The influence of categorising survival time on parameter estimates in a Cox model

Anika Buchholz[1,2], Willi Sauerbrei[2], Patrick Royston[3]

[1] Freiburger Zentrum für Datenanalyse und Modellbildung, Albert-Ludwigs-Universität Freiburg

[2] Institut für Medizinische Biometrie und Medizinische Informatik, Universitätsklinikum Freiburg

[3] MRC Clinical Trials Unit, London, UK

Funded by Deutsche Forschungsgemeinschaft

Research unit FOR 534

2. April 2007

Standard Cox model and its extension

- Standard Cox model
  $\lambda(t|X) = \lambda_0(t)\exp(\beta_1 X_1 + \ldots + \beta_p X_p)$
  with unspecified baseline hazard $\lambda_0(t)$

- Critical assumptions
  - Linear effect of continuous covariates
    $\longrightarrow$ allow for non-linear covariate effects
    $\lambda(t|X) = \lambda_0(t)\exp(\beta_1 f_1(X_1) + \ldots + \beta_p f_p(X_p))$
  - Proportional hazards (PH)
    $\longrightarrow$ allow for non-proportional hazards (time-varying effects)
    $\lambda(t|X) = \lambda_0(t)\exp(\beta_1(t) X_1 + \ldots + \beta_p(t) X_p)$

- Extended Cox model relaxing both above assumptions
  $\lambda(t|X) = \lambda_0(t)\exp(\beta_1(t) f_1(X_1) + \ldots + \beta_p(t) f_p(X_p))$

Introduction

- Cox model
- Causes for non-PH

MFPT

Categorisation

Results

Summary

Causes for non-proportional hazards

- Effect changes over time
- Incorrect modelling
  - Omission of an important covariate
  - Incorrect functional form of a covariate
  - Different survival model is appropriate

Model selection strategy

Multivariable strategy for model selection needed to

- select variables which have influence on the outcome
- model functional form of the influence of continuous variables
- model time–varying effects in case of non-PH

The **M**ultivariable **F**ractional **P**olynomial **T**ime approach combines

- backward elimination of variables
- function selection procedure to select a function from the class of fractional polynomials (non-linear if ’sufficiently’ supported by the data)
- investigation of possible time–varying effects for each variable from a multivariable proportional hazards Cox model

Introduction

MFPT
● Model selection strategy
● MFPT algorithm
● Rotterdam breast cancer series
● Kaplan-Meier estimate
● Development of the MFPT model

Categorisation

Results

Summary

Multivariable Fractional Polynomial Time (MFPT) algorithm

Stage 1: Determine time–fixed model M_0

- Select model M_0 using MFP–algorithm assuming PH (full time–period)

Multivariable Fractional Polynomial Time (MFPT) algorithm

Stage 1: Determine time–fixed model M_0

Stage 2: If necessary, add covariate with short–term effect only

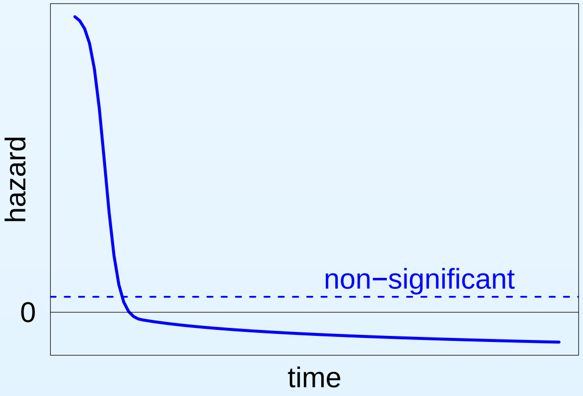
- Start with model M_0, keep variables and functions from M_0
- Restrict the time period to $(0, \tilde{t})$, e.g. $\tilde{t}$ defined by the first half of events
- Run the MFP-algorithm for $(0, \tilde{t})$ and add, if necessary, significant covariates to M_0. This gives a proportional hazards model M_1.

