subject are independent (*conditional independence assumption*)

$$p(y_i \mid b_i) = \prod_{j=1}^{n_i} p(y_{ij} \mid b_i)$$

  - ▷ Marginally (integrating out the random effects), the measurements of each subject are correlated

$$p(y_i) = \int p(y_i \mid b_i)\, p(b_i)\, db_i \quad \Rightarrow \quad y_i \sim \mathcal{N}(X_i\beta,\ Z_i D Z_i^\top + \sigma^2 \mathsf{I}_{n_i})$$

2.3 The Linear Mixed Model (cont'd)

- Hierarchical formulation
  - ▷ a model for y_i given b_i, and a model for b_i
  - ▷ D is the covariance matrix of the random effects $\Rightarrow$ needs to be positive definite

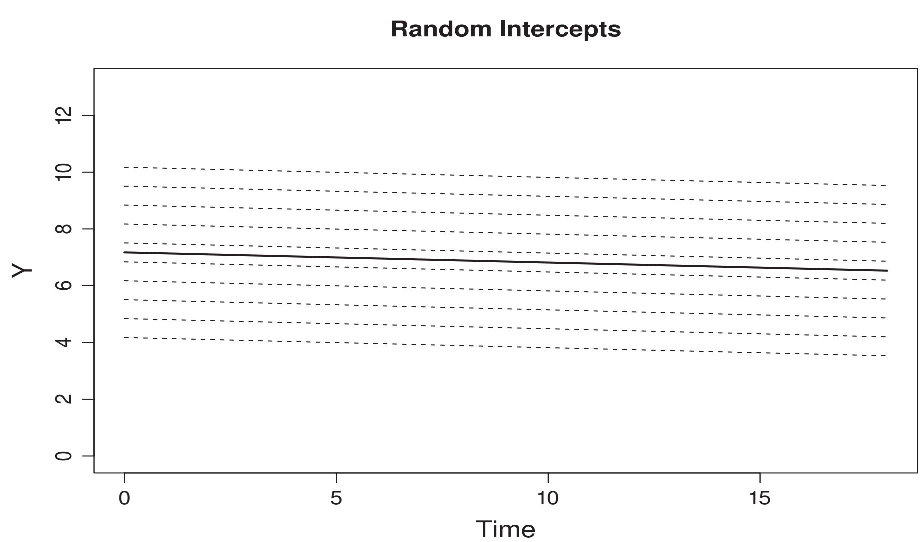
- Marginal formulation
  - ▷ a model for y_i, and a specific form of the marginal covariance matrix $V_i = Z_i D Z_i^\top + \sigma^2 \mathsf{I}_{n_i}$
  - ▷ only V_i needs to be positive definite
  - ▷ V_i can be positive definite without D being positive definite

2.3 The Linear Mixed Model (cont'd)

The hierarchical model implies the marginal one, not vice versa

- A simple example: Random-intercepts model

$$\begin{cases} y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{i0} + \varepsilon_{ij}, \\ \\ b_{i0} \sim \mathcal{N}(0, \sigma_b^2), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2). \end{cases}$$



2.3 The Linear Mixed Model (cont'd)

- Implied marginal covariance matrix has the form

$$V_i = \sigma_b^2 1_{n_i} 1_{n_i}^\top + \sigma^2 \mathsf{I}_{n_i}$$

it assumes

▷ constant variance $\sigma_b^2 + \sigma^2$ over time, and

▷ equal positive correlation $\rho = \sigma_b^2 / (\sigma_b^2 + \sigma^2)$ between the measurements of any two time points (aka *intra-class correlation*)

▷ it is known as the *compound symmetric* covariance matrix

2.3 The Linear Mixed Model (cont'd)

- Note that we could also have a compound symmetric covariance matrix with negative intra-class correlation

▷ such a matrix could never have come from a mixed model

Random intercepts **imply** compound symmetry
but
Compound symmetry **does not imply** random intercepts

2.3 The Linear Mixed Model (cont'd)

- What are the implications of this?

- Statistical software that fit mixed models under ML actually fit the implied marginal model
  - ▷ we can construct examples where two mixed models have exactly the same implied marginal model
  - ▷ based on the fitted model we **cannot** say under which model the data have been generated

- We can only do it under a Bayesian approach (because there we actually fit the hierarchical model)

2.3 The Linear Mixed Model (cont'd)

- Estimation of model parameters
  - ▷ Fixed effects: For known marginal covariance matrix $V_i = Z_i D Z_i^\top + \sigma^2 \mathsf{I}_{n_i}$, the fixed effects are estimated using generalized least squares

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i^\top V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i^\top V_i^{-1} y_i$$

  - ▷ Variance Components: The unique parameters in V_i are estimated based on either maximum likelihood (ML) or restricted maximum likelihood (REML)
    - * REML provides unbiased estimates for the variance components in small samples

2.3 The Linear Mixed Model (cont'd)

- Estimation of random effects
  - ▷ based on a fitted mixed model, estimates for the random effects are based on the posterior distribution:

$$p(b_i \mid y_i; \theta) = \frac{p(y_i \mid b_i; \theta)\, p(b_i; \theta)}{p(y_i; \theta)}$$

$$\propto p(y_i \mid b_i; \theta)\, p(b_i; \theta),$$

in which θ is replaced by its MLE $\hat{\theta}$

2.3 The Linear Mixed Model (cont'd)

- This is a whole distribution
  - ▷ measures of location ⇒ mean, mode
  - ▷ measures of dispersion ⇒ variance, local curvature at the mode

- In the linear mixed model we have seen, this posterior distribution has a closed-form:

$$[b_i \mid y_i; \theta] \sim \mathcal{N}\Big\{DZ_i^\top V_i^{-1}(y_i - X_i\beta),\ DZ_i^\top K Z_i D\Big\},$$

with

$$K = V_i^{-1} - V_i^{-1}X_i\left(\sum_{i=1}^n X_i^\top V_i^{-1} X_i\right)^{-1} X_i^\top V_i^{-1}$$

2.3 The Linear Mixed Model (cont'd)

- Example: We fit a linear mixed model for the AIDS dataset assuming
  - ▷ different average longitudinal evolutions per treatment group (fixed part)
  - ▷ random intercepts & random slopes (random part)

$$\begin{cases} y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 \{\texttt{ddI}_i \times t_{ij}\} + b_{i0} + b_{i1} t_{ij} + \varepsilon_{ij}, \\ \\ b_i \sim \mathcal{N}(0, D), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

- Note: We did not include a main effect for treatment due to randomization

2.3 The Linear Mixed Model (cont'd)

	Value	Std.Err.	t-value	p-value
β_0	7.189	0.222	32.359	< 0.001
β_1	-0.163	0.021	-7.855	< 0.001
β_2	0.028	0.030	0.952	0.342

- No evidence of differences in the average longitudinal evolutions between the two treatments

2.4 Mixed Models with Correlated Errors

- We have seen two classes of models for longitudinal data, namely
  - ▷ *Marginal Models*

$$y_i = X_i\beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, V_i), \quad \text{and}$$

  - ▷ *Conditional Models*

$$\begin{cases} y_i = X_i\beta + Z_ib_i + \varepsilon_i, \\ \\ b_i \sim \mathcal{N}(0, D), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \mathsf{I}_{n_i}) \end{cases}$$

2.4 Mixed Models with Correlated Errors (cont'd)

- It is also possible to combine the two approaches and obtain a linear mixed model with correlated error terms

$$\begin{cases} y_i = X_i\beta + Z_ib_i + \varepsilon_i, \\ \\ b_i \sim \mathcal{N}(0, D), \quad \varepsilon_i \sim \mathcal{N}(0, \Sigma_i), \end{cases}$$

  where, as in marginal models, we can consider different forms for Σ_i

- The corresponding marginal model is of the form

$$y_i \sim \mathcal{N}(X_i\beta, \; Z_iDZ_i^\top + \Sigma_i)$$

2.4 Mixed Models with Correlated Errors (cont'd)

- Features
  - ▷ both b_i and Σ_i try to capture the correlation in the observed responses y_i
  - ▷ this model does not assume conditional independence

- Choice between the two approaches is to a large extent philosophical
  - ▷ *Random Effects*: trajectory of a subject dictated by time-independent random effects ⇒ the shape of the trajectory is an inherent characteristic of this subject
  - ▷ *Serial Correlation*: attempts to more precisely capture features of the trajectory by allowing subject-specific trends to vary in time

2.4 Mixed Models with Correlated Errors (cont'd)

- It is evident that there is a contest for information between the two approaches
  - ▷ often in practice it is not possible to include both many random effects and a serial correlation term because of numerical problems

We will focus here on the Random Effects paradigm

2.5 Mixed-Effects Models in R

R> There are two primary packages in R for mixed models analysis:

- ▷ Package **nlme**
  - * fits linear & nonlinear mixed effects models, and marginal models for normal data
  - * allows for both random effects & correlated error terms
  - * several options for covariances matrices and variance functions

- ▷ Package **lme4**
  - * fits linear, nonlinear & generalized mixed effects models
  - * uses only random effects
  - * allows for nested and crossed random-effects designs

2.5 Mixed-Effects Models in R (cont'd)

R> We will only use package **nlme**

R> The basic function to fit linear mixed models is `lme()` and has three basic arguments

