LDA Interpretation

Jann Goschenhofer

January 2018

Contents

Introduction **1**

Kaplan Meier **1**

Nelson Aalen **3**

Accelerated Failure Time Transformation models **3**

Cox Regression model **3**

Model fit Analysis **5**

Prediction Error Curves (PEC) 5

Residuals 6

Semi-parametric additive Cox model **6**

Time discrete Survival models **6**

Piecewise exponential models (PEM) **6**

Piecewise additive exponential models (PAM) **6**

Piecewise additive exponential mixed models (PAMM) **7**

Frailty models **10**

Aalen model **10**

Cox-Aalen model **10**

Competing Risk models **18**

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.

Kaplan Meier

Model Equation

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.

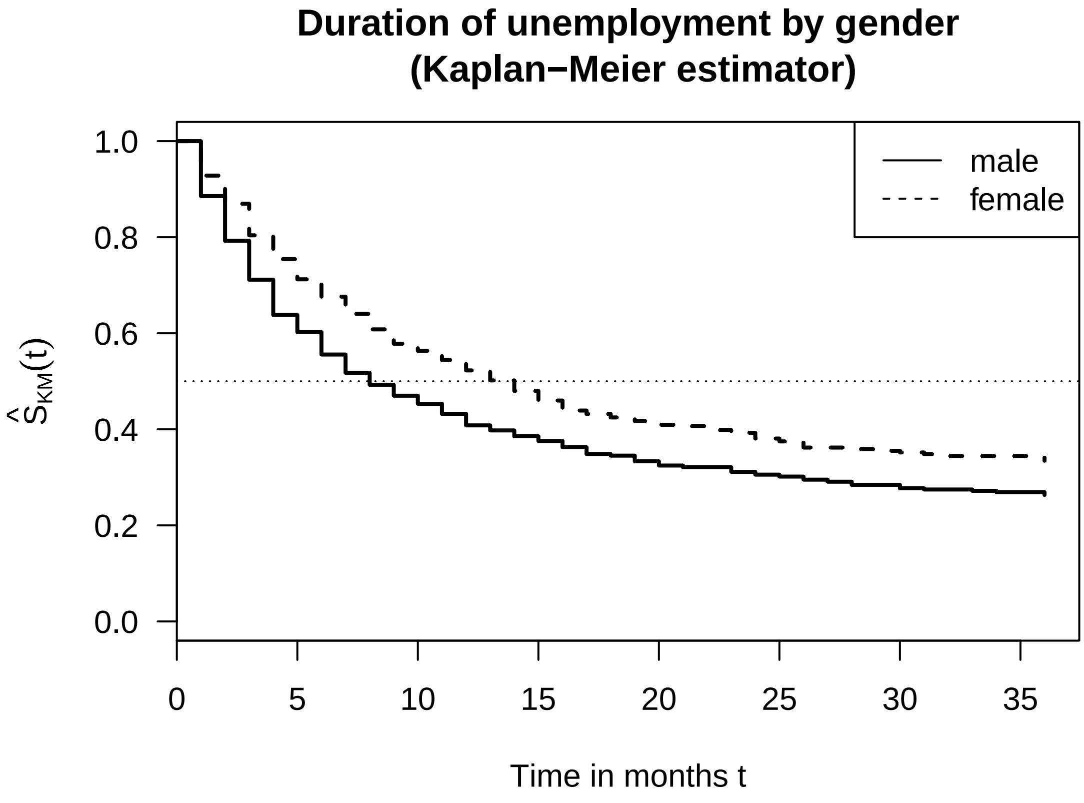
Reveals a step function with jumps at each t_k where events took place.

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
```

Model



Test

```
## 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
```

Nelson Aalen

Accelerated Failure Time Transformation models

Cox Regression model

Model equation

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

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
```

Model

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

- Person with treatment 2 has a multiplicative factor exp(0.4664) = 1.594245 higher hazard rate than a person with treatment 1
- this effect is not significant as the H0 can not be rejected at $\alpha = 0.05$, but this does not imply testing of the PH assumption DISCUSS
- log rank score test: tests for significant differencies in the survival curves for the two subpopulations divided by the 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.

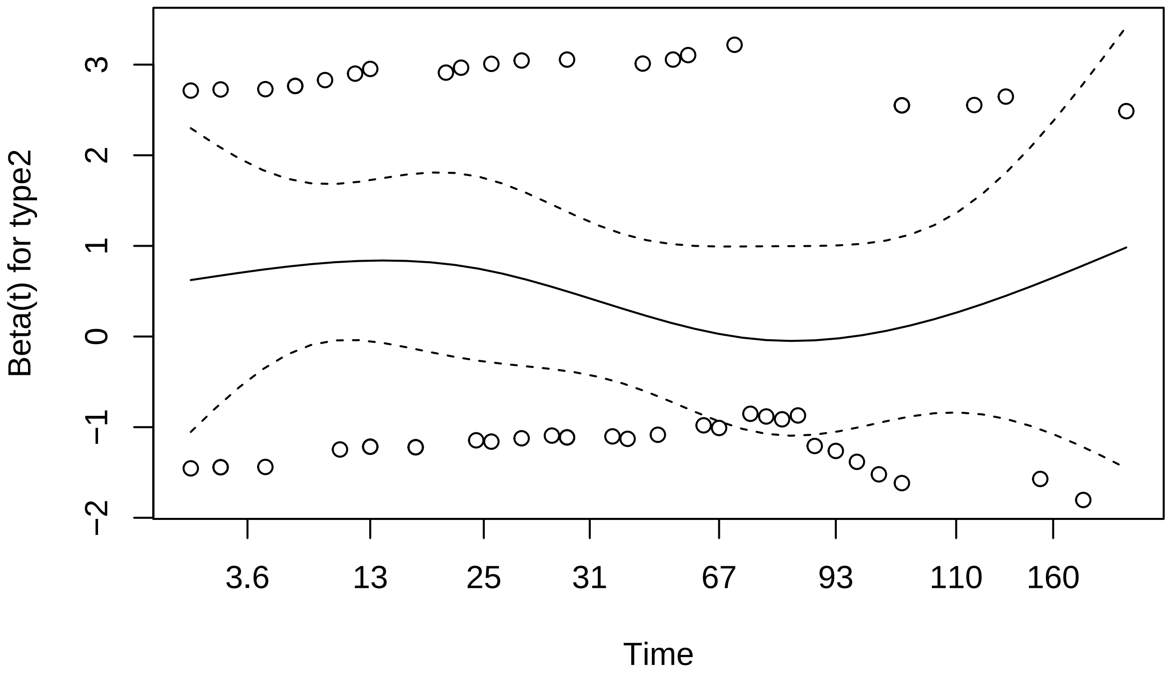
Summary of the Cox-PH model:

```
## 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
```

