Exam 2017

Exam nr. 38

16 januar 2017

Theoretical Part

1.1

Notation

Let $n = n_m + n_f$

$$
\begin{aligned}
N(t) &= \left(N_1^m(t), ..., N_{n_m}^m(t), N_1^f(t), ..., N_{n_f}^f(t)\right)^T, \quad n \times 1 \\
\lambda(t) &= \left(\lambda_1^m(t), ..., \lambda_{n_m}^m(t), \lambda_1^f(t), ..., \lambda_{n_f}^f(t)\right)^T, \quad n \times 1 \\
Y(t) &= \left(Y_1^m(t), ..., Y_{n_m}^m(t), Y_1^f(t+4), ..., Y_{n_f}^f(t+4)\right)^T, \quad n \times 1 \\
X(t) &= \left(Y_1^m(t)X_1^m(t), ..., Y_{n_m}^m(t)X_{n_m}^m(t), Y_1^f(t+4)X_1^f(t+4), ..., Y_{n_f}^f(t+4)X_{n_f}^f(t+4)\right)^T \quad n \times q \\
Z(t) &= \left(Z_1^m(t), ..., Z_{n_m}^m(t), Z_1^f(t+4), ..., Z_{n_f}^f(t+4)\right)^T \quad n \times p
\end{aligned}
$$

a)

Since females correspond to a 4 year younger male, we simply add 4 years to the time point in the female hazards. The hazard for females is thus

$$\lambda_i^f(T_i + 4) = (X_i^f(T_i + 4))^T \beta(T_i) + h(\gamma^T Z_i^f(T_i + 4))$$

b)

We can decompose the process $N(t)$ into

$$dN(t) = \lambda(t)dt + dM(t) = X(t)dB(t) + diag(Y(t))h(Z(t)\gamma) + dM(t)$$

And since the martingale increments are iid. with mean zero, we will disregard these in the estimation of B, which is done through OLS;

$$
\begin{aligned}
\frac{\partial}{\partial dB(t)} \left(dN(t) - \lambda(t)dt\right)^{\otimes 2} = -2(X(t))^T \left(dN(t) - \lambda(t)dt\right) = 0 \quad \Leftrightarrow \\
(X(t))^T dN(t) = (X(t))^T X(t)dB(t) + (X(t))^T diag(Y(t))h(Z(t)\gamma)dt \Leftrightarrow \\
dB(t) = \left((X(t))^T X(t)\right)^{-1} (X(t))^T \left(dN(t) - diag(Y(t))h(Z(t)\gamma)dt\right)
\end{aligned}
$$

where we have assumed that $(X(t))^T X(t)$ is invertible. Hence we get

$$\hat{B}(t) = \int_0^t X^-(s) \left(dN(s) - diag(Y(s))h(Z(t)\gamma)\right) ds$$

where $X^-(t) = \left((X(t))^T X(t)\right)^{-1} (X(t))^T$.

c)

We will use the same approach as above, ie. the OLS, to write up the estimating equation for γ.

$$\begin{aligned}
\frac{\partial}{\partial\gamma}\left(dN(t)-\lambda(t)dt\right)^{\otimes 2} &= -2Z(t)^T D_\gamma h(Z(t)\gamma)^T diag(Y(t))\left(dN(t)-\lambda(t)dt\right) = 0 \Leftrightarrow \\
Z(t)^T D_\gamma h(Z(t)\gamma)^T diag(Y(t))\left(dN(t)-X(t)d\hat{B}(t)+diag(Y(t))h(Z(t)\gamma)dt\right) &= 0 \Leftrightarrow \\
Z(t)^T D_\gamma h(Z(t)\gamma)^T diag(Y(t))\left[H(t)dN(t)-H(t)diag(Y(t))h(Z(t)\gamma)dt\right] &= 0
\end{aligned}$$

where $H(t) = (I - X(t)X(t)^-)$.

d1)

To estimate γ, we should solve for γ the estimating equation derived above, for a given smooth function h. To establish the assymptotic behavoir of the estimate, $\sqrt{n}(\hat{\gamma}-\gamma)$, we would rely on the decomposition of $\hat{\gamma}$. If $\sqrt{n}(\hat{\gamma}-\gamma)$ decomposes into

$$\sqrt{n}(\hat{\gamma}-\gamma) = g(U_n, A_n) + o_p(1)$$

for a continuous function g, it will suffice to show that U_n converges in distribution to a standard normal distributed variable and that A_n converges in probability to some determanistic limit. (The last term $o_p(1)$ does not matter for the convergense in distribution do to Slutsky's lemma). \

One way to get the desired convergence of U_n is to show that U_n is a local square integrable martingale at a given stopping time, and then use the martingale CLT to get convergence of U_n to U, where U is a gaussian martingale. \

To summarize: Assumne that $\sqrt{n}(\hat{\gamma}-\gamma) = g(U_n, A_n)$, where g is continuous, U_n is a local square integrable martingale, and A_n is a process that converges in probability. If

$$\begin{aligned}
\langle U^{(n)}\rangle(t) &\xrightarrow{P} V(t) \\
\langle U^{(n)}_{\epsilon l}\rangle(t) &\xrightarrow{P} 0
\end{aligned}$$

then $\sqrt{n}(\hat{\gamma}-\gamma)$ converges in distribution to a standard normal distribution. \

To show how to estimate the asymptotic variance of this estimate, assume that g is on the form $g(x,y) = xy$. Then the asymptotic variance is then given by $\hat{\Sigma} = A[U](\tau)A$ where τ is a stopping time.

d2)

To estimate B we simply plog in $\hat{\gamma}$ in the estimate from b) to obtain $\hat{B}(t)$. To estimate the asymptotic variance, obeserve the following calculations:

$$\sqrt{n}(\hat{B}(t) - B(t)) = \sqrt{n}\left(\int_0^t X^- \left(dN(t) - diag(Y(t))h(Z(t)\hat{\gamma})dt\right) ds - B(t)\right)$$
$$= \sqrt{n}\left(\int_0^t X^-(s)d(\lambda + M)(s) - \int_0^t X^- diag(Y(s))h(Z(s)\hat{\gamma})ds - B(t)\right)$$
$$= \sqrt{n}\int_0^t X^-(s)X(s)dB(s) + \sqrt{n}\int_0^t X^-(s)diag(Y(s))h(Z(s)\gamma)ds$$
$$+ \sqrt{n}\int_0^t X^-(s)dM(s) - \sqrt{n}\int_0^t X^-(s)diag(Y(s))h(Z(s)\hat{\gamma})ds - B(t)$$
$$= \sqrt{n}\int_0^t X^-(s)dM(s) - \sqrt{n}\int_0^t X^-(s)diag(Y(s))\left(h(Z(s)\hat{\gamma}) - h(Z(s)\gamma)\right) ds$$

