Survival Analysis exam

Examination number: 2

To be handed in January 20, 2017

1 Theoretical Part

1.1

a

To do this we use that we know the intensity for females (and males as well) which is given by

$$\lambda_i^f(t) = Y_i^f(t)\left(X_i^f(t)^T\beta(t-4) + h(\gamma^T Z_i^f(t))\right)$$

Where everything is like described in the exercise. From this we get that the females hazard model evaluated in $T_i^f + 4$ is given by

$$\begin{aligned}\alpha_i^f(t, X_i^f(t), Z_i^f(t)) &= X_i^f(T_i^f+4)^T\beta(T_i^f+4-4) + h(\gamma^T Z_i^f(T_i^f+4)) \\ &= X_i^f(T_i^f+4)^T\beta(T_i^f) + h(\gamma^T Z_i^f(T_i^f+4))\end{aligned}$$

So we observe that the β is no longer evaluated different men and women. We do, however, now have to "look forward" in the covariates, X and Z.

b

The model is a semiparametric additive hazard model, we therefore find inspiration in chapter 5.3 from the book about these models. We start by defining

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

Then we can write

$$\begin{aligned}
\mathrm{d}N(t) &= \lambda(t)\mathrm{d}t + \mathrm{d}M(t) \\
&= (X(t)\beta(t) + \mathrm{diag}(Y(t))h(Z(t)\gamma))\mathrm{d}t + \mathrm{d}M(t) \\
&= X(t)\mathrm{d}B(t) + \mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t + \mathrm{d}M(t)
\end{aligned}$$

Since we have that the martingale increments are independent with zero-mean, we can estimate $\mathrm{d}B(t)$ and γ from the least square equations. These are derived by taking derivatives of $(\mathrm{d}N(t) - \lambda\mathrm{d}t)^{\otimes 2}$ w.r.t. the two parameters and setting equal to 0. By doing this we obtain that

$$\begin{aligned}
\frac{\partial(\mathrm{d}N(t) - \lambda(t)\mathrm{d}t)^{\otimes 2}}{\partial \mathrm{d}B(t)} &= -2X(t)^T(\mathrm{d}N(t) - \lambda(t)\mathrm{d}t) = 0 \\
&\Leftrightarrow X(t)^T(\mathrm{d}N(t) - \lambda(t)\mathrm{d}t) = 0
\end{aligned}$$

Which is the estimating equation for $\mathrm{d}B(t)$ given a fixed γ. Since we need $\mathrm{d}\hat{B}(t)$ for the next sub question we will isolate for this here.

$$\begin{aligned}
0 = X(t)^T(\mathrm{d}N(t) - \lambda(t)\mathrm{d}t) \Leftrightarrow \\
X(t)^T\mathrm{d}N(t) = X(t)^T\left(X(t)\mathrm{d}B(t) + \mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t\right) \Leftrightarrow \\
X(t)^TX(t)\mathrm{d}B(t) = X(t)^T\left(\mathrm{d}N(t) - \mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t\right)
\end{aligned}$$

Assuming that the inverse to $X(t)^TX(t)$ exists we get that the estimator is

$$\mathrm{d}\hat{B}(t) = X^-(t)\left(\mathrm{d}N(t) - \mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t\right)$$

Where $X^-(t) = \left(X(t)^TX(t)\right)^{-1}X(t)^T$. This is very close to equation (5.31) from the book, beside that we have the function h taken on $Z(t)\gamma$. However if h is the identity function we see that we get the same result.

c

Taking the derivative with respect to γ yields

$$\begin{aligned}
\frac{\partial(\mathrm{d}N(t) - \lambda(t)\mathrm{d}t)^{\otimes 2}}{\partial \gamma} &= -2Z(t)^T\mathrm{D}h(Z(t)\gamma)^T\mathrm{diag}(Y(t))\left(\mathrm{d}N(t) - \lambda(t)\mathrm{d}t\right) = 0 \\
&\Leftrightarrow Z(t)^T\mathrm{D}h(Z(t)\gamma)^T\mathrm{diag}(Y(t))\left(\mathrm{d}N(t) - \lambda(t)\mathrm{d}t\right) = 0
\end{aligned}$$

If we now define $\hat{\lambda}(t)\mathrm{d}t = X(t)\mathrm{d}\hat{B}(t) + \mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t$ we can profile out $\mathrm{d}B(t)$ by substituting $\lambda(t)$ by $\hat{\lambda}(t)$. We get that

$$Z(t)^T\mathrm{D}h(Z(t)\gamma)^T\mathrm{diag}(Y(t))\left(\mathrm{d}N(t) - \hat{\lambda}(t)\mathrm{d}t\right) = 0$$

Which is the estimating equation for γ. I will not try isolating for γ, since γ enters the above equation two places. Through the function h and through the differential of the function h, $\mathrm{D}h$. If we reduce this, we get

$$
\begin{aligned}
0 &= Z(t)^T \mathrm{D}h(Z(t)\gamma)^T \mathrm{diag}(Y(t)) \left(\mathrm{d}N(t) - \hat{\lambda}(t)\mathrm{d}t\right) \Leftrightarrow \\
0 &= Z(t)^T \mathrm{D}h(Z(t)\gamma)^T \mathrm{diag}(Y(t)) \left(\mathrm{d}N(t) - X(t)\mathrm{d}\hat{B}(t) - \mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t\right) \Leftrightarrow \\
0 &= Z(t)^T \mathrm{D}h(Z(t)\gamma)^T \mathrm{diag}(Y(t)) \left(\mathrm{d}N(t) - X(t)X^-(t) \left(\mathrm{d}N(t)\right.\right. \\
&\quad \left.\left. -\mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t\right)\right) - \mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t) \Leftrightarrow \\
0 &= Z(t)^T \mathrm{D}h(Z(t)\gamma)^T \mathrm{diag}(Y(t)) \left(H(t)\mathrm{d}N(t) - H(t)\mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t\right) \Leftrightarrow \\
0 &= Z(t)^T \mathrm{D}h(Z(t)\gamma)^T \mathrm{diag}(Y(t))H(t) \left(\mathrm{d}N(t) - \mathrm{diag}(Y(t))h(Z(t)\gamma)\mathrm{d}t\right)
\end{aligned}
$$

Where $H(t) = I - X(t)X^-(t)$. Here it is obvious that this equation only depends on γ, and we can therefore estimate γ from this. Furthermore we observe that if h is the identity function we get exactly the same result as in the book.

d

To estimate γ one should use the result from question c. From this equation it is possible to find a estimate, $\hat{\gamma}$, by solving the equation.

To establish that $\sqrt{n}(\hat{\gamma} - \gamma)$ is asymptotically normal under regularity conditions, we would like to do as in theorem 5.3.1 from the book. This theorem says that the result follows if condition 5.2 holds. Therefore we assume that this condition must hold. To follow the proof of this theorem, we need to have that the function h is so nice that we can make a decomposition that is similar to the one in the book. If we have this, we have something that we would like to show converges in probability, let's call this C, and a martingale, $\tilde{M}$, which we can use the martingale CLT on. So we have that $C(t) \cdot \tilde{M}(t)$. We will have the convergence in probability from condition 5.2, s.t. $C \xrightarrow{P} c$. The convergence of $\tilde{M}$ comes from theorem 2.5.1. In this theorem we must have that

$$
\langle \tilde{M} \rangle(t) \xrightarrow{P} V(t)
$$

Assuming this we get that $\tilde{M}(t) \xrightarrow{\mathcal{D}}$. By applying Slutsky's Lemma we will that $\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{\mathcal{D}} \mathcal{N}(0, C^{-1}\Sigma C^{-1})$.

To estimate the variance, one could use the optional variation processes to get the following estimator for Σ

$$
\hat{\Sigma} = C^{-1}[\tilde{M}]C^{-1}
$$

d

To estimate $\mathrm{d}B(t)$ you should use the estimator found in question b together with the estimator for γ. So the estimator becomes

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

In the previous d question we couldn't make the decomposition due to the h function. Here we are a little more fortunate so we can make the following decomposition

$$
\begin{aligned}
\sqrt{n}(\hat{B}(t) - B(t)) &= \sqrt{n}\left(\int_0^t X^-(s)\left(\mathrm{d}N(s) - \mathrm{diag}(Y(s))h(Z(s)\hat{\gamma})\mathrm{d}s\right) - B(t)\right) \\
&= \sqrt{n}\left(\int_0^t X^-(s)\mathrm{d}(\lambda(s) + M(s)) - \int_0^t X^-(s)\mathrm{diag}(Y(s))h(Z(s)\hat{\gamma})\mathrm{d}s - B(t)\right) \\
&= \sqrt{n}\left(\int_0^t X^-(s)\mathrm{d}M(s) + \int_0^t X^-(s)\lambda(s)\mathrm{d}s - \int_0^t X^-(s)\mathrm{diag}(Y(s))h(Z(s)\hat{\gamma})\mathrm{d}s\right. \\
&\quad \left. -B(t)\right) \\
&= \sqrt{n}\left(\int_0^t X^-(s)\mathrm{d}M(s) + \int_0^t X^-(s)\left(X(s)\mathrm{d}B(s) + \mathrm{diag}(Y(s)h(Z(s)\gamma)\mathrm{d}s\right)\right. \\
&\quad \left. - \int_0^t X^-(s)\mathrm{diag}(Y(s))h(Z(s)\hat{\gamma})\mathrm{d}s - B(t)\right) \\
&= \sqrt{n}\left(\int_0^t X^-(s)\mathrm{d}M(s) + \int_0^t \left(X^-(s)\mathrm{diag}(Y(s))(h(Z(s)\gamma) - h(Z(s)\hat{\gamma}))\right)\mathrm{d}s\right. \\
&\quad \left. + \underbrace{\int_0^t X^-(s)X(s)\mathrm{d}B(s)}_{=B(t)} - B(t)\right) \\
&= \sqrt{n}\left(\int_0^t X^-(s)\mathrm{d}M(s) + \int_0^t \left(X^-(s)\mathrm{diag}(Y(s))(h(Z(s)\gamma) - h(Z(s)\hat{\gamma}))\right)\mathrm{d}s\right)
\end{aligned}
$$

So we have something that is a martingale, and we can use the martingale CLT like we did in the question before. Furthermore we have a term that we also would like to show converges in distribution. It is however pretty difficult to say anything about this when we don't know anything about h, except that it is sufficiently nice. What we would hopefully end up with, is something that converges in distribution to a Gaussian process.