Multivariable Fractional Polynomial Time (MFPT) algorithm

Stage 1: Determine time–fixed model M_0

Stage 2: If necessary, add covariate with short–term effect only

- Start with model M_0, keep variables and functions from M_0
- Restrict the time period to $(0, \tilde{t})$, e.g. $\tilde{t}$ defined by the first half of events
- Run the MFP-algorithm for $(0, \tilde{t})$ and add, if necessary, significant covariates to M_0. This gives a proportional hazards model M_1.



Multivariable Fractional Polynomial Time (MFPT) algorithm

Introduction

MFPT
● Model selection strategy
● MFPT algorithm
● Rotterdam breast cancer series
● Kaplan-Meier estimate
● Development of the MFPT model

Categorisation

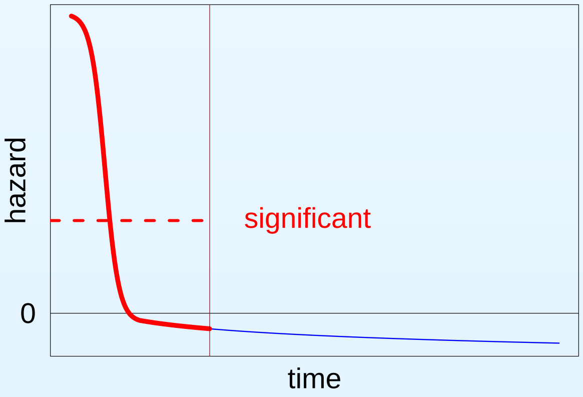
Results

Summary

Stage 1: Determine time–fixed model M_0

Stage 2: If necessary, add covariate with short–term effect only

- Start with model M_0, keep variables and functions from M_0
- Restrict the time period to $(0, \tilde{t})$, e.g. $\tilde{t}$ defined by the first half of events
- Run the MFP-algorithm for $(0, \tilde{t})$ and add, if necessary, significant covariates to M_0. This gives a proportional hazards model M_1.



Multivariable Fractional Polynomial Time (MFPT) algorithm

Stage 1: Determine time–fixed model M_0

Stage 2: If necessary, add covariate with short–term effect only

Stage 3: Add possible time–varying effects of variables in M_1

- Use a forward selection procedure to add significant time–varying effects to model M_1.
- For each covariate of M_1 in turn investigate time–varying effect $\beta(t)$ adjusting for all other covariates of M_1. This gives the final model M_2.

Rotterdam breast cancer series

Breast cancer survival data with

- 2982 patients
- 1518 events for RFS (recurrence free survival)
- 20 years max. follow–up
- 10 variables
- median uncensored survival time: 2.5 years