- ▷ `fixed`: a formula specifying the response vector and the fixed-effects structure
- ▷ `random`: a formula specifying the random-effects structure
- ▷ `data`: a data frame containing all the variables

2.5 Mixed-Effects Models in R (cont'd)

R> The data frame that contains all variables should be in the *long format*

Subject	y	time	gender	age
1	5.1	0.0	male	45
1	6.3	1.1	male	45
2	5.9	0.1	female	38
2	6.9	0.9	female	38
2	7.1	1.2	female	38
2	7.3	1.5	female	38
⋮	⋮	⋮	⋮	⋮

2.5 Mixed-Effects Models in R (cont'd)

R> Using formulas in R

- ▷ CD4 = Time + Gender
  ⇒ `cd4 ~ time + gender`

- ▷ CD4 = Time + Gender + Time*Gender
  ⇒ `cd4 ~ time + gender + time:gender`
  ⇒ `cd4 ~ time*gender` (the same)

- ▷ CD4 = Time + Time^2
  ⇒ `cd4 ~ time + I(time^2)`

R> Note: the intercept term is included by default

2.5 Mixed-Effects Models in R (cont'd)

R> The code used to fit the linear mixed model for the AIDS dataset (p. 49) is as follows

```
lmeFit <- lme(CD4 ~ obstime + obstime:drug, data = aids,
    random = ~ obstime | patient)

summary(lmeFit)
```


2.5 Mixed-Effects Models in R (cont'd)

R> The same fixed-effects structure but only random intercepts

```
lme(CD4 ~ obstime + obstime:drug, data = aids,
    random = ~ 1 | patient)
```

R> The same fixed-effects structure, random intercepts & random slopes, with a diagonal covariance matrix (using the `pdDiag()` function)

```
lme(CD4 ~ obstime + obstime:drug, data = aids,
    random = list(patient = pdDiag(form = ~ obstime)))
```

2.5 Mixed-Effects Models in R (cont'd)

R> Marginal models can be fitted using function `gls()` from the **nlme** package

R> It has four basic arguments

- ▷ `model`: a formula specifying the response vector and the covariates to include in the model
- ▷ `data`: a data frame containing all the variables
- ▷ `correlation`: an object describing the assumed correlation structure
- ▷ `weights`: an object describing the assumed describing the within-group heteroscedasticity structure

2.5 Mixed-Effects Models in R (cont'd)

R> The following code fits a marginal model for CD4 cell count with an AR1 correlation structure

```
glsFit <- gls(CD4 ~ obstime + obstime:drug, data = aids,
    correlation = corAR1(form = ~ 1 | patient))

summary(glsFit)
```

Part III

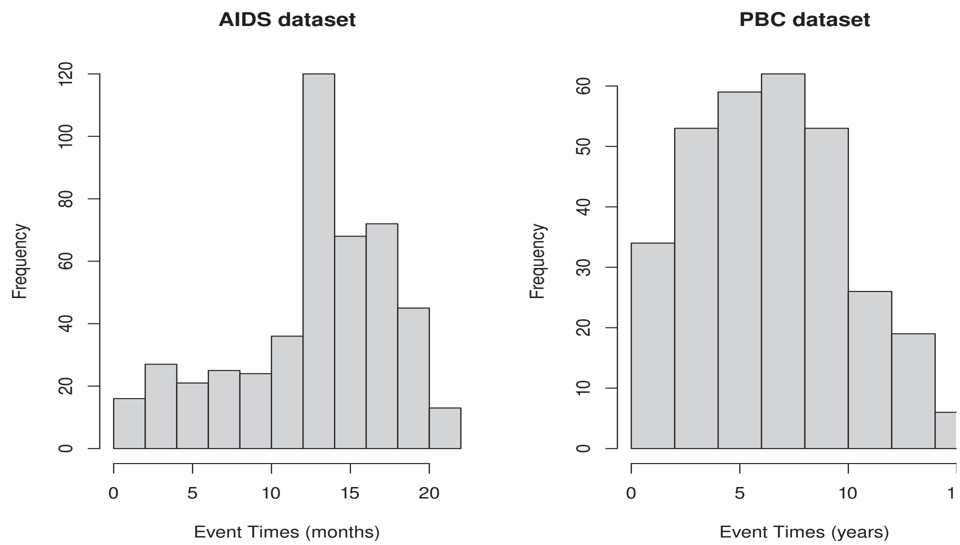
Relative Risk Models

3.1 Features of Survival Data

The statistical analysis of survival data requires special attention due the special characteristics such data have

- Let's have a look at the data...

3.1 Features of Survival Data (cont'd)



3.1 Features of Survival Data (cont'd)

- Survival times are non-negative
  - ▷ in many cases the time to failure can have unusual distribution, i.e., does <u>not</u> look like a Normal
  - ▷ skewed to the right or to the left

- Naive analysis of untransformed times may produce invalid results

3.1 Features of Survival Data (cont'd)

- The most important characteristic that distinguishes the analysis of time-to-event outcomes from other areas in statistics is **Censoring**
  - ▷ the event time of interest is not fully observed for all subjects under study

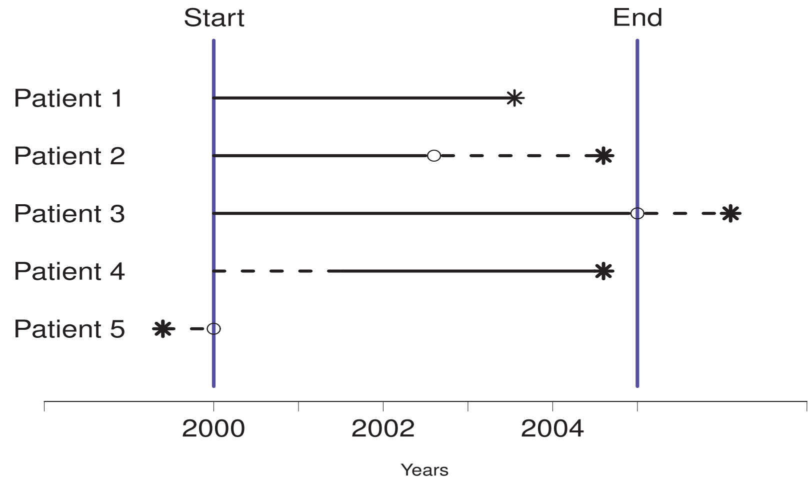
- Implications of censoring:
  - ▷ standard tools, such as the sample average, the t-test, and linear regression **cannot** be used
  - ▷ inferences may be sensitive to misspecification of the distribution of the event times

3.1 Features of Survival Data (cont'd)

- Types of censoring
  - ▷ right censoring
  - ▷ left censoring
  - ▷ interval censoring

- **Caution:** failure to take censoring into account can produce serious bias in estimates of the distribution of event times and related quantities

3.1 Features of Survival Data (cont'd)



3.1 Features of Survival Data (cont'd)

- Before talking in more detail about censoring ...

- Patients who had the event within the study period
  - ▷ Patient 1 was under observation from the start of the study until 3.5 years when she had the event ⇒ the time-to-event equals 3.5 years
  - ▷ Patient 4 enter the study after 1.5 years from the start (late entry), and she had the event at 4.6 years ⇒ the time-to-event equals $4.6 - 1.5 = 3.1$ years
    - * why can't we treat Patient 4 as observed for the full 5-year period since we know that she has survived 1.5 years?
    - * had this patient died before 1.5 years, she would not have had the opportunity to enroll the study, and the event would have never been observed ⇒ biases survival time upwards

3.1 Features of Survival Data (cont'd)

- Right censoring ⇒ the survival time is above a certain value

- Types of right censoring – Examples:

  ▷ Fixed type I: Patient 3 reached the end of the study ⇒ we know this patient had the event after 5 years

  ▷ Fixed type II: a study ends when there is a prespecified number of events

  ▷ Random: Patient 2 moved to a new location at 2.6 years ⇒ we know this patient had the event after 2.6 years

3.1 Features of Survival Data (cont'd)

- Left censoring ⇒ the survival time is below a certain value

- Example:

  ▷ Patient 5 had the event before the start of the study

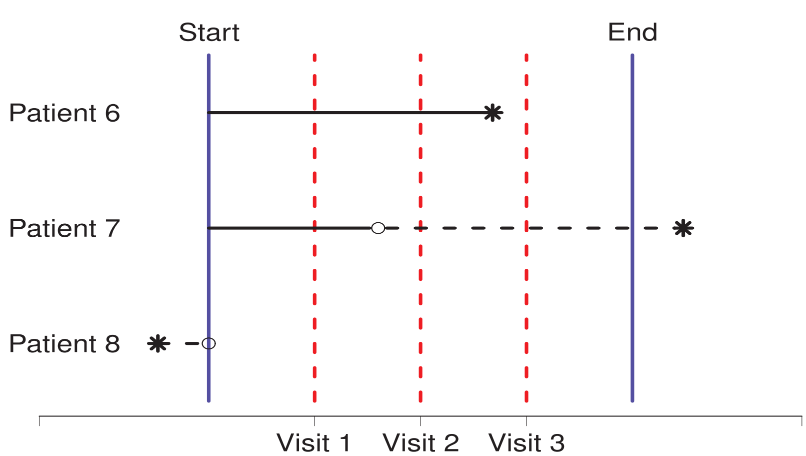
3.1 Features of Survival Data (cont'd)

- Interval censoring: $\Rightarrow$ the survival time is between two values

- Example:
  - ▷ during the study period there are 3 planned visits at which it is checked whether the event has occurred
  - ▷ Patient 6 did not yet have the event at Visit 2 but she had it at Visit 3 $\Rightarrow$ we know that she had the event in between Visits 2 and 3
  - ▷ Patient 7 did not yet have the event at Visit 1 and she left the study before Visit 2 $\Rightarrow$ we know that she had the event at some point after Visit 1
  - ▷ Patient 8 had the event before the stat of the study

Interval censoring includes left and right censoring as special cases

3.1 Features of Survival Data (cont'd)



3.1 Features of Survival Data (cont'd)

- Non-informative versus Informative Censoring
  - ▷ a patient is excluded from the study because he decided to move to a new location from which he cannot easily reach the study center
  - ▷ a patient is excluded from the study because his condition deteriorates (e.g., adverse event) and his physician decides to give him a rescue medication

- What is the substantiative difference in the above two situations?