To estimate the variance one could again use the optional variation processes. It is however a little difficult to say more about how this optional variation estimator will look like when we don't know more about the behavior of $\sqrt{n}(\hat{B}(t) - B(t))$

d

Since we transform the female intensity to look like the male, we shift the problem for the women to look like men. This is done, such that we can estimate $B(t)$ across gender. Therefore we get the range for t to be $[26; 90]$ when rebasing the female intensity to the male for the β parameter. A little more in dept we could look at it as, when we get data for a 30 year old female we can estimate. But since we have rebased it to the male "timeline" we can estimate 26 year old persons.

To sum up we can estimate $B(t)$ for $t \in [26; 90]$ based on data from 30 to 90, given how we define the model.

e

Given covariates (X_0, Z_0) for the man, we have that the survival probability is given by

$$
S_0(t) = \exp(-X_0^T B(t) - h(Z_0^T\gamma)t)
$$

for him. This can be estimated by

$$\hat{S}_0(t) = \exp(-X_0^T \hat{B}(t) - h(Z_0^T \hat{\gamma})t)$$

We therefore get that the predicted survival probability after 10 years that entered at 30 years of age is given by

$$\hat{S}_0(10) = \exp(-X_0^T \hat{B}(10) - h(Z_0^T \hat{\gamma})10)$$

This will give you the probability that he has survived 10 years from when he entered the study.

To get standard errors for this estimator, you would look at $\sqrt{n}(\hat{S}_0 - S_0)$'s distribution, which can be found from what we found in the above questions. This limits distribution will have a variance that we would like to estimate. This can again be done bu the optional variation processes. This will, however, depend heavily on the latter results, so we will not pursue this any further, except noting that, if we denote the optional variation estimate by $\hat{\Psi}(t)$, we can estimate the variance by

$$\hat{Q}(t) = \hat{S}_0(t)\hat{\Psi}(t)$$

Taking square root of this yields the desired standard errors.

f

Since the females now enter the study at a unknown time the data will be left-truncated. The way to handle this is to update the at risk function for the females such that they are at risk when they enter the study. Let's denote the time the individual female enters the study for V_i^f, so we get that the at risk indicator becomes $Y_i^f(t) = I(V_i^f < t < T_i^f)$. To get consistent estimators we assume independent left-truncation such that we an independent filtering process, which leads to consistent estimates of $B(t)$ and γ.

1.2

a

Here we would like to make use of the relationship between the survival function and the hazard. I.e. we can calculate the hazard if we know the survival function by

$$\alpha(t) = \frac{\partial}{\partial t}\left(-\log(S(t))\right)$$

So we would like to calculate the survival function given X. That is

$$\begin{aligned} P(t < T^\star | X = 0) &= \int_0^\infty P(t < T^\star | X = 1, A) \mathrm{d}A(P) \\ &= \frac{1}{2}\int_0^\infty \exp\left(-\int_0^t \beta(s) + 0 \cdot \alpha(s+a)\mathrm{d}s\right) f(a)\mathrm{d}a \\ &= \frac{1}{2}\exp\left(-\int_0^t \beta(s))\mathrm{d}s\right) \end{aligned}$$

Taking minus log yields

$$-\log(P(t < T^\star | X = 0)) = \int_0^t \beta(s))\mathrm{d}s + \log(2)$$

Differentiating w.r.t. t yields $\beta(t)$. For $X = 1$ we get

$$\begin{aligned} P(t < T^\star | X = 1) &= \int_0^\infty P(t < T^\star | X = 1, A)\mathrm{d}A(P) \\ &= \frac{1}{2}\int_0^\infty \exp\left(-\int_0^t \beta(s) + 1 \cdot \alpha(s + a)\mathrm{d}s\right) f(a)\mathrm{d}a \\ &= \frac{1}{2}\exp\left(-\int_0^t \beta(s)\mathrm{d}s\right) \underbrace{\int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\mathrm{d}s\right) f(a)\mathrm{d}a}_{=\Theta(t)} \end{aligned}$$

Taking minus log yields

$$-\log(P(t < T^\star | X = 1)) = \int_0^t \beta(s))\mathrm{d}s - \log\left(\Theta(t)\right) + \log(2)$$

Differentiating w.r.t. t yields

$$\frac{\partial}{\partial t}\left(-\log(P(t < T^\star | X = 1)\right) = \beta(t) - \frac{1}{\Theta(t)}\Theta'(t)$$

Where we have that

$$\Theta'(t) = -\int_0^\infty \alpha(t + a)\exp\left(-\int_0^t \alpha(s + a)\mathrm{d}s\right) f(a)\mathrm{d}a$$

So we have that the hazard is given by

$$\lambda(t|X) = \begin{cases} \beta(t) & X = 0 \\ \beta(t) - \frac{1}{\Theta(t)}\Theta'(t), & X = 1 \end{cases}$$

b

Here we will use the same trick as in question a

$$\begin{aligned} S(t) = P(t < T^\star) &= \int_{\{0,1\}\times(0;\infty)} P(t < T^\star | X, A)\mathrm{d}A(P) \otimes X(P) \\ &= \frac{1}{2}\int_0^\infty \exp\left(\int_0^t \beta(s)\mathrm{d}s\right) f(a)\mathrm{d}a + \frac{1}{2}\int_0^\infty \exp\left(-\int_0^t \beta(s) + \alpha(s + a)\mathrm{d}s\right) f(a)\mathrm{d}a \\ &= \frac{1}{2}\exp\left(\int_0^t \beta(s)\mathrm{d}s\right)\left(1 + \underbrace{\int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\mathrm{d}s\right) f(a)\mathrm{d}a}_{=\Theta(t)}\right) \end{aligned}$$

Taking minus log yield

$$-\log(S(t)) = \int_0^t \beta(s)\mathrm{d}s + \log(2) - \log(1+\Theta(t))$$

Differentiating w.r.t. t gives the following

$$\frac{\partial}{\partial t}\left(-\log(S(t)) = \beta(t) - \frac{1}{1+\Theta(t)}\Theta'(t)\right.$$

Where $\Theta'(t)$ is given above. This is the hazard when we doesn't observe either of X and A

2 Practical Part

1

We start by loading the data and fitting the model. See below R code for model fit. We furthermore print the summary of the model fit:

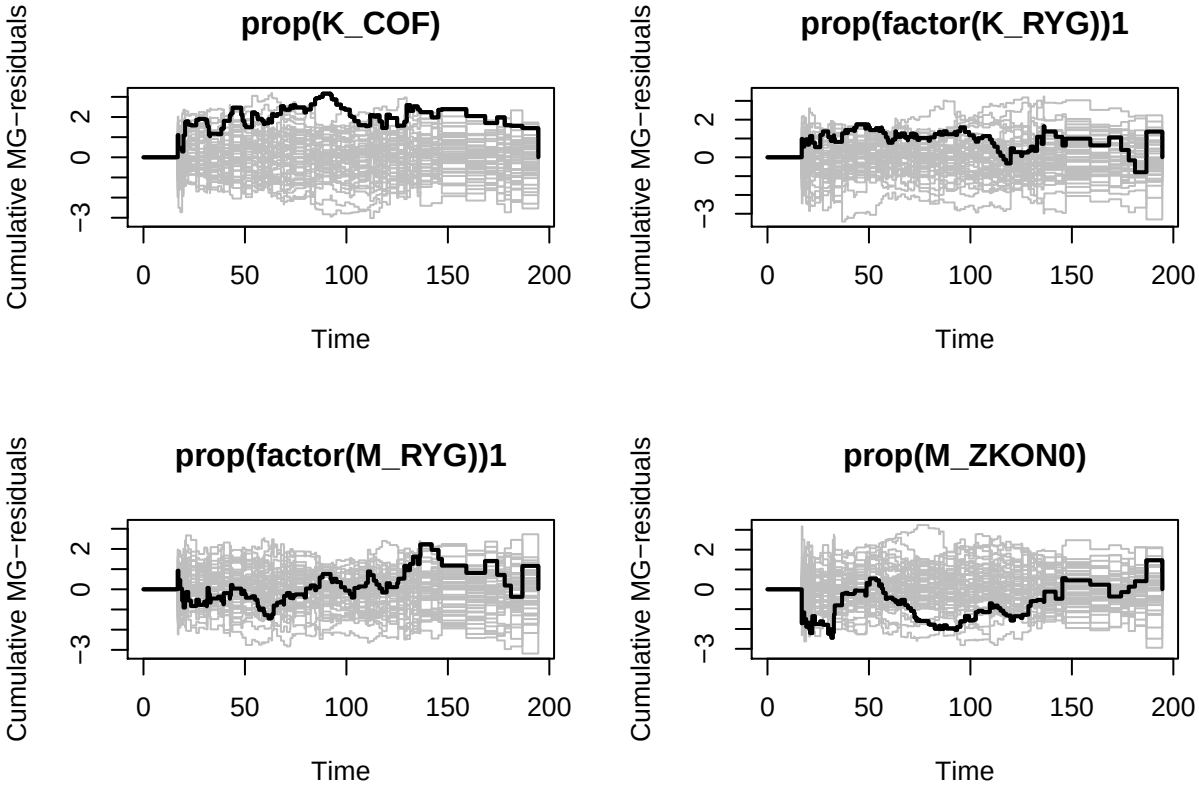
```
fit_cox_aalen <- cox.aalen(Surv(TTP, K_GRAVID) ~ prop(K_COF) + prop(factor(K_RYG))
                           + prop(factor(M_RYG)) + prop(M_ZKON0),
                           data = data, weighted.test = 1, residuals = 1, n.sim = 1000)

summary(fit_cox_aalen)
```

```
## 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 -1.030 0.304
## prop(factor(K_RYG))1 -0.134 0.173     0.180       0.175 -0.745 0.456
## prop(factor(M_RYG))1 -0.110 0.169     0.167       0.168 -0.659 0.510
## prop(M_ZKON0)         0.004 0.001     0.001       0.001  4.520 0.000
##                        lower2.5% upper97.5%
## prop(K_COF)                0.000      0.000
## prop(factor(K_RYG))1      -0.473      0.205
## prop(factor(M_RYG))1      -0.441      0.221
## prop(M_ZKON0)              0.002      0.006
## Test of Proportionality
##                        sup|  hat U(t) | p-value H_0
## prop(K_COF)                         3.16        0.020
## prop(factor(K_RYG))1                1.75        0.660
## prop(factor(M_RYG))1                2.23        0.343
## prop(M_ZKON0)                       2.44        0.246
```