Test the Cox PH assumption for the covariates

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:

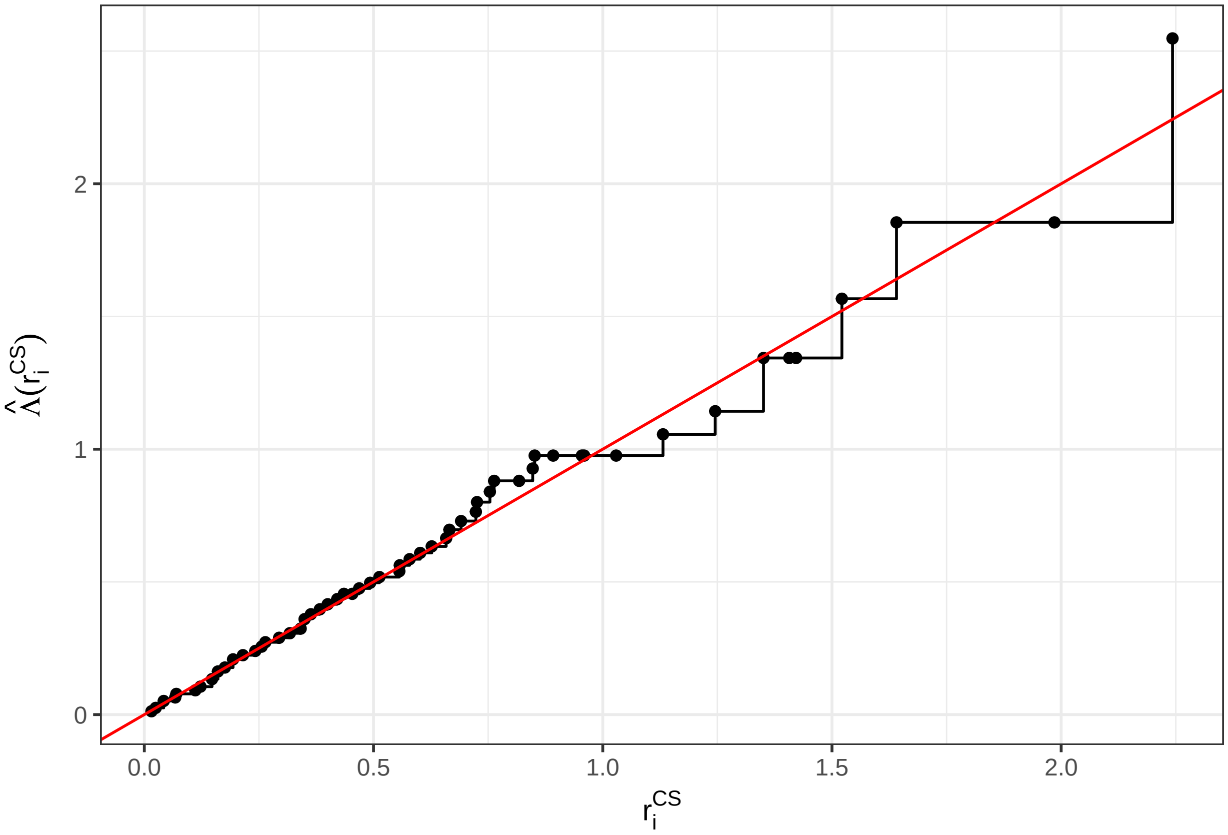


Test PH

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

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.

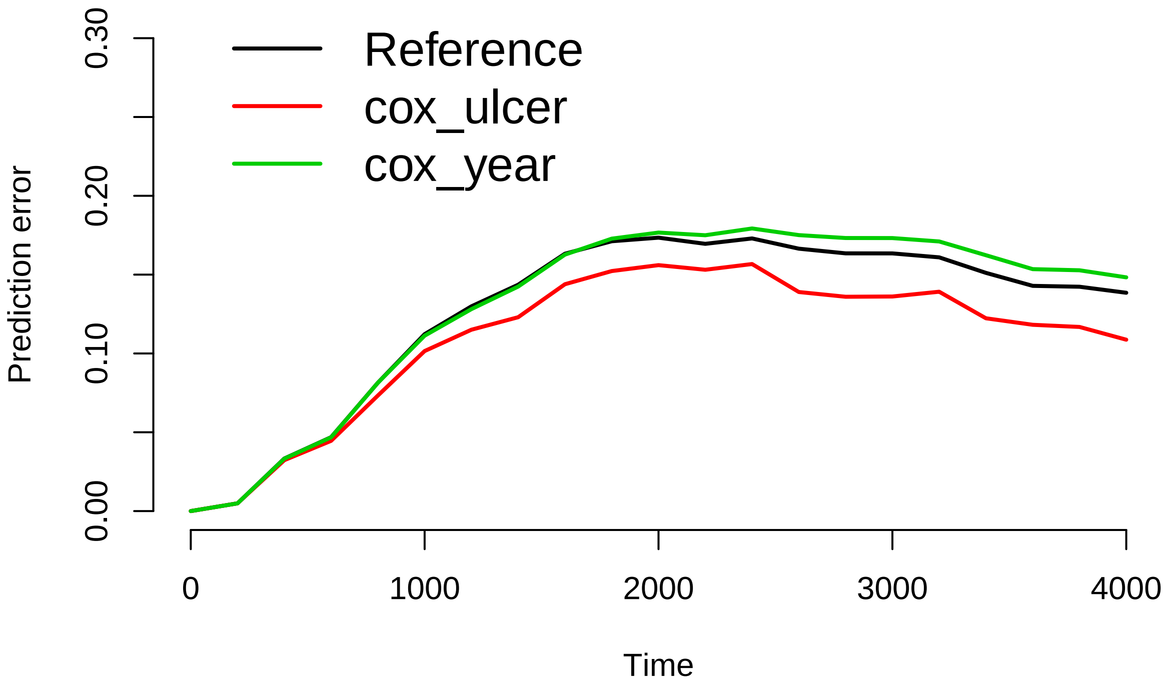


Model fit Analysis

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.



