Exam: Survival Analysis

1 Theoretical part

1.1 Part A

a) We find the hazard for the survival of a female plus 4 years by first writing up the survival function:

$$S(t) = P(T_i^f + 4 > t) = P(T_i^f > t-4) = \exp\left(-\int_0^{t-4} (X_i^f(s))^T \beta(s-4) + h(\gamma^T Z_i^f(s))ds\right).$$

Then we take log of the survival function and differentiate w.r.t. t:

$$\begin{aligned} -\frac{d}{dt}(\log(S(t)) &= -\frac{d}{dt}\left(-\int_0^{t-4} (X_i^f(s))^T \beta(s-4) + h(\gamma^T Z_i^f(s))ds\right) \\ &= (X_i^f(t-4))^T \beta(t-8) + h(\gamma^T Z_i^f(t-4)), \end{aligned}$$

which is exactly the hazard we were looking for.

b) Now we wish to find an estimating equation for the $B(t)$ given γ. To do that we define the following new variables:

$$\tilde{N}_i(t) = \begin{cases} N_i^m(t), & i = 1, ...n_m. \\ N_{i-n_m}^f(t+4), & i = n_m+1, ..., n_m+n_f. \end{cases}, \qquad \tilde{N}(t) = \begin{pmatrix} \tilde{N}_1 \\ \vdots \\ \tilde{N}_{n_m+n_f} \end{pmatrix},$$

$$\tilde{X}_i(t) = \begin{cases} Y_i^m(t)(X_i^m(t))^T, & i = 1, ...n_m. \\ Y_{i-n_m}^f(t+4)(X_{i-n_m}^f(t+4))^T, & i = n_m+1, ..., n_m+n_f. \end{cases}, \qquad \tilde{X}(t) = \begin{pmatrix} \tilde{X}_1 \\ \vdots \\ \tilde{X}_{n_m+n_f} \end{pmatrix},$$

$$\tilde{Z}_i(t) = \begin{cases} (Z_i^m(t))^T, & i = 1, ...n_m. \\ (Z_{i-n_m}^f(t+4))^T, & i = n_m+1, ..., n_m+n_f. \end{cases}, \qquad \tilde{Z}(t) = \begin{pmatrix} \tilde{Z}_1 \\ \vdots \\ \tilde{Z}_{n_m+n_f} \end{pmatrix}.$$

Furthermore we create $\tilde{h} : \mathbb{R}^n \to \mathbb{R}^n$, $\tilde{h}((x_1, ..., x_n)^T) = (h(x_1), ..., h(x_n))^T$.
With these new variables defined we can now write up the estimating equation for $B(t)$:

$$\begin{aligned} &\tilde{X}^T(t)\left(d\tilde{N}(t) - \tilde{X}(t)dB(t) - \tilde{Y}(t)\tilde{h}(\tilde{Z}(t)\gamma)dt\right) = 0 \Leftrightarrow \\ &\tilde{X}^T(t)\left(d\tilde{N}(t) - \tilde{Y}(t)\tilde{h}(\tilde{Z}(t)\gamma)dt\right) = \tilde{X}^T(t)\tilde{X}(t)dB(t) \Leftrightarrow \\ &\tilde{X}^-(t)\left(d\tilde{N}(t) - \tilde{Y}(t)\tilde{h}(\tilde{Z}(t)\gamma)dt\right) = d\hat{B}(t), \end{aligned}$$

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

c) The estimating equation for γ is given by:

$$\int \tilde{Y}(t)\left(\tilde{h}'(\tilde{Z}(t)\gamma)\tilde{Z}(t)\right)^T\left(d\tilde{N}(t) - \tilde{X}(t)d\hat{B}(t) - \tilde{Y}(t)\tilde{h}(\tilde{Z}(t)\gamma)dt\right) = 0.$$

If we insert the estimate we found in b), then $\hat{\gamma}$ will be the solution to the equation.

d) The main arguments in estimating γ we outlined above. To establish that $\sqrt{n}(\hat{\gamma}-\gamma)$ is asymptotically normal with a variance we can estimate (for fixed t) we have to use the martingale convergence theorem.