From this we can see that the only covariate that we can say is significantly different from zero is the sperm concentration of the male. We furthermore observe that the estimate is positive for this covariate. This is like we would expect it to be, since a high sperm concentration reduces the time to pregnancy. For the covariates we can't say that they are significantly different from zero on a 95% confidence. Looking closer to the estimates of the two smoking parameters, we see that they have a negative effect on the time to pregnancy. This is like expected. Furthermore we see that the estimate for womens consumption of caffeine is positive, which might be a little surprising.

Evaluating the score under the null using weighted supremum test-statistic we see that this indicates a bad fit for women caffeine intake. From the other values we can't say much about the fit. If we instead turn our attention towards plots of the score processes we get the following



From This we also see a bad fit for the caffein intake. We do, however, also see that there could be an indication of a lacking fit for sperm concentration in the start of the time-period.

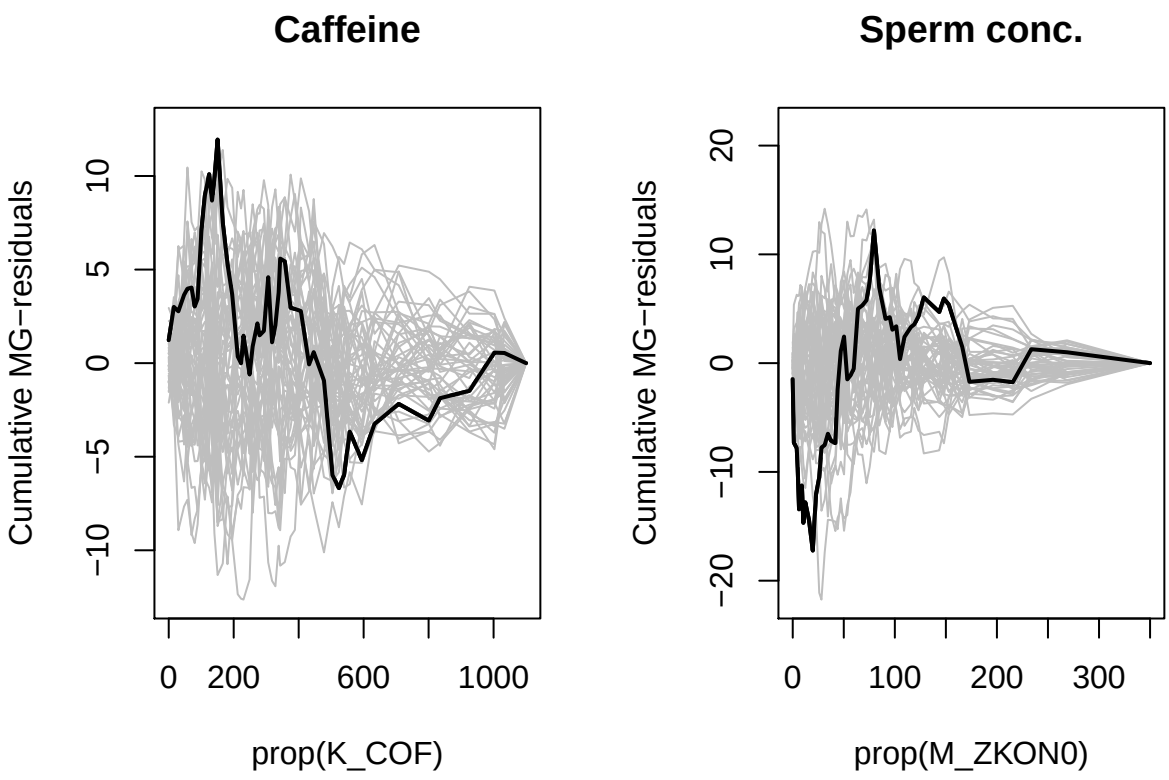
To look closer at this possible lack of fit for sperm concentration we will look at the cummulative residuals. This could indicate possible misspecifications of the functional form of the covariates. Below we run the summary for the cummulative residuals together with plots of these

```
resids <- cum.residuals(fit_cox_aalen, data = data, cum.resid = 1)
summary(resids)
```

```
## 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                0.104
## prop(M_ZKON0)            17.230                0.002
```

```
par(mfrow=c(1,2))
plot(resids, score = 2,  main = c("Caffeine", "Sperm conc."))
```



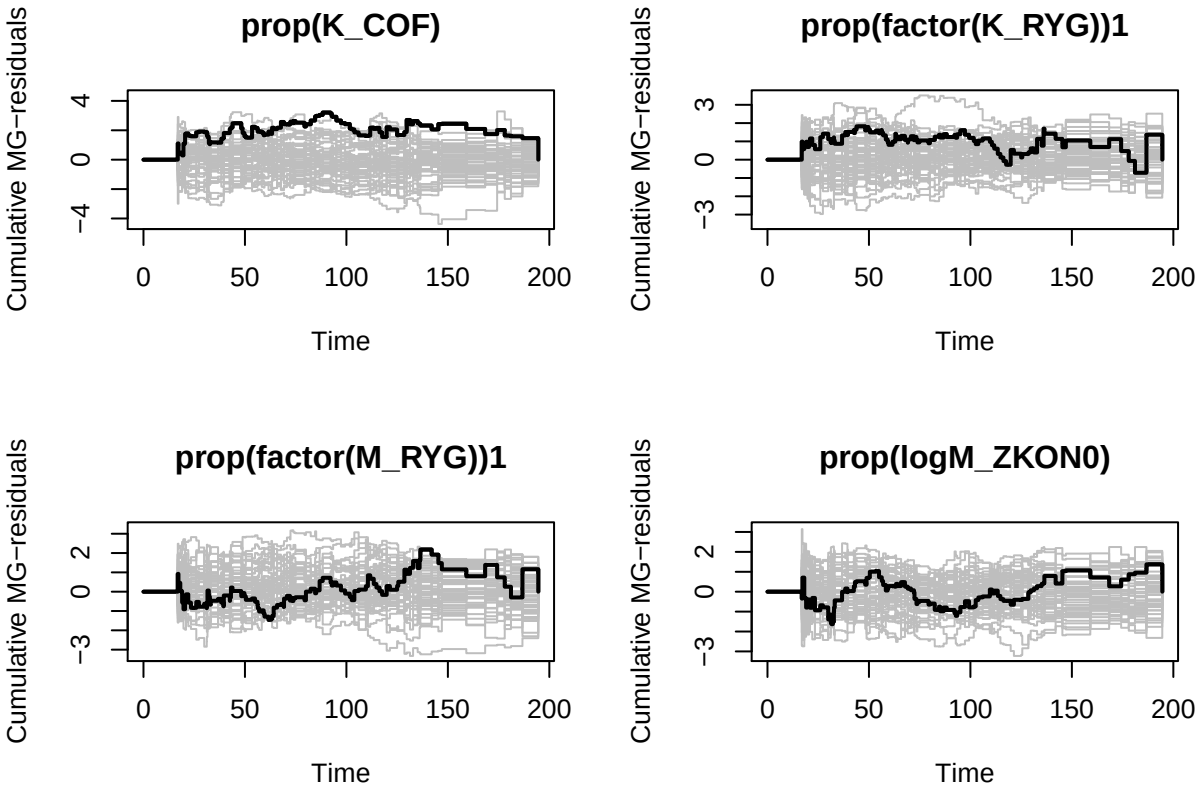
From this we see a strong indication that sperm concentration shouldn't be included in the model on this scale. Let's try to fit the model with sperm concetration log transformed and do the above analysis again

```
fit_cox_aalen <- cox.aalen(Surv(TTP, K_GRAVID) ~ prop(K_COF) + prop(factor(K_RYG))
                           + prop(factor(M_RYG)) + prop(logM_ZKON0),
                           data = data, weighted.test = 1, residuals = 1, n.sim = 1000)
summary(fit_cox_aalen)
```