3.1 Features of Survival Data (cont'd)

- Non-informative versus Informative Censoring
  - ▷ a patient is excluded from the study because he decided to move to a new location from which he cannot easily reach the study center
  - ▷ a patient is excluded from the study because his condition deteriorates (e.g., adverse event) and his physician decides to give him a rescue medication

- What is the substantiative difference in the above two situations?
  - ▷ in the second case withdrawal at time c may indicate death is likely to happen sooner than might have been expected otherwise

Informative Censoring: lost to follow-up for reasons related to the event time

3.1 Features of Survival Data (cont'd)

Here we focus on non-informative right censoring

- Note: Survival times may often be truncated; analysis of truncated samples requires similar calculations as censoring

3.1 Features of Survival Data (cont'd)

- Notation (i denotes the subject)
  - ▷ T_i^* 'true' time-to-event
  - ▷ C_i the censoring time (e.g., the end of the study or a random censoring time)

- Available data for each subject
  - ▷ observed event time: $T_i = \min(T_i^*, C_i)$
  - ▷ event indicator: $\delta_i = 1$ if event; $\delta_i = 0$ if censored

Our aim is to make valid inferences for T_i^* but using only $\{T_i, \delta_i\}$

3.2 Basic functions in Survival Analysis

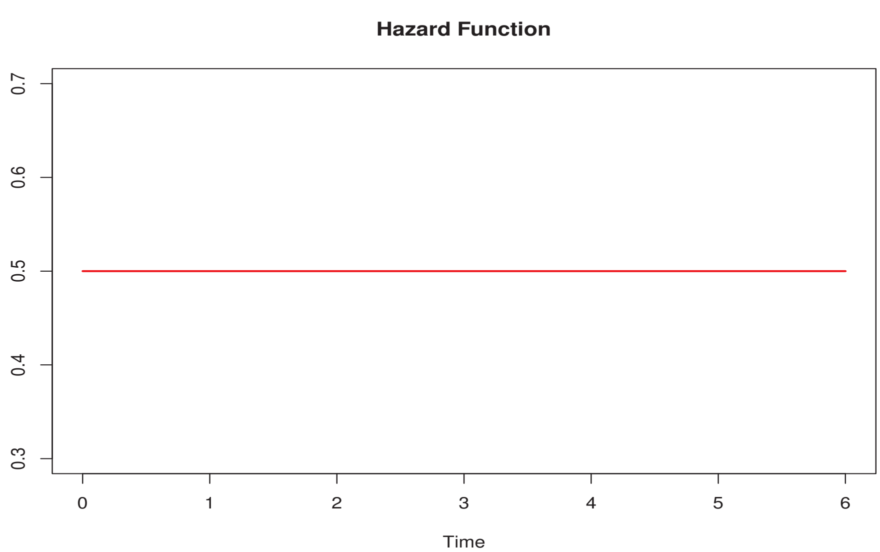
- *Hazard function:* The instantaneous risk of an event at time t, given that the event has not occurred until t

$$h(t) = \lim_{dt \to 0} \frac{\Pr(t \leq T^* < t + dt \mid T^* \geq t)}{dt}, \quad t > 0$$

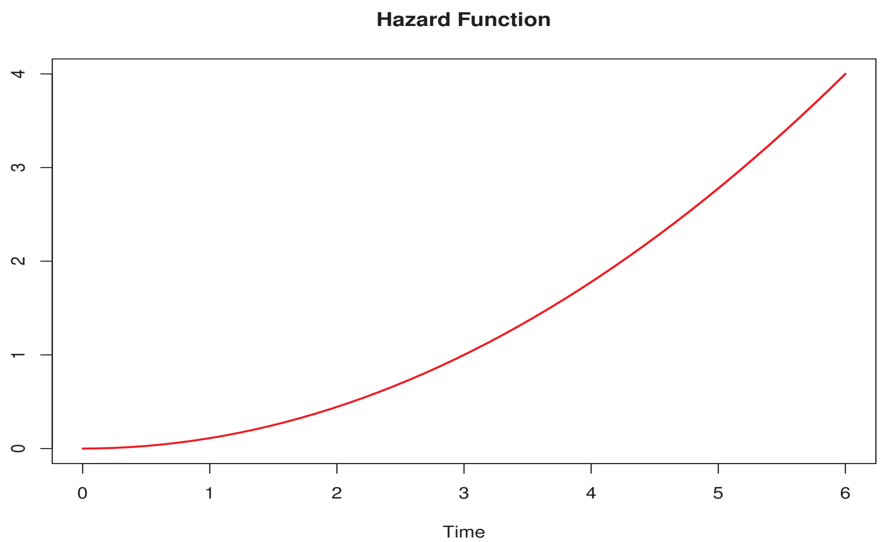
▷ it is **not** a probability, i.e., $h(t) \in (0, \infty)$

▷ can be interpreted as the expected number of events per individual per unit of time

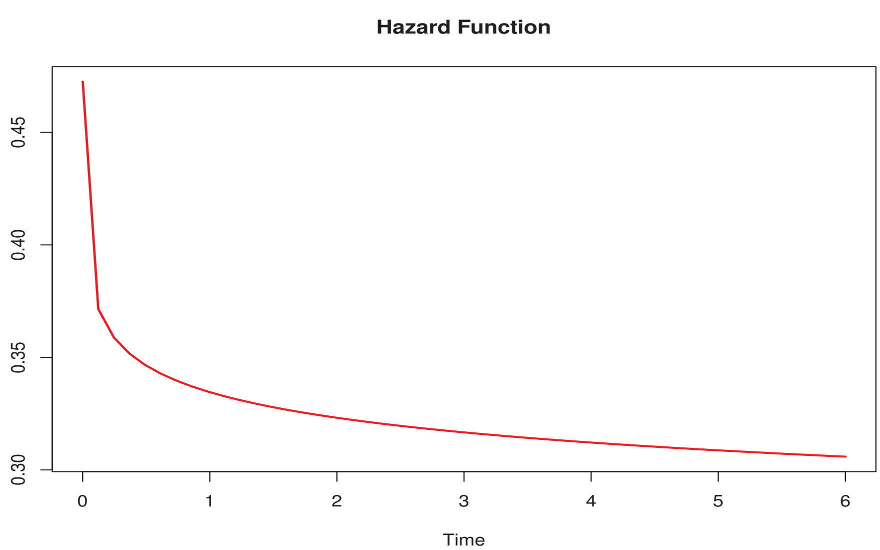
3.2 Basic functions in Survival Analysis (cont'd)



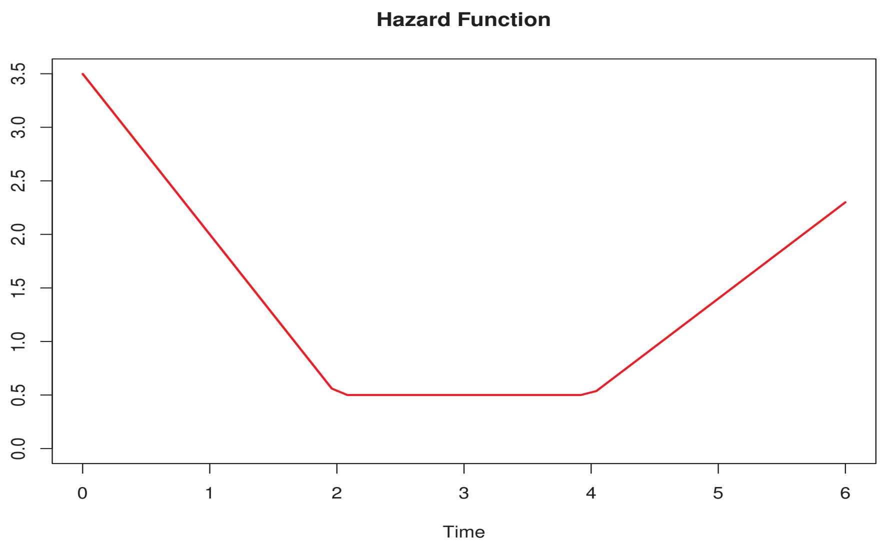
3.2 Basic functions in Survival Analysis (cont'd)