The main arguments in estimating the cumulatives, $B(t)$, is to insert the estimate for γ, and integrate from zero to infinity:

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

To find that $\sqrt{n}(\hat{B}(t)-B(t))$ is asymptotically normal, we look closer at the expression:

$$\begin{aligned}
\sqrt{n}(\hat{B}(t)-B(t)) &= \sqrt{n}\left(\int_0^t \tilde{X}^-(s)\left(d\tilde{N}(s) - \tilde{Y}(s)\tilde{h}(\tilde{Z}(s)\gamma)ds\right) - B(t)\right) \\
&= \sqrt{n}\left(\int_0^t \tilde{X}^-(s)\left(\tilde{X}(s)dB(s) + \tilde{Y}(s)\tilde{h}(\tilde{Z}(s)\gamma)ds + d\tilde{M}(s) - \tilde{Y}(s)\tilde{h}(\tilde{Z}(s)\hat{\gamma})ds\right) - B(t)\right) \\
&= \sqrt{n}\left(\int_0^t \tilde{X}^-(s)\tilde{X}(s)dB(s) + \int_0^t \tilde{X}^-(s)\tilde{Y}(s)\tilde{h}(\tilde{Z}(s)\gamma)ds + \tilde{X}^-(s)d\tilde{M}(s) - \tilde{X}^-(s)\tilde{Y}(s)\tilde{h}(\tilde{Z}(s)\hat{\gamma})ds - B(t)\right) \\
&= \sqrt{n}\left(B(t) + \int_0^t \tilde{X}^-(s)d\tilde{M}(s) + \tilde{X}^-(s)\tilde{Y}(s)\left(\tilde{h}(\tilde{Z}(s)\gamma) - \tilde{h}(\tilde{Z}(s)\hat{\gamma})\right)ds - B(t)\right) \\
&= \sqrt{n}\left(\int_0^t \tilde{X}^-(s)d\tilde{M}(s)\right) + \sqrt{n}\left(\int_0^t \tilde{X}^-(s)\tilde{Y}(s)\left(\tilde{h}(\tilde{Z}(s)\gamma) - \tilde{h}(\tilde{Z}(s)\hat{\gamma})\right)ds\right).
\end{aligned}$$

Note that $d\tilde{N}(t) = \tilde{X}(t)dB(t) + \tilde{Y}(t)\tilde{h}(\tilde{Z}(t)\gamma)dt + d\tilde{M}(t)$.
To show that the expression $\sqrt{n}(\hat{B}(t)-B(t))$ is asymptotically normal, one can use the martingale convergence theorem on the first integral, and the second integral one can show converges to zero in probability.

For the last part of this question we examine for which t we can estimate $B(t)$ based on the data. We recall that a female is like a four year younger male in this study, so therefore we must be able to estimate $B(t)$ for the females data from time $(-4, 56)$, and the males data from time $(0, 60)$. This means that for the full data $B(t)$ can be estimated from time $(-4, 60)$.

e) The predicted survival probability after 10 years for a male subject with covariates (X_0, Z_0) that enters the study at 30 years of age is:

$$\begin{aligned}
P(T > 10) &= \exp\left(-\int_0^{10} X_0^m\beta(t) + h(\hat{\gamma}^T Z_0^m)dt\right) \\
&= \exp\left(-X_0^m\hat{B}(10) - h(\hat{\gamma}^T Z_0^m)10\right).
\end{aligned}$$

Since we know the asymptotic distribution of both $\hat{\gamma}$ and $\hat{B}$ it's possible to calculate the standard errors for the estimate empirically.

f) It turns out that the females in the study are recruited at different ages, and are all older than 30 when being recruited for the study. We know their individual ages of recruitment. To still estimate $B(\cdot), \gamma$ consistently we can modify the above analysis by including a covariate the that takes the age into acount, or we could modify the at-risk indicator such that the females that enters later only are at risk from the point of entry.