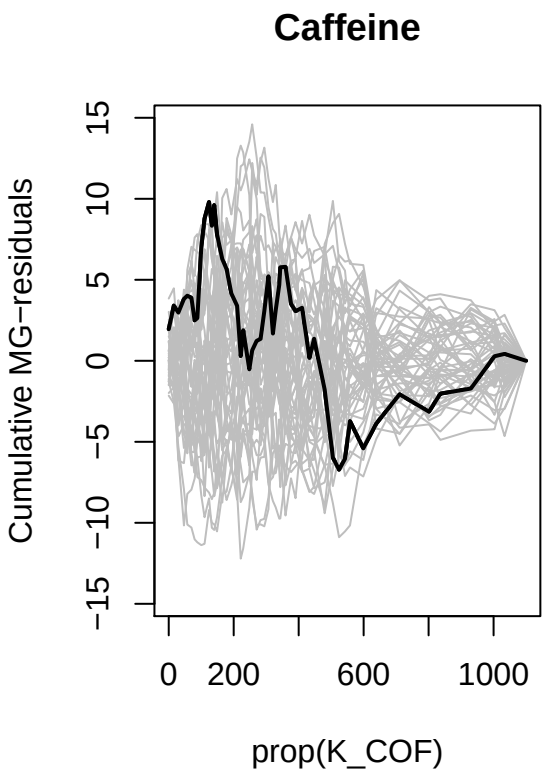
```
## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                         Coef.    SE Robust SE D2log(L)^-1      z P-val
## prop(K_COF)             0.000 0.000     0.000       0.000 -1.200 0.229
## prop(factor(K_RYG))1 -0.122 0.169     0.181       0.174 -0.676 0.499
## prop(factor(M_RYG))1 -0.097 0.166     0.166       0.167 -0.581 0.561
```

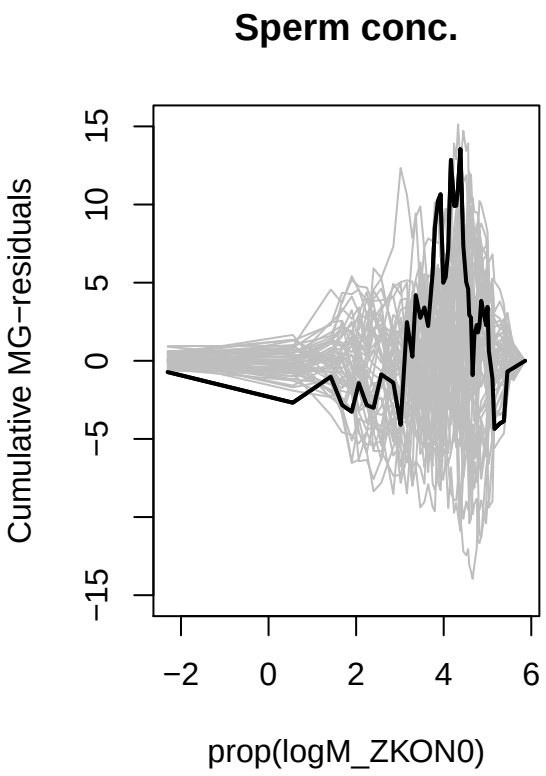
```
## prop(logM_ZKON0)      0.280 0.059     0.057       0.066  4.900 0.000
##                      lower2.5% upper97.5%
## prop(K_COF)              0.000      0.000
## prop(factor(K_RYG))1    -0.453      0.209
## prop(factor(M_RYG))1    -0.422      0.228
## prop(logM_ZKON0)         0.164      0.396
## Test of Proportionality
##                      sup|  hat U(t) | p-value H_0
## prop(K_COF)                      3.21        0.029
## prop(factor(K_RYG))1             1.82        0.603
## prop(factor(M_RYG))1             2.18        0.356
## prop(logM_ZKON0)                 1.63        0.813
```

```
par(mfrow=c(2,2))
plot(fit_cox_aalen, score = T)
```




```
resids <- cum.residuals(fit_cox_aalen, data = data, cum.resid = 1)
par(mfrow=c(1,2))
plot(resids, score = 2,  main = c("Caffeine", "Sperm conc."))
```






```
summary(resids)
```

```
## 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               0.294
## prop(logM_ZKON0)          13.556               0.046
```

From this we see a lot of the same conclusions as before, however the score process plot for sperm concentration looks better in the beginning of the time-period. We also oberve that the cummulative residuals plot has changed, but it is difficult to see if it is better than before. Looking at the summary for the cummulative residuals, we see that the p-value is just around 0.05, so not much better than before. This indicates that the log-transform is probably not the correct tranformation of sperm concentration. We will not pursue this further here.

Interactions

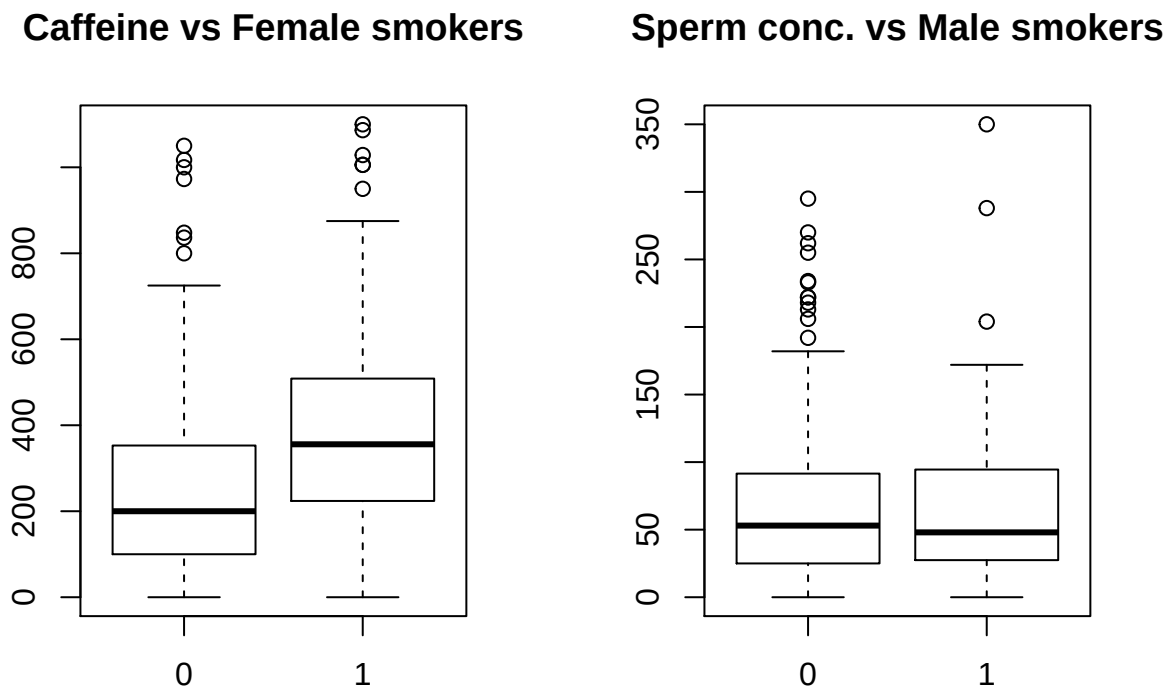
Now we will turn our attention to look for interaction terms. To do this, we will investigate which interaction terms that could be interesting. There are of course some that will not be interesting, e.g. female smokers

will not have any effect on sperm concentration. To figure out how the covariates are correlated we calculate the correlation matrix

```
cor(subset(data, select = c("K_COF",  "K_RYG", "M_RYG", "M_ZKONO")),
    method = "spearman", use = "pairwise.complete.obs")
```

```
##              K_COF       K_RYG       M_RYG     M_ZKONO
## K_COF   1.00000000  0.29145700  0.14781484  0.02405813
## K_RYG   0.29145700  1.00000000  0.39609540 -0.04306797
## M_RYG   0.14781484  0.39609540  1.00000000 -0.01363916
## M_ZKONO 0.02405813 -0.04306797 -0.01363916  1.00000000
```

From this we see that their is correlation between female smoking and male smoking. Beside this, it could look as if there is a correlation between smoking and caffeine intake for women. Maybe a little surprising, we see a small negative correlation between male smoking and the sperm concentration. The surprising thing is not that it is negative, we would definitely expect that, since it is somewhat common knowledge that smoking reduces your fertility. However this could indicate that smoking for men, doesn't lead to low sperm quality. To support these observations we plot the boxplots for female smoking vs caffeine and male smoking vs sperm concentration. This can be seen below



Here we see a indication that female non-smokers could have a lower intake of caffeine than female smokers, while there doesn't seem to be any effect of smoking on sperm concentration. However we observe that it looks like there are more higher values of sperm concentration for non-smokers than for smokers. This could therefore be interesting to investigate further, however we also need to remember that there doesn't seem to be much difference between sperm concentration for smokers and non-smokers.