Rotterdam breast cancer series

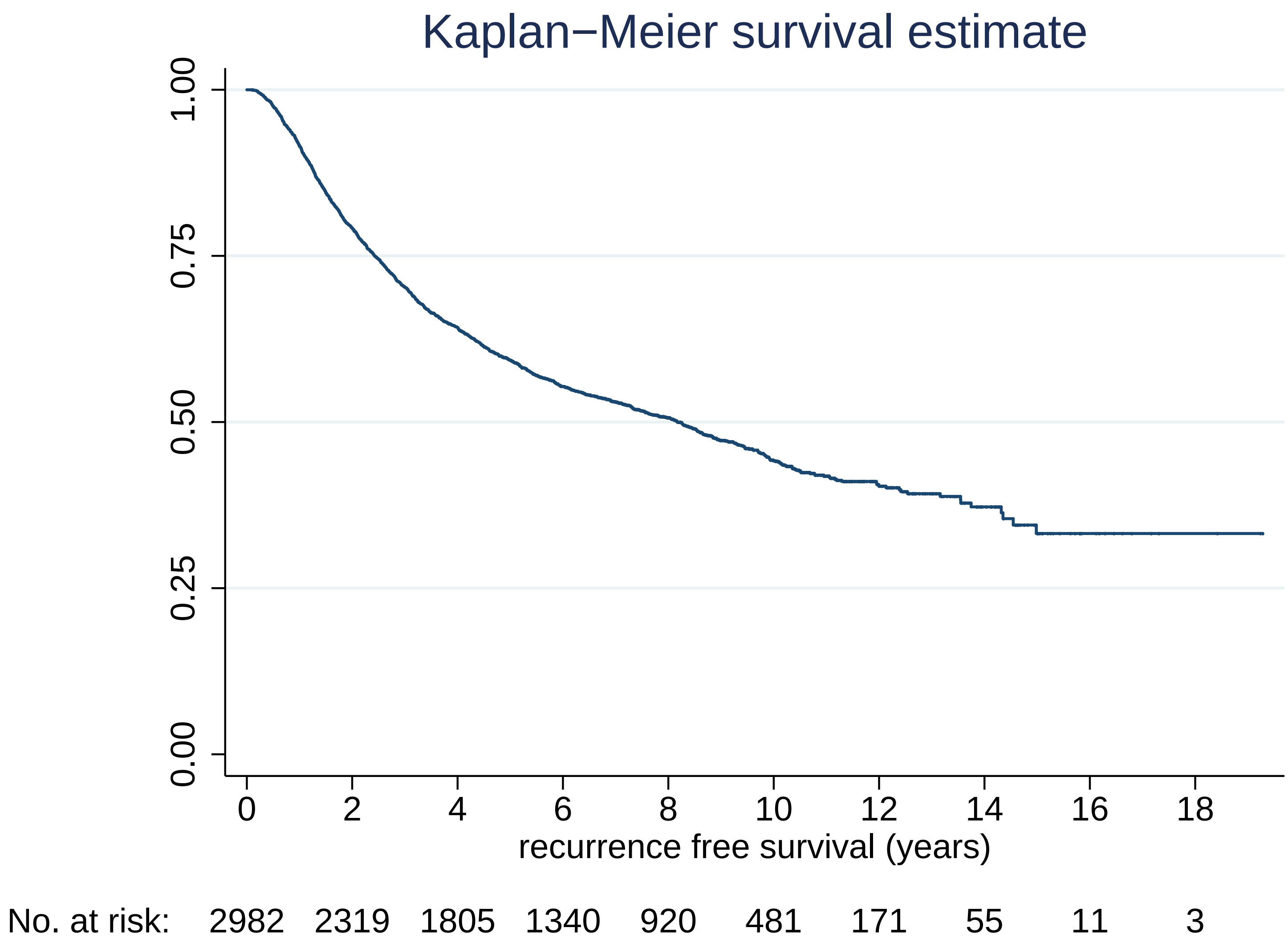
Introduction

MFPT
- ● Model selection strategy
- ● MFPT algorithm
- ● Rotterdam breast cancer series
- ● Kaplan-Meier estimate
- ● Development of the MFPT model

Categorisation

Results

Summary



Introduction

MFPT
- Model selection strategy
- MFPT algorithm
- Rotterdam breast cancer series
- Kaplan-Meier estimate
- Development of the MFPT model

Categorisation

Results

Summary

Development of the MFPT model

Variable	Model M_0	Model M_1	Model M_2
X_1 (age)	●	●	●
X_2 (menopausal status)	-	-	-
X_{3a} (tumour size $>$ 20mm)	●	●	●
X_{3b} (tumour size $>$ 50mm)	-	●	●
X_4 (tumour grade)	●	●	●
${X_5}^2$ (no. of pos. lymph nodes)	●	●	●
$\log(X_6)$ (progesterone receptor)	-	●	●
X_7 (oestrogen receptor)	-	-	-
X_8 (hormonal therapy)	●	●	●
X_9 (chemotherapy)	●	●	●
$X_{3a} \cdot (\log(t))$			●
$\log(X_6) \cdot (\log(t))$			●

Model M_0: Selected with MFP assuming PH, 4 variables eliminated
Model M_1: Add variables with short-term effect only
Model M_2: Add time-varying effects

Enlargement of the data

The analysis of time-varying effects requires

- long-term follow-up
- large sample size

Why is enlargement necessary?

$$lnL = \sum_{j=1}^{D} \left[\sum_{k \in D_j} x_k \beta(t_{(j)}) - d_j \ln \left\{ \sum_{i \in R_j} \exp(x_i \beta(t_{(j)})) \right\} \right]$$

Enlargement of such data may cause computational problems:

. *stsplit, at failures* gives about 2.2 million records in Rotterdam data

Enlarged data

- may be difficult to manage for the analysis of one data set
- is nearly impossible for simulation studies