The left term is clearly a martingale and hence we will use the martingale CLT to establish the convergense in distribution to a gaussian martingale for the left term. The right term is heavily h-dependent however, so it is difficult to state the exact procedure to establish asymptotic properties for the right term, but we would hope to establish convergense to a gaussian process (not neccesarially a martingale). In this case, $\sqrt{n}(\hat{B}(t) - B(t))$ will converge to a gaussian process. To get the variance of the asymptotic distribution we would use optional variation.

d3)

In our specification of the variables above (the notation section), we scaled the females by adding 4 years to their time points in order to correctly estimate beta across gender, since females correspond to a 4 year younger male. This means that B is estimated for t from 26 to 90. If instead of scaling the females we corrected by subtracting 4 years from the time points of the males, we would similarly get B estimated for t from 30 to 94.

e)

In order to predict the survival probability for a male after 10 years, we simply plug in to the survival function with the estimates of B and γ:

$$\hat{S}_0(40) = \exp(-\int_{30}^{40} X_0\hat{\beta}(s)ds - \int_{30}^{40} h(\hat{\gamma}^T Z_0)ds) = \exp(-(X_0\hat{B}(40) - X_0\hat{B}(30) + 10h(\hat{\gamma}^T Z_0))$$

To get standard errors for this estimate, one would derive the asymptotic distribution of $\sqrt{n}(\hat{S}_0 - S_0)$, and then utilize the martingale decomposition to find the variance through optional variation.

f)

If the women enter the study at different ages above 30 years, we simply introduce left truncation to the model. That is, we let $Y(t) = 1(V \leq t \leq T)$ where V is the left truncation time i.e. the time when females enter the study. Then we still get consistent estimators.

1.2

a)

Observe the survival function when X is observed:

$$
\begin{aligned}
P(T* > t|X = 1) &= \int P(T* > t|X = 1, A)dA(P) \\
&= \int_0^\infty f(a) \exp\left(-\int_0^t \beta(s) + \alpha(s+a)ds\right) da \\
&= \exp\left(-\int_0^t \beta(s)\right) \int_0^\infty f(a) \exp\left(-\int_0^t \alpha(s+a)ds\right) da \\
&= \exp(-B(t))\rho(t)
\end{aligned}
$$

and

$$
\begin{aligned}
P(T* > t|X = 0) &= \int P(T^* > t|X = 0, A)dA(P) \\
&= \int_0^\infty f(a) \exp\left(-\int_0^t \beta(s)ds\right) da \\
&= \exp(-B(t)).
\end{aligned}
$$

For the second case, the hazard function is given by

$$
\frac{\partial}{\partial t} - \log S(t|X = 0) = \beta(t).
$$

For the first case the hazard function is given by

$$
\begin{aligned}
\frac{\partial}{\partial t} - \log S(t|X = 1) &= \frac{\partial}{\partial t} B(t) - \frac{\partial}{\partial t} \log \rho(t) \\
&= \beta(t) - \frac{1}{\rho(t)} \rho'(t)
\end{aligned}
$$

where

$$
\rho'(t) = \int_0^\infty -f(a)\alpha(t+a) \exp\left(-\int_0^t \alpha(s+a)ds\right) da
$$

under certain regularity conditions.

b)

We will use a similar approach as in a), but here we will utilize that X and A are independent. Observe:

$$
\begin{aligned}
P(T^* > t) &= \int P(T^* > t|X, A)dA(P) \otimes X(P) \\
&= \int_{\{0,1\}\times[0,\infty)} \exp\left(-\int_0^t \lambda(s)ds\right) dA(P) \otimes X(P) \\
&= 1/2 \int_{[0,\infty)} f(a) \exp\left(-\int_0^t \beta(s)ds\right) da \\
&+ 1/2 \int_{[0,\infty)} f(a) \exp\left(-\int_0^t \beta(s) + \alpha(s+a)ds\right) da \\
&= 1/2(\exp(-B(t)) + \exp(-B(t))\rho(t)) \\
&= 1/2 \exp(-B(t))(1 + \rho(t))
\end{aligned}
$$

We then get the hazard function

$$\frac{\partial}{\partial t} - \log S(t) = \beta(t) - \frac{1}{1+\rho(t)}\rho'(t)$$

Practical part

2.1

We start by fitting a cox regression model with all four variables included in the model.

```
ttp$surv <- with(ttp, Surv(ttp, k.gravid))
fit.cox <- cox.aalen(surv ~ prop(k.cof) + prop(k.ryg) + prop(m.ryg) + prop(m.zkon0),
                     weighted.test = 0,
                     residuals = 1,
                     data = ttp)
```

