Survival Data Analysis

Jann Goschenhofer

January 2018

Contents

1 Introduction

Summary of models and especially their interpretation (graphically as well as content based) used in Survival Analysis. This document emerged throughout the exam preparation for a lecture on Survival Data Analysis at LMU in winter 2018. Most examples are based on that lecture taught by Prof. Kuechenhoff and Andreas Bender.

2 Big Picture

2.1 Estimation of $S(t)$ and $\lambda(t)$ without Covariate Effects

2.1.1 Non Parametric

- Kaplan-Meier for $S(t)$
- Nelson-Aalen for $\Lambda(t)$
- Breslow for $S(t)$
- Life-table for $\lambda(t)$
- Ramlau-Hansen for $\lambda(t)$

2.1.2 Parametric

- Assume $T_1, ..., T_n \sim$ Distribution(θ) and estimate $\hat{\theta} = argmin_\theta l(\theta)$
- **BUT** Censoring
- Random Censoring: $L(\theta) = \prod_{i=1}^n \lambda(t_i)^{\delta_i} S(t_i)$

2.2 Regression models with Covariate Effects

2.2.1 Transformation Models

- model $S(t)$ directly
- $log(T) = Y = X^T\beta + \sigma\epsilon$
- Assume $\epsilon \sim$ Distribution(θ)
- Density Transformation: get $F_T(T), f_T(T)$

2.2.2 Semi-Parametric Cox Model

- PH-Assumption: $\frac{\lambda(t|X_1)}{\lambda(t|X_1)} t$
- model $\lambda(t)$ directly
- $\lambda(t|X) = \lambda_0(t) exp(X^T\beta)$
- **parametric** Partial Likelihood Estimation: $PL(\beta) = \prod_{i=1}^m \frac{exp(x_i^T\beta)}{\sum_{j\in R(t_i)} exp(x_i^T\beta)}$
- **non-parametric** $\lambda_0(t)$ via Breslow, Profile Likelihood for *nuisance parameter* $\lambda_0(t)$
- **BUT1** effect of β_j assumed to be linear, often is not
- **BUT2** time varying covariates and effects
- **BUT3** time constant baseline hazards

2.2.2.1 Semi-parametric Additive Cox Model (BUT1:)

- $\lambda_(t|X) = \lambda_0(t)exp(f_1(x_1)\beta_1 + ... + f_k(x_k)\beta_k + x_{k+1}\beta_{k+1} + ... + x_p\beta_p)$
- Estimate $f_j(x_j)$ via splines for smooth nonlinear effects

2.2.2.2 Time Varying Covariates and Effects (BUT2:)

2.2.2.2.1 Categorical Covariates

Transform short

i	week	arrested	married	emp1	emp2	emp3
1	2	1	0	1	0	NA
2	3	0	1	1	1	1

to long format:

i	week	arrested	married	emp
1	1	0	0	1
1	2	1	0	0
2	1	0	1	1
2	2	0	1	1
2	3	0	1	1

and fit classic Cox-PH. Equivalent coefficients for both formats **without** covariates because only events in Partial Likelihood.

2.2.2.2.2 Continous Variables

- Create artificial time-dependent variable $\tilde{x}$ and **add** to classic Cox model

- e.g.: age and t:age

2.2.2.2.3 Effects ?

2.3 Extensions of Cox model

Discretisize time in intervals $[a_o, a_1[, ..., [a_{q-1}, a_q[$

2.3.1 Time Discret Survival Models via GLM's

- Transform data in long format with q time-factors
- fit GLM (logistic, cloglog, probit) **wihtout intercept** on event variable as response
- q coefficients β_{0k} for each time interval as some kind of baseline hazard
- **NICE:** GLM Toolbox
- **BUT3:** No hazard/ Survival interpretation, only Odds etc.

2.3.2 Piecewise Exponential Models (BUT3)

- Cox-Model with time varying baseline hazards λ_j for $j = 1, ..., q$
- Transform short

i	t_i	δ_i	x_{i1}	x_{i2}
1	0.25	1	0	3
2	0.13	0	1	5

to long formatted **pseudo-data** :

i	y	a	$log(\Delta)$	x_1	x_2
1	0	0.1	log(0.1)	0	3
1	0	0.2	log(0.1)	0	3
1	1	0.3	log(0.0.5)	0	3
2	0	0.1	log(0.1)	1	5
2	0	0.2	log(0.03)	1	5

- **BUT4:** exploiding parameters for small intervalls and large q
- use **Piecewise Exponential Additive Model**
- **BUT5:** random effects in the data
- use **Piecewise Exponential Additive Mixed Model** with Frailty term
- **BUT6:** only multiplicative effects of the coefficients

2.4 Additive Hazard Regression Models (BUT6)

2.4.1 Aalen Model

- **NICE:** additive effects
- **NICE:** new interpretation graphically
- Idea: model effects of covariates on baseline hazard rate λ_0 **additively**
- Formula: $\lambda(t|X) = \lambda_0(t) + \sum_{k=1}^{p} x_k(t)\beta_k(t) = \lambda_0(t) + x^T(t)\beta(t)$

2.4.2 Cox-Aalen Model

- combine best from both worlds: **additive** effects on $\lambda_0(t)$ that can be influenced by **multiplicative** coefficients
- $\lambda(t|X) = \lambda_0(t) + X(t)\beta(t)exp(Z^T(t)\gamma)$
- $\beta(t)$: time varying additive coefficients
- γ : time constant multiplicative coefficients. Interpretation: multiplicative effect on hazard if rest kept constant.
- **BUT7:** still assumption that $T_i \perp C_i$

2.4.3 Competing Risk Model (But7)

- More than one possible event (e.g.: two types of death) next to censoring of which only one can occur. The events **compete** with each other as only one of them can occur.
- Approaches:
  - Seperate "cause-specific" Cox models for each type where the competing events are subsumed in censoring.

    - Problem 1: assumption, that $T_1 \perp T_2$
    - Problem 2: Kaplan-Meier Curves are biased
  - Cumulative Incidence Curve as solution to problem 2
  - Discretization: Multinomial GLMs

3 Censoring

1. **Right**:
   1. Type 1: study ends before event occured. E.g.: fixed time study of 1 year
   2. Type 2: ...
   3. Type 3: person withdraws from study because of other event. E.g.: interest on cancer death, person is getting shot
2. **Left**: we know when event occurs but we do not know when it started. E.g.: person dies at week 4 on cancer but we don't know the time of the disease outbreak. Our observed survival time of 4 weeks is thus equal (best case) or smaller then the observed.
3. **Left Truncation**: Biased because only people that survived made it to the study. E.g.: deductible in insurances, people with losses $<$ deductibles are not getting observed.

4 Kaplan Meier

4.0.1 Model Equation

Cannot simply $1 - F(t)$ due to censoring. KM takes that into account.

Estimate the **Survival rate** non-parametrically without any covariables:

$$\hat{S}(t) = \prod_{t_k \leq t} (1 - d_k/n_k), \forall t \geq t_1$$

where d_k = number of events at time point t_k (neither dead nor censored) and $n_k = $ amount of people under risk right before time t_k.

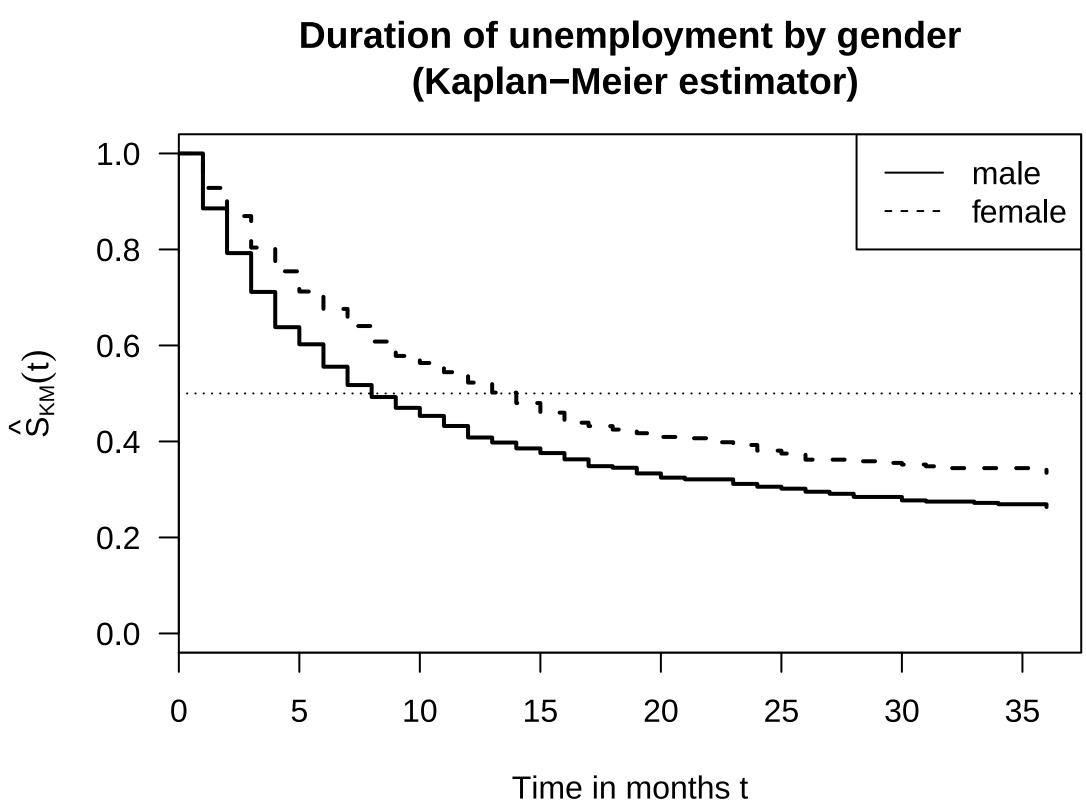
Reveals a step function with jumps at each t_k where events took place.

4.0.2 Data

This is some random SOEP data and we estimate Survival functions for both genders:

```
##   dauer status beginn.monat female male alter bild
## 1    11      0          114      0    1    47    1
## 2    30      1           83      0    1    38    2
## 3     1      1           83      0    1    44    2
## 4    36      0           85      0    1    28    2
## 5     1      1          111      0    1    38    2
## 6     7      0          104      1    0    30    1
```

4.0.3 Model



This gives incidence for the Proportional Hazards assumption as survival curves are more or less parallel.

4.0.4 Test

Plotting estimated confidence intervals **DOES NOT ENABLE** us to interpret signifiance. KI's can cross, and still there is a significant effect.

- Only interpret the p-value of the log rank test!
- log rank test resembles the score test in the cox model.
- '`surfdiff()` for p = 2 > 1 variables: H0: no differencies across 4 resulting groups. If $p < \alpha$: reject H0.

```
## Call:
## survdiff(formula = Surv(dauer, status) ~ female, data = soep)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## female=0 1206      726      651      8.62      22.1
## female=1  794      396      471     11.92      22.1
##
##  Chisq= 22.1  on 1 degrees of freedom, p= 2.6e-06
```

4.0.5 By Hand:

1. order events according to time
2. create following table:

a_{k-1}	a_k	$S_k(t)$	d_k	w_k	n_k	$SE(S(t))$	$+/-KI$
[0,	12[1	0	1	1	4 0		
[12,	29[	0.917	1	1	12	0.08	[1.06, 0.773]
[29,	31[	0.910	1	0	11	0.087	[1.065, 0.755]

3. where:
   - d_k = deaths before k
   - w_k = censored before k
   - $n_k = n_{k-1} - d_k - w_k$ people in risk set before k
   - $S(k) = 1 - d_k/n_k$
4. $Var(S(t=k)) = S(t=k)^2 \frac{d_k}{n_k(n_k-d_k)}$
5. $KI = S(k) + / - z_{1-\alpha/2} * \sqrt{Var(S(t=k))} = S(k) + / - 1.96 * \sqrt{Var(S(t=k))}$

5 Nelson Aalen

Estimate **cumulative Hazard rates** non-parametrically

5.1 Howto:

1. Split time into intervals $[t_{k-1}, t_k[$ on each event-time-point

2. Calculate estimator for cumulative Hazard in the interval via

$$\hat{\Lambda}_{NA}(t) = \sum_{t_k \leq t} \frac{d_k}{n_k} \text{ sum over all intervals until k of events / people under risk}$$

5.2 Breslow

Breslow uses the Nelson-Aalen estimates to calculate **Survival time estimates**

$$\hat{S}_{Breslow}(t) = exp(-\Lambda_{NA}(t))$$

6 Accelerated Failure Time Transformation models

6.1 General

Assumptions:

1. covariates have a multiplicative effect on the **Survival time**. E.g.: Survival time for smokers is an accelerated version of the survival time for non-smokers.

2. the survival time follows an assumed distribution that we get applying a density transformation

We model the survival time directly with the log-trafo:

$$log(T) = Y = \beta_0 + X^T\beta + \sigma\epsilon$$

$$T = exp(Y) = exp(\beta_0) * exp(X^T\beta) * exp(\sigma\epsilon)$$

with $\epsilon \sim$ Distribution e.g.: SEV, Normal, logistic, ...

Thus, the effect of the estimated coefficient $\hat{\beta}_j$ on Survival time T is $exp(\beta_j)$

Steps for the estimation:

1. calculate density for T

2. classic Maximum Likelihood Estimation

6.2 Relation to Cox PH model

In the AFT model we can show, that $T \sim WB(\alpha, \lambda)$ with $\alpha = \frac{1}{\sigma}, \lambda = exp(-X'\beta)$. The hazard from a Weibull distributed random variable is $\lambda(t|X) = \lambda\alpha(\lambda t)^{\alpha-1}$. Plugging above in, we yield:

$$\lambda(t|X) = \lambda^{\alpha}\alpha t^{\alpha-1}$$
$$\lambda(t|X) = exp(-X'\beta_{AFT})^{\frac{1}{\sigma}}\frac{1}{\sigma}t^{\frac{1}{\sigma}-1}$$
$$\lambda(t|X) = exp(-\frac{X'\beta_{AFT}}{\sigma})\frac{1}{\sigma}t^{\frac{1}{\sigma}-1}$$

Whereas in the normal Cox we yield:

$$\lambda t|X = \lambda_0(t)exp(X'\beta_{COX})$$

Thus, the relationship

$$\beta_{Cox} = -\frac{\beta_{AFT}}{\sigma}$$

, where *sigma* is our scale parameter from the AFT model equation, holds **if the AFT model satisfies the PH assumption! This interpretation goes only from AFT to CoxPH, not in both directions!**.

Therefore we compare with this baseline Cox model:

```
## Call:
## coxph(formula = Surv(futime, fustat) ~ ecog.ps + rx, data = ovarian)
##
##   n= 26, number of events= 12
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
```

```
## ecog.ps  0.3698    1.4474   0.5869  0.630    0.529
## rx      -0.5782    0.5609   0.5878 -0.984    0.325
##
##         exp(coef) exp(-coef) lower .95 upper .95
## ecog.ps    1.4474     0.6909    0.4582     4.573
## rx         0.5609     1.7829    0.1772     1.775
##
## Concordance= 0.622  (se = 0.088 )
## Rsquare= 0.054   (max possible= 0.932 )
## Likelihood ratio test= 1.45  on 2 df,   p=0.4833
## Wald test            = 1.43  on 2 df,   p=0.4897
## Score (logrank) test = 1.46  on 2 df,   p=0.4808
```

6.3 Graphical Check of Assumptions

Nice property of the Weibull AFT:

$$S(t) = exp(-(\lambda t)^{\alpha})$$

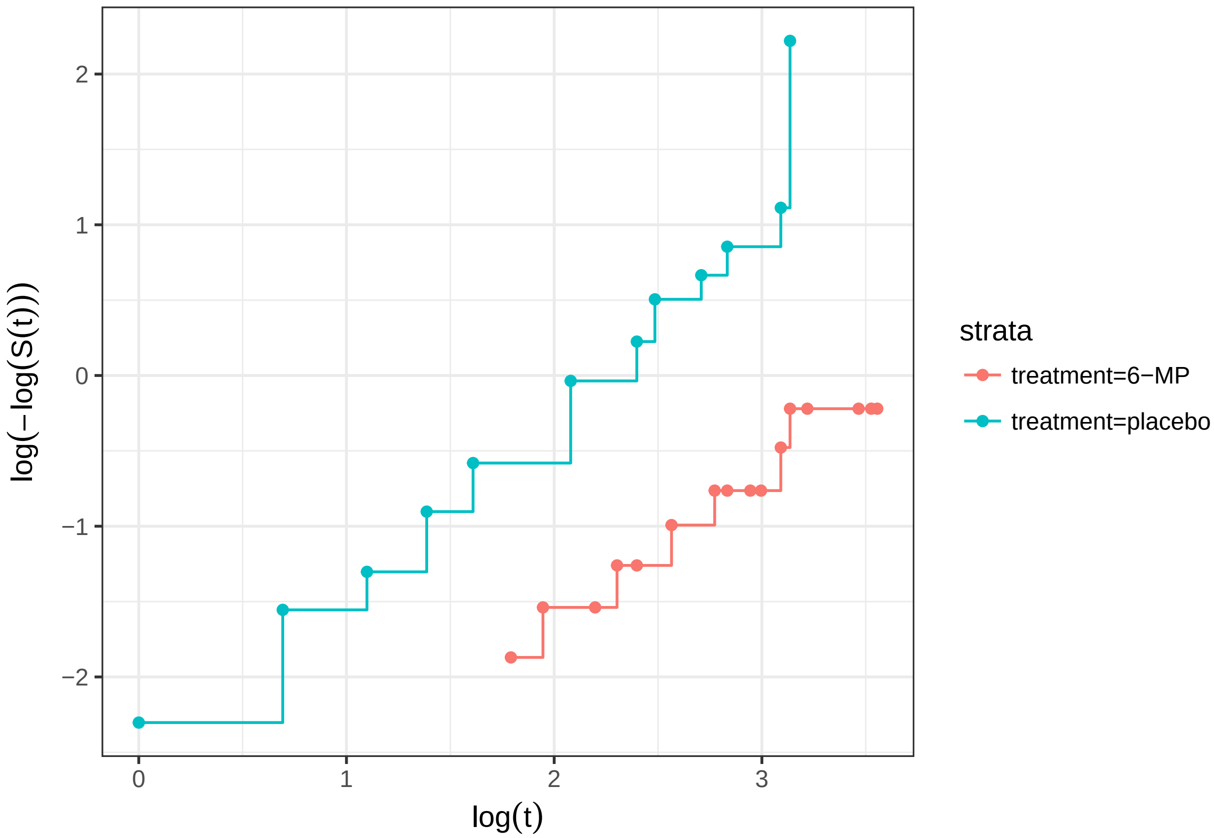
$$\Rightarrow ln[-ln[S(t)]] = \alpha ln(\lambda) + \alpha ln(t)$$

$$\Rightarrow \text{left factor is linear in } \alpha ln(t) \text{ and } \alpha ln(\lambda) \text{ is the intercept of the curves}$$

We exploit this propery to check the Weibull assumption for a binary variable graphically: plot $ln[-ln[S(t)]]$ vs. $ln(t)$ where we estimate $\hat{S}(t)$ via Kaplan-Meier (or estimate $-ln[S(t)] = \Lambda(t)$) via Nelson-Aalen. Then we can check two assumptions:

1. Are the lines for the two Strati parallel? Yes: PH assumption satisfied!
2. Lines straight? Yes: Weibulld / Exponential AFT applicable!
3. Lines parallel but not straight? Yes: PH still valid, but AFT model not applicable, use Cox PH!
4. Is the slope of the lines $\alpha = 1$? Yes: data follows $WB(\alpha = 1, \lambda) = Exp(\lambda)$ distribution!