Residuals

- Schoenfeld
- Martingale
- Deviance
- Cox-Snell

Semi-parametric additive Cox model

Time discrete Survival models

Piecewise exponential models (PEM)

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.

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

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:

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

Piecewise additive exponential mixed models (PAMM)

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:

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

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 vs. 400.

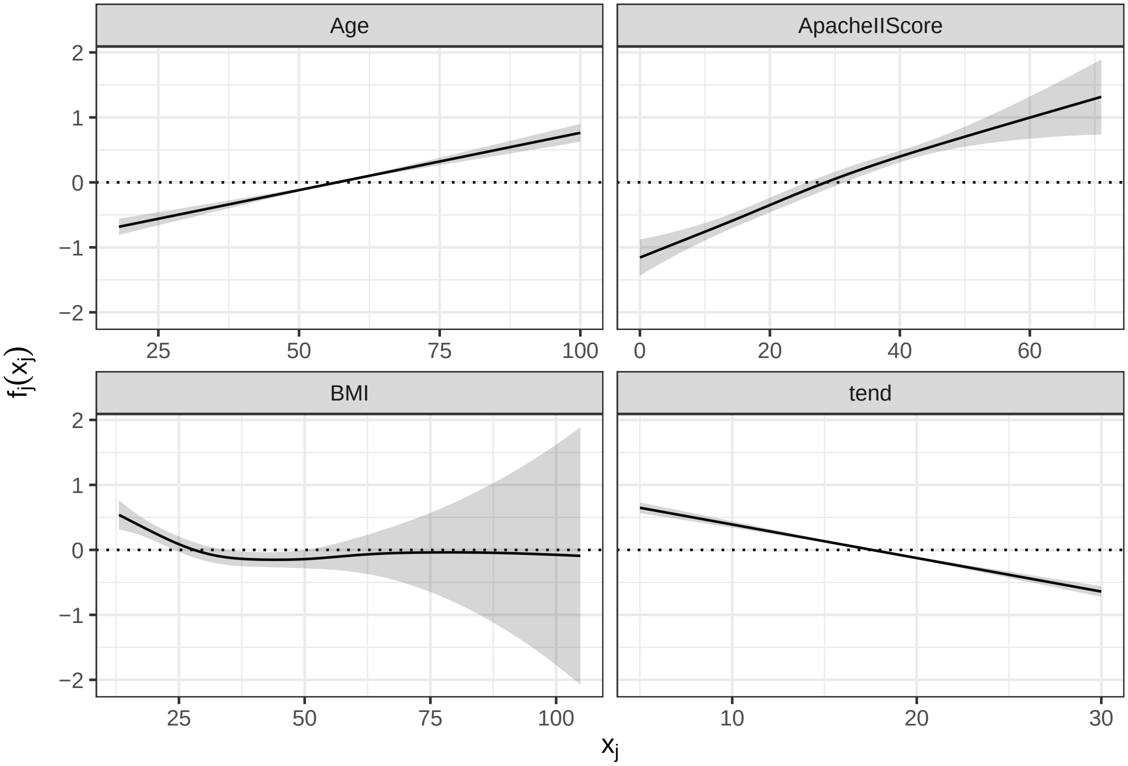
We model the PAM as a Poisson model with log link on the death-indicator `ped_status`

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
```

What can we say?

- smooth terms for continuos 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
  - 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)
  - time (tend) has a falling slope aka a decreasing effect on the log(hazard) -> has 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
```

Frailty models

Aalen model

model equation

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

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

Cox-Aalen model

model equation

$$\lambda(t) = \lambda_0(t) + X(t)\beta(t) \cdot exp(Z(t)'\gamma)$$

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.