To conclude the above discussion we will look closer into interaction terms between smokers, i.e. if one part of the couple smokes then this might affect the other part, male smoking and sperm concentration and female

smoking and caffeine intake. Below we fit a model with an interaction term between smokers and get the following summary

```
summary(cox.aalen(Surv(TTP, K_GRAVID) ~ prop(K_COF) + prop(factor(K_RYG))
                + prop(factor(M_RYG)) + prop(M_ZKON0) + prop(factor(K_RYG))*prop(factor(M_RYG)),
                data = data, weighted.test = 1, residuals = 1, n.sim = 1000))
```

```
## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                                          Coef.    SE Robust SE
## prop(K_COF)                              0.000 0.000     0.000
## prop(factor(K_RYG))1                    -0.143 0.238     0.248
## prop(factor(M_RYG))1                    -0.117 0.219     0.208
## prop(M_ZKON0)                            0.004 0.001     0.001
## prop(factor(K_RYG))1:prop(factor(M_RYG))1  0.019 0.367     0.368
##                                         D2log(L)^-1      z P-val
## prop(K_COF)                                   0.000 -0.977 0.328
## prop(factor(K_RYG))1                          0.235 -0.575 0.565
## prop(factor(M_RYG))1                          0.214 -0.564 0.573
## prop(M_ZKON0)                                 0.001  4.510 0.000
## prop(factor(K_RYG))1:prop(factor(M_RYG))1     0.355  0.052 0.958
##                                         lower2.5% upper97.5%
## prop(K_COF)                                 0.000      0.000
## prop(factor(K_RYG))1                       -0.609      0.323
## prop(factor(M_RYG))1                       -0.546      0.312
## prop(M_ZKON0)                               0.002      0.006
## prop(factor(K_RYG))1:prop(factor(M_RYG))1  -0.700      0.738
## Test of Proportionality
##                                         sup|  hat U(t) | p-value H_0
## prop(K_COF)                                        3.17        0.037
## prop(factor(K_RYG))1                               1.75        0.662
## prop(factor(M_RYG))1                               2.23        0.345
## prop(M_ZKON0)                                      2.44        0.229
## prop(factor(K_RYG))1:prop(factor(M_RYG))1          1.23        0.937
```

From this we see that we can't reject the hypothesis of this being zero, since it has a p-value of 0.958. Furthermore we see that the only variable still being significantly different from 0 is sperm concentration. Looking at the supremum test we do however see a high p-value which could indicate a good fit, but we need to remember that when doing the supremum test for one variable it is assumed that the model is correct for the variables, which might not be so true.

Now we will look at interaction between smoking and sperm concentration. The model is fitted below and the summary is run

```
summary(cox.aalen(Surv(TTP, K_GRAVID) ~ prop(K_COF) + prop(factor(K_RYG))
                 + prop(factor(M_RYG)) + prop(M_ZKON0) + prop(M_ZKON0)*prop(factor(M_RYG)),
                 data = data, weighted.test = 1, residuals = 1, n.sim = 1000))
```

```
## Cox-Aalen Model
```

```
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                                      Coef.    SE Robust SE D2log(L)^-1
## prop(K_COF)                          0.000 0.000     0.000       0.000
## prop(factor(K_RYG))1                -0.149 0.176     0.182       0.177
## prop(factor(M_RYG))1                 0.162 0.246     0.239       0.244
## prop(M_ZKON0)                        0.005 0.001     0.001       0.001
## prop(factor(M_RYG))1:prop(M_ZKON0) -0.004 0.002     0.002       0.002
##                                          z P-val lower2.5% upper97.5%
## prop(K_COF)                         -1.030 0.304     0.000      0.000
## prop(factor(K_RYG))1                -0.820 0.412    -0.494      0.196
## prop(factor(M_RYG))1                 0.678 0.498    -0.320      0.644
## prop(M_ZKON0)                        4.730 0.000     0.003      0.007
## prop(factor(M_RYG))1:prop(M_ZKON0) -1.670 0.094    -0.008      0.000
## Test of Proportionality
##                                     sup|  hat U(t) | p-value H_0
## prop(K_COF)                                    3.15        0.042
## prop(factor(K_RYG))1                           1.85        0.606
## prop(factor(M_RYG))1                           2.38        0.282
## prop(M_ZKON0)                                  2.54        0.197
## prop(factor(M_RYG))1:prop(M_ZKON0)             3.33        0.021
```

Here we again see that we can't reject the null hypothesis on a 95% confidence. We futhermore see a low p-value for the supremum test, so this model get rejected, which might be expected from the analysis above.

The last interaction we will look at is the interaction between female smokers and caffeine intake. The model is fitted and the summary is run below

```
summary(cox.aalen(Surv(TTP, K_GRAVID) ~ prop(K_COF) + prop(factor(K_RYG))
                  + prop(factor(M_RYG)) + prop(M_ZKON0) + prop(factor(K_RYG))*prop(K_COF),
                  data = data, weighted.test = 1, residuals = 1, n.sim = 1000))
```

```
## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                                       Coef.    SE Robust SE D2log(L)^-1      z
## prop(K_COF)                          -0.001 0.000     0.000       0.000 -1.570
## prop(factor(K_RYG))1                 -0.470 0.336     0.331       0.311 -1.420
## prop(factor(M_RYG))1                 -0.063 0.178     0.173       0.171 -0.362
## prop(M_ZKON0)                         0.004 0.001     0.001       0.001  4.500
## prop(K_COF):prop(factor(K_RYG))1      0.001 0.001     0.001       0.001  1.220
##                                      P-val lower2.5% upper97.5%
## prop(K_COF)                          0.116    -0.001     -0.001
## prop(factor(K_RYG))1                 0.155    -1.130      0.189
## prop(factor(M_RYG))1                 0.717    -0.412      0.286
## prop(M_ZKON0)                        0.000     0.002      0.006
## prop(K_COF):prop(factor(K_RYG))1 0.222    -0.001      0.003
## Test of Proportionality
```

```
##                                     sup|  hat U(t) | p-value H_0
## prop(K_COF)                                     3.19        0.032
## prop(factor(K_RYG))1                            1.77        0.644
## prop(factor(M_RYG))1                            2.17        0.388
## prop(M_ZKON0)                                   2.46        0.235
## prop(K_COF):prop(factor(K_RYG))1                2.42        0.193
```

Looking at this we see that we can't reject the null hypothesis for the interaction term, we do however see a p-value that could indicate a somewhat better fit. Furthermore we see a change in the estimates of the variables, such that the caffeine now is negative. This effect does however seem to be cancelled out by the interaction term, if you also smoke. The interpretation of this must be that, if you doesn't smoke, caffeine will have a negative effect on how long time it takes to get pregnant. On the other hand, if you smoke then the smoking will be biggest effect on the time to pregnancy and the caffeine intake doesn't matter to much.

2

Here we will fit an additive model and analyse this. The model fit and summary is run below

```
fit_add <- aalen(Surv(TTP, K_GRAVID) ~ K_COF + factor(K_RYG) + factor(M_RYG) + M_ZKON0,
                 data = data, residuals = 1, weighted.test = 1, silent = 0)
```

```
## Error in invert: estimated reciprocal condition number = 6.4722260e-33
##  X'X not invertible at time 194.541000 0
```

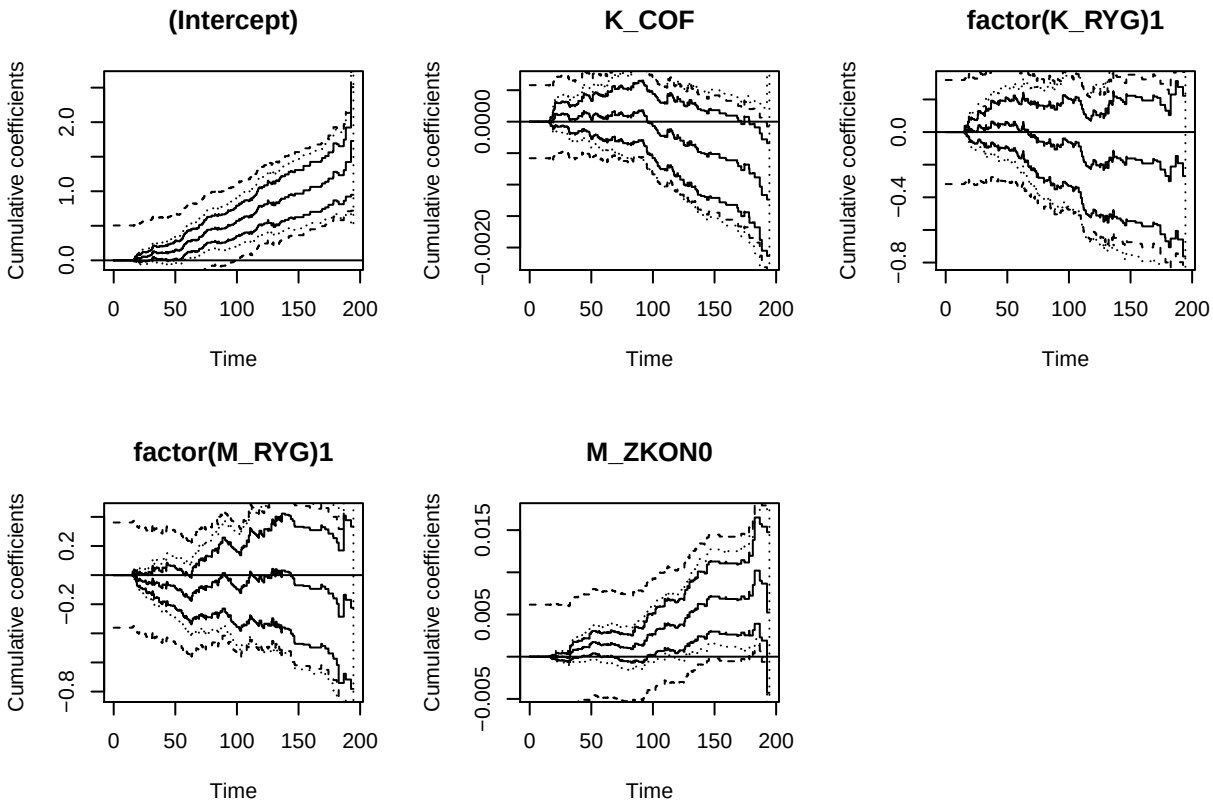
```
summary(fit_add)
```

```
## Additive Aalen Model
##
## Test for nonparametric terms
##
## Test for non-significant effects
##                 Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                               Inf                   0
## K_COF                                     Inf                   0
## factor(K_RYG)1                            Inf                   0
## factor(M_RYG)1                            Inf                   0
## M_ZKON0                                   Inf                   0
##
## Test for time invariant effects
##                       Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                             35.90                       0.000
## K_COF                                    8.00                       0.000
## factor(K_RYG)1                           3.28                       0.017
## factor(M_RYG)1                           3.06                       0.040
## M_ZKON0                                 23.00                       0.000
##                         Cramer von Mises test p-value H_0:constant effect
## (Intercept)                              3440                       0.000
## K_COF                                     910                       0.003
## factor(K_RYG)1                            135                       0.440
## factor(M_RYG)1                             80                       0.725
## M_ZKON0                                  1130                       0.000
##
```

```
##
##
##   Call:
## aalen(formula = Surv(TTP, K_GRAVID) ~ K_COF + factor(K_RYG) +
##     factor(M_RYG) + M_ZKON0, data = data, residuals = 1, weighted.test = 1,
##     silent = 0)
```

From this we observe that there is a problem with data, since at some specific time points we get something that is not invertible. Therefore we get infinity as estimates for the variables, and we can't of course not calculate any p-values from this. We also see that the two tests testing for time invariance gives somewhat same results, however the p-value is very different for sperm concentration. In the Kolmogorov-Smirnov test we see that it is significantly varying over time, i.e. time dependent, while the Cramer von Mises test makes the other conclusion. The overall conclusion from these tests, must be that the smoking variables could be constant over time. This makes quite good sense, since you are either smoking or not, and not a lot changes their mind when trying to get pregnant. They might first stop when they get pregnant (hepofully).