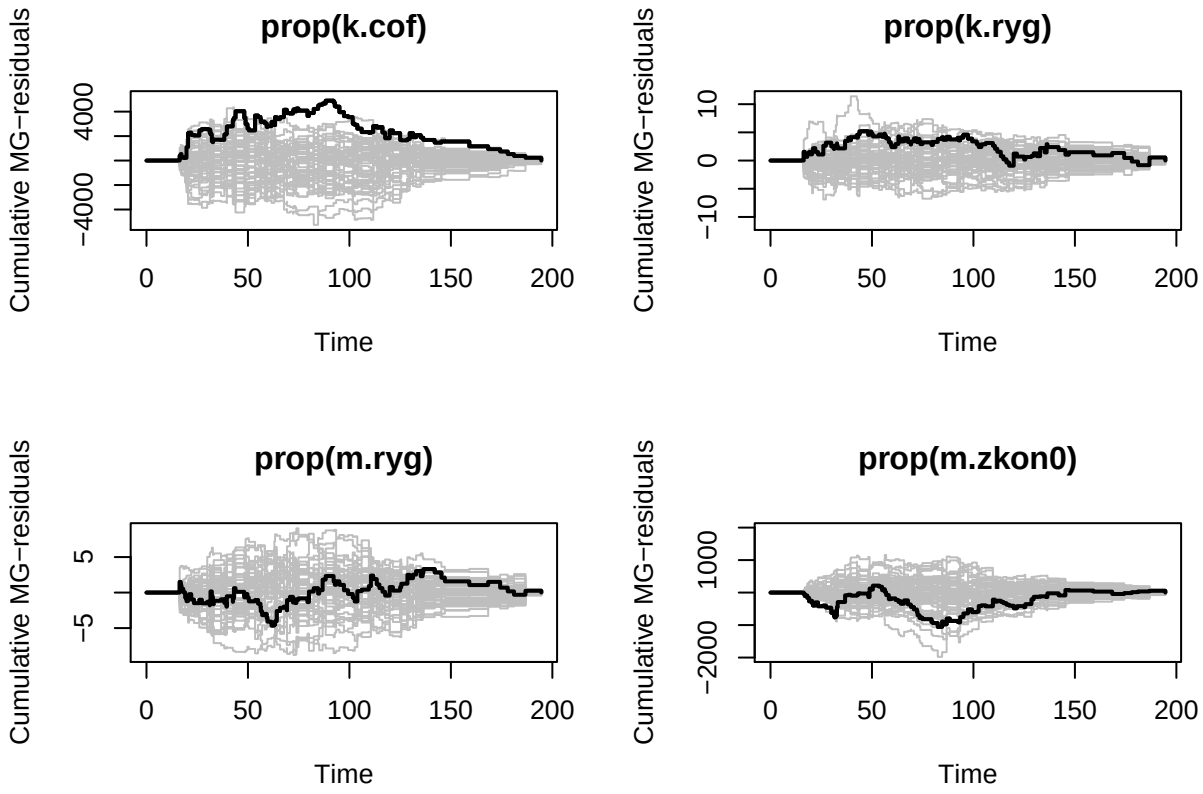
To acces the model fit, we look at the score process test for proportionality. In the table below we look at the score processes evaluated in the unweighted supremum test under the null. We see that there are a lack of fit for the caffein intake of females.

```
(summary(fit.cox))
```

```
## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                Coef.    SE Robust SE D2log(L)^-1      z P-val lower2.5%
## prop(k.cof)    0.000 0.000     0.000       0.000 -1.030 0.304     0.000
## prop(k.ryg)   -0.134 0.173     0.180       0.175 -0.745 0.456    -0.473
## prop(m.ryg)   -0.110 0.169     0.167       0.168 -0.659 0.510    -0.441
## prop(m.zkon0)  0.004 0.001     0.001       0.001  4.520 0.000     0.002
##               upper97.5%
## prop(k.cof)        0.000
## prop(k.ryg)        0.205
## prop(m.ryg)        0.221
## prop(m.zkon0)      0.006
## Test of Proportionality
##               sup|  hat U(t) | p-value H_0
## prop(k.cof)           4900.00        0.022
## prop(k.ryg)              5.19        0.392
## prop(m.ryg)              4.65        0.596
## prop(m.zkon0)         1060.00        0.104

## NULL
```

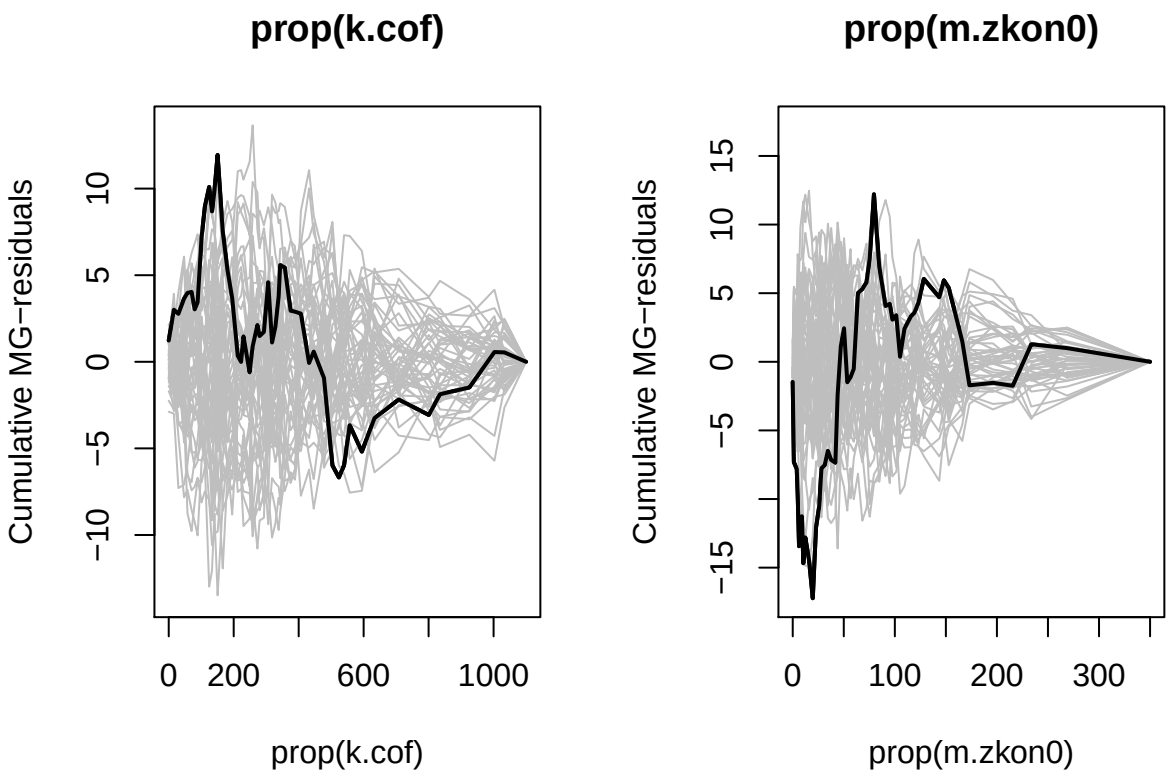
To support this finding, we look at the score processes plottet against simulated tests. Below we see the process for the caffein intake lays outside of the range of the simulated process or in the outmost extreme. This support our statement of lack of model fit for caffein intake of the females.

```
par(mfrow = c(2,2))
plot(fit.cox, score = TRUE)
```