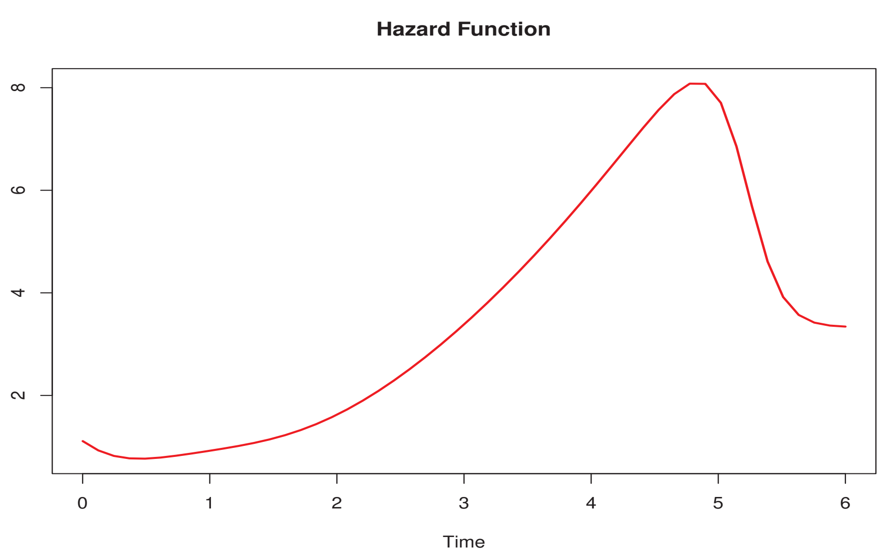
3.2 Basic functions in Survival Analysis (cont'd)



3.2 Basic functions in Survival Analysis (cont'd)



3.2 Basic functions in Survival Analysis (cont'd)



3.2 Basic functions in Survival Analysis (cont'd)

- *Survival function:* The probability of being alive up to time t

$$S(t) \;=\; \Pr(T^* > t)$$

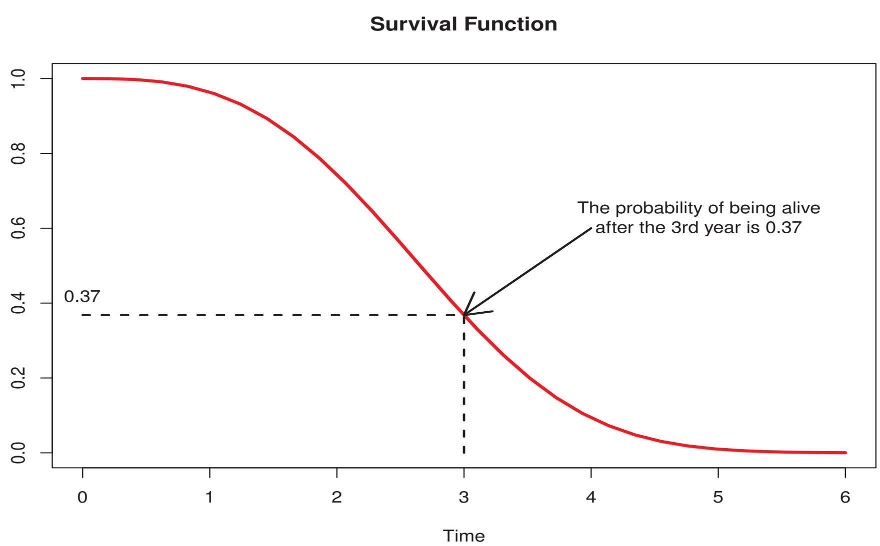
  ▷ decreasing function of time

  ▷ connected to the hazard via

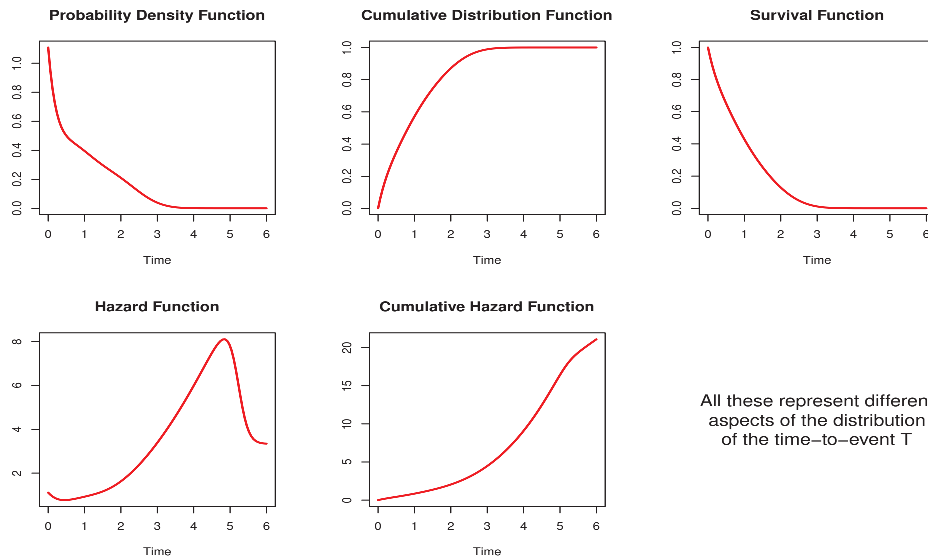
$$S(t) \;=\; \exp\Big\{-\int_0^t h(s)\, ds\Big\}$$

  $\mathcal{H}(t) = \int_0^t h(s)ds$ is known as the *cumulative hazard function*

3.2 Basic functions in Survival Analysis (cont'd)



3.2 Basic functions in Survival Analysis (cont'd)



3.2 Basic functions in Survival Analysis (cont'd)

- To estimate these functions we need to account for censoring
  $\Rightarrow$ we **cannot** use standard tools such as
  - ▷ empirical cumulative distribution function
  - ▷ kernel density estimation
  - ▷ ...

- To account for censoring we suitably adjust the *risk set*
  - ▷ at any particular time point t, the risk set contains the patients who have not died or were not censored before t
  - ▷ that is, the risk set contains the patients who can still have the event and <u>we</u> are able to record it

3.2 Basic functions in Survival Analysis (cont'd)

- Consistent estimates for the survival and cumulative hazard functions that account for censoring are provided by the non-parametric
  - ▷ Kaplan-Meier estimator

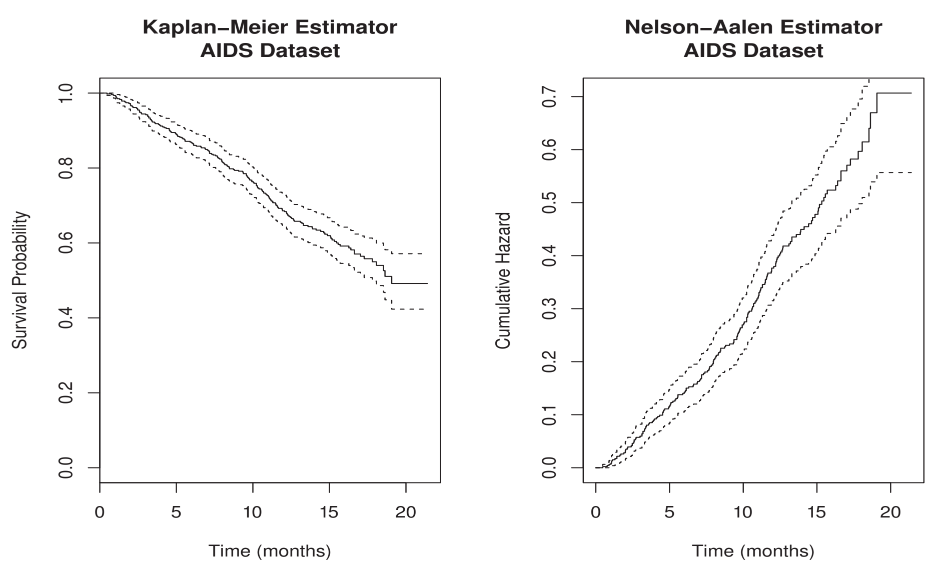
$$\widehat{S}_{KM}(t) = \prod_{i:t_i \leq t} \frac{r_i - d_i}{r_i}$$

  - ▷ Nelson-Aalen estimator

$$\widehat{\mathcal{H}}_{NA}(t) = \sum_{i:t_i \leq t} \frac{d_i}{r_i},$$

with r_i # subjects still at risk at t_i, and d_i # events at t_i

3.2 Basic functions in Survival Analysis (cont'd)



3.2 Basic functions in Survival Analysis (cont'd)

- The variance of $\hat{S}_{KM}(t)$ can be estimated using Greenwood's formula

- Using the formula and asymptotic normality of $\hat{S}_{KM}(t)$, we can derive a 95% confidence interval

- Problem: This can exceed 1 or fall below 0!