Data

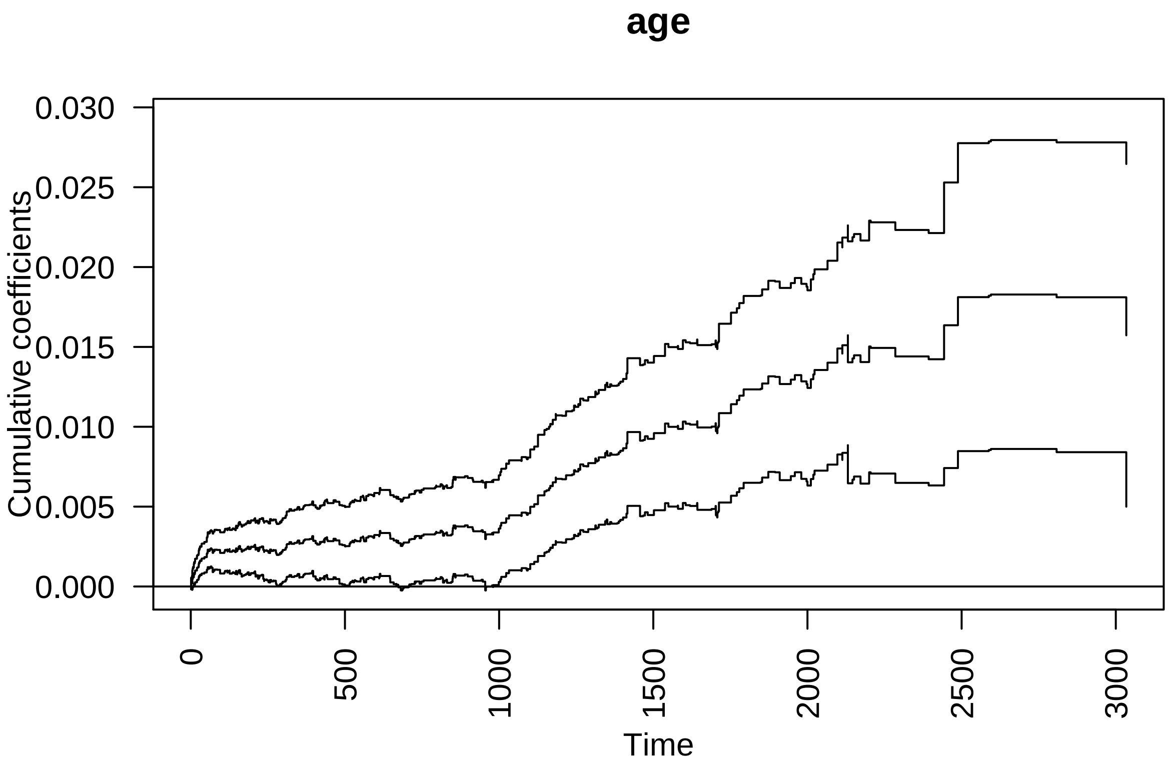
looks like that

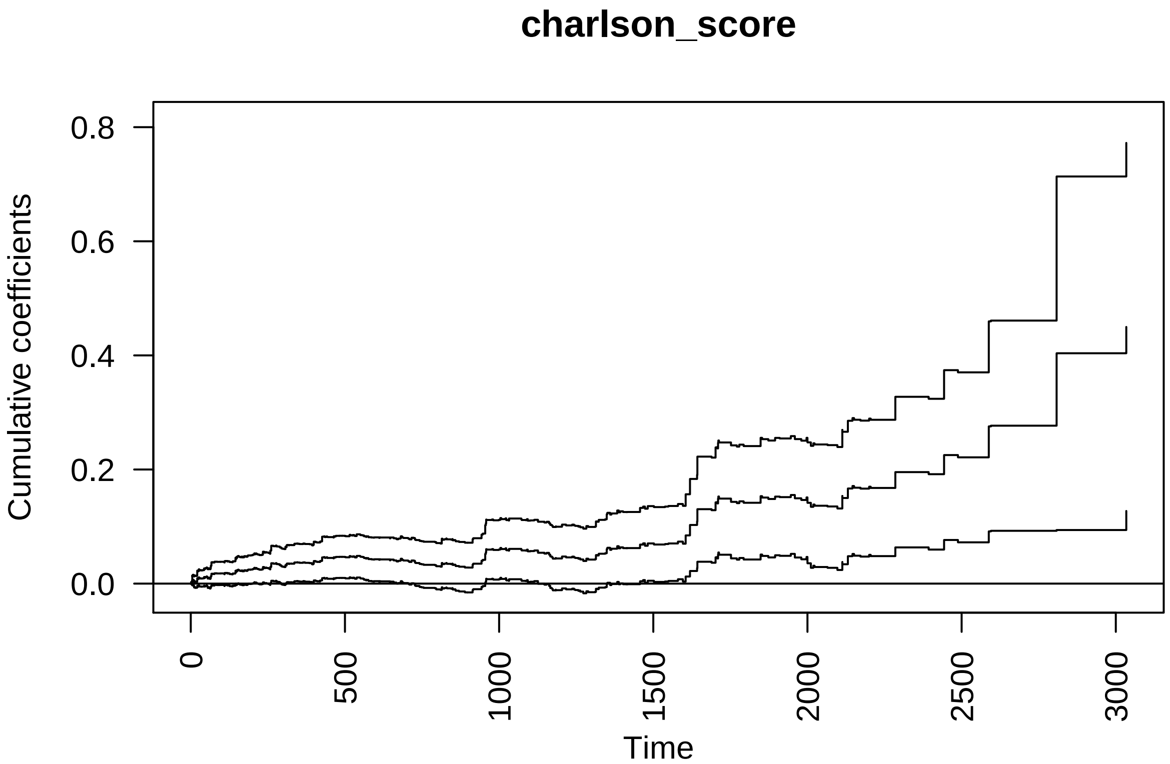
```
##   major_complications age charlson_score sex transfusion metastasesYN
## 1                  no  58              2   f         yes            1
## 2                 yes  52              2   m          no            1
## 3                  no  74              2   f         yes            1
## 4                 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
```

What can we say from the graphic?

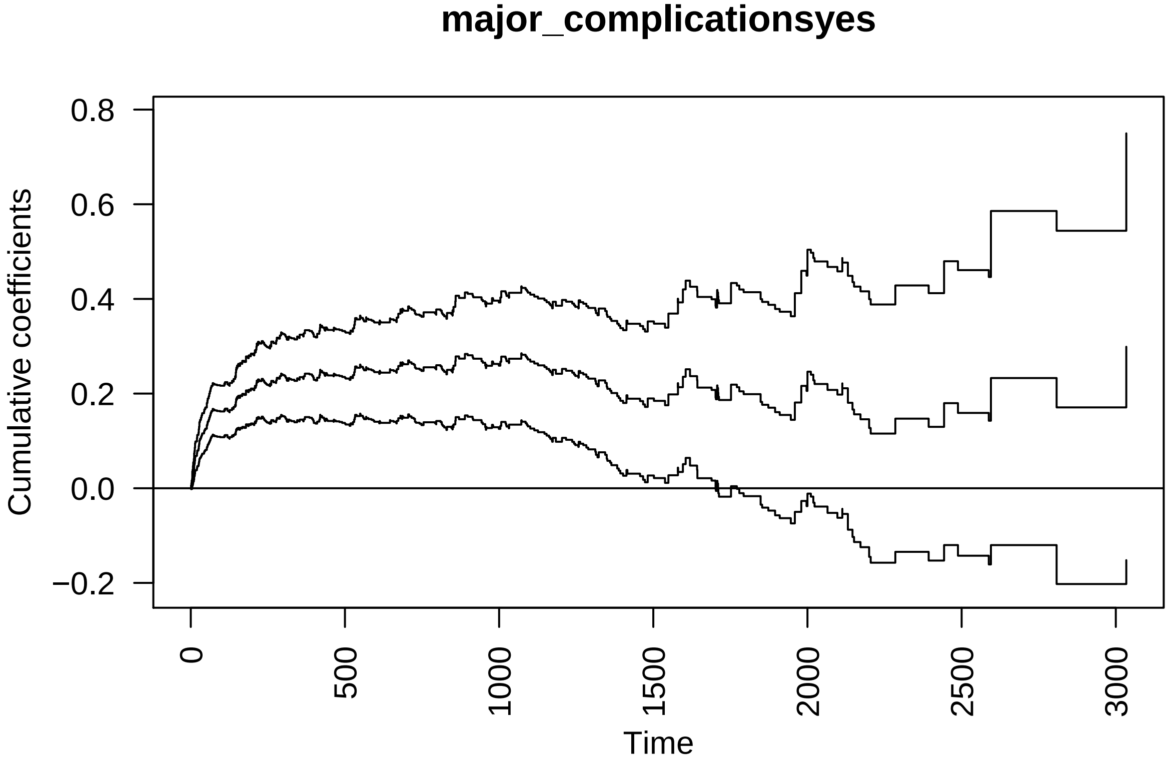
- Age:
  - the cumulative Hazard of a person aged A+1 at time point t = 1500 is 0.01 higher than that of a person aged A
  - the effect of metastases on the cumulative hazard rate starts to increase t = 1000 after the surgery and is approx. constant before
- Complications:
  - the cumulative Hazard of 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)`