The tests processes above are processes over time. For the continuous variables, we might also look at the model fit across the covariate levels. Below we see plots similar to the ones above, but here it is the residuals against the covariates that are plottet. Here m.zkon0 is the variable of concern, since it has some large abreviations for small values of m.zkon0. Below the plots we see the processes evaluated in the supremum test, where we get a small p-value for m.zkon0, which support our statement of lack of model fit for m.zkon0 across different concentrations.

```
resid <- cum.residuals(fit.cox, data = ttp, cum.resid = 1)
par(mfrow = c(1,2))
plot(resid, score = 2)
```




```
summary(resid)
```

```
## Test for cumulative MG-residuals
##
## Grouped cumulative residuals not computed, you must provide
## modelmatrix to get these (see help)
##
## Residual versus covariates consistent with model
##
##                sup|  hat B(t) | p-value H_0: B(t)=0
## prop(k.cof)              11.944                0.094
## prop(m.zkon0)            17.230                0.002
```

To circumvent this, we will try to look at m.zkon0 at a log scale.

```
ttp$log_zkon <- log(ttp$m.zkon0)
ttp$log_cof <- log(ttp$k.cof)
ttp$log_zkon[ttp$log_zkon == -Inf] <- NA
ttp$log_cof[ttp$log_cof == -Inf] <- NA
fit.cox2 <- cox.aalen(surv ~ prop(k.cof) + prop(k.ryg) + prop(m.ryg) + prop(log_zkon),
                      weighted.test = 0,
                      residuals = 1,
                      data = ttp)
summary(fit.cox2)
```

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

```
## Test not computed, sim=0
##
## Proportional Cox terms :
##                  Coef.    SE Robust SE D2log(L)^-1      z P-val lower2.5%
## prop(k.cof)      0.000 0.000     0.000       0.000 -1.200 0.229     0.000
## prop(k.ryg)     -0.122 0.169     0.181       0.174 -0.676 0.499    -0.453
## prop(m.ryg)     -0.097 0.166     0.166       0.167 -0.581 0.561    -0.422
## prop(log_zkon)   0.280 0.059     0.057       0.066  4.900 0.000     0.164
##                 upper97.5%
## prop(k.cof)          0.000
## prop(k.ryg)          0.209
## prop(m.ryg)          0.228
## prop(log_zkon)       0.396
## Test of Proportionality
##                 sup|  hat U(t) | p-value H_0
## prop(k.cof)             4950.00        0.012
## prop(k.ryg)                5.37        0.388
## prop(m.ryg)                4.72        0.586
## prop(log_zkon)             8.24        0.732
```

```
resid2 <- cum.residuals(fit.cox2, data = ttp, cum.resid = 1)
summary(resid2)
```

```
## Test for cumulative MG-residuals
##
## Grouped cumulative residuals not computed, you must provide
## modelmatrix to get these (see help)
##
## Residual versus covariates consistent with model
##
##                 sup|  hat B(t) | p-value H_0: B(t)=0
## prop(k.cof)               9.797               0.324
## prop(log_zkon)           13.556               0.058
```

Above we see that the p-value for the m.zkon0 process on time-scale increased dramatically. The transformation did however not have the biggest effect on covariate scale, where it did increase the p-value, but only to a borderline significant level.

To check for interaction, we first look at which interactions to include. We might think that there would be a correlation between the sperm concentration and smoking for men. We look at the empirical correlations between the variables in the model:

```
cor(ttp[, c("k.cof", "m.zkon0", "m.ryg", "k.ryg")], method = "spearman", use = "pairwise.complete.obs")
```

```
##              k.cof     m.zkon0       m.ryg       k.ryg
## k.cof   1.00000000  0.02405813  0.14781484  0.29145700
## m.zkon0 0.02405813  1.00000000 -0.01363916 -0.04306797
## m.ryg   0.14781484 -0.01363916  1.00000000  0.39609540
## k.ryg   0.29145700 -0.04306797  0.39609540  1.00000000
```

We see that skmoking for males and females are correlated, and that smoking for females and their caffein intake seems to be correlated. Below we test for intacertion between smoking for males and females.

```
fit.cox3 <- cox.aalen(surv ~ prop(k.cof) + prop(k.ryg) * prop(m.ryg) + prop(log_zkon),
                      weighted.test = 1,
                      data = ttp)

summary(fit.cox3)
```

```
## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                           Coef.    SE Robust SE D2log(L)^-1      z P-val
## prop(k.cof)               0.000 0.000     0.000       0.000 -1.110 0.266
## prop(k.ryg)              -0.166 0.235     0.249       0.233 -0.667 0.505
## prop(m.ryg)              -0.134 0.216     0.207       0.213 -0.647 0.517
## prop(log_zkon)            0.281 0.059     0.057       0.066  4.900 0.000
## prop(k.ryg):prop(m.ryg)   0.101 0.364     0.371       0.355  0.274 0.784
##                           lower2.5% upper97.5%
## prop(k.cof)                   0.000      0.000
## prop(k.ryg)                  -0.627      0.295
## prop(m.ryg)                  -0.557      0.289
## prop(log_zkon)                0.165      0.397
## prop(k.ryg):prop(m.ryg)      -0.612      0.814
## Test of Proportionality
##                           sup|  hat U(t) | p-value H_0
## prop(k.cof)                            3.21        0.018
## prop(k.ryg)                            1.82        0.652
## prop(m.ryg)                            2.20        0.328
## prop(log_zkon)                         1.63        0.798
## prop(k.ryg):prop(m.ryg)                1.24        0.944
```

We see that the interaction is insignificant, and we will therefore not include it in the model. Below we test for an interaction between the caffein intake of females and smoking.

```
fit.cox4 <- cox.aalen(surv ~ prop(k.cof) * prop(k.ryg) + prop(m.ryg) + prop(log_zkon),
                      weighted.test = 1,
                      data = ttp)

summary(fit.cox4)
```

```
## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                           Coef.    SE Robust SE D2log(L)^-1      z P-val
## prop(k.cof)              -0.001 0.000     0.000       0.000 -1.850 0.064
## prop(k.ryg)              -0.528 0.332     0.334       0.310 -1.580 0.114
## prop(m.ryg)              -0.044 0.173     0.172       0.169 -0.254 0.800
## prop(log_zkon)            0.286 0.059     0.057       0.066  4.990 0.000
```