- A better asymmetric 95% confidence interval for $\hat{S}_{KM}(t)$ that respects the boundaries is derived from a symmetric 95% confidence interval for either

$$\hat{H}_{KM}(t) = -\log \hat{S}_{KM}(t) \quad \text{or} \quad \log \hat{H}_{KM}(t) = \log\{-\log \hat{S}_{KM}(t)\}$$

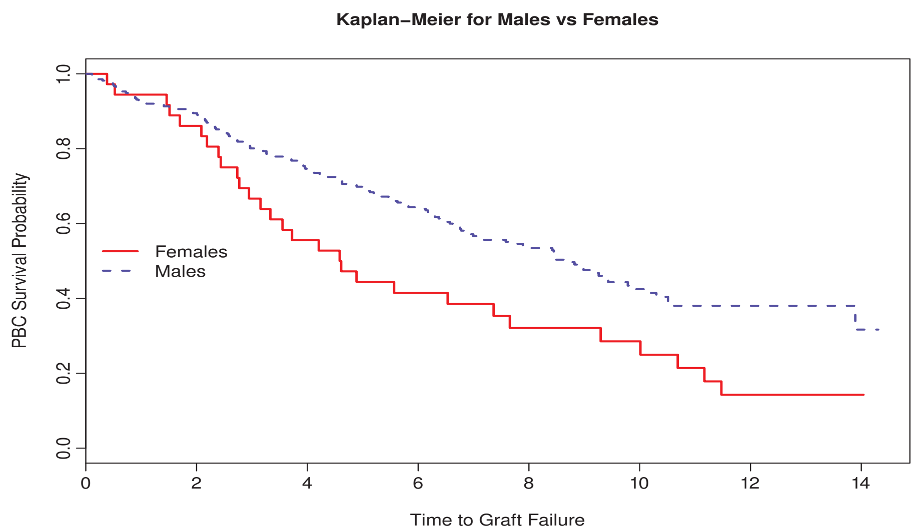
3.2 Basic functions in Survival Analysis (cont'd)

- Comparing survival functions: We have 2 groups of patients
  - ▷ treatment vs placebo
  - ▷ females vs males
  - ▷ history of diabetes, Yes vs No
  - ▷ ...

- Question of Interest: how can we compare these groups with respect to survival

- We can estimate separate survival curves for the 2 groups,

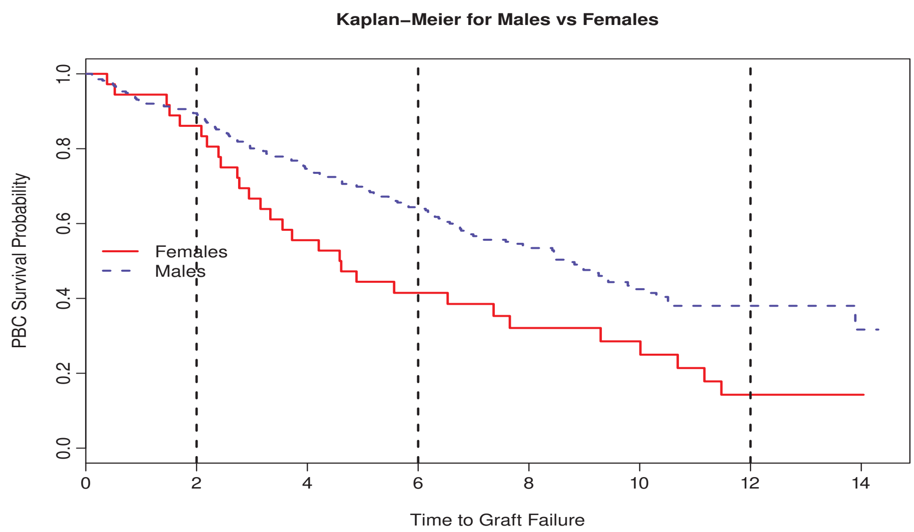
3.2 Basic functions in Survival Analysis (cont'd)



3.2 Basic functions in Survival Analysis (cont'd)

- But how to compare these survival curves?
- We could compare at a specific time point
- At which time point?
  - ▷ start of follow-up
  - ▷ end of follow-up
  - ▷ intermediate points
  - ▷ . . .

3.2 Basic functions in Survival Analysis (cont'd)



3.2 Basic functions in Survival Analysis (cont'd)

- Not very informative because the difference between the survival curves can be greater at some time points than others

- Alternatively, it seems more appropriate to compare the 2 survival curves over the whole follow-up period

- Formally, we are interested in testing the following set of hypotheses

$$\begin{array}{lcl} H_0 & : & \text{the distribution of survival times is the same for the two groups} \\ H_a & : & \text{it is not the same} \end{array}$$

3.2 Basic functions in Survival Analysis (cont'd)

- The most famous statistical test to test this hypothesis is the *Mantel-Haenszel Test (aka Log-Rank Test)*

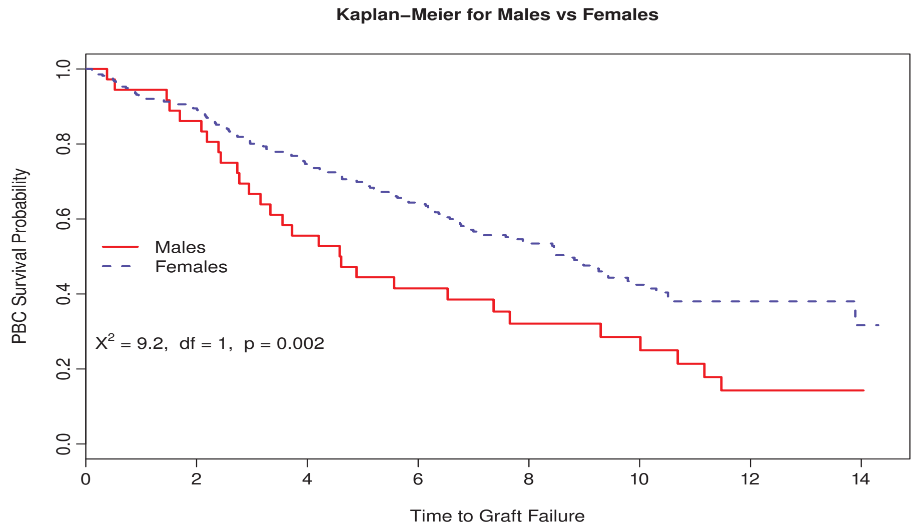
- This is a nonparametric test
  - ▷ no distributional assumption is made for the survival times of the 2 groups

- The philosophy behind it is to construct 2×2 contingency tables for each unique event time, and compare observed with expected numbers of events.

3.2 Basic functions in Survival Analysis (cont'd)

- $\boxed{\text{Example:}}$ For the PBC data we are interested in testing whether the survival curve of males is different from the one of females

3.2 Basic functions in Survival Analysis (cont'd)



3.3 Relative Risk Models

- We have seen how we can compare the survival curves of groups of patients
  - ▷ log-rank test

- However, in many cases we may have more complex research questions – for example,
  - ▷ what is the effect of weight on survival (continuous covariate which we do not want to categorize)
  - ▷ what is the effect of treatment if we control for other variables (e.g., age at baseline, history of other diseases, etc.)

3.3 Relative Risk Models (cont'd)

- **Relative Risk Models** assume a multiplicative effect of covariates on the hazard scale, i.e.,

$$\begin{aligned} h_i(t) &= h_0(t) \exp(\gamma_1 w_{i1} + \gamma_2 w_{i2} + \ldots + \gamma_p w_{ip}) \;\Rightarrow \\ \log h_i(t) &= \log h_0(t) + \gamma_1 w_{i1} + \gamma_2 w_{i2} + \ldots + \gamma_p w_{ip}, \end{aligned}$$

  where

  ▷ $h_i(t)$ denotes the hazard for an event for patient i at time t

  ▷ $h_0(t)$ denotes the baseline hazard

  ▷ $w_{i1}, \ldots, w_{ip}$ a set of covariates

3.3 Relative Risk Models (cont'd)

- The baseline hazard $h_0(t)$ represents the hazard for an event when all the covariates or all the γs are 0

- That is, $h_0(t)$ represents the instantaneous risk of experiencing the event at time t, without the influence of any covariate

- Therefore,

  ▷ if a covariate has a beneficial effect, decreases $h_0(t) \rightarrow \boxed{\gamma < 0}$

  ▷ if it has a harmful effect, increases $h_0(t) \rightarrow \boxed{\gamma > 0}$

3.3 Relative Risk Models (cont'd)

- In general, one-unit change in covariate W_j, $(j = 1, \ldots, p)$ corresponds to
  - ▷ a γ_j change of $\log\{h_i(t)/h_0(t)\}$
  - ▷ increases $h_i(t)/h_0(t)$ by a factor of $\exp(\gamma_j)$ (if $\gamma_j < 0$, then $\exp(\gamma_j) < 1$ and therefore the risk is decreased)

- Hence, parameters from a relative risk model have a log hazard ratio interpretation

$$\Downarrow$$

Care in the (mis)interpretation of the hazard ratio

3.3 Relative Risk Models (cont'd)

- Estimation: Standard MLE can be applied based on the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \log p(T_i; \theta) + (1 - \delta_i) \log S_i(T_i; \theta),$$

which also can be re-expressed in terms of the hazard function

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \log h_i(T_i; \theta) - \int_0^{T_i} h_i(s; \theta) \, ds$$