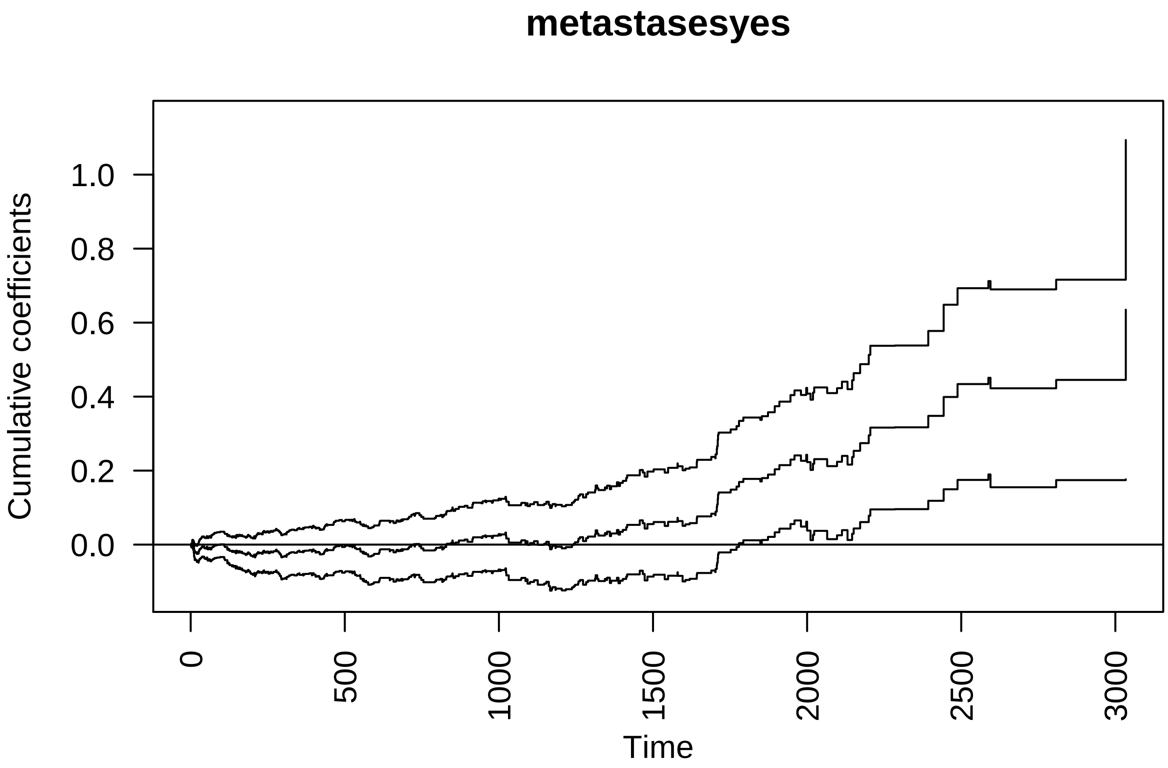




charlson_score



major_complicationsyes



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              0.004
## age                                              4.18              0.000
## charlson_score                                   4.04              0.000
## major_complicationsyes                           6.07              0.000
## metastasesyes                                    3.85              0.002
##
## 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)                                   0.200
## age                                           0.408
## charlson_score                                0.084
## major_complicationsyes                        0.150
## metastasesyes                                 0.020
##
```