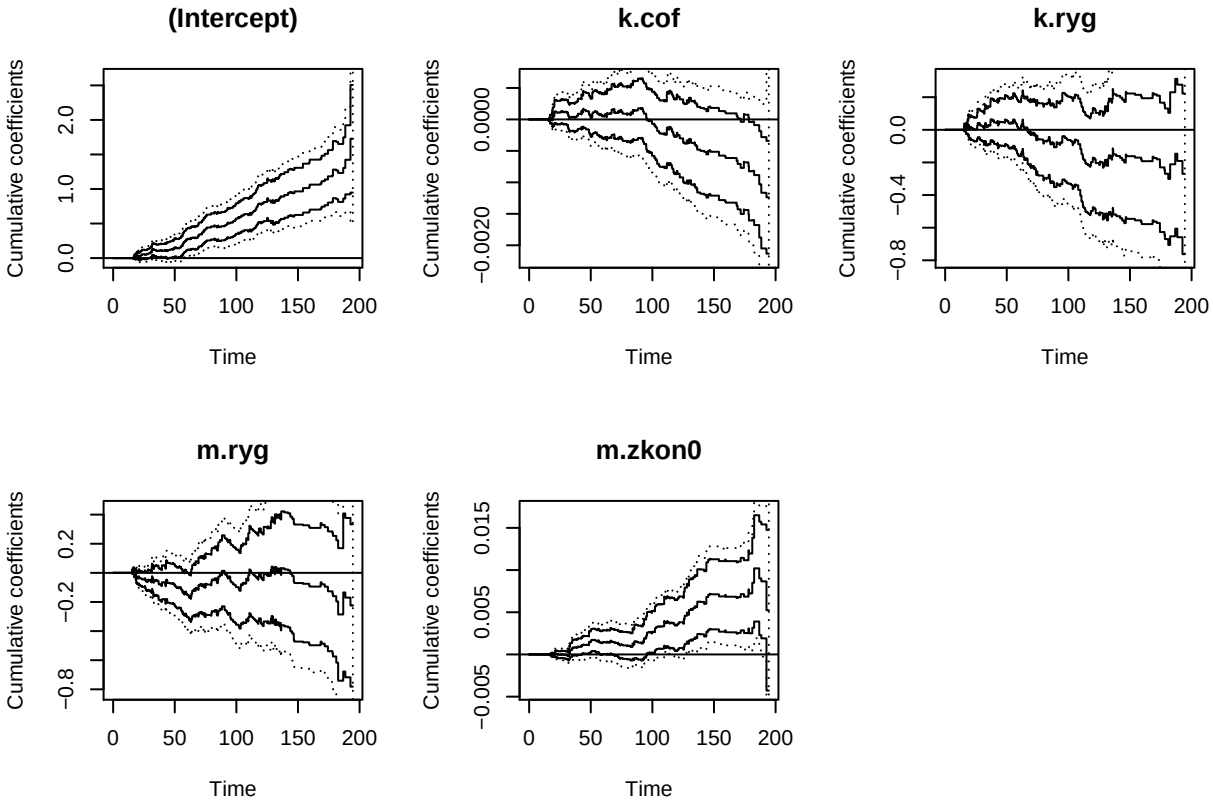
```
## prop(k.cof):prop(k.ryg)  0.001 0.001     0.001       0.001  1.480 0.139
##                          lower2.5% upper97.5%
## prop(k.cof)                 -0.001     -0.001
## prop(k.ryg)                 -1.180      0.123
## prop(m.ryg)                 -0.383      0.295
## prop(log_zkon)               0.170      0.402
## prop(k.cof):prop(k.ryg)     -0.001      0.003
## Test of Proportionality
##                          sup|  hat U(t) | p-value H_0
## prop(k.cof)                          3.24        0.028
## prop(k.ryg)                          1.84        0.592
## prop(m.ryg)                          2.13        0.366
## prop(log_zkon)                       1.66        0.758
## prop(k.cof):prop(k.ryg)              2.46        0.188
```

This interaction also seems insignificant, and will therefore not be included in the model either.

2.2

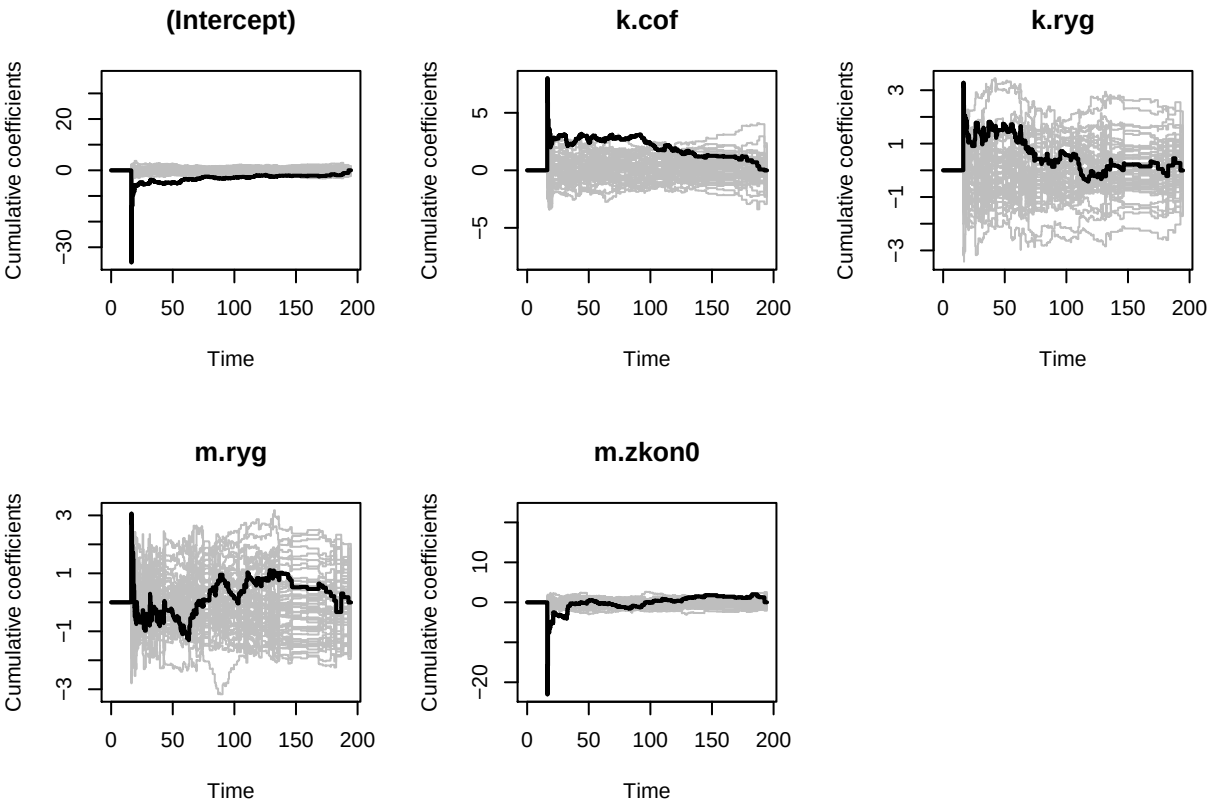
We fit an additive hazards model with the four above mentioned variables and plot the cumulative coefficients, to see if any of the variables can be set constant.

```
fit.add <- aalen(surv ~
                   k.cof + k.ryg + m.ryg + m.zkon0,
                 data = ttp,
                 weighted.test = 1,
                 residuals = 1)

par(mfrow = c(2,3))
plot(fit.add, sim.ci = 3)
```