Sensitivity to distributional assumptions due to censoring

3.3 Relative Risk Models (cont'd)

- **Cox Model:** We make no assumptions for the baseline hazard function

- Parameter estimates and standard errors are based on the log partial likelihood function

$$p\ell(\gamma) = \sum_{i=1}^{n} \delta_i \Bigl[\gamma^\top w_i - \log \Bigl\{ \sum_{j: T_j \geq T_i} \exp(\gamma^\top w_j) \Bigr\} \Bigr],$$

where only patients who had an event contribute

3.3 Relative Risk Models (cont'd)

- The obtained Maximum Partial Likelihood Estimates, which are usually denoted as $\widehat{\gamma}$, are asymptotically (i.e., when the number of events is large) normally distributed

$$\widehat{\gamma} \sim \mathcal{N}\bigl(\gamma_0, \{\mathcal{I}_p(\gamma_0)\}^{-1}\bigr)$$

where

▷ γ_0 denotes the true values of parameters γ

▷ $\{\mathcal{I}_p(\gamma_0)\}$ expected information matrix based on the partial likelihood

3.3 Relative Risk Models (cont'd)

- $\boxed{\text{Example:}}$ For the PBC dataset were interested in the treatment effect while correcting for sex and age effects

$$h_i(t) \;=\; h_0(t)\exp(\gamma_1 \texttt{D-penic}_i + \gamma_2 \texttt{Female}_i + \gamma_3 \texttt{Age}_i)$$

	Value	HR	Std.Err.	z-value	p-value
γ_1	-0.138	0.871	0.156	-0.882	0.378
γ_2	-0.493	0.611	0.207	-2.379	0.017
γ_3	0.021	1.022	0.008	2.784	0.005

3.4 Relative Risk Models in R

R> The primary package in R for the analysis of survival data is the **survival** package

R> A key function in this package that is used to specify the available event time information in a sample at hand is `Surv()`

R> For right censored failure times (i.e., what we will see in this course) we need to provide the observed event times `time`, and the event indicator `status`, which equals 1 for true failure times and 0 for right censored times

```
Surv(time, status)
```

3.4 Relative Risk Models in R (cont'd)

R> Cox models are fitted using function `coxph()`. For instance, for the PBC data the following code fits the Cox model that contains the main effects of 'drug', 'sex' and 'age':

```
coxFit <- coxph(Surv(years, status2) ~ drug + sex + age,
    data = pbc2.id)

summary(coxFit)
```

R> The two main arguments are a formula specifying the design matrix of the model and a data frame containing all the variables

Part IV
Practical

4.1 Practical 1: Linear Mixed Models with R

- We will illustrate some basic linear mixed models analysis

- We will use the PBC dataset; this is available as the object `pbc2` in the R workspace you have received

- We will need the following variables
  * `id`: patient id number
  * `serBilir`: serum bilirubin (the response variable of interest)
  * `year`: follow-up times in years
  * `drug`: the randomized treatment
  * `sex`: the gender of the patients
  * `age`: the age of the patients

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- The response variable we will use will be the natural logarithm of `serBilir`

- We start with some descriptive plots; load the **lattice** package using: `library("lattice")` (or your favorite graphics package, e.g., **ggplot2**)

- T1: Plot the average longitudinal evolutions of the two treatment groups using loess. Should we or should we not trust this plot?

- T2: Do the same plot for sex

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- $\boxed{\text{T3:}}$ Create the plot of the subject-specific longitudinal trajectories
  - ▷ it will be useful to save the plots in a pdf, using `pdf()` before executing the plot and `dev.off()` afterwards

- $\boxed{\text{T4:}}$ As an initial analysis we will test for a treatment effect using the AUC
  - ▷ calculate the AUC for each subject (see p. 26)
  - ▷ do a t-test for the difference in the AUC between the two treatment groups

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- We will proceed by fitting appropriate linear mixed models to the data

- One approach to graphically investigate the variance function over time is to smooth the squared OLS residuals
  - ▷ in order the OLS residuals to correctly reflect the properties of the marginal covariance matrix of the response variable, it is important to remove all systematic trends
  - ▷ hence we want to fit an elaborate mean structure linear model
  - ▷ we will allow for nonlinear time evolutions using natural cubic splines
  - ▷ correct for `sex`, `drug` and `age` + interactions of the time effect with `sex` and `drug`

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- A bit of motivation and background for splines: When modeling continuous covariates it is customary to assume that such covariates affect linearly the response

- However, this assumption is very restrictive, and in many real applications it may not hold
  - ▷ increasing age from 20y to 25y does not increase the risk in the same amount as increasing age from 60y to 65y
  - ▷ similar conjectures also can be made for the time effect in a longitudinal setting

- Wrongly assuming linearity may affect the resulting inference for such covariates as well as the predictive ability of the model

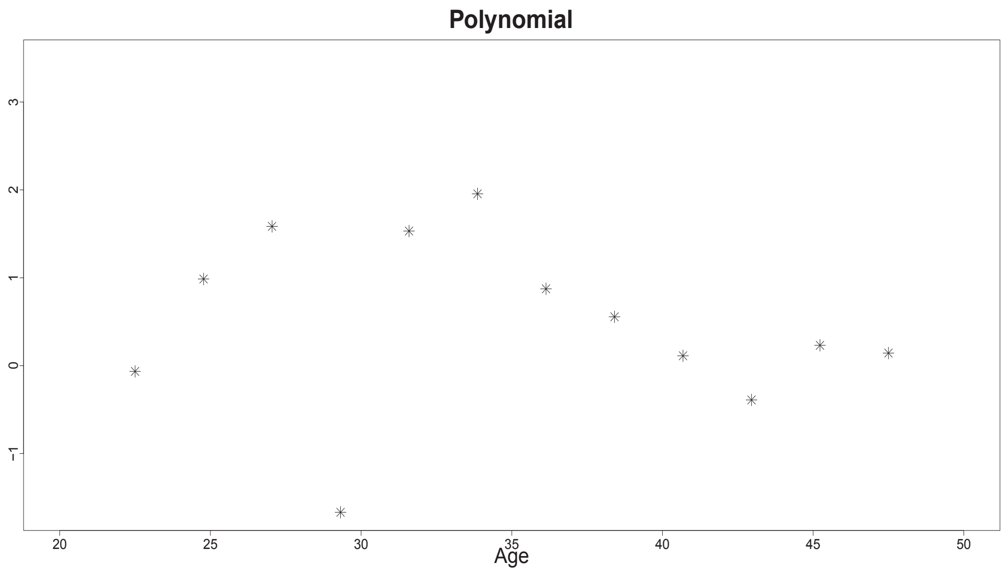
- Therefore, it is highly advisable not to restrict a priori the effects of continuous predictors to be linear and let the data tell you the true story

- The easiest way to relax linearity is to assume polynomial effects

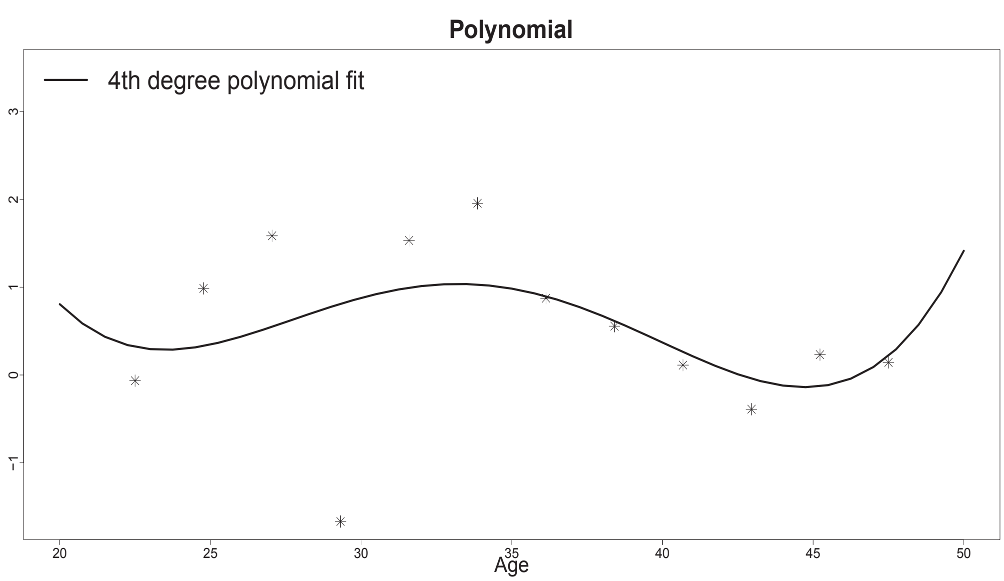
$$\beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \ldots$$

- However, polynomials have some disadvantages, namely
  - ▷ they are not local $\Rightarrow$ changing one data point will affect the overall fit
  - ▷ numerically ill-conditioned (however, not too worrisome with modern software)