Example:



1. Curves are parallel and straight: PH assumption holds and AFT is applicable!
2. Slope ≈ 1: Exponential AFT is applicable

6.4 Exponential

```
##
## Call:
## survreg(formula = Surv(futime, fustat) ~ ecog.ps + rx, data = ovarian,
##     dist = "exponential")
##              Value Std. Error      z        p
## (Intercept)  6.962      1.322  5.267 1.39e-07
## ecog.ps     -0.433      0.587 -0.738 4.61e-01
## rx           0.582      0.587  0.991 3.22e-01
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -97.2   Loglik(intercept only)= -98
##  Chisq= 1.67 on 2 degrees of freedom, p= 0.43
## Number of Newton-Raphson Iterations: 4
## n= 26
```

6.4.1 AFT interpretation:

- Geometric mean of survival time: 1055

- 1 unit change in ecog.ps shortens survival time by exp(-0.433) = 0.65
- 1 unit change in rx increases survival time by exp(0.582) = 1.79
- though, both effects are non significant

```
##  (Intercept)      ecog.ps           rx
## 1055.5715021    0.6484732    1.7887244
```

6.4.2 PH interpretation (invert the coefficients)

- 1 unit change in ecog.ps increases the hazard h(t) by 1/exp(0.433) = 1.54
- 1 unit change in rx decreases h(t) by 0.56

```
##  (Intercept)      ecog.ps           rx
## 0.0009473541 1.5420838523 0.5590576235
```

6.5 Weibull

$$T = exp(Y) = exp(X^T\beta)exp(\sigma\epsilon), \text{ with } \epsilon \sim SEV$$

using the **density transformation rule**

$$f_T(T) = f_\epsilon(g^{-1}(T))det|\frac{\partial g^{-1}(T)}{\partial T}|$$

we can show that $T \sim Weibull(\alpha, \lambda)$ with $\alpha = \frac{1}{\sigma}$ and $\lambda = exp(-X^T\beta)$. Thus:

- $\lambda(t|X) = \frac{1}{\sigma}t^{\frac{1}{\sigma}-1}exp(X^T\beta)$
- $S(t|X) = exp(-exp(-X^T\beta)^{\frac{1}{\sigma}}t^{\frac{1}{\sigma}})$

```
##
## Call:
## survreg(formula = Surv(futime, fustat) ~ ecog.ps + rx, data = ovarian,
##     dist = "weibull")
##              Value Std. Error      z        p
## (Intercept)  6.897      1.178  5.857 4.72e-09
## ecog.ps     -0.385      0.527 -0.731 4.65e-01
## rx           0.529      0.529  0.999 3.18e-01
## Log(scale)  -0.123      0.252 -0.489 6.25e-01
##
## Scale= 0.884
##
## Weibull distribution
## Loglik(model)= -97.1   Loglik(intercept only)= -98
##  Chisq= 1.74 on 2 degrees of freedom, p= 0.42
## Number of Newton-Raphson Iterations: 5
## n= 26
```

6.5.1 AFT interpretation:

- Geometric mean of survival time: 988

- 1 unit change in ecog.ps shortens survival time by exp(-0.385) = 0.68

- 1 unit change in rx increases survival time by exp(0.529) = 1.70
- Scale parameter gives flexibility to the model:
  - **transform to shape parameter from Weibull distribution:** shape = 1/scale aka $\alpha = \frac{1}{\sigma}$
  - shape < 1: hazard decreases over time
  - shape = 1: constant hazard over time (= Exponential AFT)
  - shape > 1: hazard increases over time
  - Here: $shape = \alpha = \frac{1}{\sigma} = 1/0.844 = 1.13$
- though, both effects are non significant

- If scale parameter

```
## [1] 0.8838731
```

was close to 1 we would yield an exponential model. * coefficients:

```
## (Intercept)     ecog.ps          rx
## 988.9988102   0.6804217   1.6966327
```

6.5.2 PH interpretation (multiply by -1 and the shape parameter before exp())

- 1 unit change in ecog.ps increases the hazard h(t) by $exp(- -0.385/0.844) = 1.55$
- 1 unit change in rx decreases h(t) by 0.55

```
##  (Intercept)      ecog.ps           rx
## 0.0004085855 1.5459383069 0.5498547398
```

6.6 Log Normal

Only AFT Interpretation!

- 1 unit incrase in ecog.ps shortens survival time by exp(-.229) = 0.79
- 1 unit incrase in rx increases survival time by exp(0.813) = 2.25
- Can we interpret the scale parameter? Yes, but how?
- MORE TO ADD! DISCUSS

```
##
## Call:
## survreg(formula = Surv(futime, fustat) ~ ecog.ps + rx, data = ovarian,
##     dist = "lognormal")
##              Value Std. Error      z        p
## (Intercept)  5.878      1.094  5.373 7.72e-08
## ecog.ps     -0.229      0.537 -0.427 6.70e-01
## rx           0.813      0.537  1.514 1.30e-01
## Log(scale)   0.167      0.228  0.731 4.65e-01
##
## Scale= 1.18
##
## Log Normal distribution
## Loglik(model)= -95.9   Loglik(intercept only)= -97.1
##  Chisq= 2.35 on 2 degrees of freedom, p= 0.31
## Number of Newton-Raphson Iterations: 3
## n= 26
```

6.7 Log logistic

```
##
## Call:
## survreg(formula = Surv(futime, fustat) ~ ecog.ps + rx, data = ovarian,
##     dist = "loglogistic")
##              Value Std. Error      z        p
## (Intercept)  6.161      1.134  5.435 5.49e-08
## ecog.ps     -0.336      0.537 -0.626 5.32e-01
## rx           0.705      0.539  1.308 1.91e-01
## Log(scale)  -0.363      0.248 -1.466 1.43e-01
##
## Scale= 0.695
##
## Log logistic distribution
## Loglik(model)= -96.3   Loglik(intercept only)= -97.4
##  Chisq= 2.07 on 2 degrees of freedom, p= 0.36
## Number of Newton-Raphson Iterations: 4
## n= 26
```

7 Cox Regression model

Estimates coefficents β that have multiplicative effect on time-dependent hazard $\lambda_0(t)$. The baseline hazard is estimated non-parametrically via Breslow estimate. Thus, we yield step-functions for visualization, estimation, …

super sweet R-bloggers post on Cox models

7.0.1 Model equation

$$\lambda_i(t) = \lambda_0(t) exp(x_i'\beta)$$

To get the estimator for the cumulative Hazard and the Survival rate:

1. estimate βs via Cox **parametrically**

2. estimate non-parametrically baseline hazards $\lambda_0(t)$ e.g. via Breslow **non-parametrically**

3. calculate for each t $\lambda(t) = \lambda_0(t) exp(x_i'\beta)$

4. cumulate the $\lambda(t)$ to the cumulative Hazards $\Lambda_t = \sum_{i=1}^{t} \lambda_i$. `basehaz()` plottet Λ_0

5. calculate estimated Survival $S(t) = exp(-\Lambda_t)$
   Therfore Cox PH model is termed **semi parametric**.

7.0.2 Data

where delta depicts the event indicator (delta = 1: non-censored, delta = 0: censored)

```
##   type time delta
## 1    1    1     1
## 2    1    3     1
## 3    1    3     1
## 4    1    4     1
## 5    1   10     1
## 6    1   13     1
```

7.0.3 Model

We are searching for the effect of the binary treatment type.

- Person with type 2 has a multiplicative factor exp(0.4664) = 1.594245 higher hazard rate than a person with type 1 (ceteris paribus in case of other covariates)

- this effect is not significant as the H0 can not be rejected at $\alpha = 0.05$, REMIND but this does not imply testing of the PH assumption

- (log rank-) score test: tests for significant differencies in the survival curves for the two subpopulations seperated by the **categorical variable** of interest (here: treatment). This means that the probability of an event occurring at any time point is the same for each subpopulation. H0: they do not differ -> p > 0.05: H0 cannot be rejected -> no significant effect of treatment. If there are more than 1 categorical variablewe have the H0: no effect of no covariate at all. Reject again if $p < \alpha$
- ‘`surfdiff()` for p = 2 > 1 variables: H0: no differencies across 4 resulting groups. If $p < \alpha$: reject H0.

- Partial likelihood test: for **continuos** variables! WHAT HAPPENS WITH MORE COVARIATES? E.G.: one significant, the other not

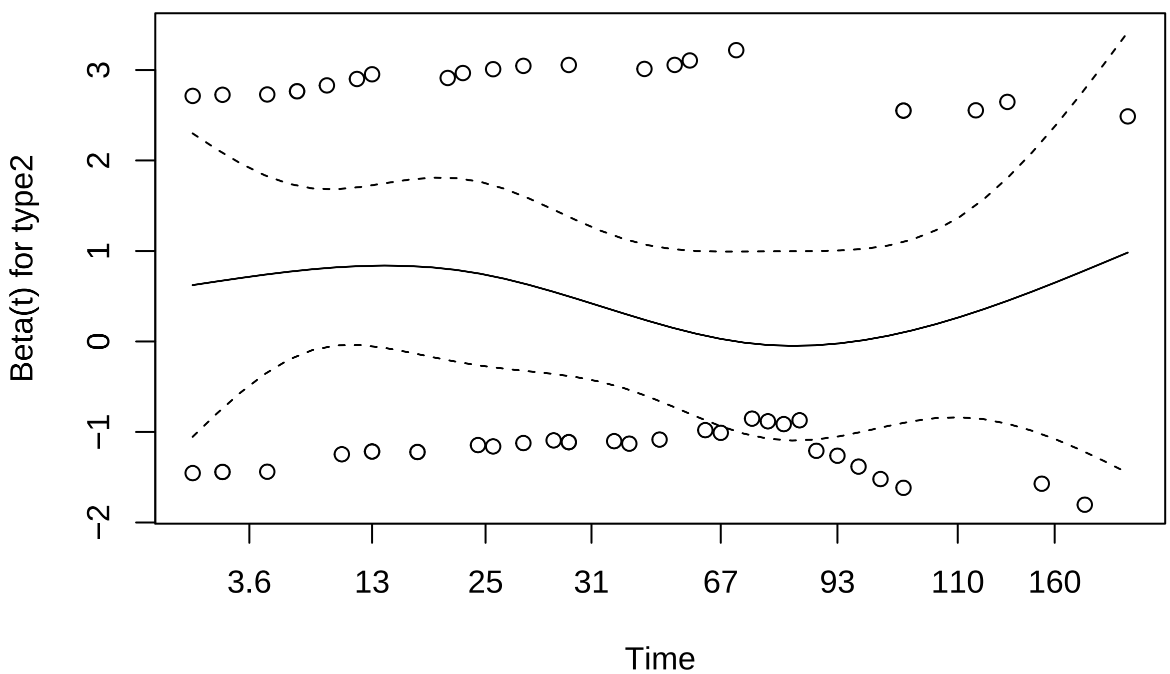
Summary of the Cox-PH model:

```
## Call:
## coxph(formula = Surv(time, delta) ~ type, data = tongue)
##
##   n= 80, number of events= 53
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## type2 0.4664    1.5942   0.2804 1.663   0.0963 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## type2     1.594     0.6273    0.9201     2.762
##
## Concordance= 0.564  (se = 0.036 )
## Rsquare= 0.033   (max possible= 0.993 )
## Likelihood ratio test= 2.67  on 1 df,   p=0.102
## Wald test            = 2.77  on 1 df,   p=0.09632
## Score (logrank) test = 2.81  on 1 df,   p=0.09343
```

7.0.4 Test the Cox PH assumption for the covariates

7.0.4.1 Graphically

The scaled Schoenfeld residuals are used for that test and plotted against the time. Do this for each covariate to check the PH assumption for each covariate. If they **randomly and unstructured** center around zero: PH assumption holds! If not, not. The plot estimates a smooth function of the residuals over time for better visualization. Holds here:

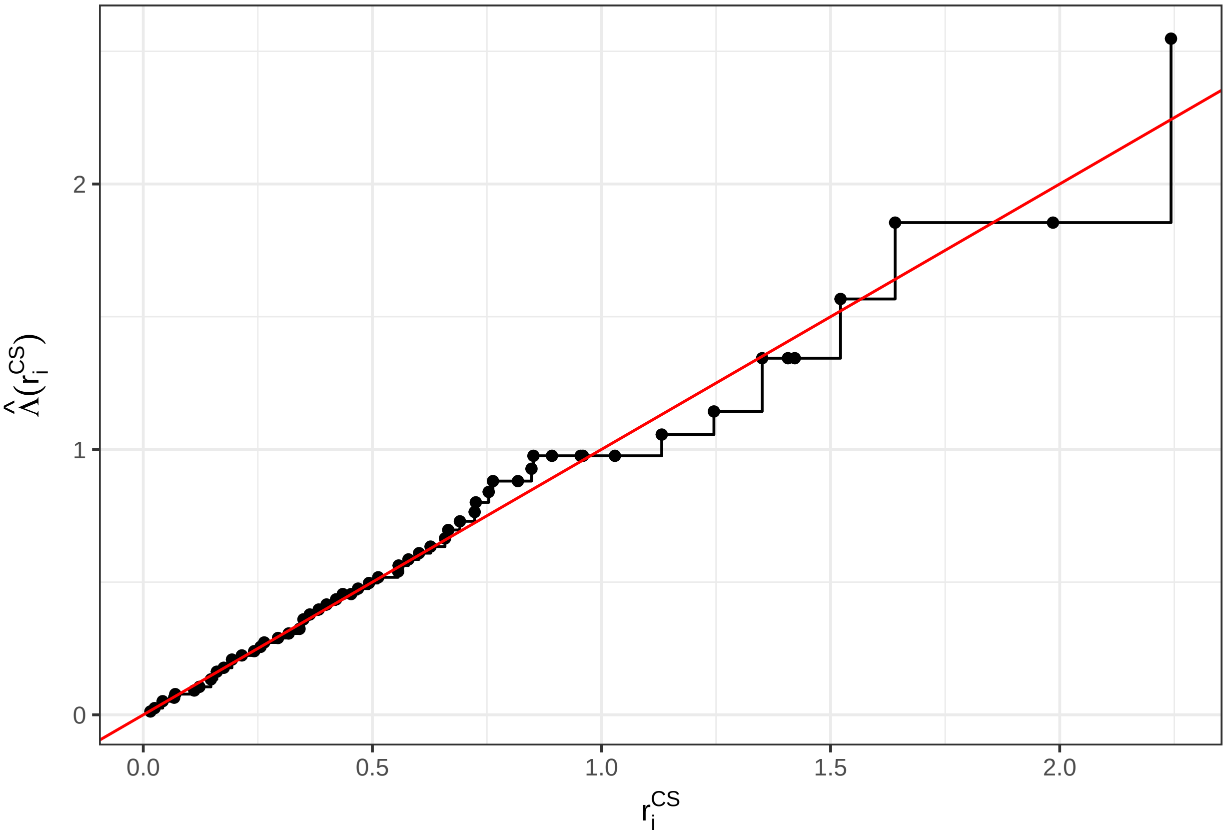


7.0.4.2 Test PH

Also based on Schoenfeld residuals, not exam-relevant. If $p >> 0.05$ there is no violation of the PH.

7.0.5 Test overall fit

Plot Cox-Snell residuals vs. Cumulated Hazard. If they share the diagonal, everything is fine and we have a good overall model fit.



8 Stratified Cox Model

Assumptions:

1. Separate Baseline Hazard for each of the two Strati
2. Effects of covariates on hazard rates identical in both Strati

Requirements:

1. Binary, categorical variables
2. Makes sense if PH for covariate not satisfied
3. No interest on interpretation of the variable

Estimation:

1. Coefficients via sum of Partial Likelihoods $\sum_{z_0} PL(\beta) + \sum_{z_1} PL(\beta)$
2. Baseline hazard___s___ via Breslow

9 Semi-parametric additive Cox model

- $\lambda_(t|X) = \lambda_0(t)exp(f_1(x_1)\beta_1 + ... + f_k(x_k)\beta_k + x_{k+1}\beta_{k+1} + ... + x_p\beta_p)$
- Estimate $f_j(x_j)$ via splines for smooth nonlinear effects
- Example: age has non-linear effect, smooth age variable via Splines

10 Cox model: time varying covariates

Effect remains constant, but covariate varies over time: $\beta x(t)$. For instance, the effect of being employed or unemployed stays constant over time, but the employment status of the people varies over time.

We convert

```
##   week arrest fin age race wexp mar paro prio educ emp1 emp2 emp3 emp4
## 1   20      1   0  27    1    0   0    1    3    3    0    0    0    0
## 2   17      1   0  18    1    0   0    1    8    4    0    0    0    0
## 3   25      1   0  19    0    1   0    1   13    3    0    0    0    0
## 4   52      0   1  23    1    1   1    1    1    5    0    0    0    0
##   emp5 emp6 emp7 emp8 emp9 emp10 emp11 emp12 emp13 emp14 emp15 emp16 emp17
## 1    0    0    0    0    0     0     0     0     0     0     0     0     0
## 2    0    0    0    0    0     1     1     1     1     1     0     0     0
## 3    0    0    0    0    0     0     0     0     0     0     0     0     1
## 4    1    1    1    1    1     1     1     1     1     1     1     1     1
##   emp18 emp19 emp20 emp21 emp22 emp23 emp24 emp25 emp26 emp27 emp28 emp29
## 1     0     0     0    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 2    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 3     0     0     0     0     0     0     0     0    NA    NA    NA    NA
## 4     1     1     1     1     0     0     0     0     0     0     0     0
##   emp30 emp31 emp32 emp33 emp34 emp35 emp36 emp37 emp38 emp39 emp40 emp41
## 1    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 2    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 3    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 4     0     0     1     1     1     1     1     1     1     1     1     1
##   emp42 emp43 emp44 emp45 emp46 emp47 emp48 emp49 emp50 emp51 emp52
## 1    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 2    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 3    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
## 4     1     1     1     1     1     1     1     1     1     1     1
##   subject
## 1       1
## 2       2
## 3       3
## 4       4
```

to long format

```
##   subject calendar.week start stop arrest employed fin age race wexp mar
## 1       1             1     0    1      0        0   0  27    1    0   0
## 2       1             2     1    2      0        0   0  27    1    0   0
## 3       1             3     2    3      0        0   0  27    1    0   0
## 4       1             4     3    4      0        0   0  27    1    0   0
##   paro prio educ
## 1    1    3    3
## 2    1    3    3
## 3    1    3    3
## 4    1    3    3
```

We yield the same Coefficients for both data sets if we include only the time-constant predictors:

```
## Call:
## coxph(formula = Surv(week, arrest) ~ fin + age + mar + prio,
##     data = prison.short, method = "efron")
##
```

```
##         coef exp(coef) se(coef)     z       p
## fin  -0.3602    0.6975   0.1905 -1.89 0.05864
## age  -0.0604    0.9414   0.0209 -2.90 0.00376
## mar  -0.5331    0.5868   0.3728 -1.43 0.15266
## prio  0.0975    1.1024   0.0272  3.58 0.00034
##
## Likelihood ratio test=31.4  on 4 df, p=2.53e-06
## n= 432, number of events= 114
```

```
## Call:
## coxph(formula = Surv(start, stop, arrest) ~ fin + age + mar +
##     prio, data = prison.long)
##
##         coef exp(coef) se(coef)     z       p
## fin  -0.3602    0.6975   0.1905 -1.89 0.05864
## age  -0.0604    0.9414   0.0209 -2.90 0.00376
## mar  -0.5331    0.5868   0.3728 -1.43 0.15266
## prio  0.0975    1.1024   0.0272  3.58 0.00034
##
## Likelihood ratio test=31.4  on 4 df, p=2.53e-06
## n= 19809, number of events= 114
```

Now we include the time-varying employment variable and yield a constant effect coefficient for the time varying variable.

```
## Call:
## coxph(formula = Surv(start, stop, arrest) ~ fin + age + prio +
##     mar + employed, data = prison.long)
##
##             coef exp(coef) se(coef)     z      p
## fin      -0.3390    0.7125   0.1904 -1.78 0.0750
## age      -0.0460    0.9551   0.0206 -2.23 0.0255
## prio      0.0842    1.0878   0.0278  3.03 0.0024
## mar      -0.3612    0.6968   0.3733 -0.97 0.3333
## employed -1.3290    0.2647   0.2498 -5.32  1e-07
##
## Likelihood ratio test=67.2  on 5 df, p=3.87e-13
## n= 19809, number of events= 114
```

11 Cox model: time varying effects

11.1 Idea

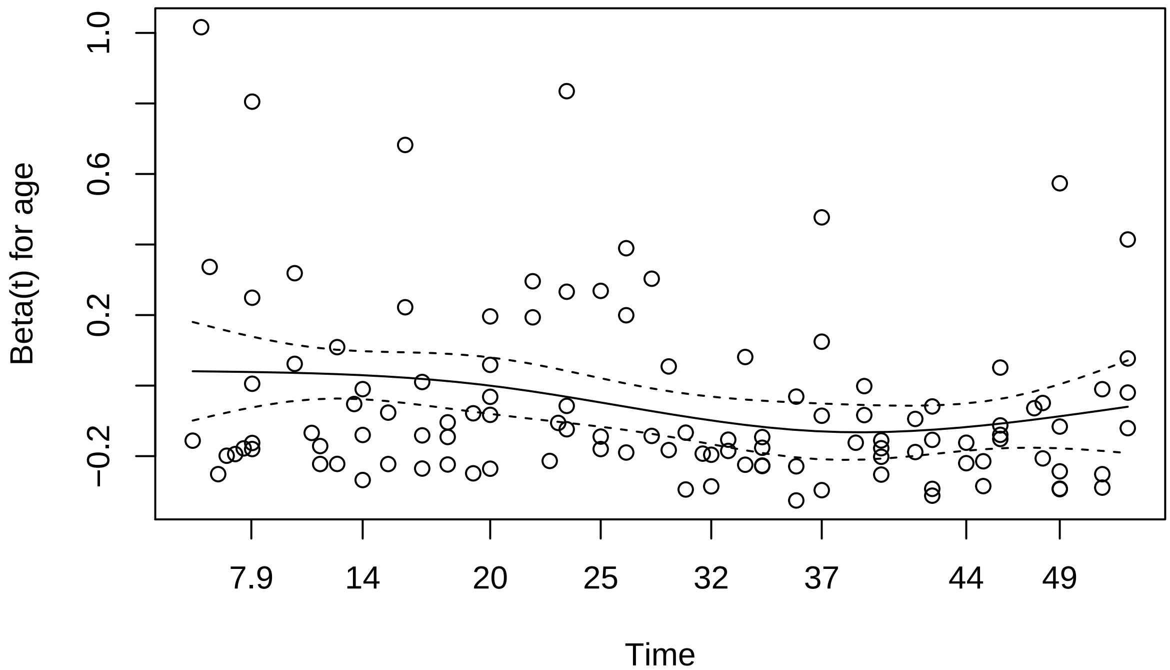
Covariate remains constant but the effect changes over time: $\beta(t) \times x$. The same value of x has a different effect at t_1 than an t_2. Create a fake-time-varying variable, e.g. $log(t)$ and model the coefficients for the interaction of the real variable with the time-dummy. No clear interpretation as coefficients are bound to time-fake-dummy but plot of the effect over time possible.