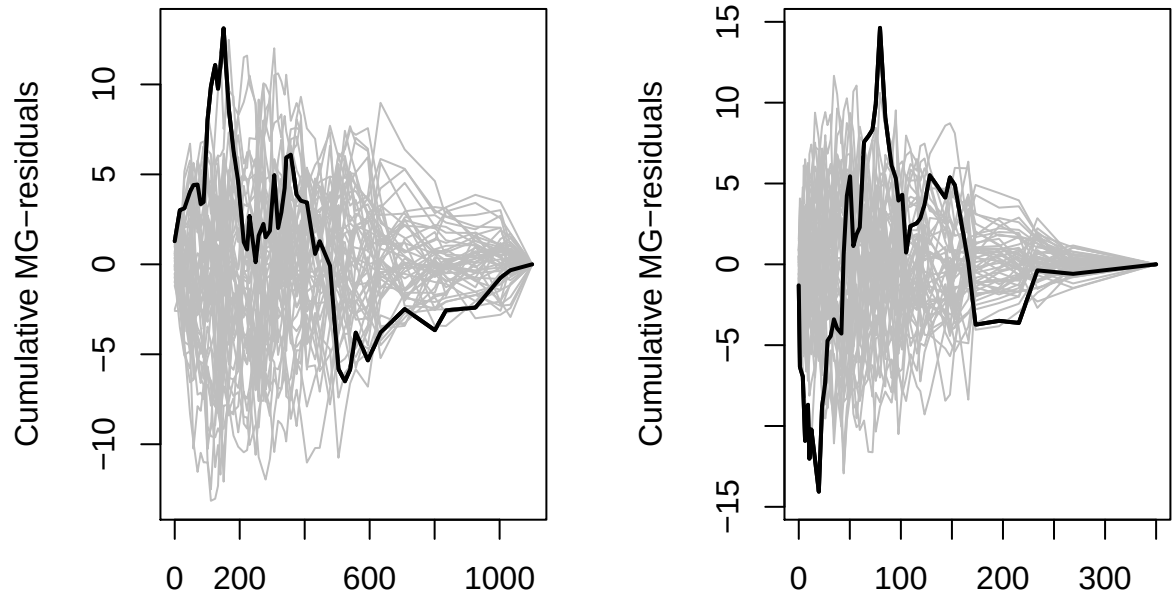
We see that it seems reasonable to assume the effects of smoking to be constant. To check the model fit, we look at the score test proccesses as in 2.1. Below we see that caffein and sperm concentration do not seem to fit the model.

```
par(mfrow = c(2,3))
plot(fit.add, score = T)
```



As in 2.1 we can also look at the cumulative residuals across the covarites. Below we see the residuals plottet across the covariates, and we see that sperm concentration seem to fit the model poorly (as we also stated in 2.1).

```
par(mfrow = c(1,2))
resid3 <- cum.residuals(fit.add, data = ttp, cum.resid = 1)
plot(resid3, score = 2)
```

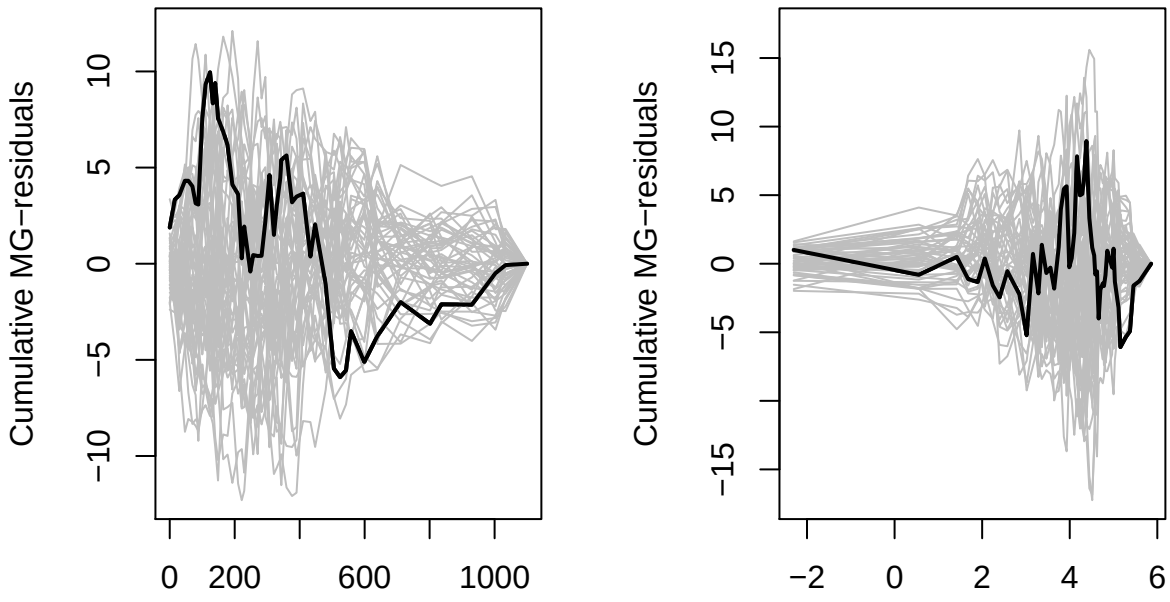


```
summary(resid3)
```

```
## Test for cumulative MG-residuals
##
## Grouped cumulative residuals not computed, you must provide
## modelmatrix to get these (see help)
##
## Residual versus covariates consistent with model
##
##  sup|  hat B(t) | p-value H_0: B(t)=0
##             13.114               0.060
##             14.626               0.014
```