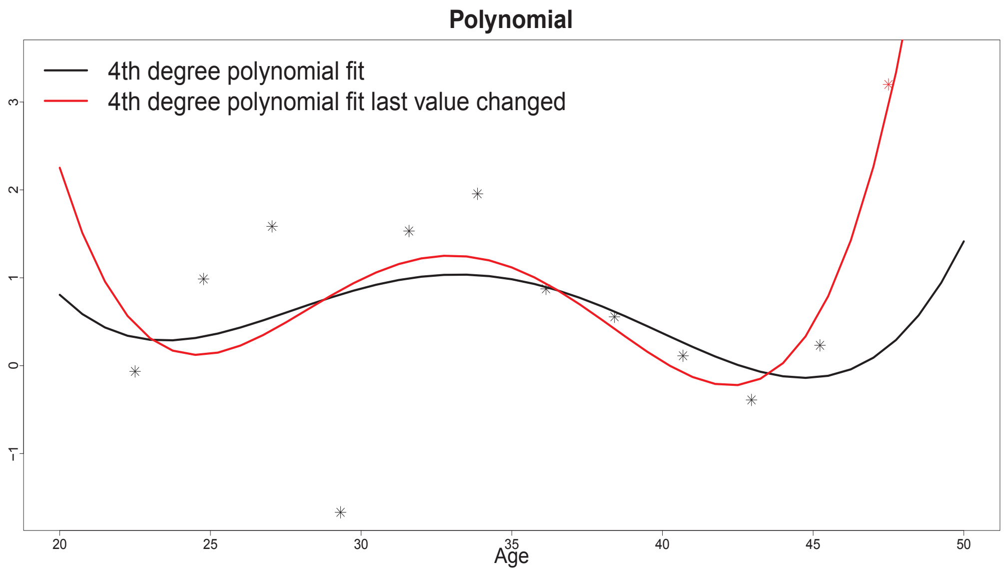
4.1 Practical 1: Lin. Mixed Models with R (cont'd)



4.1 Practical 1: Lin. Mixed Models with R (cont'd)



4.1 Practical 1: Lin. Mixed Models with R (cont'd)

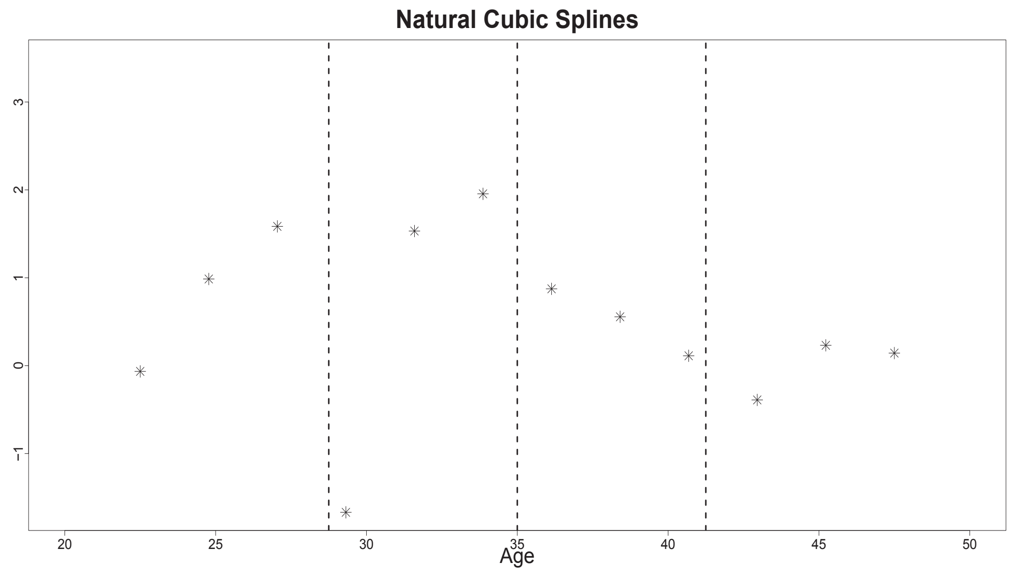


4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- An alternative approach to relax the linearity assumption of continuous predictors is to use regression splines

- Idea behind regression splines: use polynomials but locally
  - ▷ split the range of values of the continuous predictor into subintervals using a series of knots
  - ▷ within each subinterval assume that the effect of the predictor is nonlinear and can be approximated by a cubic polynomial
  - ▷ put extra smoothness assumptions, i.e., the cubic polynomial fits between neighboring subintervals must be connected

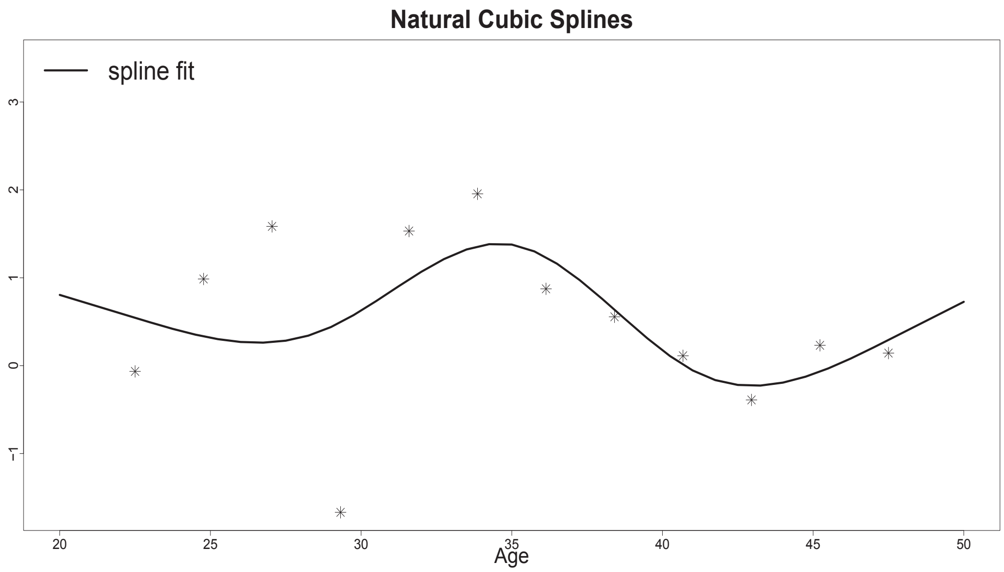
4.1 Practical 1: Lin. Mixed Models with R (cont'd)



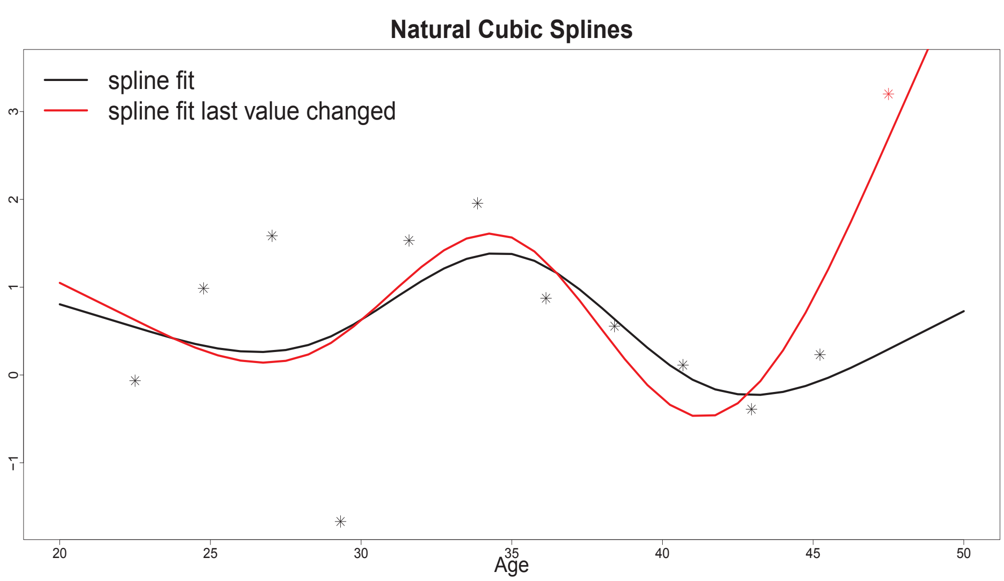
- There are several types of regression splines available
  - ▷ advisable to use natural cubic splines, which assume linearity outside the boundary knots – better statistical properties

- Other approaches (we are not going to discuss them here)
  - ▷ penalized splines
  - ▷ local regression
  - ▷ wavelets
  - ▷ . . .