11.2 Example

Old model with wrong intake of age variable, detected by the schoenfeld test. We assume age to have a time-varying effect on the hazard.

```
## Call:
## coxph(formula = Surv(start, stop, arrest) ~ fin + age + prio +
##     mar + employed.lag1, data = prisss)
##
##   n= 19809, number of events= 114
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## fin           -0.34753   0.70643  0.19039 -1.825 0.067956 .
## age           -0.05106   0.95022  0.02075 -2.461 0.013862 *
## prio           0.08894   1.09302  0.02759  3.223 0.001268 **
## mar           -0.41749   0.65870  0.37373 -1.117 0.263958
## employed.lag1 -0.80028   0.44920  0.21663 -3.694 0.000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## fin              0.7064     1.4156    0.4864    1.0260
## age              0.9502     1.0524    0.9124    0.9897
## prio             1.0930     0.9149    1.0355    1.1538
## mar              0.6587     1.5181    0.3166    1.3703
## employed.lag1    0.4492     2.2262    0.2938    0.6868
##
## Concordance= 0.668  (se = 0.027 )
## Rsquare= 0.002   (max possible= 0.066 )
## Likelihood ratio test= 46.49  on 5 df,   p=7.209e-09
## Wald test            = 42.22  on 5 df,   p=5.326e-08
## Score (logrank) test = 45.11  on 5 df,   p=1.378e-08
```

```
##                   rho  chisq       p
## fin            0.0122 0.0174 0.89513
## age           -0.2150 6.8214 0.00901
## prio          -0.0714 0.6283 0.42798
## mar            0.1295 2.0109 0.15617
## employed.lag1 -0.0455 0.2352 0.62769
## GLOBAL             NA 8.7605 0.11900
```

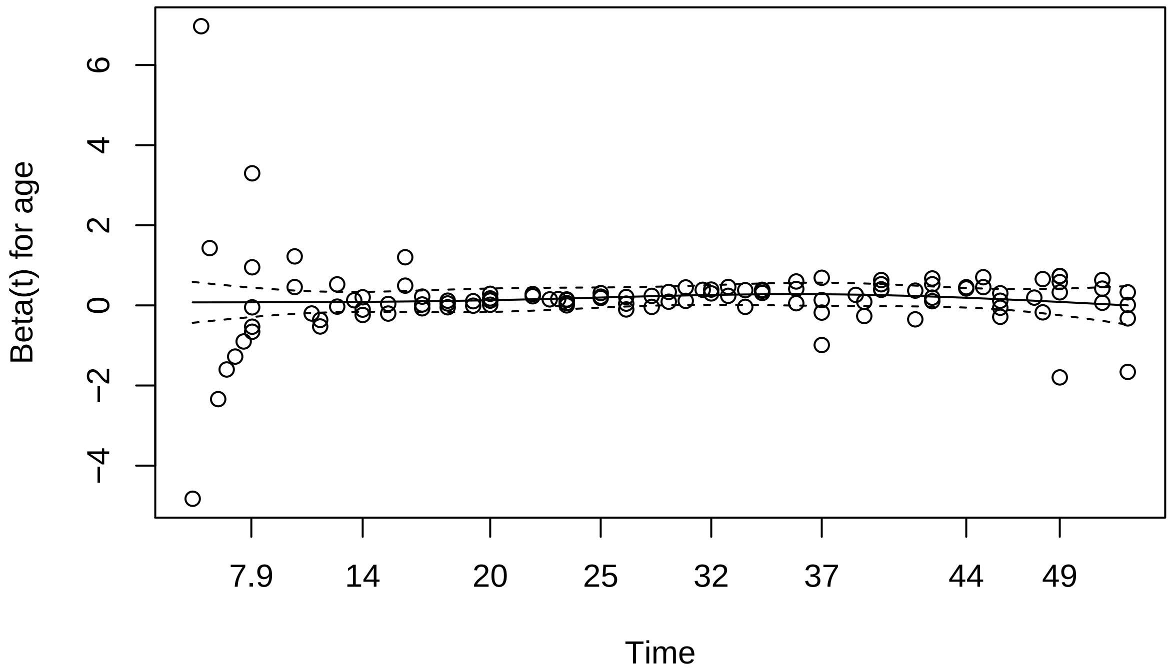


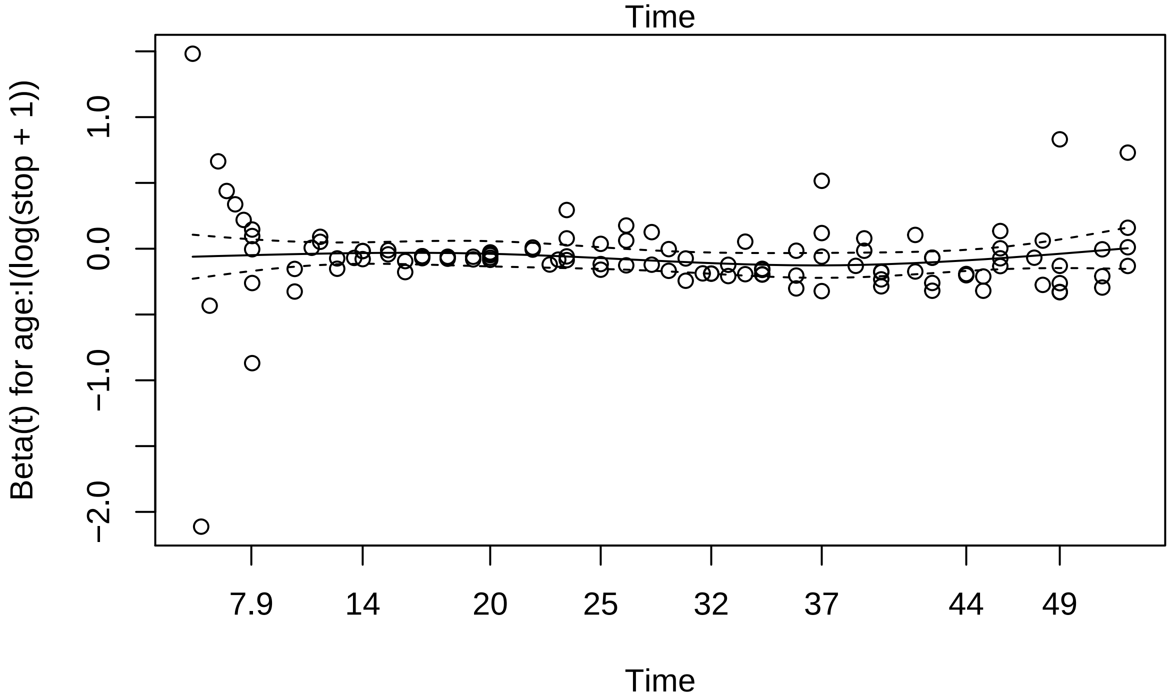
Model age explicitly via an interaction term with time such that $x_{age2} = x_{age}log(t+1)$. This yields perfect Schoenfeld-plots and an increase in the global p-values of the zph test:

```
## Call:
## coxph(formula = Surv(start, stop, arrest) ~ fin + age + age:I(log(stop +
##     1)) + prio + mar + employed.lag1, data = prisss)
##
##   n= 19809, number of events= 114
##
##                          coef exp(coef) se(coef)      z Pr(>|z|)
## fin                  -0.35220   0.70314  0.19038 -1.850  0.06432 .
## age                   0.15824   1.17145  0.07580  2.088  0.03682 *
## prio                  0.09021   1.09440  0.02763  3.265  0.00110 **
## mar                  -0.38543   0.68016  0.37409 -1.030  0.30286
## employed.lag1        -0.79313   0.45243  0.21655 -3.663  0.00025 ***
## age:I(log(stop + 1)) -0.06768   0.93455  0.02485 -2.724  0.00645 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## fin                     0.7031     1.4222    0.4842    1.0212
## age                     1.1715     0.8536    1.0097    1.3591
## prio                    1.0944     0.9137    1.0367    1.1553
## mar                     0.6802     1.4702    0.3267    1.4159
## employed.lag1           0.4524     2.2103    0.2960    0.6916
## age:I(log(stop + 1))    0.9346     1.0700    0.8901    0.9812
##
## Concordance= 0.678  (se = 0.027 )
## Rsquare= 0.003   (max possible= 0.066 )
## Likelihood ratio test= 53.21  on 6 df,   p=1.063e-09
## Wald test            = 46.62  on 6 df,   p=2.229e-08
## Score (logrank) test = 50.13  on 6 df,   p=4.419e-09
```

```
##                          rho  chisq     p
```

```
## fin                     0.0268 0.0839 0.772
## age                     0.0246 0.1066 0.744
## prio                   -0.0722 0.6419 0.423
## mar                     0.1313 2.0545 0.152
## employed.lag1          -0.0486 0.2666 0.606
## age:I(log(stop + 1)) -0.0344 0.2006 0.654
## GLOBAL                       NA 3.0166 0.807
```





We can also test the **PH Assumption** for a variable this way: if the time varying effect is not significant, the PH is valid.

11.3 Interpretation

```
##                  age age:I(log(stop + 1))
##           0.15824326          -0.06768484
```

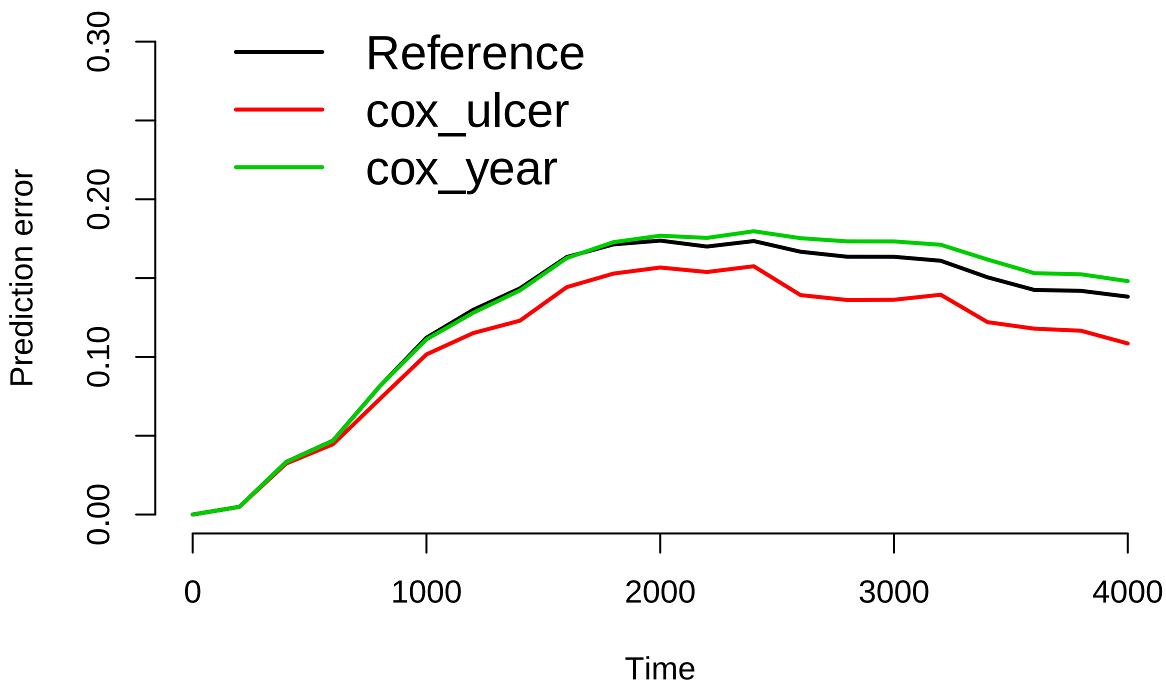
1. At t=0 is hazard of a 20-year old $\frac{exp(0.16\times20-0.07\times20\times0}{exp(0.16\times30-0.07\times30\times0}\approx 0.2$ times as high as the one of a 30-year old ceteris paribus.
2. At t=15 is hazard of a 20-year old $\frac{exp(0.16\times20-0.07\times20\times log(15+1)}{exp(0.16\times30-0.07\times30\times log(15+1)}\approx 1.41$ times as high as the one of a 30-year old ceteris paribus.

12 Model fit Analysis

12.1 Prediction Error Curves (PEC)

The predicted survival time for each time point is compared with the true survival time within the **Brier Score**. Some magic is added such as *inverse probability of censoring weights (IPCW)* to account for right censoring. Then scores for each time point are computed using Cross-Validation and the Brier Scores over time are plotted for all desired models. The lower the score, the better. This method is **model agnostic**.

For Melanoma compare predictive performance of Cox model with only variable ulcer as predictor with the reference Kaplan-Meier estimates and a Cox-PH model that uses year as a linear predictor. We see, that our cox-model outperforms the simple Kaplan-Meier estimator (which does not use any variables) and both outperform the stupid Cox model with time as linear predictor.



12.2 Residuals

12.2.1 Schoenfeld Residuals $s_{i,j}$

Use case: test PH assumption for each covariate

Idea: compute Schoenfeld residuals for Variable k and m observations. Those residuals should be independent of the survival time. This is the test that `cox.zph()` performs.

PH: effects of covariates are proportional and thus, time invariant. Thus, check for timely structure in residuals, if some timely structure *is left in the residuals*, the models assumption failed.

H0: Corr(Schoenfeld-Residuals, Ranking of event times) = 0. If $p < 0.05$, reject that H0 and the corresponding feature violates the PH assumption as there is still timely structure left in the residuals.

12.2.1.1 Test

```
##                  rho  chisq       p
## fin           0.0267 0.0836 0.77247
## age          -0.2263 7.5599 0.00597
```

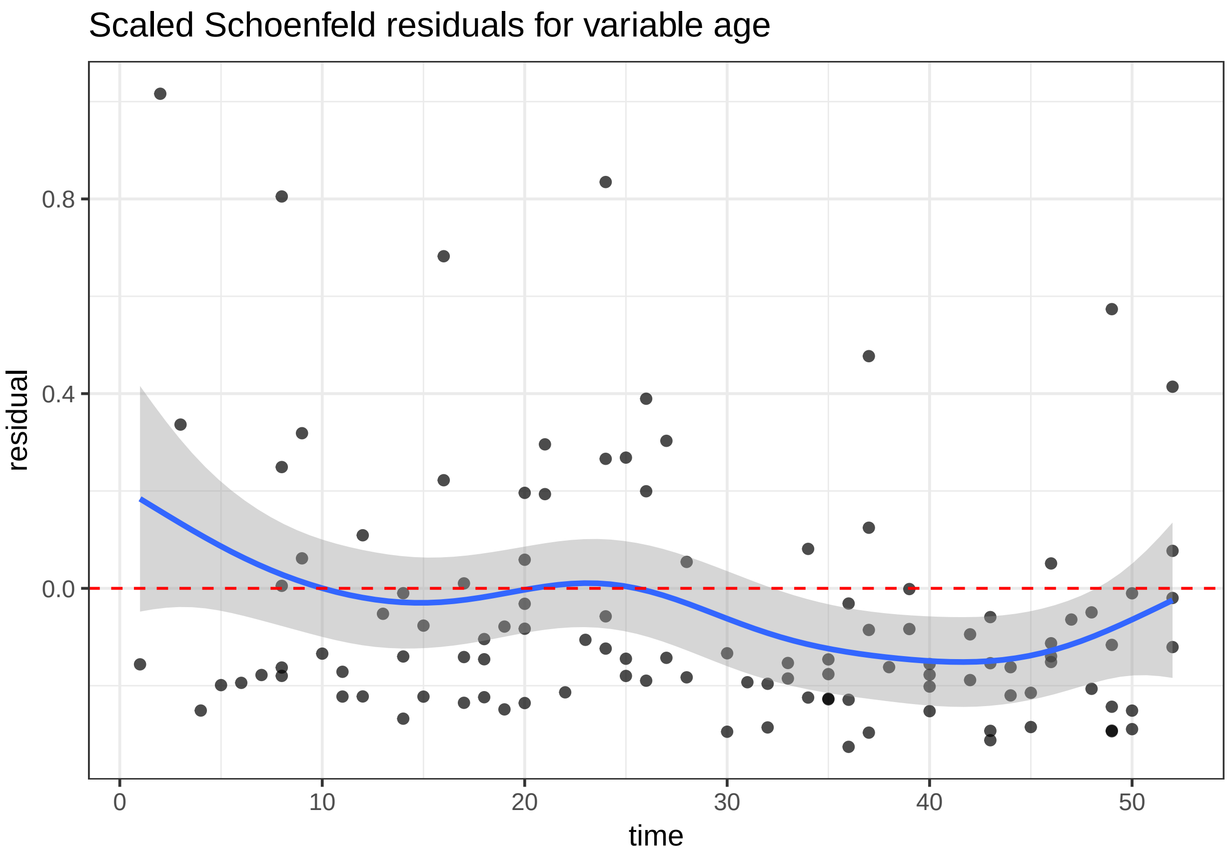
```
## prio          -0.0658 0.5333 0.46523
## mar            0.1327 2.1141 0.14594
## employed.lag1 -0.0427 0.2074 0.64879
## GLOBAL             NA 9.4127 0.09369
```

Small p-value for variable age indicates problem with the PH assumption here. High value for employed.lag1 indicates nice fulfillment of PH assumption.

Can we observe this graphically?