As in 2.1 we will try a log-transformation of the sperm concentration to circumvent this issue. Below we fit the model with sperm concentration log-transformed.

```
fit.add2 <- aalen(surv ~
                    k.cof + k.ryg + m.ryg + log_zkon,
                  data = ttp,
                  weighted.test = 1,
                  residuals = 1)

resid4 <- cum.residuals(fit.add2, data = ttp, cum.resid = 1)
par(mfrow = c(1,2))
plot(resid4, score = 2)
```




```
summary(resid4)
```

```
## Test for cumulative MG-residuals
##
## Grouped cumulative residuals not computed, you must provide
## modelmatrix to get these (see help)
```

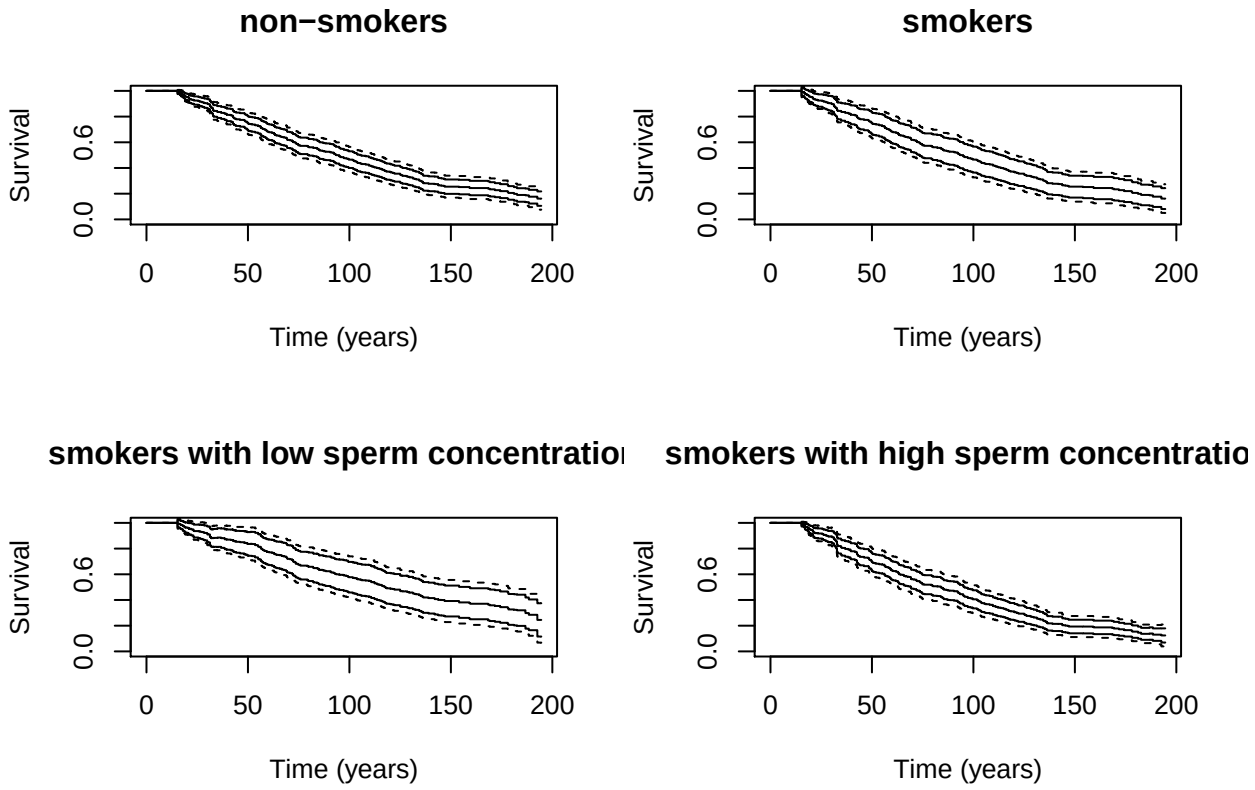
```
##
## Residual versus covariates consistent with model
##
##  sup|  hat B(t) | p-value H_0: B(t)=0
##             9.954               0.254
##             8.937               0.490
```

Now we a much better fit, with a p-value of 0.5 for sperm concentration, and hence we will state the model with a log-transformed sperm concentration. To Summarize the effects of the model, we would estimate the survival function for different choices of covariates, and plot them against each other.

Below we see the estimated survival functions for the median of caffein intake and sperm concentration, but across smokers and non-smokers. We also see the effects of low and high sperm concentration. It is hard to see a big difference between smokers and non-smokers, but fertility for couples with high sperm concentration seems to be better compared to those with low sperm concentration.

```
fit <- aalen(surv ~
               k.cof + const(k.ryg) + const(m.ryg) + log_zkon,
             data = ttp,
             weighted.test = 1,
             residuals = 1,
             resample.iid = 1)


survEst <- function(x0, z0, title){
  delta <- matrix(0, 223, 308)
  for (i in 1:308){
    delta[, i] <- x0 %*% t(fit$B.iid[[i]]) +
    fit$cum[, 1] * sum(z0 * fit$gamma.iid[i, ])
  }
  S0 <- exp(- x0 %*% t(fit$cum[,-1]) - fit$cum[, 1] * sum(z0 * fit$gamma.iid[i, ]))
  se <- apply(delta^2, 1, sum)^.5

  plot(fit$cum[, 1], S0, type="s", ylim=c(0, 1), xlab="Time (years)",
       ylab = "Survival", main =     title)
  lines(fit$cum[, 1], S0 - 1.96 * S0 * se, type = "s")
  lines(fit$cum[,1],S0+1.96*S0*se,type="s")

  mpt <- c()
  for (i in 1:308) {
   g<-rnorm(308)
   pt<-abs(delta %*% g)/se
   mpt<-c(mpt,max(pt[-1]));
  }
  Cband <- percen(mpt, 0.95)
  lines(fit$cum[, 1], S0 - Cband * S0 * se, lty = 2, type = "s")
  lines(fit$cum[, 1], S0 + Cband * S0 * se, lty = 2, type = "s")
}

par(mfrow = c(2,2))
survEst(c(1, median(ttp$k.cof), median(na.omit(ttp$log_zkon))), c(0,0), "non-smokers")
survEst(c(1, median(ttp$k.cof), median(na.omit(ttp$log_zkon))), c(1,1), "smokers")
survEst(c(1, median(ttp$k.cof), quantile(na.omit(ttp$log_zkon), 0.1)),
        c(1,1), "smokers with low sperm concentration")
```