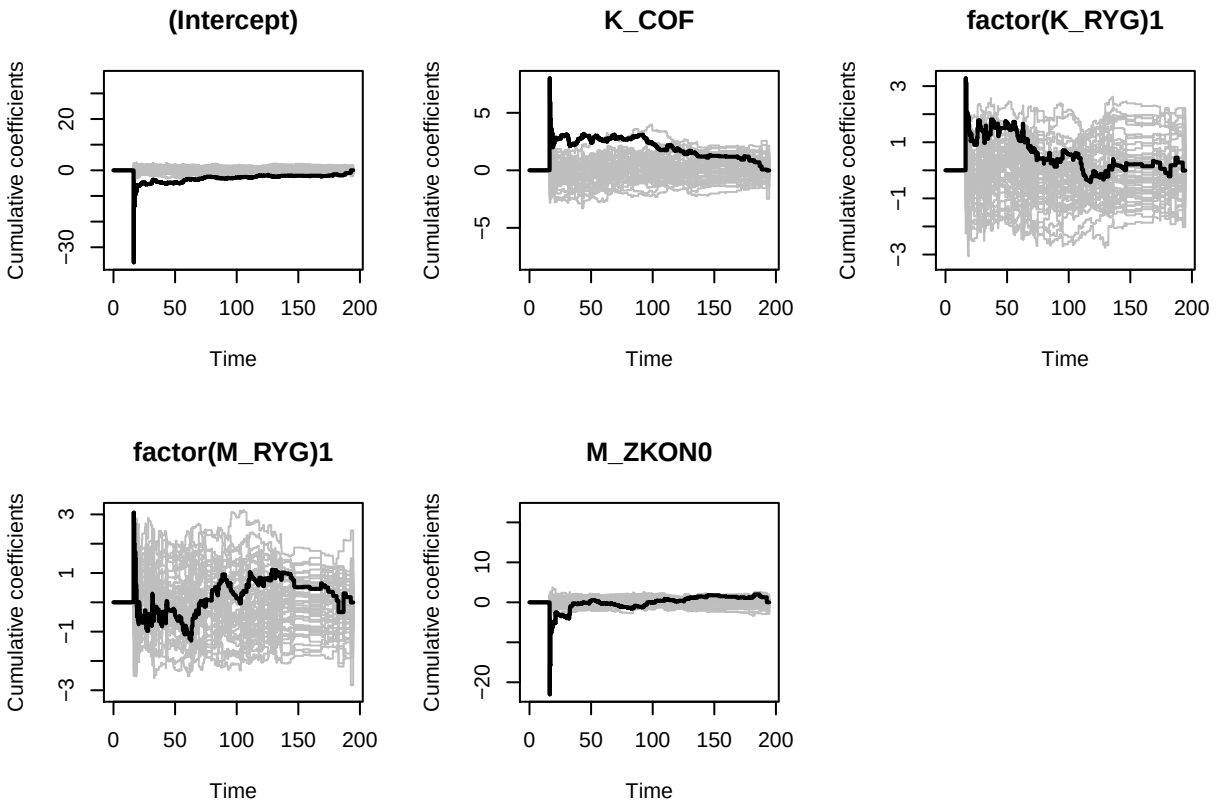
It is a problem that we doesn't get any estimates from the summary, and we are interested in these. Fortunate for us there are other ways in which we can investigate the estimates. We can plot the cummulative regression coefficients to see if we can say something about the estimates being different from 0. These plot can be seen be-



low

For Caffeine and the two smoking variables it is difficult to say that they should be different from 0, since 0 is within the bands. For the intercept and sperm concentration there might be indications that they are different from 0, since 0 is mostly only within the Hall-Wellner bands and almost not within the 95% confidence intervals.

As mentioned earlier we are interested to find out whether any of the variables are constant over time. If we can conclude this, we can reduce our model. From earlier we found some small contradictions, but we also saw that there were indications that the smoking variables could be constant over time. To investigate further we will plot the score processes. This is done below



From this we see that smoking for men and women look constant over time, while tthe intercept and caffeine intake for females seems to be time dependent. It is, however, a bit more difficult to say anything about the sperm concentration from this, just like it was from the two tests we looked at earlier. Based on this we will make the smoking variables constant in the model. See below for the reduced fitted model and its summary

```
fit_redu <- aalen(Surv(TTP, K_GRAVID) ~ K_COF + const(factor(K_RYG)) + const(factor(M_RYG)) + M_ZKON0,
                  data = data, residuals = 1, weighted.test = 1, resample.iid = 1)
summary(fit_redu)
```

```
## Additive Aalen Model
##
## Test for nonparametric terms
##
## Test for non-significant effects
##             Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                          6.47                0.000
## K_COF                                2.65                0.123
## M_ZKON0                              4.08                0.002
##
## Test for time invariant effects
##                   Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                         4.67                        0.000
## K_COF                               2.60                        0.078
## M_ZKON0                             1.53                        0.374
##                     Cramer von Mises test p-value H_0:constant effect
## (Intercept)                        793.0                        0.022
```

```
## K_COF                                440.0                       0.091
## M_ZKON0                               78.3                       0.531
##
## Parametric terms :
##                          Coef.    SE Robust SE      z P-val lower2.5%
## const(factor(K_RYG))1 -0.001 0.001     0.001 -0.743 0.457    -0.003
## const(factor(M_RYG))1 -0.001 0.001     0.001 -0.770 0.441    -0.003
##                         upper97.5%
## const(factor(K_RYG))1        0.001
## const(factor(M_RYG))1        0.001
##
##   Call:
## aalen(formula = Surv(TTP, K_GRAVID) ~ K_COF + const(factor(K_RYG)) +
##     const(factor(M_RYG)) + M_ZKON0, data = data, residuals = 1,
##     weighted.test = 1, resample.iid = 1)
```

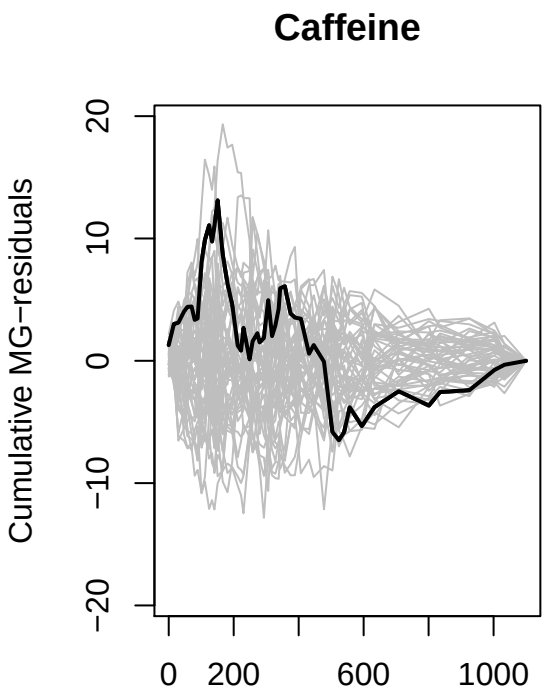
From this we now get estimates and p-values. These confirm the things we saw ealier, namely that intercept and sperm concentration looked significantly different from 0, while we can't say anything about the female intake of caffeine.