12.2.1.2 Graphically

Plot the Schoenfeld residuals for variable age:

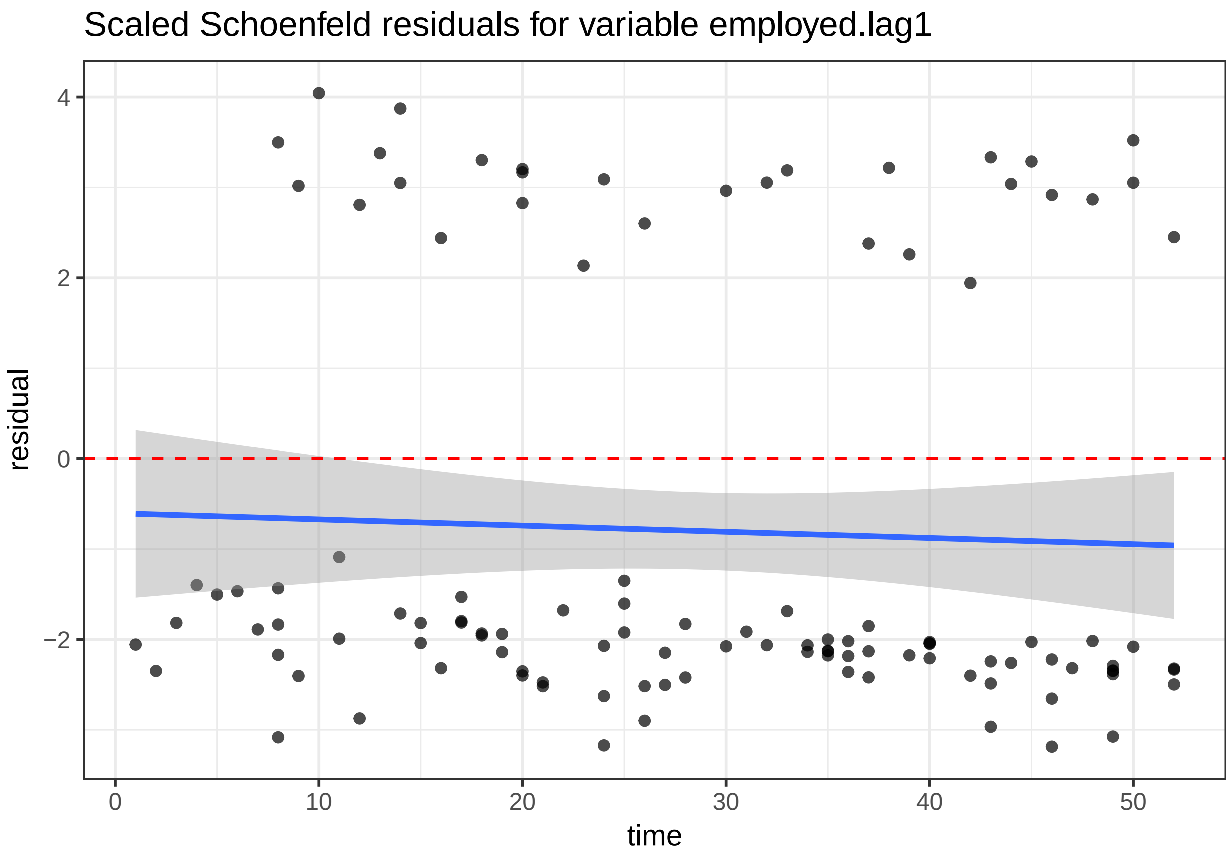


PH assumption violated because there is non linear structure in the data.

What can we do?

1. Exclude variable
2. **additionally** model time varying effect as e.g. $x_{age} \cdot log(1+t)$
3. non-linearly e.g. using splines

Check variable `employed lag1` that had huge p-value in zph test (good sign for PH):



We see what we expected: there seems to be no PH violation. Sweet!

12.2.2 Martingale residuals m_i

$m_i = \delta_i - r_i$ Use case: determine the functional form to be used for a given covariate.

Measure of the difference between expected and observed events. Takes values between $-\infty$ and $+1$. $\sum_{i=1}^{n} m_i = 0$ plot residuals against a single covariate. A smoothed fit is used. If the plot is linear no transformation of the covariate is needed. If there is a threshold, a discretization may be in order.

12.2.3 Deviance residuals d_i

Use case: Check for outliers. Assess the effect of a given individual on the model.

Idea: One could use martingale residuals , but they are highly skewed. The logarithm inflates values of the martingale residual close to 1 and shrinks large negative values. This leads to a more normally shaped distribution.

Plot d_i versus the risk scores $x'\beta$ (it's from the book page 381. the slides say something different, but it doesn't make a lot of sense imo...)

12.2.4 Cox Snell residuals r_i

Use case: Check overall goodness of fit

12.2.4.1 Graphically

H0: Model works - Cox-snell residuals should follow an Exp(1) distribution. If the cox-snell-residuals distribution deviates strongly from the Exp(1), the model does not fit well.

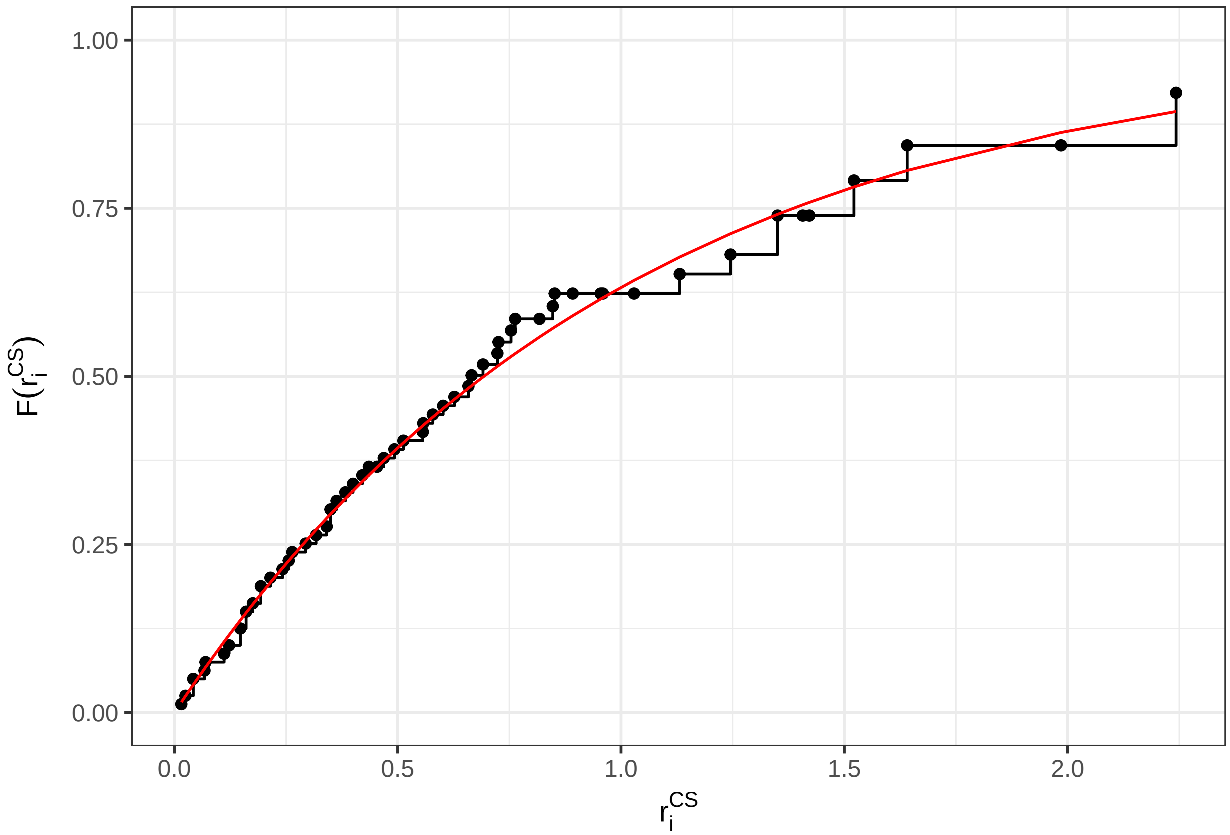
12.2.4.2 Test

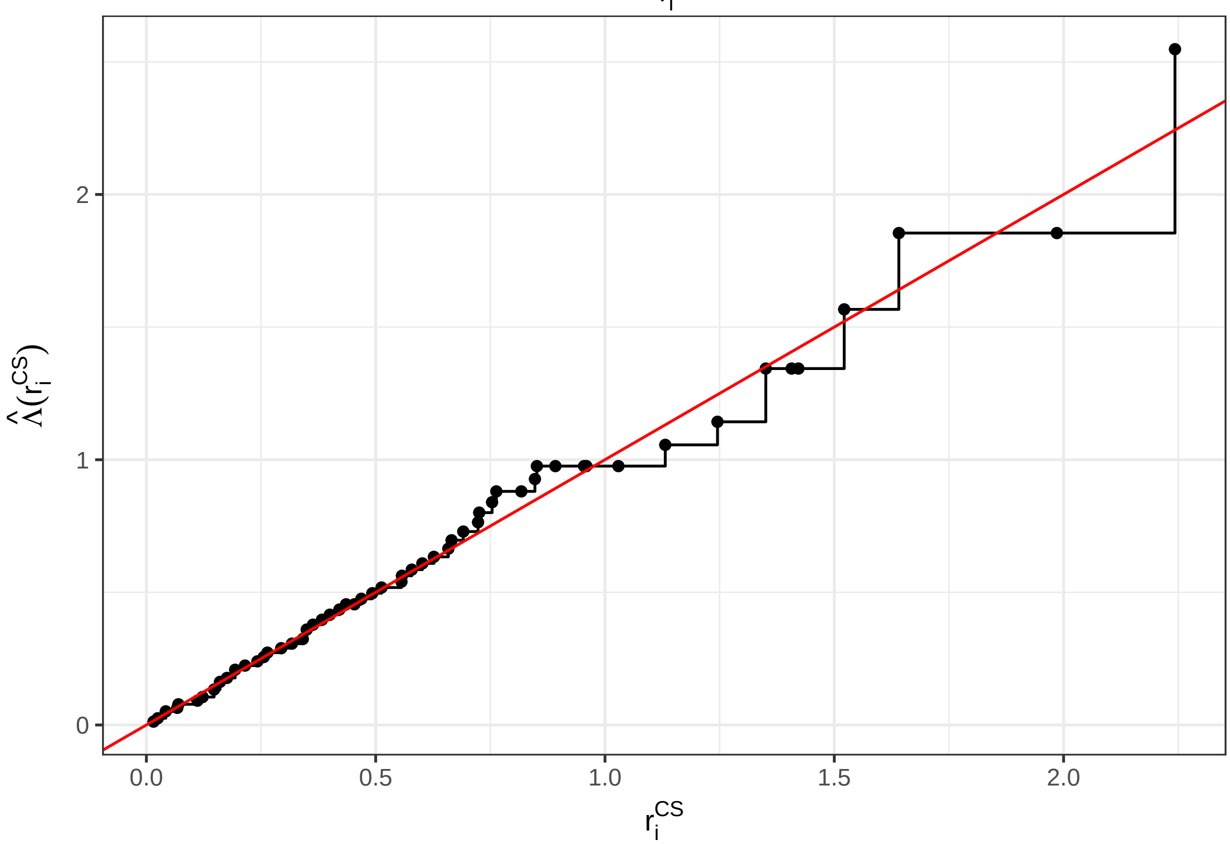
12.2.4.3 Model

12.2.4.4 Example

Check the overall goodness of fit for a simple cox model:

```
## Call:
## coxph(formula = Surv(time, delta) ~ type, data = tongue)
##
##   n= 80, number of events= 53
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## type2 0.4664    1.5942   0.2804 1.663   0.0963 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## type2     1.594     0.6273    0.9201     2.762
##
## Concordance= 0.564  (se = 0.036 )
## Rsquare= 0.033   (max possible= 0.993 )
## Likelihood ratio test= 2.67  on 1 df,   p=0.102
## Wald test            = 2.77  on 1 df,   p=0.09632
## Score (logrank) test = 2.81  on 1 df,   p=0.09343
```





Two options: 1. Plot cs-residuals against estimated distribution Function values. Their distribution should then follow a standard exponential distribution if the model is fit correctly. 2. Plot against estimted cumulative hazard function. This should result in a straight line if the model fits the data.

12.3 (Partial) Log Likelihood Ratio Test

Idea: Test reduced model β_0 against full model β and check, which fits better.

Formally: $H0 : C\beta = d$ and $H1 : C\beta \neq d$.

In standard R output: reduced model is model with all $\beta_0^T = 0^T$ and the full model is the fitted model. Formally this means $H0 : C\beta = 0$ and $H1 : C\beta \neq 0$.

Test statistics:

$$lq = 2(logPL(\hat{\beta}) - logPL_{H0}(\hat{\beta})) \sim \chi^2_{df}$$

H0: all coefficients are insignificant.

- $lq > \chi^2_{df}(1-\alpha) \rightarrow$ reject H0
- $p < \alpha \rightarrow$ reject H0 aka $\hat{\beta}$ **is not insignificant**.

Example:

```
## Call:
## coxph(formula = Surv(time, delta) ~ type, data = tongue)
##
##   n= 80, number of events= 53
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## type2 0.4664    1.5942   0.2804 1.663   0.0963 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## type2     1.594     0.6273    0.9201     2.762
##
## Concordance= 0.564  (se = 0.036 )
## Rsquare= 0.033   (max possible= 0.993 )
## Likelihood ratio test= 2.67  on 1 df,   p=0.102
## Wald test            = 2.77  on 1 df,   p=0.09632
## Score (logrank) test = 2.81  on 1 df,   p=0.09343
```

The p-value of the Likelihood ratio test is $0.102 > \alpha = 0.05$: we cannot reject the H0 that the coefficient vector β (here with only one coefficient for `type2`) is equal to 0. This goes in line with the p-value for this coefficient. We have 1df as there is only one coefficient to be tested. Works the same way with additional coefficients.

12.4 (Log rank) Score test

Idea: Tests for significant differencies in the survival curves for the two subpopulations seperated by the **categorical variable** of interest (above: type2). This means that the probability of an event occurring at any time point is the same for each subpopulation.

- H0: they do not differ
- p > 0.05: H0 cannot be rejected -> no significant effect of type2.
- If there are more than 1 categorical variable we yield the H0: no effect of no covariate at all. Reject H0 again in favor of significant effects if $p < \alpha$.
- Reject the H0 if $W_L := \frac{U_L k^2}{V_l} \geq \chi^2_{1-\alpha}(1)$

13 Time discrete Survival models

Discretize time in intervals $[a_0, a_1[, ..., [a_{q-1}, a_q[$ and fit classic GLM's **without** an intercept on the transformed data with the event variable as response. The coefficients of the time variables are used as intercepts.

13.1 Data

We add the time variable `t` as a factor to our data frame

```
##   subject calendar.week start stop arrest employed fin age race wexp mar
## 1       1             1     0    1      0        0   0  27    1    0   0
## 2       1             2     1    2      0        0   0  27    1    0   0
## 3       1             3     2    3      0        0   0  27    1    0   0
## 4       1             4     3    4      0        0   0  27    1    0   0
## 5       1             5     4    5      0        0   0  27    1    0   0
## 6       1             6     5    6      0        0   0  27    1    0   0
##   paro prio educ t
## 1    1    3    3 1
## 2    1    3    3 2
## 3    1    3    3 3
## 4    1    3    3 4
## 5    1    3    3 5
## 6    1    3    3 6
```

13.2 Model

13.2.1 Logit

$$P(T = t|T \geq t, x) = \frac{exp(\eta)}{1 + exp(\eta)}$$

$$P(T > t|ăT \geq t, x) = 1 - P(T = t|ăT \geq t, x) = \frac{1}{1 + exp(\eta)}$$

$$\frac{P(T = t|ăT \geq t, x)}{P(T > t|ăT \geq t, x)} = exp(\eta) = exp(\beta_0 + X'\beta) = exp(\beta_0)exp(X'\beta) \text{ \textbf{Continuation ratio}}$$

e.g. $beta_{female} = 0.5 \rightarrow$ Risk of woman to have event in time $T = t$ instead of later ih $T > t$ is factor $exp(0.5) = 1.68$ higher than for men ($x_{female} = 0$).

$\eta = X'\beta$ linear predictor.

13.2.1.1 Example

```
##
## Call:
## glm(formula = arrest ~ -1 + t + fin + age + mar + prio, family = binomial(link = "logit"),
##     data = prison.long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3929  -0.1236  -0.0930  -0.0669   3.8088
##
## Coefficients:
##       Estimate Std. Error z value Pr(>|z|)
```

```
## t1     -4.78099    1.11448  -4.290 1.79e-05 ***
## t2     -4.77850    1.11458  -4.287 1.81e-05 ***
## t3     -4.77779    1.11446  -4.287 1.81e-05 ***
## t4     -4.77595    1.11435  -4.286 1.82e-05 ***
## t5     -4.77283    1.11452  -4.282 1.85e-05 ***
## t6     -4.76924    1.11466  -4.279 1.88e-05 ***
## t7     -4.76440    1.11479  -4.274 1.92e-05 ***
## t8     -3.13959    0.66561  -4.717 2.39e-06 ***
## t9     -4.05164    0.86167  -4.702 2.58e-06 ***
## t10    -4.74558    1.11454  -4.258 2.06e-05 ***
## t11    -4.03949    0.86129  -4.690 2.73e-06 ***
## t12    -4.02106    0.86105  -4.670 3.01e-06 ***
## t13    -4.71502    1.11415  -4.232 2.32e-05 ***
## t14    -3.60635    0.75813  -4.757 1.97e-06 ***
## t15    -4.00111    0.86115  -4.646 3.38e-06 ***
## t16    -3.99587    0.86127  -4.640 3.49e-06 ***
## t17    -3.58399    0.75803  -4.728 2.27e-06 ***
## t18    -3.56763    0.75836  -4.704 2.55e-06 ***
## t19    -3.96549    0.86152  -4.603 4.17e-06 ***
## t20    -3.03273    0.66548  -4.557 5.18e-06 ***
## t21    -3.94828    0.86169  -4.582 4.60e-06 ***
## t22    -4.64070    1.11449  -4.164 3.13e-05 ***
## t23    -4.63504    1.11454  -4.159 3.20e-05 ***
## t24    -3.23726    0.70175  -4.613 3.97e-06 ***
## t25    -3.52033    0.75851  -4.641 3.47e-06 ***
## t26    -3.49480    0.75737  -4.614 3.94e-06 ***
## t27    -3.89779    0.86017  -4.531 5.86e-06 ***
## t28    -3.89186    0.86015  -4.525 6.05e-06 ***
## t29   -18.19320  547.40235  -0.033 0.973487
## t30    -3.88282    0.86034  -4.513 6.39e-06 ***
## t31    -4.57306    1.11385  -4.106 4.03e-05 ***
## t32    -3.87112    0.86090  -4.497 6.90e-06 ***
## t33    -3.86440    0.86123  -4.487 7.22e-06 ***
## t34    -3.85163    0.86151  -4.471 7.79e-06 ***
## t35    -3.14353    0.70153  -4.481 7.43e-06 ***
## t36    -3.42146    0.75910  -4.507 6.57e-06 ***
## t37    -3.12107    0.70258  -4.442 8.90e-06 ***
## t38    -4.50319    1.11496  -4.039 5.37e-05 ***
## t39    -3.80263    0.86223  -4.410 1.03e-05 ***
## t40    -3.09650    0.70259  -4.407 1.05e-05 ***
## t41   -18.16530  568.23431  -0.032 0.974498
## t42    -3.77967    0.86300  -4.380 1.19e-05 ***
## t43    -3.07561    0.70369  -4.371 1.24e-05 ***
## t44    -3.75914    0.86375  -4.352 1.35e-05 ***
## t45    -3.75108    0.86412  -4.341 1.42e-05 ***
## t46    -3.04004    0.70544  -4.309 1.64e-05 ***
## t47    -4.42745    1.11689  -3.964 7.37e-05 ***
## t48    -3.72618    0.86470  -4.309 1.64e-05 ***
## t49    -2.78833    0.66969  -4.164 3.13e-05 ***
## t50    -3.29221    0.76310  -4.314 1.60e-05 ***
## t51   -18.14517  589.76136  -0.031 0.975455
## t52    -2.98335    0.70667  -4.222 2.42e-05 ***
## fin    -0.36333    0.19143  -1.898 0.057692 .
## age    -0.06071    0.02092  -2.902 0.003706 **
```

```
## mar   -0.53659    0.37384  -1.435 0.151186
## prio   0.09836    0.02745   3.584 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27461  on 19809  degrees of freedom
## Residual deviance:  1325  on 19753  degrees of freedom
## AIC: 1437
##
## Number of Fisher Scoring iterations: 18
```

The risk to go back to jail for a married person in $T = t$ instead of $T > t$ is $exp(-0.53) = 0.58$ lower than for a not-married person. We always **interpret the continutation ratio in the logit time discrete model**!