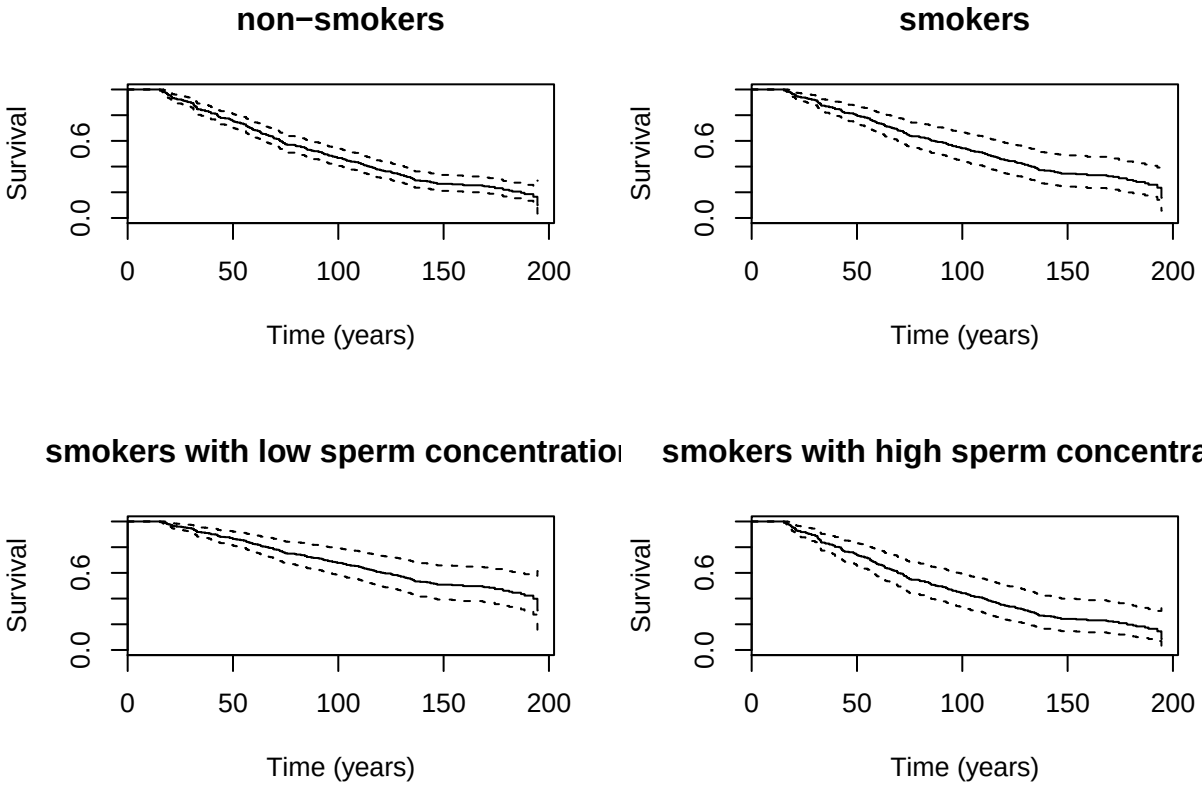
```
survEst(c(1, median(ttp$k.cof), quantile(na.omit(ttp$log_zkon), 0.9)),
        c(0,0), "smokers with high sperm concentration")
```




```
x0 <- c(1, median(ttp$k.cof), median(na.omit(ttp$log_zkon)))
z0 <- c(0, 0)
```

To compare with the results in 2.1 we estimate the survival function with the cox model stated in 2.1 with the same covariate levels as above. Below we see the estimated survival functions from the cox model, where the effect of smoking seems to be a bit bigger, but again it is the level of sperm concentration that drastically determines the fertility.

```
test.cox <- coxph(surv ~ k.cof + k.ryg + m.ryg + log_zkon, data = ttp)
covs1 <- data.frame(k.cof = median(ttp$k.cof), k.ryg = 0, m.ryg = 0, log_zkon = median(na.omit(ttp$log_z
covs2 <- data.frame(k.cof = median(ttp$k.cof), k.ryg = 1, m.ryg = 1, log_zkon = median(na.omit(ttp$log_z
covs3 <- data.frame(k.cof = median(ttp$k.cof), k.ryg = 1, m.ryg = 1, log_zkon = quantile(na.omit(ttp$log
covs4 <- data.frame(k.cof = median(ttp$k.cof), k.ryg = 1, m.ryg = 1, log_zkon = quantile(na.omit(ttp$log

est1 <- survfit(test.cox, newdata = covs1, type = "aalen")
est2 <- survfit(test.cox, newdata = covs2, type = "aalen")
est3 <- survfit(test.cox, newdata = covs3, type = "aalen")
est4 <- survfit(test.cox, newdata = covs4, type = "aalen")

par(mfrow = c(2,2))
plot(est1, xlab="Time (years)", ylab = "Survival", main = "non-smokers")
plot(est2, xlab="Time (years)", ylab = "Survival", main = "smokers")
plot(est3, xlab="Time (years)", ylab = "Survival", main = "smokers with low sperm concentration")
plot(est4, xlab="Time (years)", ylab = "Survival", main = "smokers with high sperm concentration")
```



2.3

The findings of the models derived above is summarized in 2.2. It it hard to determine, if one of the models is superior to other in terms of prediction, but the additive hazards model does not require as strict assumptions as the cox model (proportionality), and we obtained a reaonable model fit for the additive model, which was not the case for the cox model. The additive model would thus be the preferred one in this case.