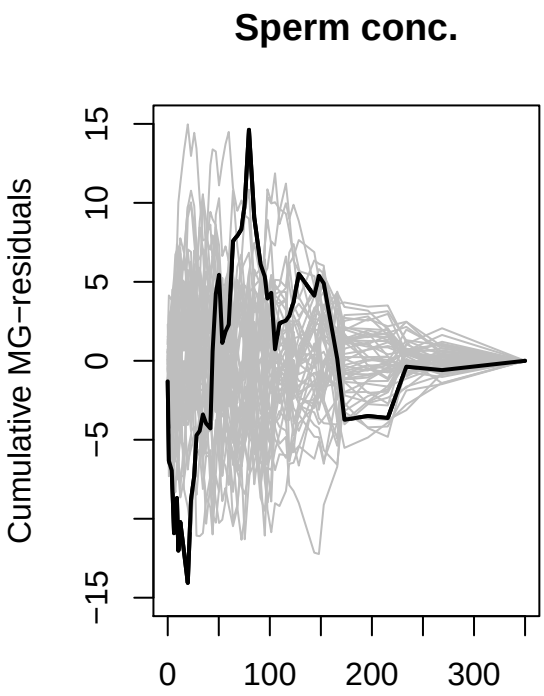
Now we look closer at GOF. To this analysis we find the cummulative residuals and plot these. This can be seen below

```
resids <- cum.residuals(fit_add, data = data, cum.resid = 1)
summary(resids)
```

```
## 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               0.062
##            14.626               0.018
```

```
par(mfrow=c(1,2))
plot(resids, score = 2,  main = c("Caffeine", "Sperm conc."))
```





From this the model looks like a pretty bad fit under the null hypothesis. Looking at the plots (the same as in the last sub question) we also see the same pattern. They doesn't seem fit into the selected model. This could maybe suggest some kind of transformation. In the last sub question we saw that log-transforming the sperm concentration didn't result in a much better fit. If we try to do this here we get the following

```
fit_Logredu <- aalen(Surv(TTP, K_GRAVID) ~ K_COF + const(factor(K_RYG))
                     + const(factor(M_RYG)) + logM_ZKON0,
                   data = data, residuals = 1, weighted.test = 1, resample.iid = 1)
summary(fit_Logredu)
```

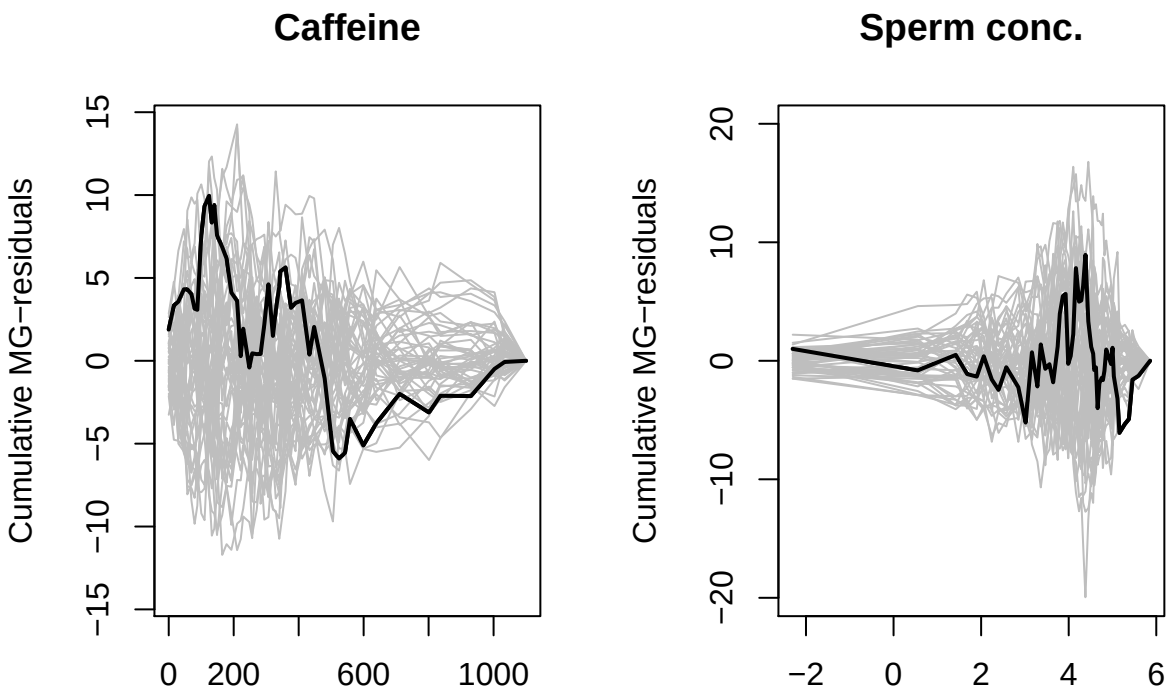
```
## Additive Aalen Model
##
## Test for nonparametric terms
##
## Test for non-significant effects
##             Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                          3.21                0.030
## K_COF                                2.57                0.149
## logM_ZKON0                           6.05                0.000
##
## Test for time invariant effects
##                   Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                         2.68                        0.046
## K_COF                               2.93                        0.037
## logM_ZKON0                          2.30                        0.113
##                     Cramer von Mises test p-value H_0:constant effect
```

```
## (Intercept)                              528                       0.073
## K_COF                                    700                       0.032
## logM_ZKON0                               109                       0.462
##
## Parametric terms :
##                          Coef.    SE Robust SE      z P-val lower2.5%
## const(factor(K_RYG))1 -0.001 0.001     0.001 -0.766 0.444    -0.003
## const(factor(M_RYG))1 -0.001 0.001     0.001 -0.497 0.619    -0.003
##                         upper97.5%
## const(factor(K_RYG))1      0.001
## const(factor(M_RYG))1      0.001
##
##   Call:
## aalen(formula = Surv(TTP, K_GRAVID) ~ K_COF + const(factor(K_RYG)) +
##     const(factor(M_RYG)) + logM_ZKON0, data = data, residuals = 1,
##     weighted.test = 1, resample.iid = 1)
```

```
resids <- cum.residuals(aalen(Surv(TTP, K_GRAVID) ~ K_COF + factor(K_RYG)
                    + factor(M_RYG) + logM_ZKON0,
                  data = data, residuals = 1, weighted.test = 1, resample.iid = 1),
                  data = data, cum.resid = 1)
summary(resids)
```

```
## 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               0.284
##              8.937               0.492
```

```
par(mfrow=c(1,2))
plot(resids, score = 2, main = c("Caffeine", "Sperm conc."))
```



From this we see that we get a much better fit than before. We therefore stick with this model for the remainder of the questions.

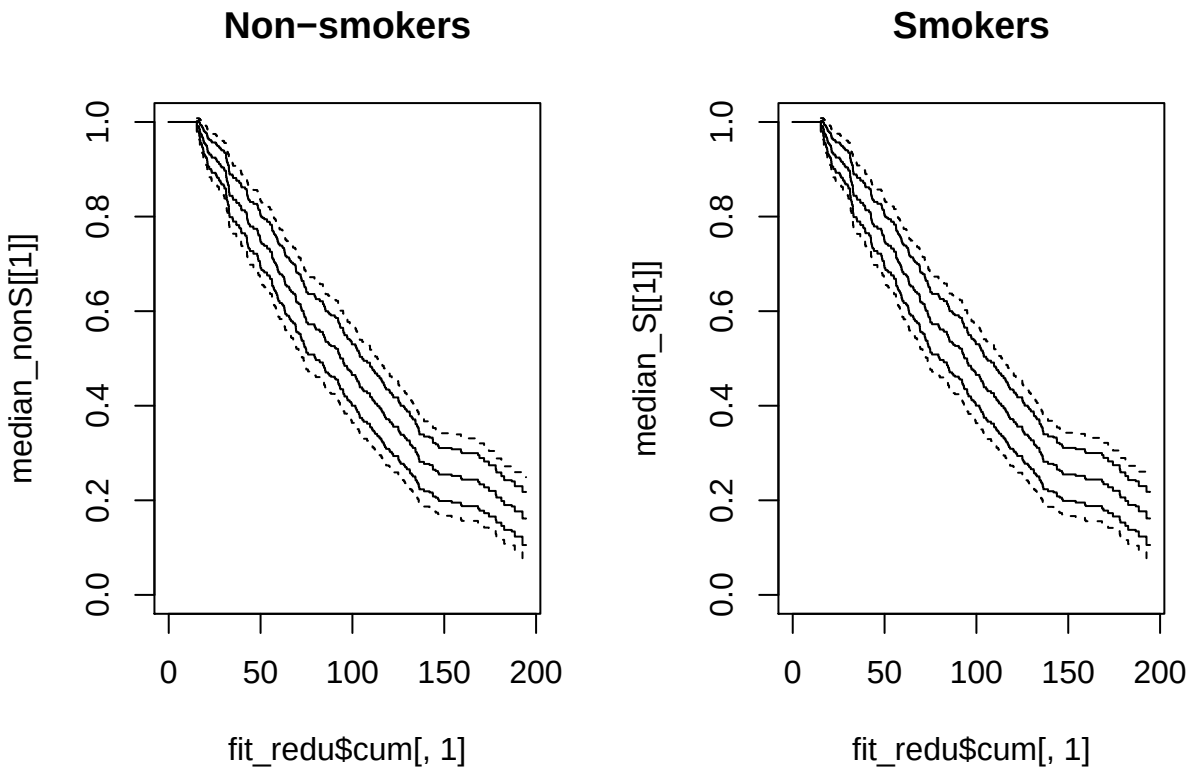
Estimating survival function

To answer the question on how you can summarize the effects as simply as possible we will look into estimating the survival function. The reason for this is that it is easy to interpret and we can turn on the covariates here to see how this will change the survival function. To estimate the survival function we follow an example from the book, example 5.8.1. What we basically is doing is using the formula we also looked at in the theoretical part of the exam, and then implementing this in R. Below you find a helper function to estimate the survival function. This takes 3 inputs, two vectors of covariates (time varying and constant covariates) and a dimension vector. Then it returns a list with our estimate of the survival function, standard errors and a helper matrix.