The hazard in the logit model follows:

$$\lambda(t|X_{it}) = P(y_{it} = 1|\text{ă}X_{it}) = \frac{exp(\beta_{0t} + X_{it}^T\beta)}{1 + exp(\beta_{0t} + X_{it}^T\beta)}$$

and thus the baseline hazard (all other covariates than the time dummy-variable of interest):

$$\lambda_0(t; X_{it} = 0) = P(y_{it} = 1; \text{ă}X_{it} = 0) = \frac{exp(\beta_{0t})}{1 + exp(\beta_{0t})}$$

13.2.2 Cloglog

13.2.2.1 Model equation

$$\lambda(t|x_i) = P(y_{it} = 1|\text{ă}x_{it}) = 1 - exp(-exp(\eta_i)) \text{ equivalent-moves: } log(), -1$$

$$exp(\eta_i) = -log(1-\lambda(t|x_i)) \text{ with } 1-\lambda(t|x_i) = S(t_i) exp(\eta_i) = -log(S(t_i)) exp(\eta_i) = \Lambda(t) \text{ with } \frac{\partial\Lambda(t)}{\partial t} = \lambda(t) \Rightarrow \lambda(t|x_i) = \frac{\partial exp(}{\partial t}$$

Neat thing: this form allows for Interpetation with respect to the hazard rate. Exactly as shown here:

```
##
## Call:
## glm(formula = arrest ~ -1 + t + fin + age + mar + prio, family = binomial(link = "cloglog"),
##     data = prison.long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3968  -0.1237  -0.0930  -0.0669   3.8084
##
## Coefficients:
##       Estimate Std. Error z value Pr(>|z|)
## t1    -4.78603    1.11240  -4.302 1.69e-05 ***
## t2    -4.78443    1.11290  -4.299 1.72e-05 ***
## t3    -4.78316    1.11252  -4.299 1.71e-05 ***
## t4    -4.78069    1.11213  -4.299 1.72e-05 ***
## t5    -4.77735    1.11219  -4.295 1.74e-05 ***
## t6    -4.77315    1.11204  -4.292 1.77e-05 ***
## t7    -4.76967    1.11277  -4.286 1.82e-05 ***
## t8    -3.15214    0.66189  -4.762 1.91e-06 ***
```

```
## t9    -4.06012    0.85929  -4.725 2.30e-06 ***
## t10   -4.74720    1.11085  -4.273 1.92e-05 ***
## t11   -4.03998    0.85659  -4.716 2.40e-06 ***
## t12   -4.02951    0.85873  -4.692 2.70e-06 ***
## t13   -4.71966    1.11183  -4.245 2.19e-05 ***
## t14   -3.61233    0.75433  -4.789 1.68e-06 ***
## t15   -4.00826    0.85841  -4.669 3.02e-06 ***
## t16   -4.00409    0.85883  -4.662 3.13e-06 ***
## t17   -3.58894    0.75395  -4.760 1.93e-06 ***
## t18   -3.57540    0.75487  -4.736 2.18e-06 ***
## t19   -3.97225    0.85858  -4.627 3.72e-06 ***
## t20   -3.04742    0.66195  -4.604 4.15e-06 ***
## t21   -3.95633    0.85906  -4.605 4.12e-06 ***
## t22   -4.64461    1.11168  -4.178 2.94e-05 ***
## t23   -4.64117    1.11271  -4.171 3.03e-05 ***
## t24   -3.25029    0.69846  -4.653 3.26e-06 ***
## t25   -3.52123    0.75334  -4.674 2.95e-06 ***
## t26   -3.50551    0.75437  -4.647 3.37e-06 ***
## t27   -3.90514    0.85728  -4.555 5.23e-06 ***
## t28   -3.89770    0.85679  -4.549 5.39e-06 ***
## t29  -18.10872  523.83825  -0.035 0.972423
## t30   -3.89005    0.85737  -4.537 5.70e-06 ***
## t31   -4.57738    1.11110  -4.120 3.79e-05 ***
## t32   -3.87824    0.85786  -4.521 6.16e-06 ***
## t33   -3.86854    0.85731  -4.512 6.41e-06 ***
## t34   -3.85824    0.85829  -4.495 6.95e-06 ***
## t35   -3.15460    0.69768  -4.522 6.14e-06 ***
## t36   -3.43121    0.75567  -4.541 5.61e-06 ***
## t37   -3.13158    0.69854  -4.483 7.36e-06 ***
## t38   -4.50854    1.11251  -4.053 5.07e-05 ***
## t39   -3.81072    0.85930  -4.435 9.22e-06 ***
## t40   -3.10618    0.69832  -4.448 8.66e-06 ***
## t41  -18.08094  543.77455  -0.033 0.973475
## t42   -3.78902    0.86035  -4.404 1.06e-05 ***
## t43   -3.08569    0.69940  -4.412 1.02e-05 ***
## t44   -3.76634    0.86041  -4.377 1.20e-05 ***
## t45   -3.75728    0.86047  -4.367 1.26e-05 ***
## t46   -3.05388    0.70169  -4.352 1.35e-05 ***
## t47   -4.43288    1.11427  -3.978 6.94e-05 ***
## t48   -3.73278    0.86107  -4.335 1.46e-05 ***
## t49   -2.80434    0.66549  -4.214 2.51e-05 ***
## t50   -3.29935    0.75862  -4.349 1.37e-05 ***
## t51  -18.06089  564.37810  -0.032 0.974471
## t52   -2.99697    0.70265  -4.265 2.00e-05 ***
## fin   -0.36091    0.19048  -1.895 0.058134 .
## age   -0.06051    0.02085  -2.903 0.003700 **
## mar   -0.53367    0.37265  -1.432 0.152114
## prio   0.09779    0.02721   3.594 0.000325 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 39495  on 19809  degrees of freedom
```

```
## Residual deviance:  1325  on 19753  degrees of freedom
## AIC: 1437
##
## Number of Fisher Scoring iterations: 18
```

The hazard rate for a married person is $exp(-0.53) = 0.59 $ times as high as for a an unmarried person with all other covariates held constant.

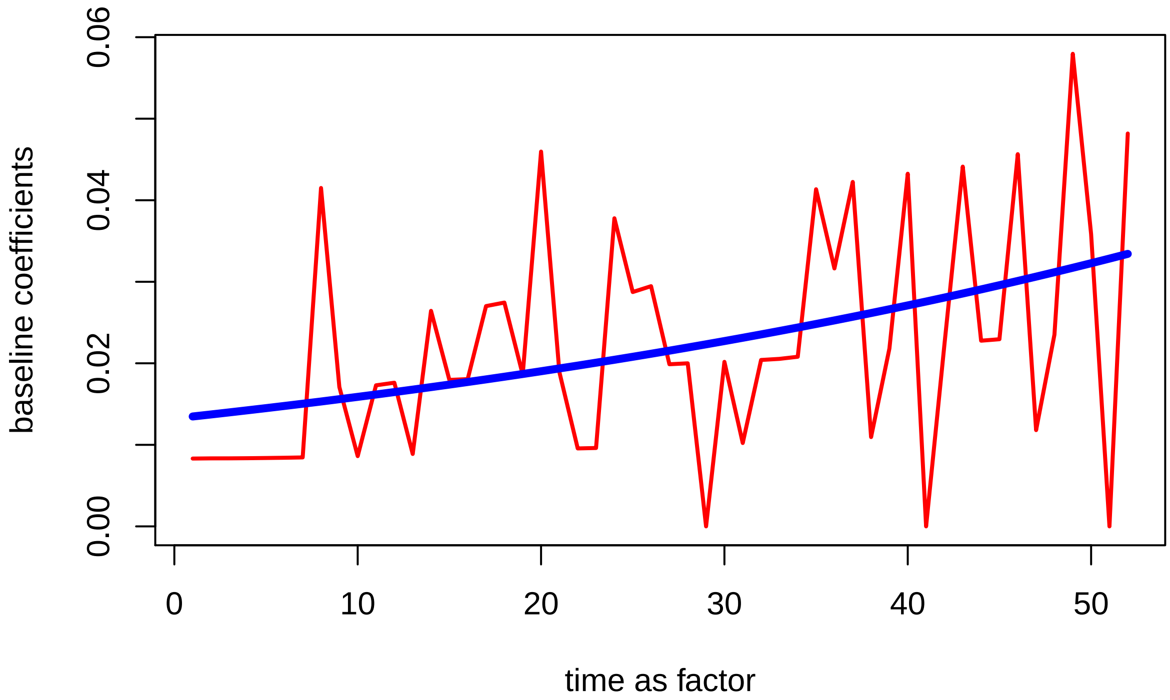
13.3 Smooth time variables

We include the time variable via a smoothing spline and yield a sparser model with more or less the same coefficients for our covariates:

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## arrest ~ s(stop) + fin + age + mar + prio
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.85034    0.49838  -7.726 1.11e-14 ***
## fin         -0.36353    0.19120  -1.901 0.057269 .
## age         -0.06052    0.02090  -2.896 0.003780 **
## mar         -0.53563    0.37359  -1.434 0.151646
## prio         0.09800    0.02739   3.578 0.000346 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df Chi.sq p-value
## s(stop) 1.024  1.047  8.271 0.00473 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.00223   Deviance explained =  2.7%
## UBRE = -0.93047  Scale est. = 1         n = 19809
```

We cannot interpret those baseline hazards in a reasonable manner.

Graphically, we see that the baseline hazards from the blue, spline curve is much smoother and less "outlier-sensitive" (e.g.: time points without events), than from the simple logit model in red:


baseline coefficients
0.00
0.02
0.04
0.06
0
10
20
30
40
50
time as factor

14 Piecewise exponential models (PEM)

14.0.1 Model equation:

$$\lambda_i(t|x_i) = \lambda_j exp(x^T\beta), \forall t \in]a_{j-1}, a_j]$$

with constant baseline hazards in each of the J intervals.

14.0.2 Data

```
##   id tstart tend interval offset ped_status treatment pair
## 1  1      0    1    (0,1]      0          1   placebo    1
## 2  2      0    1    (0,1]      0          0      6-MP    1
## 3  2      1    2    (1,2]      0          0      6-MP    1
## 4  2      2    3    (2,3]      0          0      6-MP    1
## 5  2      3    4    (3,4]      0          0      6-MP    1
## 6  2      4    5    (4,5]      0          0      6-MP    1
```

We fit a model for many intervals and the treatment variable resulting in many baseline intercepts:

```
##
## Call:
## glm(formula = ped_status ~ interval - 1 + treatment, family = poisson(link = log),
##     data = leuk.ped, offset = offset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0618  -0.3950  -0.2622  -0.1675   2.5615
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## interval(0,1]     -2.5511     0.7110  -3.588 0.000333 ***
## interval(1,2]     -2.4699     0.7117  -3.471 0.000519 ***
## interval(2,3]     -3.0747     1.0039  -3.063 0.002192 **
## interval(3,4]     -2.3342     0.7131  -3.273 0.001063 **
## interval(4,5]     -2.2323     0.7144  -3.125 0.001780 **
## interval(5,6]     -1.7134     0.5885  -2.911 0.003601 **
## interval(6,7]     -2.7574     1.0048  -2.744 0.006064 **
## interval(7,8]     -1.3570     0.5086  -2.668 0.007632 **
## interval(8,9]    -18.3864  1755.3284  -0.010 0.991643
## interval(9,10]    -2.4262     1.0072  -2.409 0.015998 *
## interval(10,11]   -1.6932     0.7153  -2.367 0.017932 *
## interval(11,12]   -1.4648     0.7182  -2.040 0.041392 *
## interval(12,13]   -1.8951     1.0132  -1.870 0.061443 .
## interval(13,15]   -2.5544     1.0119  -2.524 0.011592 *
## interval(15,16]   -1.6923     1.0167  -1.665 0.096003 .
## interval(16,17]   -1.6508     1.0150  -1.626 0.103869
## interval(17,19]  -18.8379  2645.1651  -0.007 0.994318
## interval(19,20]  -18.1753  2763.5867  -0.007 0.994753
## interval(20,22]   -1.2663     0.7293  -1.736 0.082525 .
## interval(22,23]   -0.1512     0.7447  -0.203 0.839068
## interval(23,25]  -18.4865  4215.7112  -0.004 0.996501
## interval(25,32]  -19.7393  4713.3084  -0.004 0.996658
## interval(32,34]  -18.4865  6665.6247  -0.003 0.997787
```

```
## interval(34,35]  -17.7934  9426.6169  -0.002 0.998494
## treatment6-MP     -1.5092     0.4096  -3.685 0.000229 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1017.8  on 475  degrees of freedom
## Residual deviance:  132.9  on 450  degrees of freedom
## AIC: 242.9
##
## Number of Fisher Scoring iterations: 17
```

- we fit way too many parameters
- intervals which did not face events have super high standard errors and strange coefficients
- -> Two reasons for fitting PAM's with smooth baseline hazards

15 Piecewise additive exponential models (PAM)

New compared to PEM: smooth modeling of the piecewise constant baseline hazards e.g. via splines. Cool because:

- PEM constrained by use of intervals as high J leads to parameter explosion
- Smoother curves due to penalization of splines on the overlaps of the intervals
- Problem PEM: no data in interval $]a_{l-1}, a_l]$ -> $\lambda_l = 0$, wiggely hazard rate curves

15.0.1 Model equation:

$$\lambda_i(t|x_i) = exp(f_0(t_j) + x^T \beta)$$

with spline for time dependent baseline hazard:

$$f_0(t_j) = log(\lambda_0(t_j)) = \sum_{k=1}^{K} \gamma_k B_k(t_j)$$

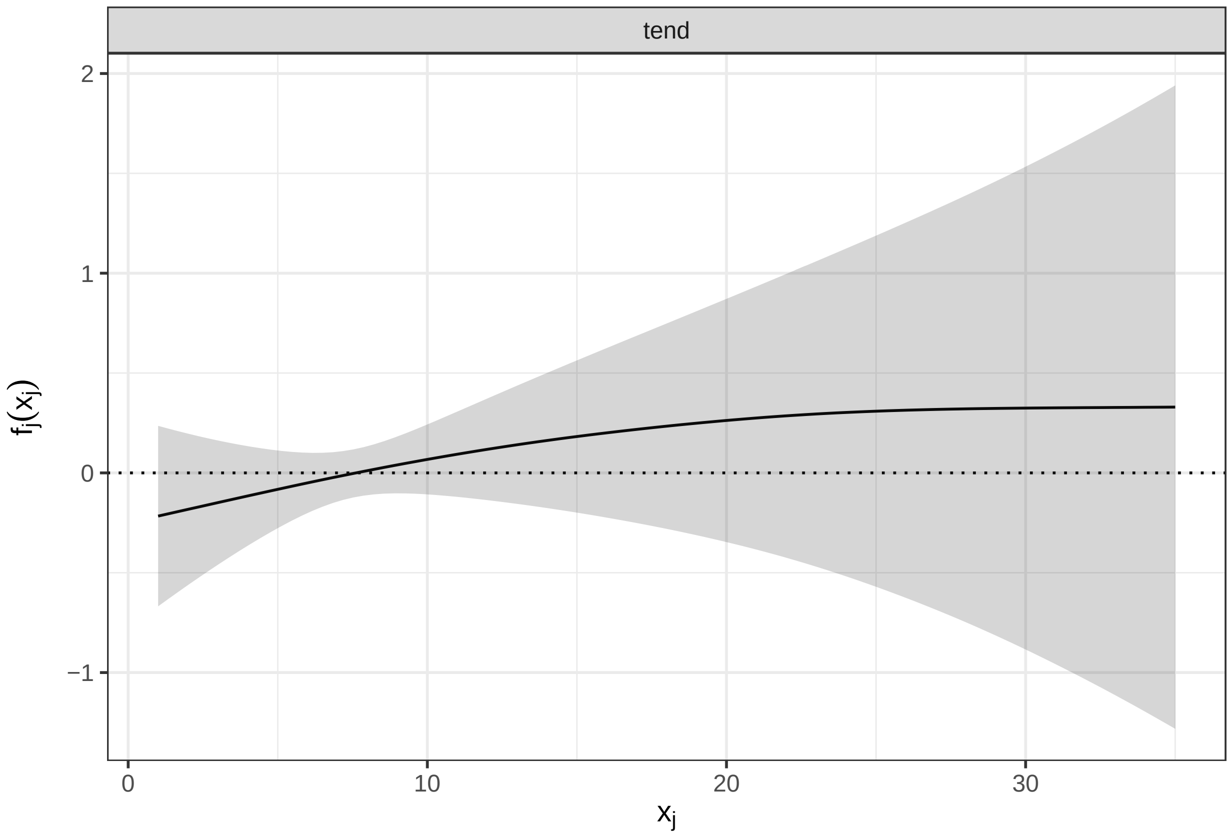
and for time varying covariates:

$$\lambda_i(t|x_i) = exp(f_0(t_j) + \sum_{j=1}^{p} f_k(x_i, k))$$

15.1 Model

```
##
## Family: poisson
## Link function: log
##
## Formula:
## ped_status ~ s(tend) + treatment
##
## Parametric coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.1389     0.2188  -9.773  < 2e-16 ***
## treatment6-MP  -1.6288     0.4259  -3.824 0.000131 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df Chi.sq p-value
## s(tend) 1.265  1.483  0.479   0.544
##
## R-sq.(adj) =  0.0297   Deviance explained = 10.5%
## UBRE = -0.66702  Scale est. = 1         n = 475
```

we can also plot the splined piecewise smoothed hazards:


tend
2
1
0
−1
$f_j(x_j)$
0
10
20
30
x_j

16 Piecewise additive exponential mixed models (PAMM)

Extension of the PAM with **Frailty terms**. Here: 400 different ICU's and we do not want to control for each of it. Therefore, we fit a smooth, splined version of the frailty. Nice: need only to fit a distribution with shape parameter

16.0.1 Data

looks like that:

```
##   CombinedID tstart tend interval offset ped_status CombinedicuID Year Age
## 1       1101      4    5    (4,5]      0          0          1114 2007  71
## 2       1101      5    6    (5,6]      0          0          1114 2007  71
## 3       1101      6    7    (6,7]      0          0          1114 2007  71
## 4       1101      7    8    (7,8]      0          0          1114 2007  71
## 5       1101      8    9    (8,9]      0          0          1114 2007  71
## 6       1101      9   10   (9,10]      0          0          1114 2007  71
##        BMI AdmCatID DiagID2 ApacheIIScore DaysInICU
## 1 38.97392  Medical  Sepsis            13  6.743056
## 2 38.97392  Medical  Sepsis            13  6.743056
## 3 38.97392  Medical  Sepsis            13  6.743056
## 4 38.97392  Medical  Sepsis            13  6.743056
## 5 38.97392  Medical  Sepsis            13  6.743056
## 6 38.97392  Medical  Sepsis            13  6.743056
```

Fit a PAMM with a smooth spline term for time (tend) and the other continous variables using this formula:

```
pamm_icu <- bam(ped_status ~ s(tend) + Year + AdmCatID + DiagID2 + s(Age) + s(BMI) +
      s(ApacheIIScore) + s(CombinedicuID, bs="re"), offset=offset, data = ped,
    family=poisson(), discrete = TRUE)
```

We include the variable CombinedicuID as a random effect aka as a **frailty term**. Therefore wie use `bs = "re"`. We control for the random effects of the ICU units without having to model a dummy for each of the ICU's. The frailty model just estimates a Gaussian over the different ICU's for which we only have to estimate the variance: 1 parameter instead of 400.

We model the PAM as a Poisson model with log link on the death-indicator `ped_status`

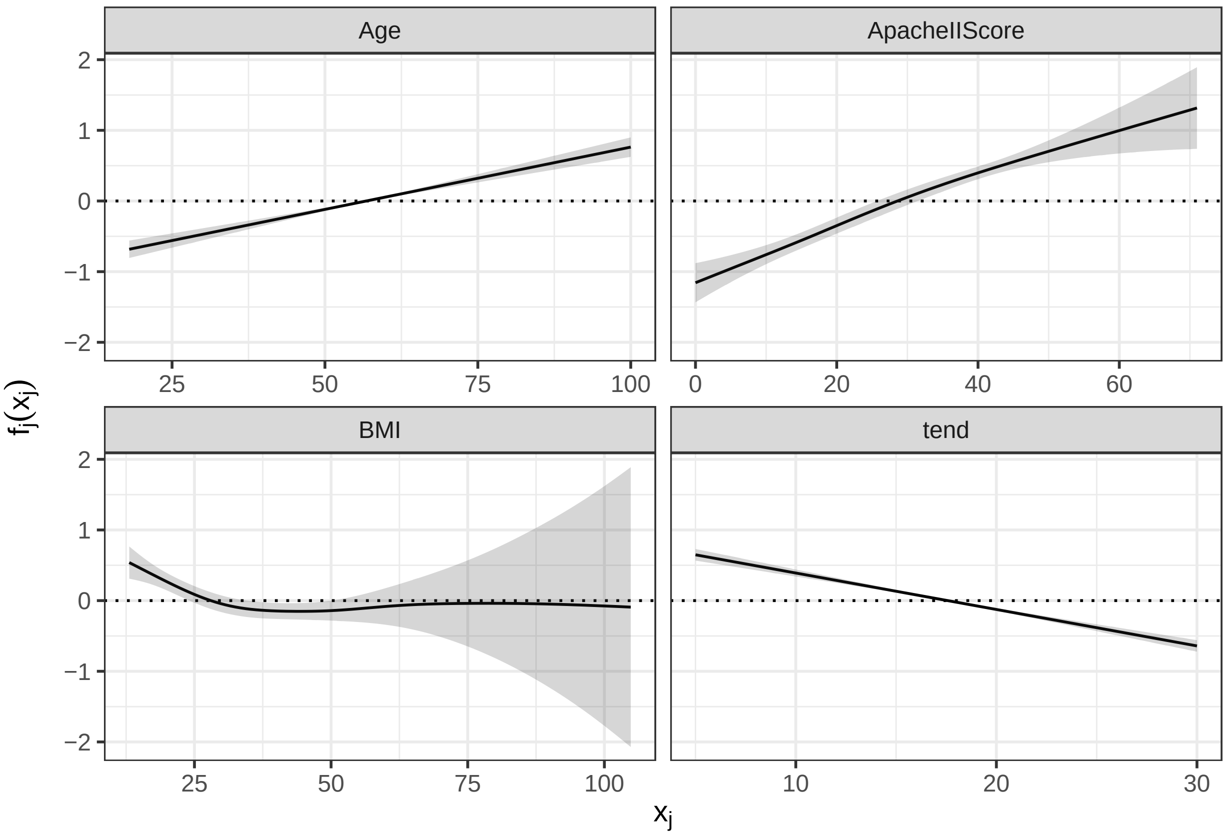
This is the model summary:

```
##
## Family: poisson
## Link function: log
##
## Formula:
## ped_status ~ s(tend) + Year + AdmCatID + DiagID2 + s(Age) + s(BMI) +
##     s(ApacheIIScore) + s(CombinedicuID, bs = "re")
##
## Parametric coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -4.59863    0.11388 -40.383  < 2e-16 ***
## Year2008                    0.02718    0.07425   0.366 0.714314
## Year2009                   -0.08622    0.07466  -1.155 0.248156
## Year2011                   -0.02329    0.06966  -0.334 0.738144
## AdmCatIDSurgical Elective  -0.47450    0.09297  -5.104 3.33e-07 ***
## AdmCatIDSurgical Emergency -0.25668    0.07228  -3.551 0.000384 ***
## DiagID2Cardio-Vascular      0.12439    0.08721   1.426 0.153774
```

```
## DiagID2Other                0.10391    0.12855   0.808 0.418914
## DiagID2Metabolic           -0.92768    0.25552  -3.631 0.000283 ***
## DiagID2Neurologic           0.01267    0.09508   0.133 0.893972
## DiagID2Orthopedic/Trauma   -0.26816    0.11560  -2.320 0.020354 *
## DiagID2Renal               -0.02734    0.21580  -0.127 0.899183
## DiagID2Respiratory         -0.13289    0.08618  -1.542 0.123091
## DiagID2Sepsis               0.05627    0.09895   0.569 0.569587
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf  Ref.df Chi.sq  p-value
## s(tend)           1.000   1.001 248.94  < 2e-16 ***
## s(Age)            1.002   1.003 122.98  < 2e-16 ***
## s(BMI)            3.061   3.879  40.61 3.55e-08 ***
## s(ApacheIIScore)   1.890   2.422 163.17  < 2e-16 ***
## s(CombinedicuID) 101.279 363.000 152.16 3.35e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  -0.00897   Deviance explained =  -15%
## fREML = 2.0196e+05  Scale est. = 1         n = 208536
```

16.0.2 What can we say?

- **smooth terms for continous variables:**
  - if the edf (estimated degress of freedom) = 1, our spline smoother estimated the variable as a linear effect on the hazard rate. This is the case for Age and time
  - BMI, ApacheIIScore and CombinedicuID (only frailty effect) seem to have a non-linear effect on the hazard rate
  - We can and will not interpret the frailty term in detail
  - those effects can also be seen graphically which shows the effect of the variable's values on the linear predictor aka the log(hazard-rate). This is the exact value that enters our linear predictor, e.g. 75 year old person enters 0.3
  - time (tend) has a falling slope aka a decreasing effect on the log(hazard) -> hazard decreases also
  - ApacheIIScore has almost linear effect: (log-) hazard increases with increasing Apache Scores though this increase is getting lower with higher values of the score
  - increasing linear age effect, the older, the higher the (log-)hazard
  - typical shape of the BMI effect, very low BMIs have increased hazard, that decreases toward "normal" BMIs, high uncertainty with respect to effect of very high BMIs as number of patients with respective BMIs decreases (few persons with very high obesity)



- **non-smooth terms for categorical variables:**
  - exponentiate the coefficients `exp(beta)` and interpret their **mulitplicative** effect on the hazard rate w.r.t the reference category
  - example 1: hazard rate for a person treated in 2009 is exp(-0.08622441) = 0.9173883 times as high as the hazard rate for similar person treated in 2007 (reference category)
  - example 2: hazard rate for a person with Metabolic cancer is exp(-0.92767602) = 0.3954717 times as high as the hazard rate for similar person with Gastrointestinal cancer (reference category)
  - For more, interpret this table:

```
##                                    beta        HR
## Year2008                      0.02718222 1.0275550
## Year2009                     -0.08622441 0.9173883
## Year2011                     -0.02328905 0.9769801
## AdmCatIDSurgical Elective    -0.47449956 0.6221964
## AdmCatIDSurgical Emergency   -0.25667793 0.7736173
## DiagID2Cardio-Vascular        0.12438947 1.1324568
## DiagID2Other                  0.10391129 1.1095020
## DiagID2Metabolic             -0.92767602 0.3954717
## DiagID2Neurologic             0.01267184 1.0127525
## DiagID2Orthopedic/Trauma     -0.26815998 0.7647854
## DiagID2Renal                 -0.02733998 0.9730304
## DiagID2Respiratory           -0.13289109 0.8755604
## DiagID2Sepsis                 0.05627062 1.0578839
```

17 Frailty models

18 Aalen model

Super flexible with **additive** and not multiplicative (Cox) effects on the hazard. Why flexibel? Because all covariates and coefficients can be time-varying!

18.0.1 model equation

$$\lambda(t) = \lambda_0(t) + x'(t)\beta(t) = \lambda_0(t) + \sum_{k=1}^{p} x_k(t)\beta_k(t)$$

with additive effects of time-varying covariates on baseline hazard rate

18.0.2 Data

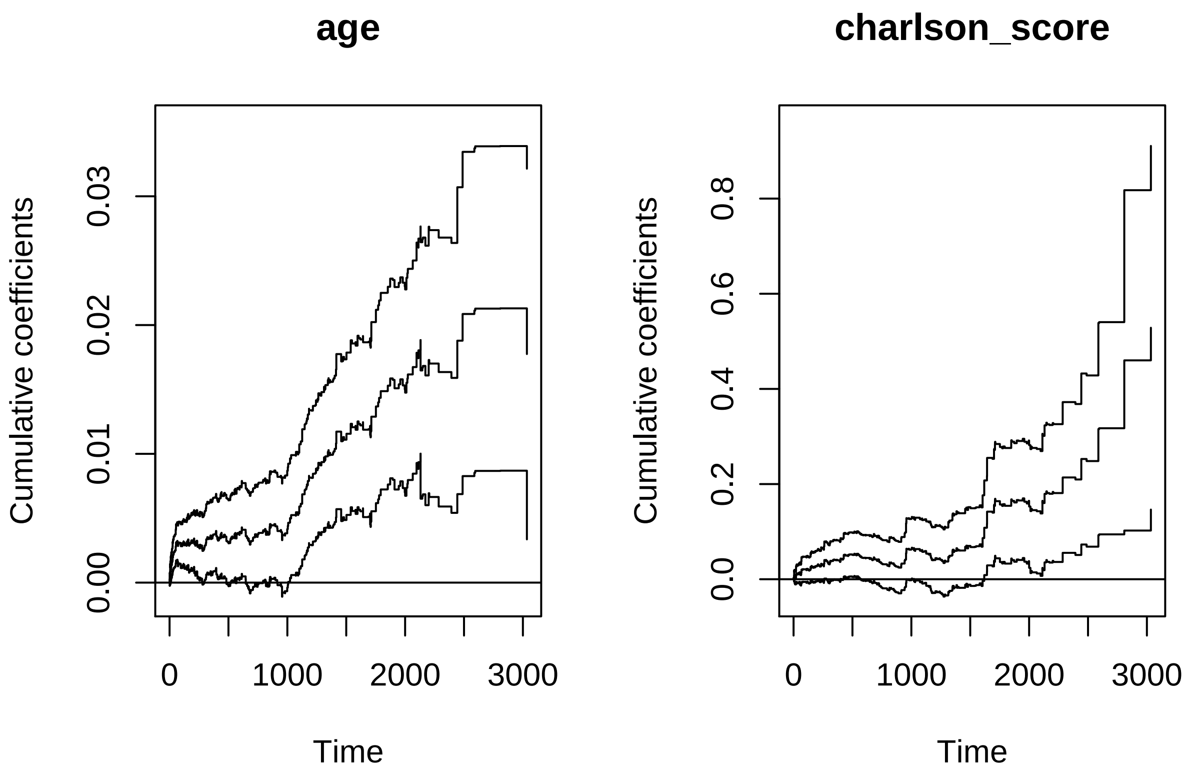
```
##   major_comp age charlson_score sex transfusion metastasesYN
## 1         no  58              2   f         yes            1
## 2        yes  52              2   m          no            1
## 3         no  74              2   f         yes            1
## 4        yes  57              2   m         yes            1
## 5         no  30              2   f         yes            1
## 6         no  66              2   f         yes            1
##   major_resection days status id metastases
## 1              no  579      0  1        yes
## 2              no 1192      0  2        yes
## 3              no  308      1  3        yes
## 4             yes   33      1  4        yes
## 5             yes  397      1  5        yes
## 6             yes 1219      0  6        yes
```

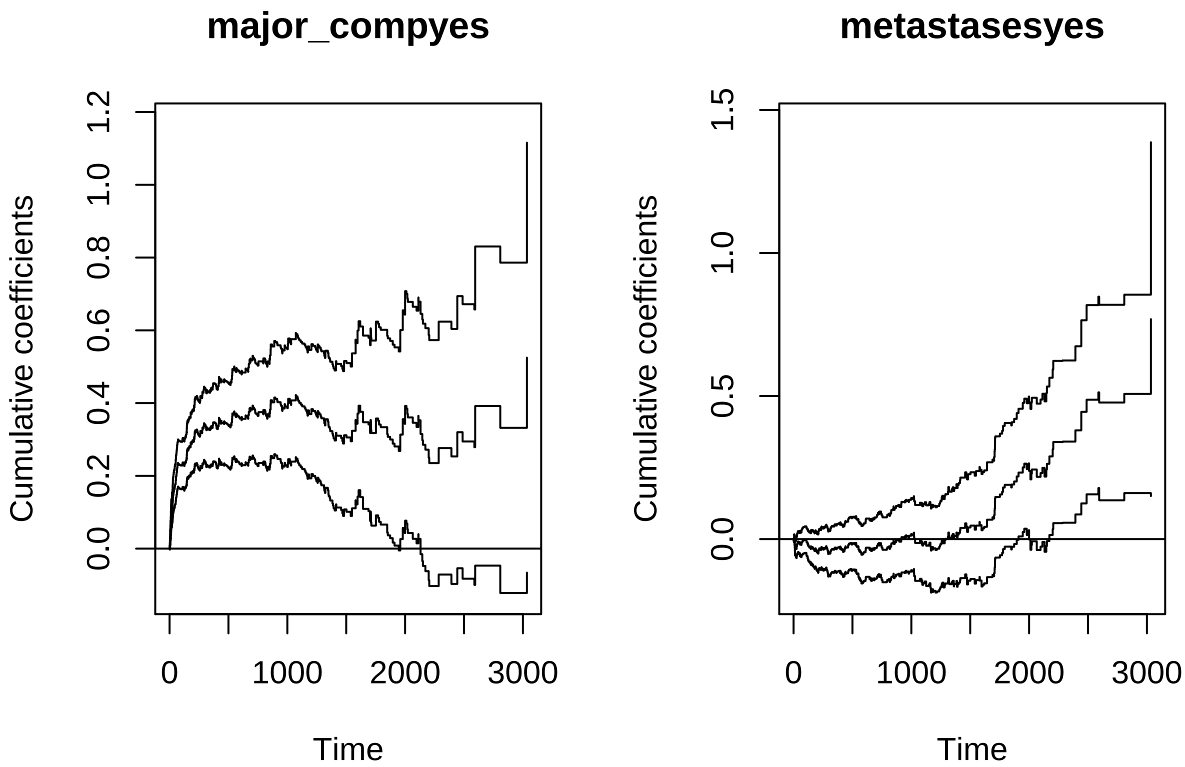
18.0.3 Model

```
## Additive Aalen Model
##
## Test for nonparametric terms
##
## Test for non-significant effects
##               Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                            4.24               0.001
## age                                    4.35               0.000
## charlson_score                         4.21               0.001
## major_compyes                          7.14               0.000
## metastasesyes                          3.41               0.025
##
## Test for time invariant effects
##                     Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                         0.60900                       0.196
## age                                 0.00637                       0.552
## charlson_score                      0.22700                       0.054
## major_compyes                       0.29400                       0.136
## metastasesyes                       0.37300                       0.048
##                       Cramer von Mises test p-value H_0:constant effect
## (Intercept)                         419.000                       0.086
```

```
## age                                        0.027                        0.506
## charlson_score                            54.700                        0.023
## major_compyes                             89.000                        0.076
## metastasesyes                            166.000                        0.023
##
##
##
##   Call:
## aalen(formula = Surv(days, status) ~ age + charlson_score + major_comp +
##     metastases, data = liver, residuals = 1)
```

- **Supremum-Test**
  - if $p < 0.05$, H0: additive coefficient = 0 can be rejected (all 4 variables **do have significant effect different from 0**, though metastases is not as super significant as the others, check Graphics)
  - if $p > 0.05$: H0: additive coefficient = 0 **can not** be rejected
- **Kolmogorov Smirnoff Test**
  - if $p < 0.05$, H0: constant effect can be rejected (charlson score, mestastasesyes have **time varying, non-constant effect**)
  - if $p > 0.05$, H0: constant effect can not be rejected (age, major_complicationyes have ___time constant effect_)
- **Graphically:**
  - linear trend over time: constant effect (age, complications)
  - non-linear, disrupted trend over time: time varying effect (metastases, charlson)
  - if 0 in Confidence Interval: insiginificant effect (explains, why $p_{metastases} > p_{others}$)





18.0.4 Semi parametric Aalen:

As we now know, that age and major_comp have constant effects on the hazard, we can model them explicitly like that and yield fixed coefficients:

```
## Additive Aalen Model
##
## Test for nonparametric terms
##
## Test for non-significant effects
##               Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                            4.20              0.000
## charlson_score                         3.70              0.004
## metastasesyes                          3.16              0.046
##
## Test for time invariant effects
##                     Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                           0.744                       0.012
## charlson_score                        0.224                       0.080
## metastasesyes                         0.345                       0.066
##                       Cramer von Mises test p-value H_0:constant effect
## (Intercept)                           681.0                       0.004
## charlson_score                         53.5                       0.027
## metastasesyes                         136.0                       0.026
##
## Parametric terms :
##                         Coef.        SE Robust SE    z    P-val lower2.5%
## const(age)           7.22e-06 1.91e-06  1.95e-06 3.70 2.15e-04  3.48e-06
## const(major_comp)yes 3.59e-04 6.88e-05  7.92e-05 4.53 5.95e-06  2.24e-04
```