4.1 Practical 1: Lin. Mixed Models with R (cont'd)



4.1 Practical 1: Lin. Mixed Models with R (cont'd)



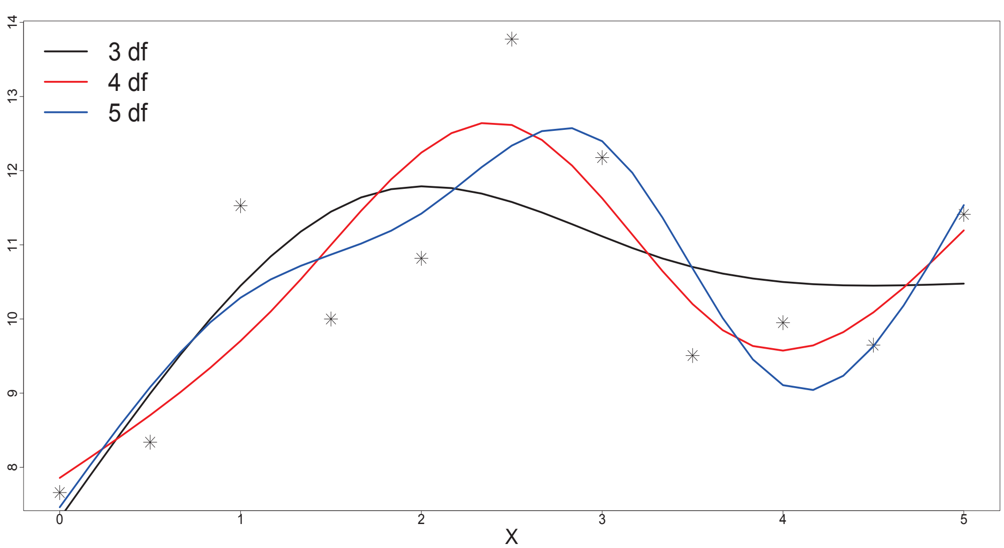
4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- As also in the case of the polynomials, we can tune the degree of nonlinearity by specifying the degrees of freedom for the spline

    ▷ increasing the degrees of freedom results in more flexible modeling

    ▷ bias-variance tradeoff

4.1 Practical 1: Lin. Mixed Models with R (cont'd)



4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- T5: Calculate the squared OLS residuals for the above defined linear regression model, and do the loess plot
  - ▷ load package **splines** using `library("splines")` in order to make the spline functions available
  - ▷ the function that can be used to fit natural cubic splines is `ns()` and it can be directly included in a model formula
  - ▷ fit the above defined model using function `lm()`
  - ▷ extract the residuals using function `resid()`
  - ▷ make the plot of the squared residuals using `xyplot()` (or your favorite plotting function)

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- We will start our model-building exercise...

- General recipe: First model the covariance structure and then the mean structure
  - ▷ start with an elaborate mean model (i.e., in order to be more or less certain that we have removed all systematic trends)
  - ▷ build up the random-effects structure, starting from random intercepts, random intercepts and random slopes, etc. until you find a satisfying model
  - ▷ then return to the mean structure and simplify it if required

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- T6: Fit a linear mixed model with mean structure the same as the one you used in the simple linear model to calculate the OLS residuals in T5, and random intercepts – you will need to load package **nlme** first using `library("nlme")`

- T7: Continue on elaborating the random-effects structure and perform likelihood ratio tests (using function `anova()`) to see if the additional random effects are required
  - ▷ random intercepts & random slopes
  - ▷ random intercepts & splines for the time effect

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- Technical/Theorical Issue: Consider the hypothesis test between the random intercepts and the random intercepts & random slopes models
  - ▷ random intercepts model

$$y_{ij} = X\beta + b_{i0} + \varepsilon_{ij}, \qquad b_{i0} \sim \mathcal{N}(0, \sigma_{b1}^2)$$

  - ▷ random intercepts & random slopes model

$$y_{ij} = X\beta + b_{i0} + b_{i1}t + \varepsilon_{ij}, \qquad b_{i0} \sim \mathcal{N}(0, D)$$

with

$$D = \begin{bmatrix} \sigma_{b1}^2 & \sigma_{b12} \\ \sigma_{b12} & \sigma_{b2}^2 \end{bmatrix}$$

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- Hence, the hypotheses to be tested are

$$H_0: \quad \sigma_{b2}^2 = \sigma_{b12} = 0$$

$$H_a: \quad \sigma_{b2}^2 \neq 0 \text{ or } \sigma_{b12} \neq 0$$

- What is the problem? The null hypothesis for σ_{b2}^2 is on the boundary of its corresponding parameter space
  - ▷ statistical tests derived from standard ML theory assume the H_0 is an interior point of the parameter space
  - ▷ the classical asymptotic χ^2 distribution for the likelihood ratio test statistic does not apply

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- For simple settings (as the one above), it has been proposed to use a mixture of χ^2 distributions
  - ▷ nonetheless, it has been suggested that this does not always work satisfactorily (e.g., see package **RLRsim** and the references therein)

- Here we will just use the χ^2 distribution and be a bit conservative

- $\boxed{\text{T8:}}$ Continue by relaxing the fixed-effects structure
  - ▷ start be checking if all interaction terms can be dropped using a likelihood ratio test
  - ▷ due to a numerical problem, fit first again the final model of $\boxed{\text{T7}}$ assuming a diagonal matrix for the random effects – this can be done by using function `pdDiag()` in the `random` argument of `lme()`

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- Technical/Theorical Issue: By default `lme()` fits linear mixed models using REML
  - ▷ REML estimation proceeds by transforming the response variable using the design matrix X
  - ▷ hence, by comparing linear mixed models with different fixed-effect structures, we are actually comparing models with different response variables $\Rightarrow$ LRT is not valid in models with different response variables

- T9: Re-fit the mixed model you ended up with in T8 using maximum likelihood instead of REML, and redo the LRT (check argument `method` of `lme()`)
  - ▷ continue by checking if any main effects may be dropped

4.1 Practical 1: Lin. Mixed Models with R (cont'd)

- T10: For the final model use function `summary()` to obtain a detailed output and interpret the results

4.2 Practical 2: Cox Models

- We will perform some basic survival analysis calculations and fit a series of Cox models for the AIDS dataset

- Start R and load package **survival**, using `library("survival")`

- Load the R workspace with the AIDS dataset

4.2 Practical 2: Cox Models (cont'd)

- We will need the following variables
  * `Time`: observed event times in years
  * `death`: the death indicator
  * `drug`: the randomized treatment
  * `gender`: the sex of the patients
  * `AZT`: intolerance or failure
  * `CD4`: the square root CD4 cell count at baseline

- T1: Calculate and plot the Kaplan-Meier estimator for the time to death
  - ▷ to compute the Kaplan-Meier estimator you will need function `survfit()`
  - ▷ to plot it, just use the `plot()` function on the resulting object

4.2 Practical 2: Cox Models (cont'd)

- T2: Calculate and plot the Kaplan-Meier estimator for the time to death, separately for the two treatment groups
  - ▷ what do you observe?

- T3: Calculate and plot the Kaplan-Meier estimator for the time to death, separately for males and females
  - ▷ what do you observe?

- T4: Calculate the log-rank tests for the two treatment groups and for males versus females
  - ▷ you will need function `survdiff()`, which has a very similar syntax as `survfit()`

4.2 Practical 2: Cox Models (cont'd)

- T5: We are interesting in studying the relationship between the hazard for death, and `drug`, `gender`, `AZT`, and `CD4`. Fit a Cox model that relaxes the linearity assumption for the effect of `CD4` using natural cubic splines (you need function `ns()`). In addition, assume that there is an effect `drug`, `gender` and `AZT` on the hazard for death, but the effect of these predictors is different for different levels of `CD4` cell count
  - ▷ use the `summary()` method and try to interpret the results

- T6: Use a likelihood ratio test to test whether the model can be reduced by dropping all interaction terms
  - ▷ use the `anova()` function

4.2 Practical 2: Cox Models (cont'd)

- $\boxed{\text{T7:}}$ Use the `summary()` method to obtain a detailed summary of the second fitted model. What is the interpretation of the estimated coefficient for `drug`? In addition, in the output you have values for `exp(coef)` and `exp(-coef)`. What do these values represent?

- The main motivation to introduce the semiparametric Cox model was to avoid the impact of a possibly wrong assumption for the distribution of the event times

- However, all statistical models make assumptions – in the Cox model we make no assumption for the distribution of T_i^* but we do make other assumptions:
  - ▷ **proportional hazards (PH)**

4.2 Practical 2: Cox Models (cont'd)

- If PH is seriously violated, then the results we obtain from the Cox model may not be trustworthy!

- In practice, PH means that the effect of a covariate in the risk for an event is **constant over time**

- Some times the PH assumption may not be reasonable, e.g.,
  - ▷ the new treatment requires a time period to start working $\Rightarrow$ at the beginning of follow-up the risk for the treatment group is the same as in the control group, however we expect that later the risk for the treatment group will decrease
  - ▷ . . .

4.2 Practical 2: Cox Models (cont'd)

- To check the PH assumption we will (hypothetically) consider an extension of the Cox model, namely the Cox model with a *time-dependent coefficient*

$$h_i(t) = h_0(t) \exp\{X_i \beta(t)\}$$

where, the effect of X on the hazard *varies* with time

- Grambsch and Therneau (Biometrika, 1994) have shown that, if $\widehat{\beta}$ is the estimated coefficient from the ordinary (time-independent) Cox model, then

$$\beta(t) \approx \widehat{\beta} + E\{s^*(t)\}$$

where $s^*(t)$ is the scaled Schoenfeld residual

4.2 Practical 2: Cox Models (cont'd)

- The formula and rationale behind the scaled Schoenfeld residuals is rather technical
  - ▷ we will not give them here (see Therneau & Grambsch (2000) for more info)

- Plotting scaled Schoenfeld residuals against time or suitable transformation of time, reveals violations of the PH assumption

- An additional advantage of the scaled Schoenfeld residuals is that they can be used to statistically test PH (though this is not advisable)

4.2 Practical 2: Cox Models (cont'd)

- T8: In R, plots of the Schoenfeld residuals are calculated by function `cox.zph()`
  - ▷ use this function on the final Cox model you fitted above
  - ▷ use the `plot()` function to produce the plots (before running `plot()`, run `par(mfrow = c(3, 3))`)
  - ▷ we will interpret together the results. . .

- T9: Check if conclusions change by using other transformations of the time variable (i.e., argument `transform` of `cox.zph()`)