```
## 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 0.0371
## prop(transfusion)yes     0.233 0.111     0.113       0.112 2.07 0.0386
## 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         0.196
## prop(transfusion)yes                6.51         0.580
## prop(major_resection)yes            8.99         0.176
```

- Aalen part:
  - Supremum-test: for all 4 variables the H0: no effect can be rejected
  - Kolmogorov Smirnov for time variant effects: H0: constant effect can only clearly be rejected for metastases DISCUSS THIS
- 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
  - same for transfusion ($\exp(0.233) = 1.262381$) and major_resection ($\exp(0.254) = 1.289172$)
  - DISCUSS

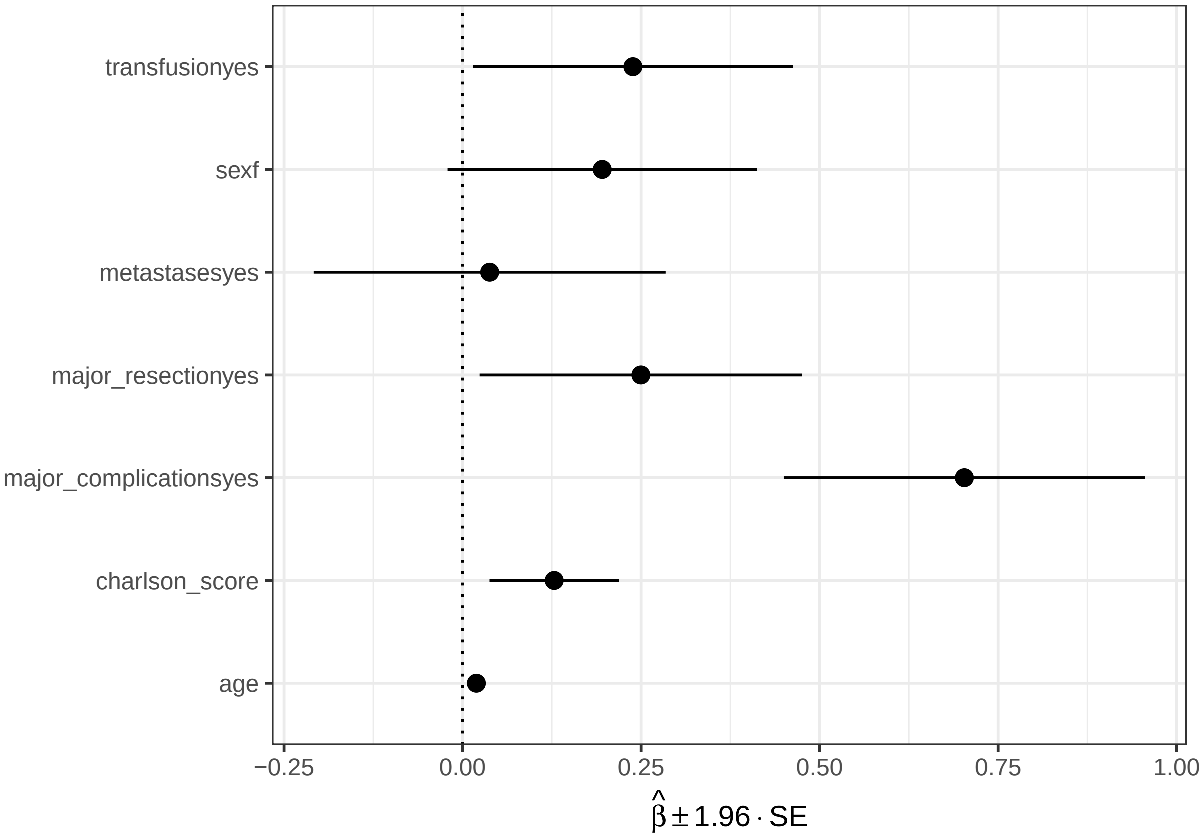
Cox-Aalen vs. PAM

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:

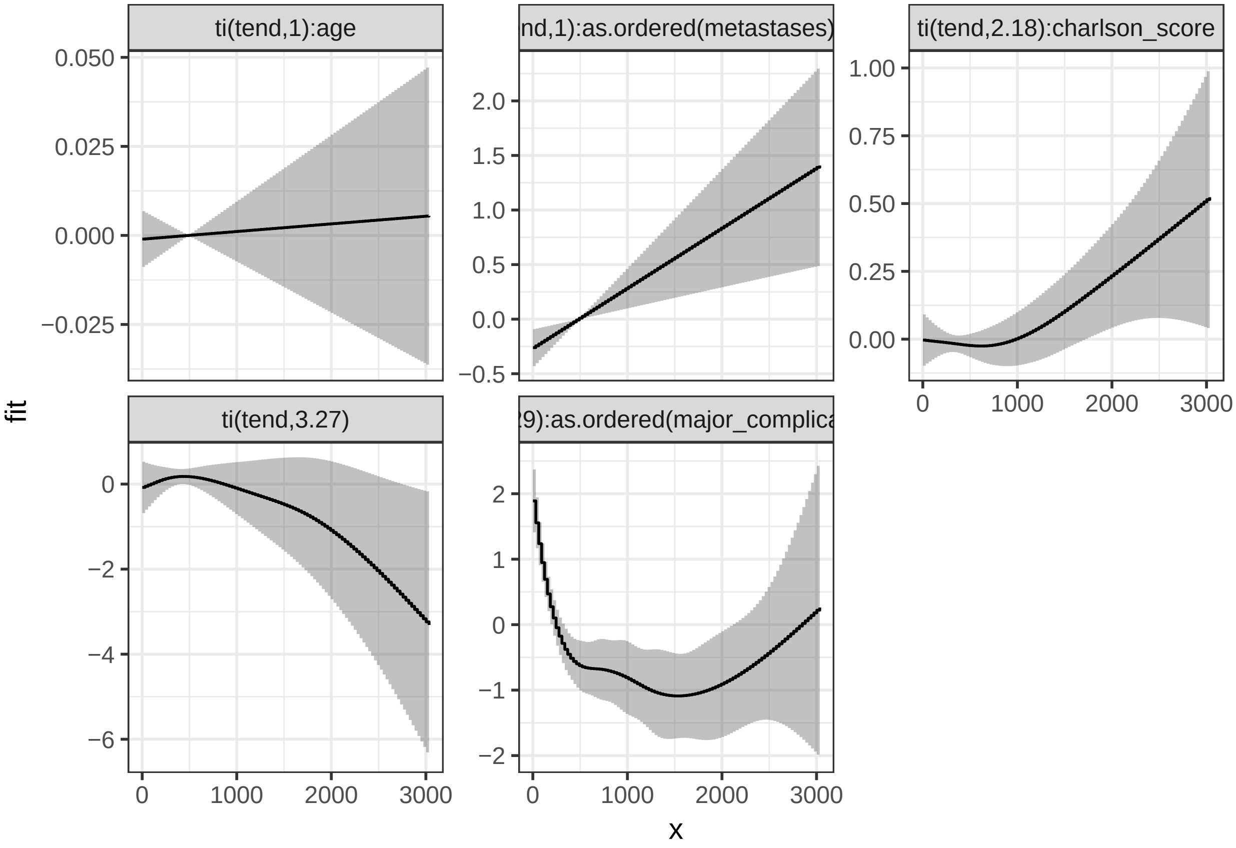
- NOTE: Constant contributions to time-varying can be interpreted as effects at t=0. Check the model equation and DISCUSS
- sex: Compared to males, females have a 1.22 times increased risk of experiencing an event (c.p.)
- transfusion: Compared to patients without transfusion, patients with transfustion 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:

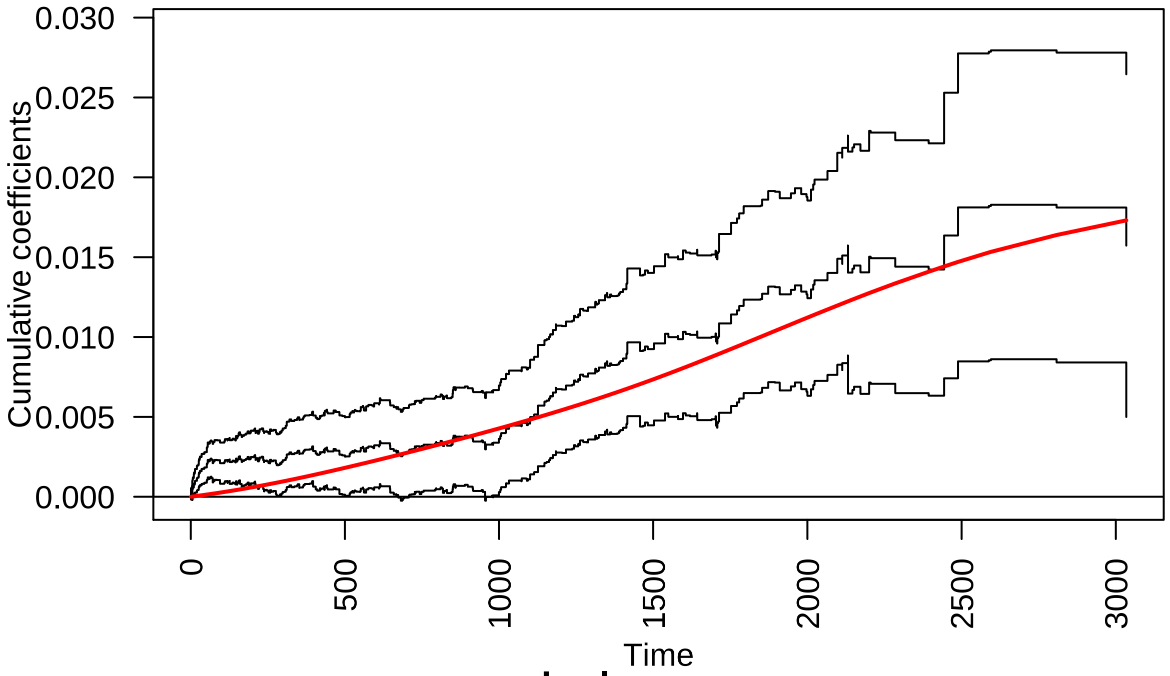
```
##
## Family: poisson
## Link function: log
##
## Formula:
## ped_status ~ ti(tend, k = 10) + metastases + ti(tend, by = as.ordered(metastases),
##     k = 10, mc = c(1, 0)) + major_complications + ti(tend, by = as.ordered(major_complications),
##     k = 10, mc = c(1, 0)) + age + ti(tend, by = age, 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:

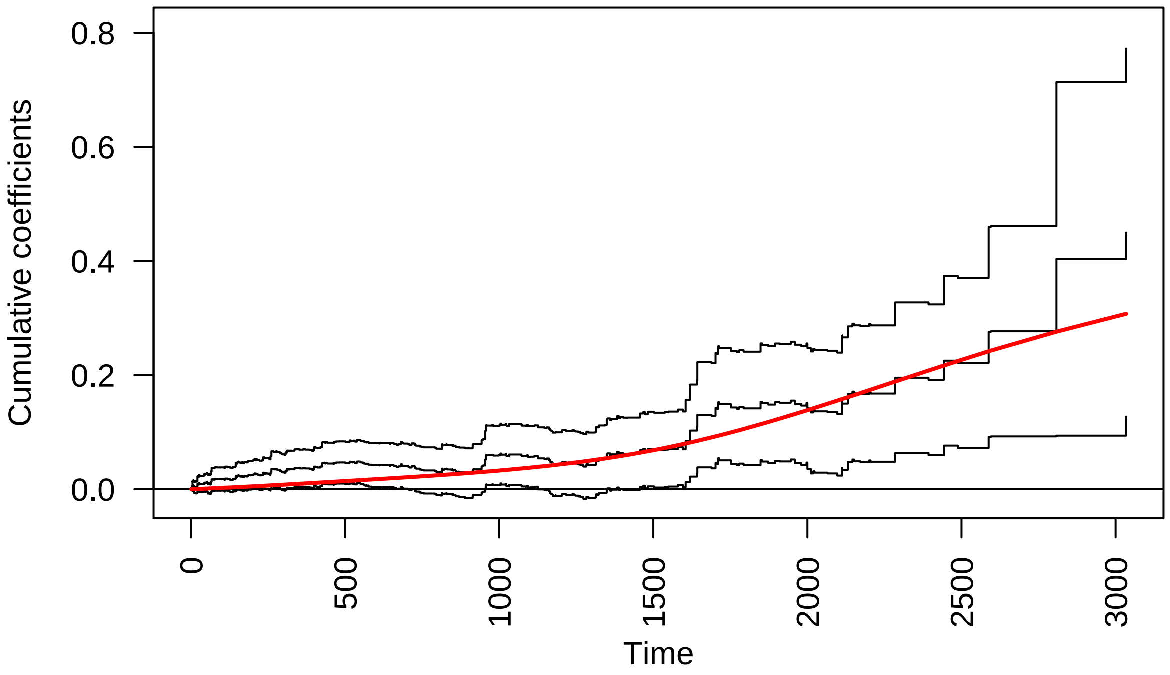


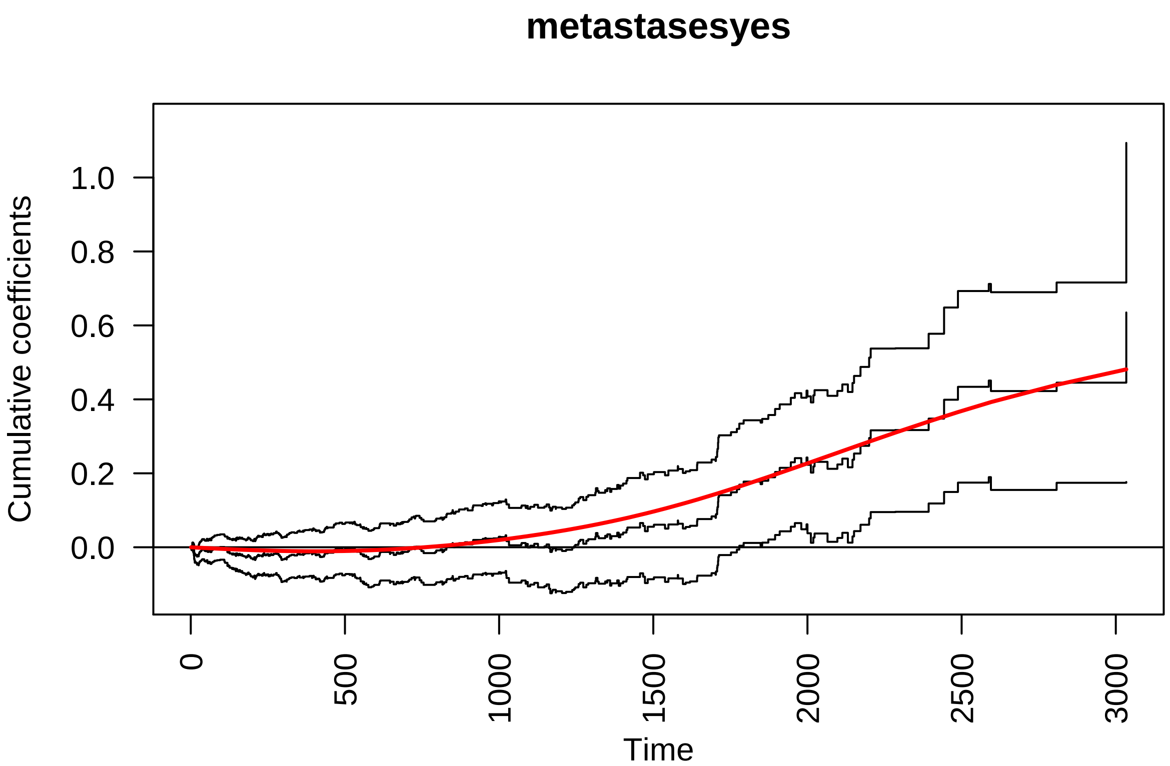
Visual comparison of the time-varying effects from Cox-Aalen model on the cumulated Hazard over time (black) vs. the smooth multiplivative effects of the PAM model (red).

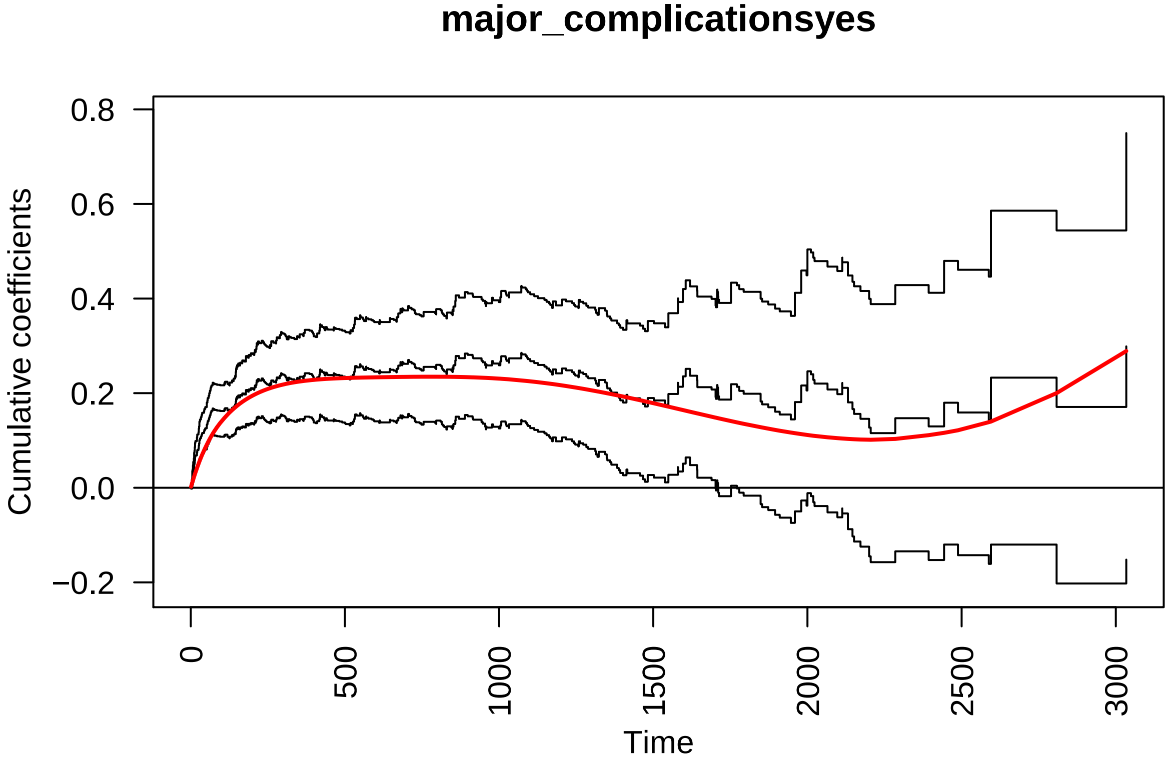
age



charlson_score







Competing Risk models