```
##                        upper97.5%
## const(age)              0.000011
## const(major_comp)yes    0.000494
##
##   Call:
## aalen(formula = Surv(days, status) ~ const(age) + charlson_score +
##     const(major_comp) + metastases, data = liver, residuals = 1)
```

19 Cox-Aalen model

19.0.1 model equation

$$\lambda(t) = X(t)\beta(t) \cdot exp(Z(t)'\gamma)$$

with additive effects of time-varying covariates on baseline hazard rate which are also multiplicatively affected via Cox part of the model. γ are time-constant coefficients, PH-assumption, and β are time varying additive coefficients by the Aalen-part.

19.0.2 Data

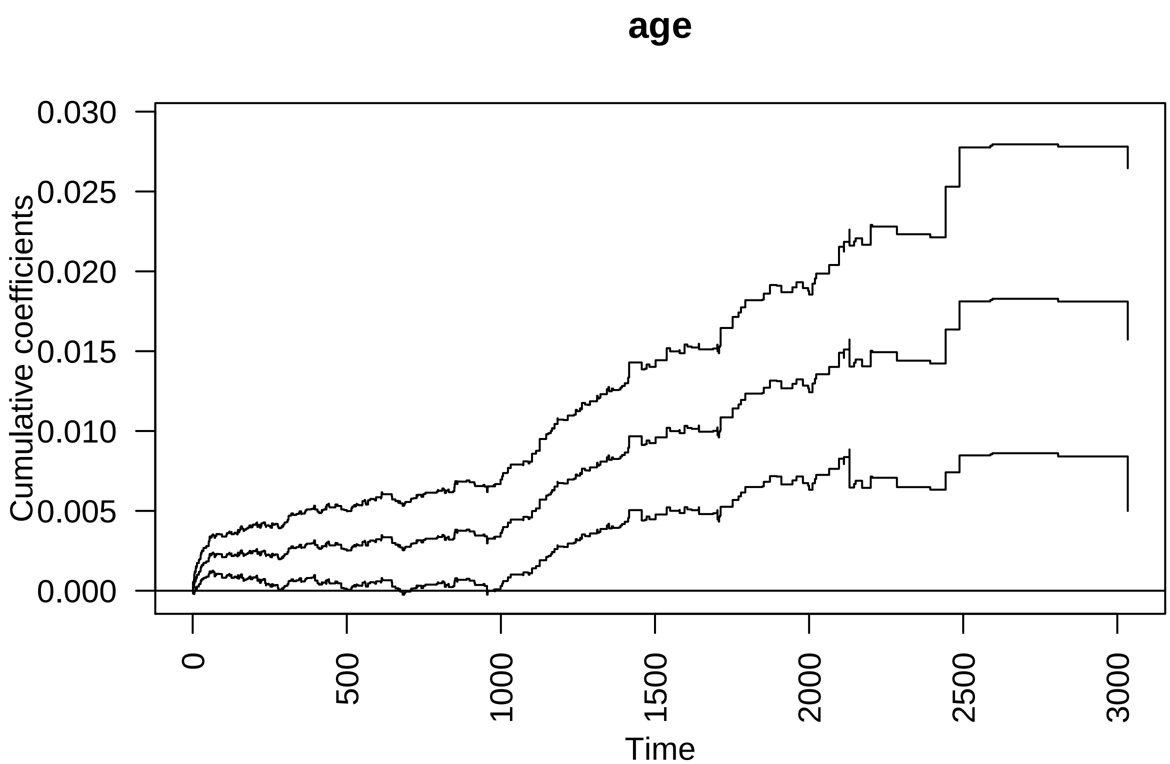
looks like that

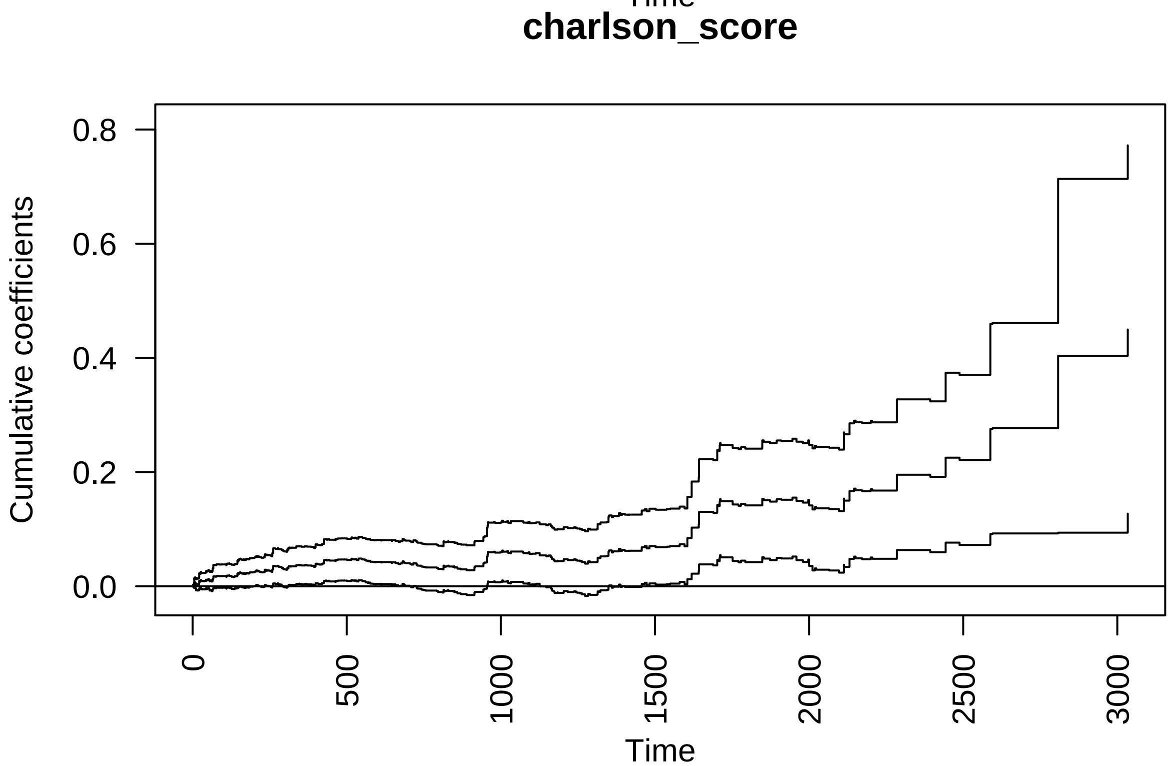
```
##   major_complications age charlson_score sex transfusion metastasesYN
## 1                  no  58              2   f         yes            1
## 2                 yes  52              2   m          no            1
## 3                  no  74              2   f         yes            1
## 4                 yes  57              2   m         yes            1
## 5                  no  30              2   f         yes            1
## 6                  no  66              2   f         yes            1
##   major_resection days status id metastases
## 1              no  579      0  1        yes
## 2              no 1192      0  2        yes
## 3              no  308      1  3        yes
## 4             yes   33      1  4        yes
## 5             yes  397      1  5        yes
## 6             yes 1219      0  6        yes
```

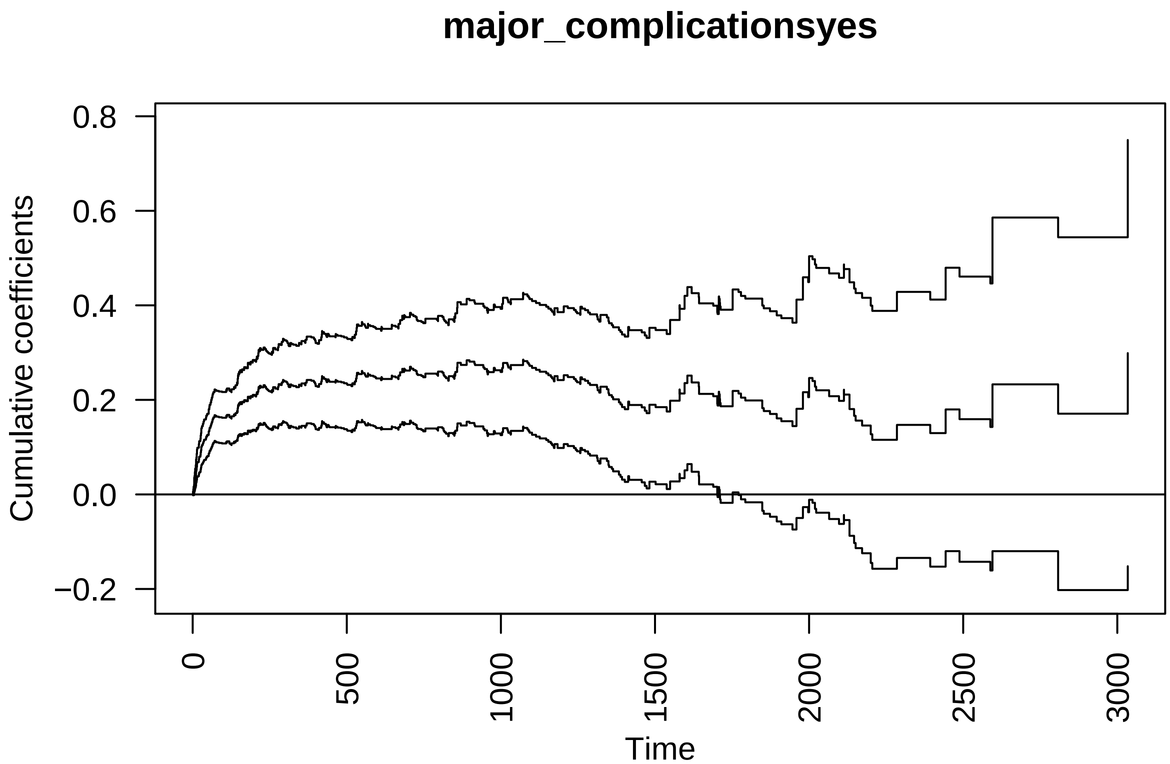
19.0.3 What can we say from the graphic?

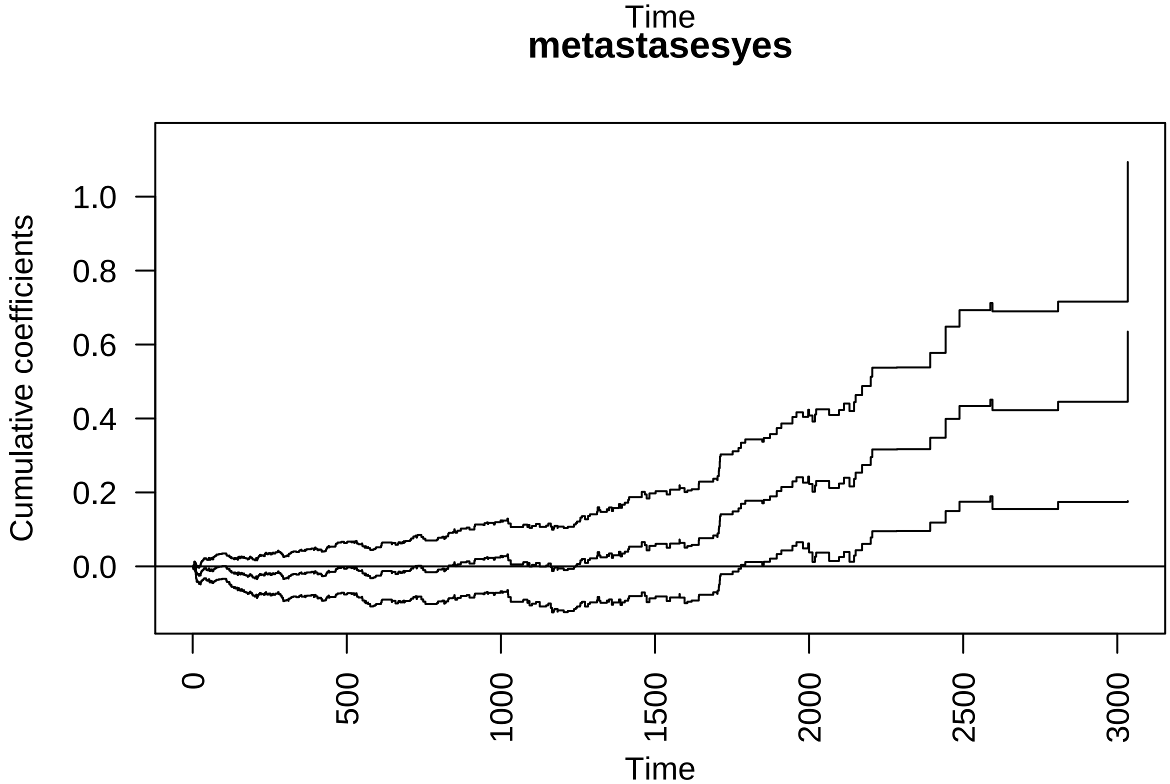
- Age:
  - the cumulative Hazard of a person aged A+1 at time point t = 1500 is 0.01 higher than that of a person aged A
  - the effect of metastases on the cumulative hazard rate starts to increase t = 1000 after the surgery and is approx. constant before
- Complications:
  - the cumulative Hazard of a person with major complications at time point t = 1500 is 0.2 higher than that of a person without complications
  - the effect of complications on the cumulative hazard rate decreases over time
- Metastases:
  - the cumulative Hazard of a person with metastases at time point t = 2500 is 0.4 higher than that of a person without metastases
  - the effect of metastases on the cumulative hazard rate starts to matter only after t = 1500 and then increases more or less linearly
  - before t = 1500 the effect is non siginificant as the 0 is part of the confidence intervals

Effects for the continous variables estimated as additive via the Aalen-part of the model using the formula `Surv(days, status) ~ age +  charlson_score + major_complications + metastases + prop(sex) + prop(transfusion) + prop(major_resection), data = liver, residuals = 1, basesim   = 1)`


age
Cumulative coefficients
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
500
1000
1500
2000
2500
3000
Time


charlson_score
Cumulative coefficients
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
3000
Time





19.0.4 What can we say from the model summary?

```
## Cox-Aalen Model
##
## Test for Aalen terms
```

```
## Test for nonparametric terms
##
## Test for non-significant effects
##                        Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                                     4.00              0.000
## age                                             4.18              0.000
## charlson_score                                  4.04              0.000
## major_complicationsyes                          6.07              0.000
## metastasesyes                                   3.85              0.006
##
## Test for time invariant effects
##                              Kolmogorov-Smirnov test
## (Intercept)                                  0.43700
## age                                          0.00522
## charlson_score                               0.16400
## major_complicationsyes                       0.21200
## metastasesyes                                0.28100
##                        p-value H_0:constant effect
## (Intercept)                                  0.202
## age                                          0.388
## charlson_score                               0.070
## major_complicationsyes                       0.146
## metastasesyes                                0.010
##
## Proportional Cox terms :
##                          Coef.    SE Robust SE D2log(L)^-1    z  P-val
## prop(sex)f               0.224 0.111     0.107       0.109 2.08 0.0372
## prop(transfusion)yes     0.233 0.111     0.113       0.112 2.07 0.0385
## prop(major_resection)yes 0.254 0.113     0.110       0.113 2.31 0.0207
##                          lower2.5% upper97.5%
## prop(sex)f                 0.00644      0.442
## prop(transfusion)yes       0.01540      0.451
## prop(major_resection)yes   0.03250      0.475
## Test of Proportionality
##                          sup|  hat U(t) | p-value H_0
## prop(sex)f                           9.53       0.166
## prop(transfusion)yes                 6.52       0.608
## prop(major_resection)yes             8.99       0.212
```

- Aalen part:
  - Supremum-test: for all 4 variables the *H0: no effect* can be rejected, all covariables have significant influence
  - Kolmogorov Smirnov for time variant effects: H0: constant effect can only clearly be rejected for metastases, others have constant effects
- Cox part:
  - sexf: the additive, time-varying effects $\beta(t) = (\beta_{age}(t), \beta_{charlson}(t), \beta_{complications}(t), \beta_{metastases}(t))^T$ from the Aalen model is getting multiplied by factor $exp(0.224) = 1.251071$ for a female compared with a similar man **for fixed time t**.
  - same for transfusion (exp(0.233) = 1.262381) and major_resection (exp(0.254) = 1.289172)

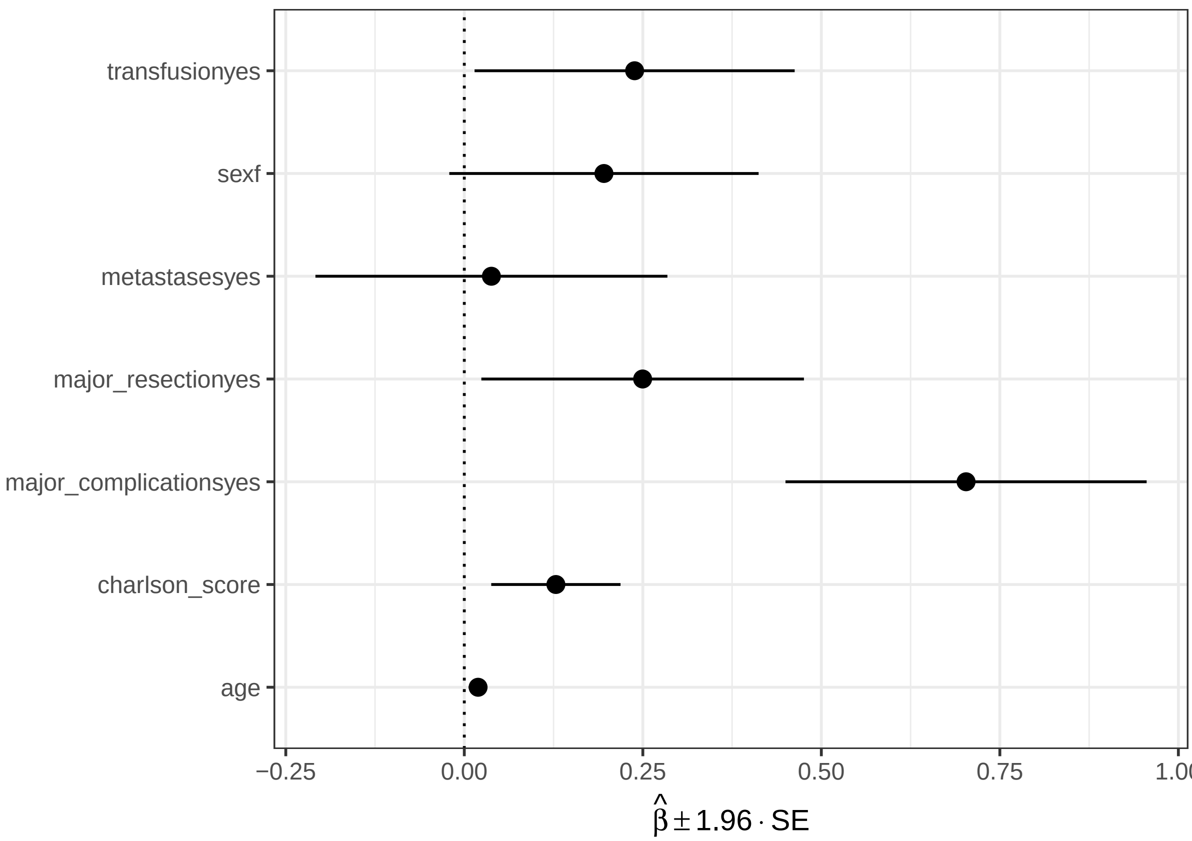
19.0.5 Cox-Aalen vs. PAM

Compare this with the PAM fitted on the data using the below formula. We explicitly model time varying effects of the 4 variables (metastases, marjo_complications, age, charlson) as in the Aalen model via ti().