Possible solution: categorisation of survival time

- Categorisation scheme
  - Equidistant intervals (e.g. 6 month length)
    . *stsplit period, at(.5(.5)20)* results in only 35747 records
  - Other categorisation schemes are possible, e.g. categorisation in quantiles
- How to code categorised survival times
  - Here: represent intervals by integers
  - In clinical investigations e.g. use the mean survival time within each interval

Example for categorised time in the enlarged data

```
. stsplit period, at(.5(.5)20)
(32765 observations (episodes) created)

. egen categorised_EFS = group(period)

. list categorised_EFS EFS_yrs event Patient_ID if Patient_ID==1

     +--------------------------------------------+
     | catego~S   EFS_yrs   event   Patien~D |
  1. |        1        .5       .          1 |
  2. |        2         1       .          1 |
  3. |        3       1.5       .          1 |
  4. |        4         2       .          1 |
  5. |        5       2.5       .          1 |
     |--------------------------------------------|
  6. |        6         3       .          1 |
  7. |        7       3.5       .          1 |
  8. |        8         4       .          1 |
  9. |        9       4.5       .          1 |
 10. |       10     4.925       0          1 |
     +--------------------------------------------+
```

Issues under investigation

Categorisation of survival time raises issues as to

- the number and position of cutpoints
- the loss of information
- the increased number of ties

Introduction

MFPT

Categorisation

- Enlargement
- Categorisation
- Example
- Issues under investigation
- Handling ties

Results

Summary

Issues under investigation

Categorisation of survival time raises issues as to

- the number and position of cutpoints
- the loss of information
- the increased number of ties

We will

- consider interval lengths 1.5, 3, 6, 12 and 24 months

Issues under investigation

Categorisation of survival time raises issues as to

- the number and position of cutpoints
- the loss of information
- the increased number of ties

We will

- consider interval lengths 1.5, 3, 6, 12 and 24 months
- compare parameter estimates obtained by *stcox* for the different interval lengths

Issues under investigation

Categorisation of survival time raises issues as to

- the number and position of cutpoints
- the loss of information
- the increased number of ties

We will

- consider interval lengths 1.5, 3, 6, 12 and 24 months
- compare parameter estimates obtained by *stcox* for the different interval lengths
- compare parameter estimates using the four methods of handling ties provided by *stcox*

Methods for handling ties in Stata

breslow: approximation of exact marginal log likelihood; fast but least accurate (default)

efron: approximation of the exact marginal log likelihood; slower than breslow but more accurate

exactm: exact marginal log likelihood; very slow

exactp: exact partial log likelihood; very slow

Influence of the length of categorisation interval

- Differences of parameter estimates in percent relative to the original time
- Breslow method for handling ties

		Interval length (months)				
	Original time	1.5	3	6	12	24
No. of records	**2982***	138502	70000	35747	18649	10086
No. of distinct observed times	**2183**	155	78	39	20	10
X_1 (age)	**-0.013**	-0.4	-1.1	-1.9	-3.5	-8.5
X_{3a} (tumour size $>$ 20mm)	**0.289**	-0.5	-1.2	-2.5	-5.0	-11.2
X_4 (tumour grade)	**0.390**	-0.8	-1.4	-2.5	-6.0	-12.9
X_5^2 (# pos. lymph nodes)	**-1.713**	-1.0	-2.1	-3.9	-7.8	-15.6
X_8 (hormonal therapy)	**-0.386**	-1.3	-2.3	-3.4	-7.8	-13.9
X_9 (chemotherapy)	**-0.454**	-1.3	-2.7	-4.7	-10.1	-22.0

* 2982 is the original number of observations (1419 ties), splitting the data at each event time gives approximately 2.2 million records

Comparison of the methods of handling ties (original time)

- Parameter estimates using the original time

	Method			
	Breslow	Efron	exactm	exactp
X_1 (age)	-0.0132	-0.0132	-0.0132	-0.0132
X_{3a} (tumour size $>$ 20mm)	0.2885	0.2886	0.2886	0.2886
X_4 (tumour grade)	0.3900	0.3900	0.3900	0.3901
X_5^2 (# pos. lymph nodes)	-1.7128	-1.7132	-1.7132	-1.7136
X_8 (hormonal therapy)	-0.3857	-0.3859	-0.3858	-0.3859
X_9 (chemotherapy)	-0.4539	-0.4540	-0.4540	-0.4541