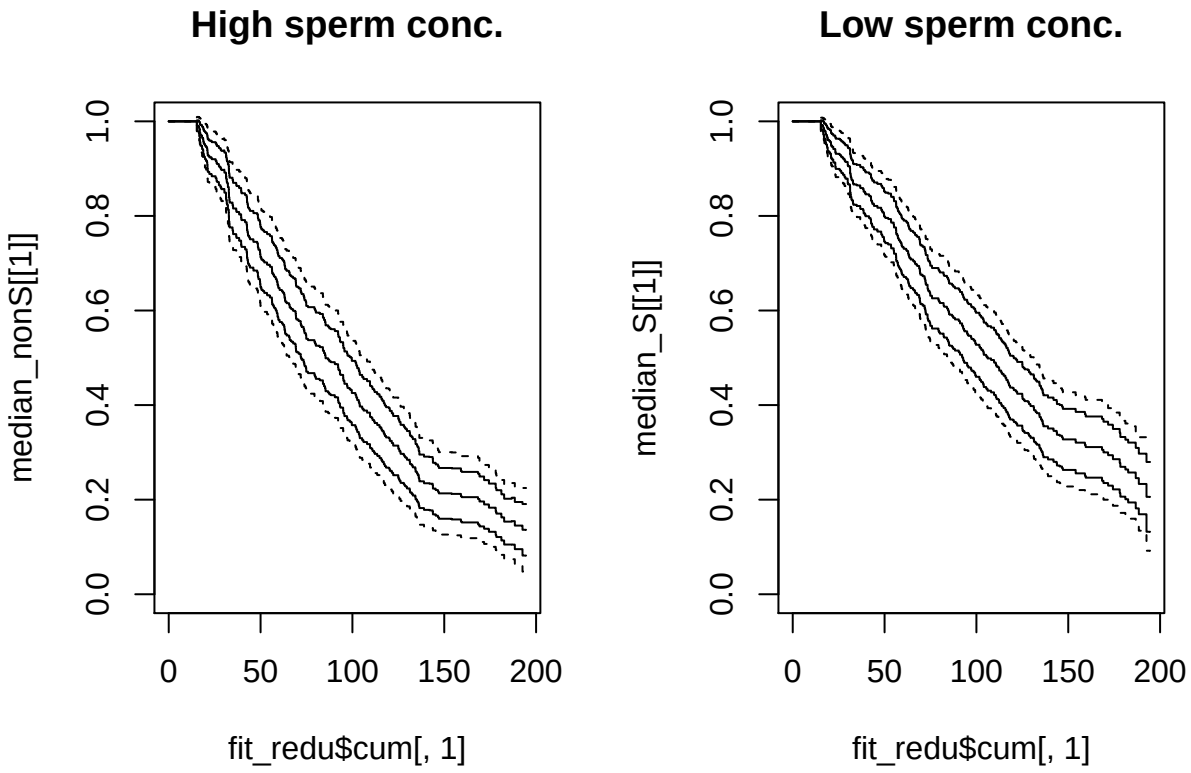
```
estSurvival <- function(x,z, dim)
{
  delta <- matrix(0, nrow = dim[1] , ncol = dim[2])
  for (i in 1:dim[2])
  {
    delta[, i] <- x0 %*% t(fit_Logredu$B.iid[[i]])
    + fit_Logredu$cum[, 1]*sum(z0*fit_Logredu$gamma.iid[i,])
  }
  S0 <- exp(-x0 %*% t(fit_Logredu$cum[,-1])
            - fit_Logredu$cum[, 1]*sum(z0*fit_Logredu$gamma.iid[i,]))
  se <- apply(delta^2, 1, sum)^0.5
  res <- list(S0,se, delta)
```

```
  return(res)
}
```

Now we would like to use this function to estimate the survival function and make confidence bands. We start by examining how a couple that is non-smokers and has a median intake of caffein and median sperm concentration survival function looks like compared to an identical couple that smokes. Plots of these survival functions and their confidence bands can be seen below

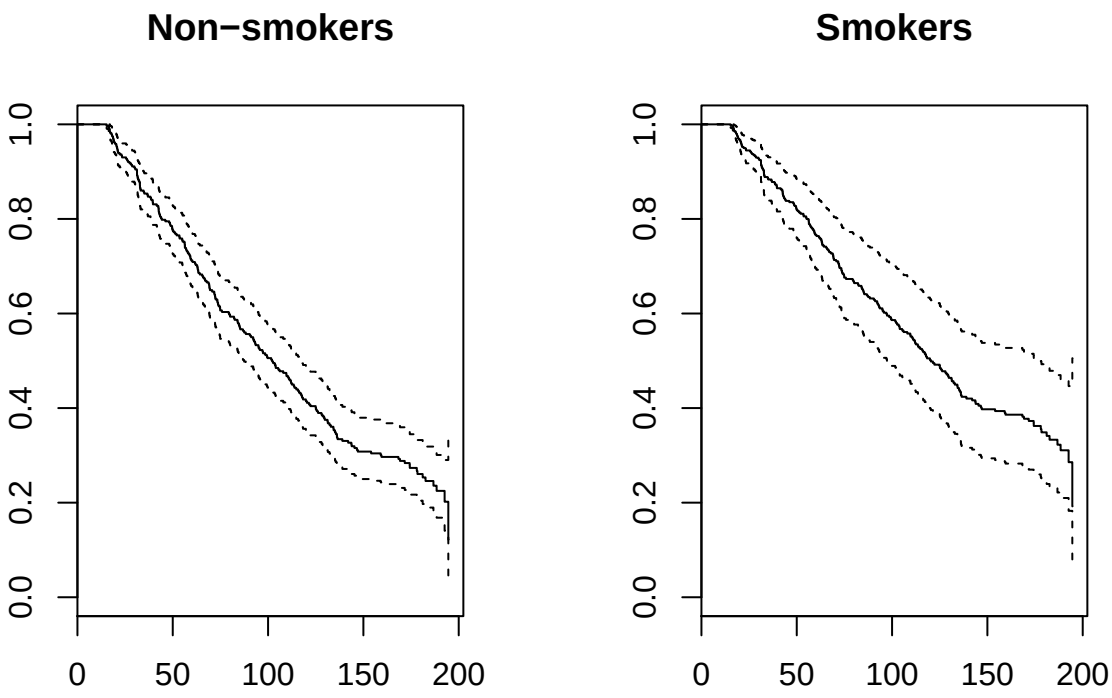


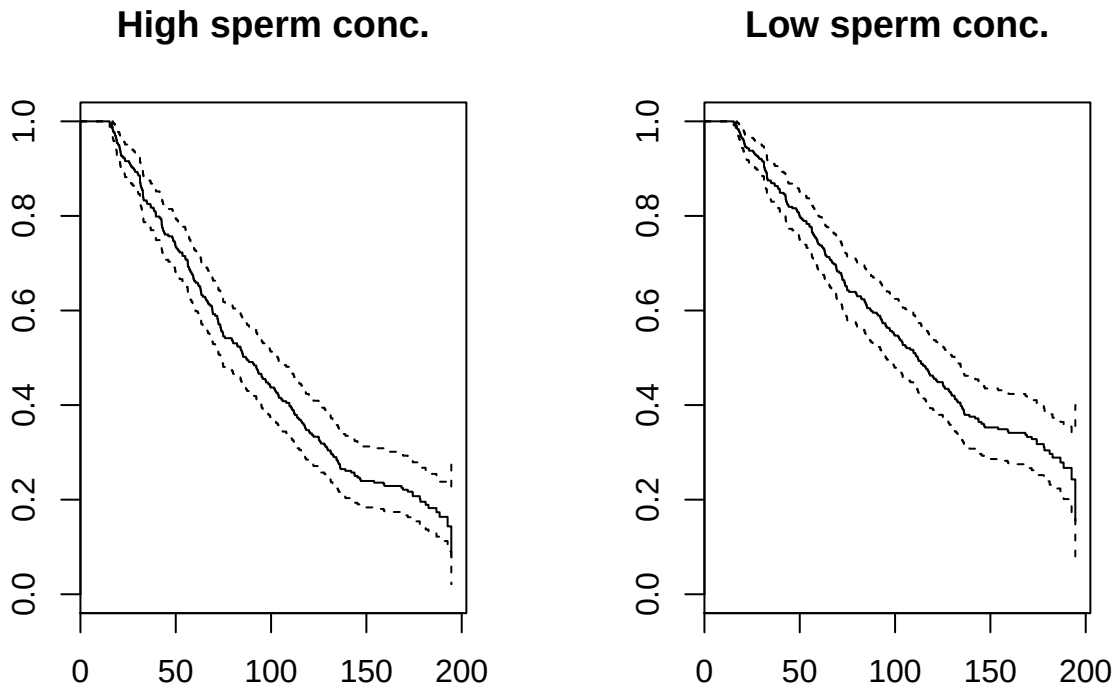
From this we see a tendency that the estimated survival function for the non-smoking couple lies beneath the smoking couples, however not by much. This indicates that there is a tendency that non-smoking couples get pregnant faster than smoking couples. Furthermore we could look at what effect the sperm concentration has. Assuming two non-smoking couples with median female caffein intake, and one of the males having a sperm concentration of 100 and the other a sperm concentration of 20, yields the following survival functions with confidence bands



Here we see a clear indication that identical couples only differing in sperm concentraion number, the couple with the male with the high sperm concentration gets pregnant faster. We also observe that the confidence bands for the couple with low sperm concentration are wider, meaning more uncertainty.

Now we will estimate the survival function for the cox model discussed in question 1. We do this by fitting the model with coxph and then using survfit. We generate the same plots as you saw above. These can be seen below


Non−smokers
Smokers
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200



From these we see that for non-smoking couple the survival function is prettymuch similar for the cox model and the addtive model. For smoking couples we see that the survival function is more flat for the cox model than for the additive model, i.e. it takes longer for smoking couples to become pregnant if we use the cox model. Comparing the plots for high and low sperm concentration for the two models doesn't show much difference. The estimat for the cox model seems to be a little less rocky.

3

In both models we see problems with the goodnes-of-fit. There are clear indications that some of the covariates isn't modelled well by these models. We furthermore see indications that sperm concentration could use a transformation, however, it is difficult to see exactly what transformation is needed to make the model fit better. In the additive model the log transformation makes the model fit better, but in the cox a log transformation doen't do much for the fit of the model. Looking at the estimates of the covariates we observe that only sperm concentration has a p-value such that we can reject the null hypothesis. Taking into account that the bad model fit from the cox model, independent of what scale sperm concentration is included on, the additive model with the log transformed sperm concentration seems to be the best fit.