```
bam(
  formula = ped_status ~ ti(tend,k=10) +
    # use ti() for non-identifiability issue
    metastases + ti(tend, by = as.ordered(metastases),k=10, mc = c(1,0)) +
   major_complications + ti(tend,by = as.ordered(major_complications),k=10, mc = c(1,0)) +
    age + ti(tend, by = age,k=10, mc = c(1,0)) +
    charlson_score + ti(tend, by = charlson_score,k=10, mc = c(1,0)) +
    sex + transfusion + major_resection,
  data   = ped_liver,
  offset = offset,
  family = poisson())
```

The figure below shows the effect of the **time constant variables** which allow some interpretation:

- NOTE: Constant contributions to time-varying can be interpreted as effects at t=0. Check the model equation and DISCUSS
- sex: Compared to males, females have a 1.22 times increased risk of experiencing an event (c.p.)
- transfusion: Compared to patients without transfusion, patients with transfustion have a 1.27 times increased risk of experiencing an event (c.p.)
- major resection: A major resection increases the risk of event by a factor of 1.28, compared to patients without a major resection
- DISCUSS If above interpretation holds, this would fit nicely the effect of the time-constant factors in the Cox-part of above Cox-Aalen model

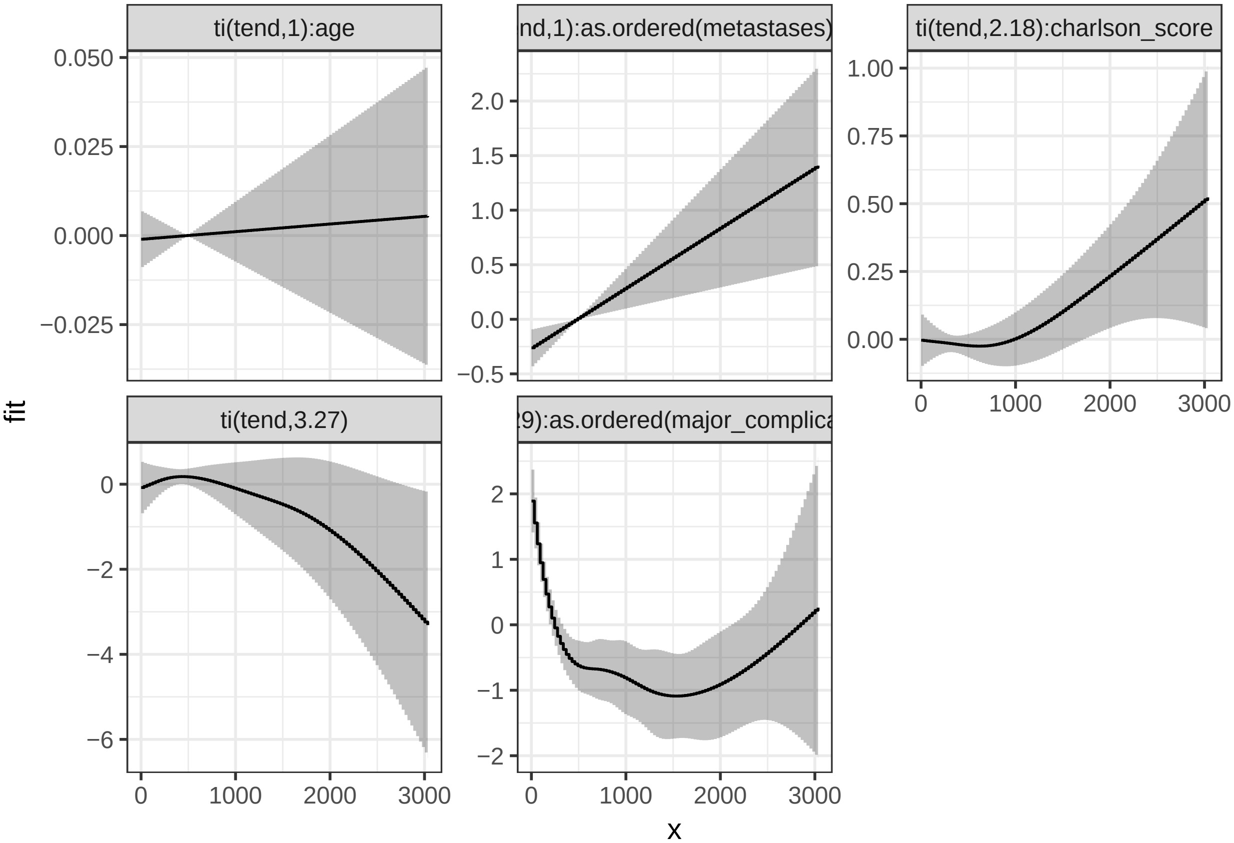


Model summary:

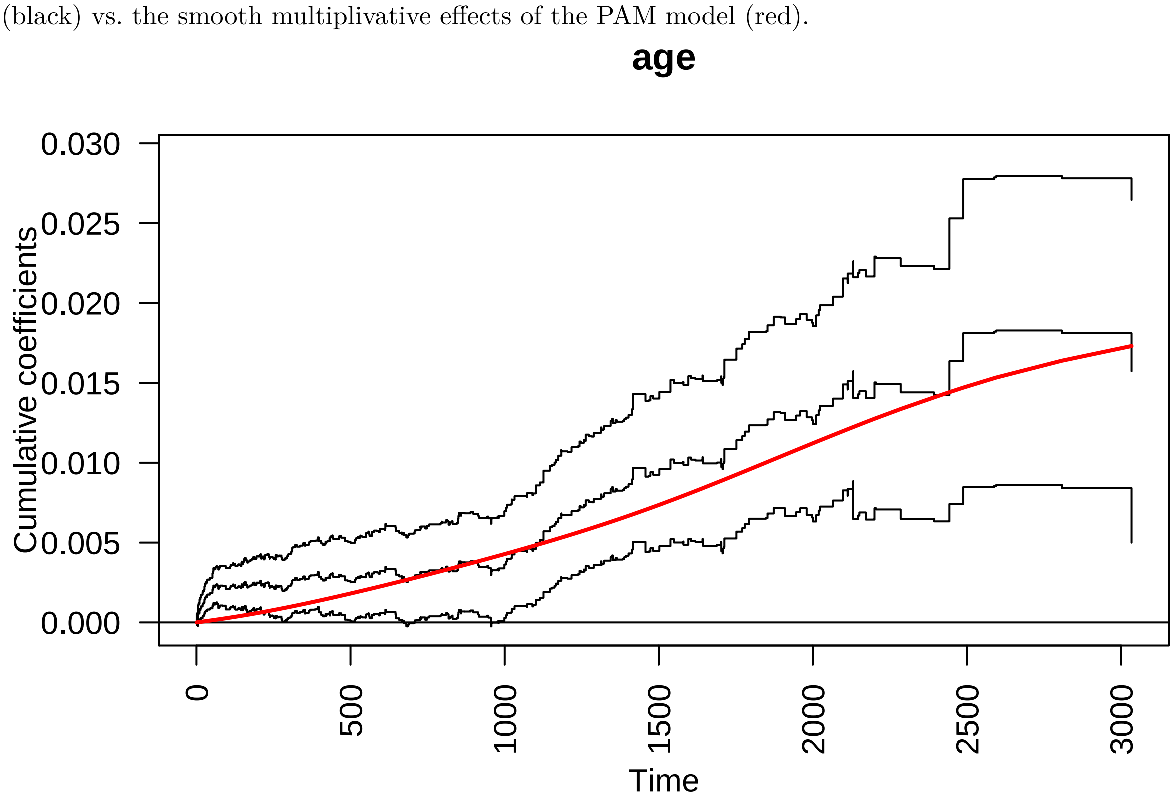
```
##
## Family: poisson
## Link function: log
##
```

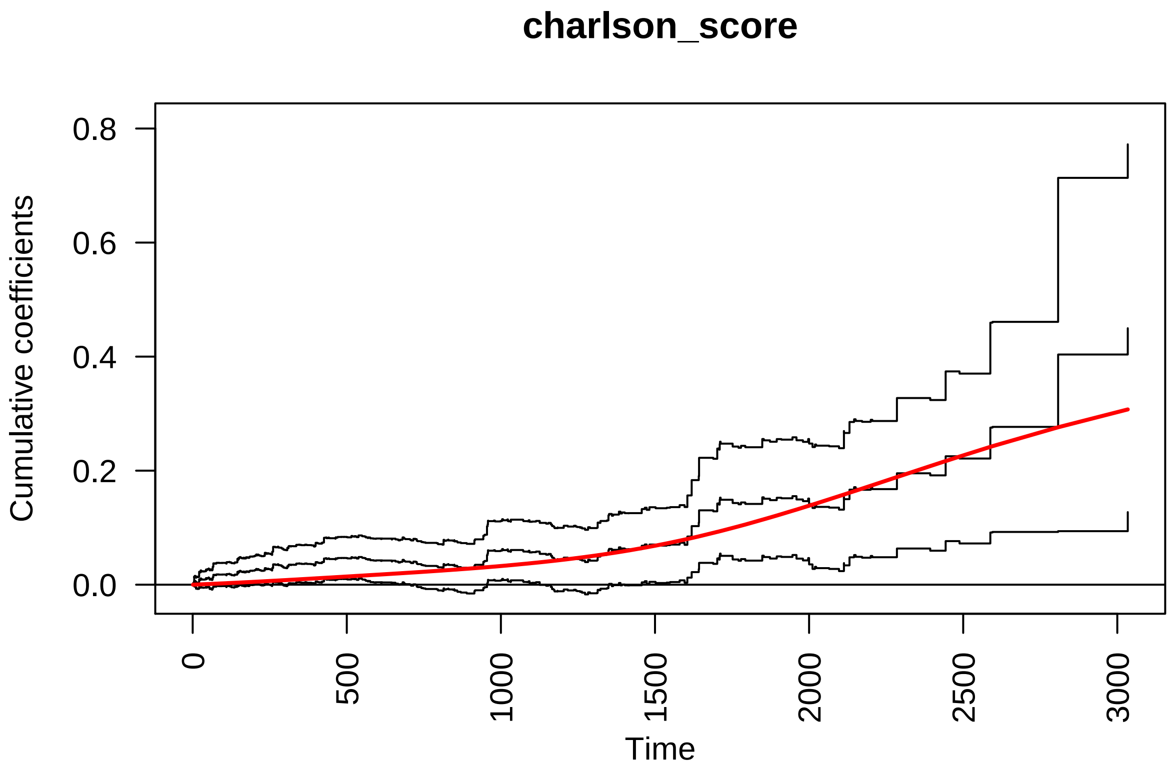
```
## Formula:
## ped_status ~ ti(tend, k = 10) + metastases + ti(tend, by = as.ordered(metastases),
##   k = 10, mc = c(1, 0)) + major_complications + ti(tend, by = as.ordered(major_complications),
##     k = 10, mc = c(1, 0)) + age + ti(tend, by = age, k = 10,
##     mc = c(1, 0)) + charlson_score + ti(tend, by = charlson_score,
##     k = 10, mc = c(1, 0)) + sex + transfusion + major_resection
##
## Parametric coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -9.756319   0.384061 -25.403  < 2e-16 ***
## metastasesyes            0.037949   0.123233   0.308 0.758122
## major_complicationsyes  0.702678   0.126452   5.557 2.75e-08 ***
## age                      0.019308   0.005269   3.664 0.000248 ***
## charlson_score           0.128265   0.045268   2.833 0.004604 **
## sexf                     0.195558   0.108301   1.806 0.070967 .
## transfusionyes           0.238512   0.112066   2.128 0.033311 *
## major_resectionyes       0.249730   0.112940   2.211 0.027024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                            edf Ref.df Chi.sq  p-value
## ti(tend)                                 3.266  3.960  9.103  0.05775
## ti(tend):as.ordered(metastases)yes       1.003  1.005  9.513  0.00208
## ti(tend):as.ordered(major_complications)yes 5.289  6.165 70.698 5.55e-13
## ti(tend):age                             1.000  1.001  0.068  0.79468
## ti(tend):charlson_score                  2.183  2.682  7.672  0.05013
##
## ti(tend)                                 .
## ti(tend):as.ordered(metastases)yes       **
## ti(tend):as.ordered(major_complications)yes ***
## ti(tend):age
## ti(tend):charlson_score                  .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.000679   Deviance explained = -10.1%
## fREML = 2.7942e+05  Scale est. = 1         n = 147896
```

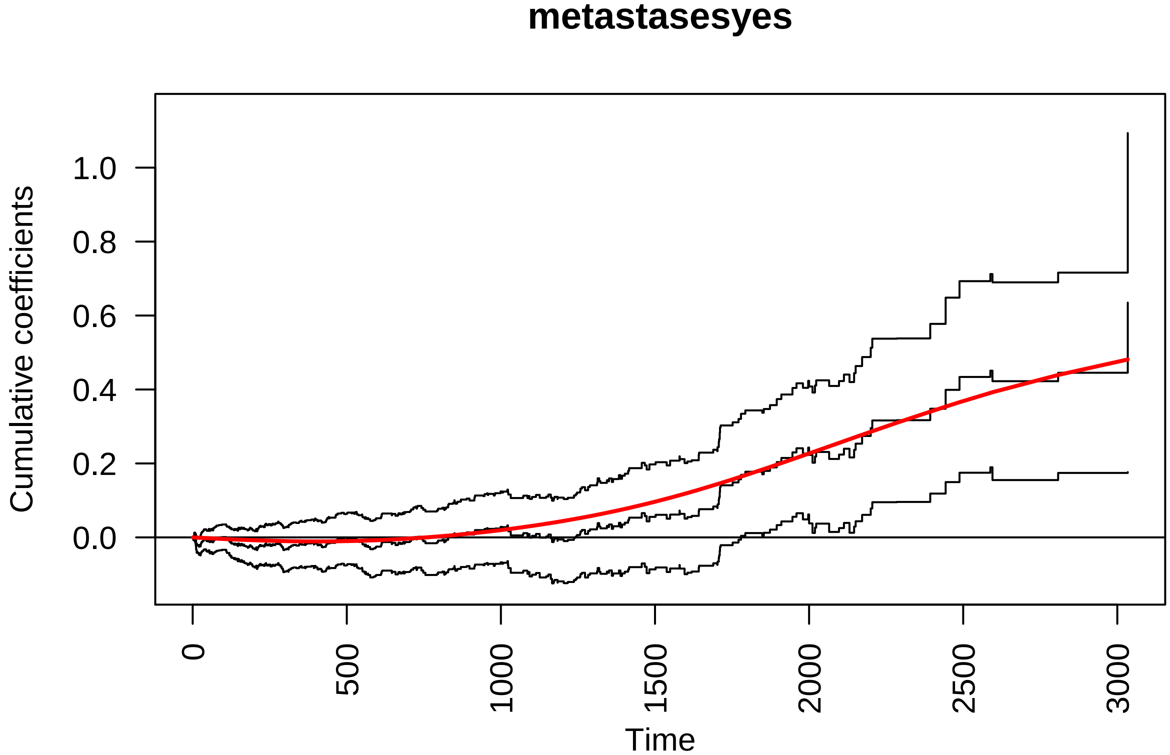
This is the effect estimated for the smooth terms. The total effect of x at time point t is $\beta_x * x + f_x(t)$ where $\beta_x * x$ are the constant effects from the previous graphic and $f_x(t)$ models the effect of the smooth time varying term. Recap the PAM model equation $\lambda_i(t|x_i) = exp(f_0(t_j) + x^T\beta)$ and DISCUSS. They look like that:

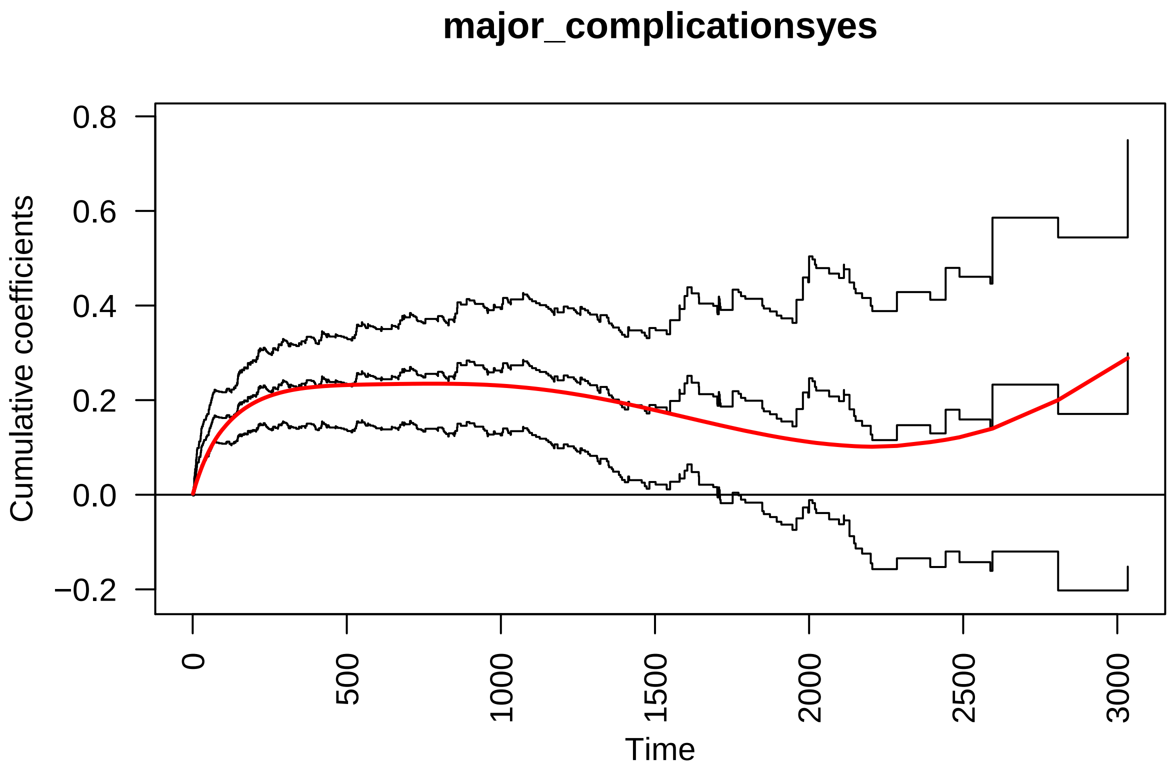


Visual comparison of the time-varying effects from Cox-Aalen model on the cumulated Hazard over time (black) vs. the smooth multiplivative effects of the PAM model (red).




charlson_score
Cumulative coefficients
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
3000
Time


metastasesyes
Cumulative coefficients
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
3000
Time



major_complicationsyes
Cumulative coefficients
0.8
0.6
0.4
0.2
0.0
−0.2
0
500
1000
1500
2000
2500
3000
Time

20 Competing Risk models

- More than one possible event (e.g.: two types of death) next to censoring of which only one can occur. The events **compete** with each other as only one of them can occur.
- Problem with Survival rate estimates (such as KM):
  - Soldiers can die in combat or by accident
  - All 100 soldiers die in helicopter accident at time t **before** they could take part in combat
  - Nobody died in combat at t $\rightarrow S_{Combat}(t) = P(T_{Combat} > t) = 1$ though no combat took place
  - for Kaplan Meier: P(T_{Combat} = t) undefined because nobody at risk at time t.
  - $\Rightarrow$ difficult interpretation of Survival Curves in competing risk scenario
- Approaches:
  - Seperate "cause-specific" Cox models for each type where the competing events are subsumed in censoring.
    * Problem 1: assumption, that $T_1 \perp T_2$
    * Problem 2: Kaplan-Meier Curves are biased
  - Cumulative Incidence Curve as solution to problem 2
  - Discretization: Multinomial GLMs

20.1 Cause-specific Cox PH Models

- One Cox model for each cause.
- Interpretation based on non-occurence of competing events
- Estimate via Partial Likelihood
- Treat competing events $-j$ as being censored which is again the unrealistic independence assumption

$$\lambda_j(t) = \lambda_{0j}(t)exp(X^T\beta_j) \text{ with possibly cause-specific coefficients } \beta_j$$

20.2 Cumulative Incidence Curves

20.2.1 Problem

Study with 100 people over 5 months. Two possible deats: Virus or Cancer. 99 patients die t = 3 on V, 1 dies at t = 5 on C. What is survival rate at t=5 $S(t = 5)$? Depending on the interpretation of V:

1. they represent the C-subpopulation and would have died on Cancer also: $S(t = 5)$ = (1-1)/1 = 0? Thus $Risk_{C1}(T = 5) = 1$ which is the classic Kaplan-Meier way
2. they would have survived Cancer: $S(t = 5) = 1 - 0.01 = 0.99$. Thus $Risk_{C2}(T = 5) = 0.01$ also termed **marginal probability** as V-patients are understood as Cancer-Survivors

We would like to know, who of the V-deaths would have died on Cancer in case they survived V. Which of both Risks is more informative?

20.2.2 Howto CIC

1. Estimate hazard at ordered failure times t_f for event-type j of interest:

$$\hat{\lambda}(t_f)) = \frac{m_{jf}}{n_f} = \frac{\text{\# events j at } t_f}{\text{\# subjects at risk at } st_f}$$

2. Estimate **overall Survival Probability for all event-types** $\hat{S}(t_{f-1})$
3. Compute estimated incidence of failing at time t_f from event type c:

$$\hat{I}_{jf} = \hat{S}(t_{f-1}) \times \hat{\lambda}_j(t_f)$$

4. Cumulate:

$$\text{CIC}_j(t_f) = P(T \leq f|C = j) = \sum_{l=1}^{f} \hat{S}(t_l) \times \hat{\lambda}_j(t_l)$$

Also termed **Aalen-Johannsen Estimator of cumulative incidence**. Kaplan Meier would use event dependent $\hat{S}_c(t_{l-1})$ instead of overall $\hat{S}(t_{l-1})$

20.3 Multinomial time-discret models

Discretize time in q intervals $[a_0, a_1[, ..., [a_{q-1}, a_q[$

$$\lambda_j(t|X) = P(T = t, C = j, T \geq t, X) = \frac{exp(\beta_{0tj} + X^T\beta_j)}{1 + \sum_{i=1}^{k} exp(\beta_{0ti} + X^T\beta_i)} \text{ with \# different events = k + 1}$$

$$\lambda_0(t|X) = P(T > t, C = j, T \geq t, X) = \frac{1}{1 + \sum_{i=1}^{k} exp(\beta_{0ti} + X^T\beta_i)}$$

Interpretation by cause specific log odds w.r.t. reference event type:

$$log\frac{\lambda_j(t|X)}{\lambda_0(t|X)} = \beta_{0tj} + X^T\beta_j$$

$$\frac{\lambda_j(t|X)}{\lambda_0(t|X)} = exp(\beta_{0tj})exp(X^T\beta_j)$$

and $exp(\beta_l j)$ = the effect of covariate x_l on cause specific hazard w.r.t nothing else happens

21 Function calls

21.0.0.1 Cox PH

```
coxph(Surv(start, stop, arrest) ~ fin + age + mar + employed.lag1, data = prison.long)
```

$\lambda(t|x) = \lambda_0(t)exp(\beta_{fin}x_{fin} + \beta_{age}x_{age} + \beta_{mar}x_{mar} + ...)$

21.0.0.2 Cox PH with time varying effect

```
coxph(Surv(start, stop, arrest) ~ fin + age + mar + age:I(log(t)) + employed.lag1, data =
prison.long)
```

$\lambda(t|x) = \lambda_0(t)exp(\beta_{fin}x_{fin} + \beta_{age1}x_{age} + \beta_{age2}(x_{age}log(t)) + \beta_{mar}x_{mar} + ...)$

21.0.0.3 Aalen Model

21.0.0.3.1 All time varying

```
aalen(Surv(days, status) ~ age +  charlson_score + major_complications + metastases, data
= liver)
```

$\lambda(t|X) = \beta_0(t) + \beta_{age}(t) * x_{age}(t) + \beta_{charlson}(t) * x_{charlson}(t) + \beta_{complications}(t) * x_{complications}(t) + \beta_{metastases}(t) * x_{metastases}(t)$

Semi parametric

```
aalen(Surv(days, status) ~ const(age) +  charlson_score + const(major_complications) +
metastases, data = liver)
```

$\lambda(t|X) = \beta_0(t) + \beta_{age} * x_{age}(t) + \beta_{charlson}(t) * x_{charlson}(t) + \beta_{complications} * x_{complications}(t) + \beta_{metastases}(t) * x_{metastases}(t)$

21.0.0.3.2 Cox Aalen model

```
cox.aalen(formula   = Surv(days, status) ~ age +  charlson_score + major_complications +
metastases + prop(sex) + prop(transfusion) + prop(major_resection), data = liver, residuals
= 1, basesim   = 1)
```

$\lambda(t|X) = [\beta_0(t) + \beta_{age} * x_{age}(t) + \beta_{charlson}(t) * x_{charlson}(t) + \beta_{complications} * x_{complications}(t) + \beta_{metastases}(t) * x_{metastases}(t)]exp(x_{se}$

21.0.0.3.3 PEM

```
glm(ped_status ~ interval -1 + treatment, offset = offset, data = leuk.ped, family =
poisson(link = log))
```

$$\lambda_i(t|x_i) = exp(\lambda_{0j} + x_{i,treatment} * \beta_{treatment})$$

21.0.0.3.4 PAM

```
gam(ped_status ~ s(tend, k = 10) + treatment, offset = offset, data = leuk.ped, family =
poisson(link = log))
```

$$\lambda_i(t|x_i) = exp(f_0(t_j) + x_{i,treatment} * \beta_{treatment}) = exp(\sum_{k=1}^{K=10} \gamma_k B_k(t_j) + x_{i,treatment} * \beta_{treatment})$$

PAMM

```
pamm_icu <- bam(ped_status ~ s(tend) + Year + AdmCatID + DiagID2 + s(Age) + s(BMI) +
s(ApacheIIScore) + s(CombinedicuID, bs="re"), offset=offset, data = ped, family=poisson(),
discrete = TRUE)
```

$$\lambda_i(t|x_i) = exp(f_0(t_j)+x_{i,year}*\beta_{year})+x_{i,AdmCatID}*\beta_{AdmCatID})++x_{i,DiagID2}*\beta_{DiagID2})+f(x_{i,age},t_j)+f(x_{i,bmi},t_j)++f(x_{i,ap}$$

22 Random Stuff

22.0.1 Attest significance only based on β and $se(\beta)$

1. compute z-score: $z = \beta/se(\beta)$
2. thresholds for $alpha = 0.05$:
3. one sided: $z_{thresh} = 1.64$
4. two sided: $z_{thresh} = 1.96$
5. Reject H0 (coefficient = 0 aka is not significant) if $z > z_{thresh}$

check their explanation