- Method of handling ties has no influence on parameter estimates

Introduction

MFPT

Categorisation

Results

- Length of interval
- Ties (original time)
- Ties (3 months)
- Ties (6 months)
- Run time
- MFP

Summary

Comparison of the methods of handling ties (3 months)

- Difference in parameter estimates in percent relative to analysis using original time
- Interval length: 3 months

	Method			
	Breslow	Efron	exactm	exactp
X_1 (age)	-1.1	$+1.4$	$+2.3$	$+3.4$
X_{3a} (tumour size $>$ 20mm)	-1.2	$+0.4$	$+1.1$	$+1.6$
X_4 (tumour grade)	-1.4	-0.2	$+0.2$	$+1.7$
X_5^2 (# pos. lymph nodes)	-2.1	0.0	$+0.7$	$+2.2$
X_8 (hormonal therapy)	-2.3	$+0.4$	$+1.7$	$+3.4$
X_9 (chemotherapy)	-2.7	0.0	$+1.1$	$+2.6$

Introduction

MFPT

Categorisation

Results

- Length of interval
- Ties (original time)
- Ties (3 months)
- Ties (6 months)
- Run time
- MFP

Summary

Comparison of the methods of handling ties (6 months)

- Difference in parameter estimates in percent relative to analysis using original time
- Interval length: 6 months

	Method			
	Breslow	Efron	exactm	exactp
X_1 (age)	-1.9	$+3.1$	-79.2	–
X_{3a} (tumour size $>$ 20mm)	-2.5	$+0.7$	-86.9	–
X_4 (tumour grade)	-2.5	-0.1	-84.1	–
X_5^2 (# pos. lymph nodes)	-3.9	0.0	-76.9	–
X_8 (hormonal therapy)	-3.4	$+1.9$	-79.5	–
X_9 (chemotherapy)	-4.7	$+0.4$	-81.1	–

Comparison of run time of the methods of handling ties

- Run time relative to Breslow using original time

	Run time			
	Breslow	Efron	exactm	exactp
original time (2982 records)	1	1.11	19.56	1237.85
3 month intervals (70000 records)	40.18	44.88	85.80	1675.15
split at failures (2.2 mio records)	1926.25	1926.26	2067.26	3367.71

- Breslow and Efron similar
- exactm and exactp much more computationally demanding

Selection of model using *mfp*

Do stage 1 of MFPT algorithm with all 10 candidate variables

. *mfp stcox x1 x2 x3a x3b x4b x5e x6 x7 x8 x9, select(0.01)*

(breslow and efron only)

Identical model as in original data

- for interval lengths 1.5, 3, and 6 months
- for both breslow and efron method for ties
- with similar parameter estimates

Summary of results

- For a single analysis in one data set categorisation is usually not required
- Categorisation may be sufficient for computer intensive methods (simulations, bootstrap, cross validation etc.)

In case of categorisation:

- Categorising long-term survival into 40-100 distinct values seems sensible:
  - Parameter estimates nearly identical
  - Loss of information seems negligible
- Handling ties:
  - Many distinct values: nearly identical results
  - Small(er) number of distinct values:
    - Exact methods break down
    - Breslow and Efron give acceptable results
  - Breslow and Efron suitable for simulation studies, Efron slightly preferable

References

- Sauerbrei, W., Royston, P. and Look, M. (2007). *A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation.* Biometrical Journal, in press
- Buchholz, A., Sauerbrei, W. and Royston, P. (2006). *Investigation of time-varying effects in survival analysis may require categorisation of time: does it matter?* submitted

Program *stmfpt* available upon request from Patrick Royston (pr@ctu.mrc.ac.uk)

Thanks for your attention.

Introduction

MFPT

Categorisation

Results

Summary

- Summary
- References

Required amount of memory for different interval lengths

Interval length (months)	No. of records	Amount of memory* (bytes)
Original time	2,982	146,118
Split at failures	2,220,499	115,465,948
1.5	138,502	8,310,120
3	70,000	4,200,000
6	35,747	2,144,820
12	18,649	988,397
24	10,086	534,558

*data only, without overhead