1.2 Part B

a) To find the hazard function given only X we start by writing up the survival function:

$$P(T > t|X, A) = \int_0^\infty P(T > t|X)f(a)da$$

$$= \int_0^\infty \exp\left(-\int_0^t \beta(s) + X\alpha(s+a)ds\right) f(a)da$$
$$= \exp\left(-\int_0^t \beta(s)ds\right) \int_0^\infty \exp\left(-\int_0^t X\alpha(s+a)ds\right) f(a)da.$$

Like in part A exercise a) we now look at

$$-\frac{d}{dt}\log(S(t)) = -\frac{d}{dt}\left(\left(-\int_0^t \beta(s)ds\right) + \log\left(\int_0^\infty \exp\left(-\int_0^t X\alpha(s+a)ds\right) f(a)da\right)\right)$$
$$= \beta(t) - \frac{\int_0^\infty -X\exp\left(-\int_0^t X\alpha(s+a)ds\right)\alpha(t+a)f(a)da}{\int_0^\infty \exp\left(-\int_0^t X\alpha(s+a)ds\right) f(a)da},$$

which is the hazard given X.

b) The hazard function given both X and A unobserved is found using the same procedure as in the previous question. The hazard we end up with will just be some what more complicated, since now X is also unobserved. Like before let's start by looking at the survival function:

$$S(t) = P(T > t)$$
$$= \int_0^\infty \exp\left(-\int_0^t \beta(s) + 0\cdot\alpha(s+a)ds\right)\frac{1}{2}f(a)da + \int_0^\infty \exp\left(-\int_0^t \beta(s) + 1\cdot\alpha(s+a)ds\right)\frac{1}{2}f(a)da$$
$$= \int_0^\infty \exp\left(-\int_0^t \beta(s)ds\right)\frac{1}{2}f(a)da + \int_0^\infty \exp\left(-\int_0^t \beta(s) + \alpha(s+a)ds\right)\frac{1}{2}f(a)da.$$

Note that

$$\frac{d}{dt}\log(S(t)) = \frac{\frac{d}{dt}S(t)}{S(t)}.$$

To find the hazard we, once again, take

$$-\frac{d}{dt}\log(S(t)) = \frac{\int_0^\infty \frac{1}{2}\beta(t)\exp\left(-\int_0^t \beta(s)ds\right)f(a)da + \int_0^\infty \frac{1}{2}\left(\beta(t)+\alpha(t+a)\right)\exp\left(-\int_0^t \beta(s)+\alpha(s+a)ds\right)f(a)da}{\int_0^\infty \exp\left(-\int_0^t \beta(s)ds\right)\frac{1}{2}f(a)da + \int_0^\infty \exp\left(-\int_0^t \beta(s)+\alpha(s+a)ds\right)\frac{1}{2}f(a)da}$$
$$= \frac{\frac{1}{2}\beta(t)\exp\left(-\int_0^t \beta(s)ds\right)\int_0^\infty f(a)da + \frac{1}{2}\int_0^\infty \left(\beta(t)+\alpha(t+a)\right)\exp\left(-\int_0^t \beta(s)+\alpha(s+a)ds\right)f(a)da}{\frac{1}{2}\exp\left(-\int_0^t \beta(s)ds\right)\int_0^\infty f(a)da + \frac{1}{2}\int_0^\infty \exp\left(-\int_0^t \beta(s)+\alpha(s+a)ds\right)f(a)da}$$
$$= \frac{\frac{1}{2}\beta(t)\exp\left(-\int_0^t \beta(s)ds\right) + \frac{1}{2}\int_0^\infty \left(\beta(t)+\alpha(t+a)\right)\exp\left(-\int_0^t \beta(s)+\alpha(s+a)ds\right)f(a)da}{\frac{1}{2}\exp\left(-\int_0^t \beta(s)ds\right) + \frac{1}{2}\int_0^\infty \exp\left(-\int_0^t \beta(s)+\alpha(s+a)ds\right)f(a)da}$$
$$= \beta(t) + \frac{\int_0^\infty \alpha(t+a)\exp\left(-\int_0^t \beta(s)+\alpha(s+a)ds\right)f(a)da}{\exp\left(-\int_0^t \beta(s)ds\right) + \int_0^\infty \exp\left(-\int_0^t \beta(s)+\alpha(s+a)ds\right)f(a)da}.$$

2 Practical part

For this part we consider a data set on 423 first pregnancy planners and study risk factors that may affect the "time to pregnancy"(TTP). The couples were followed for 6 months until conception was achieved. We will consider the following risk factors in a model:

`k_cof`: intake of caffeine for the female (mg per day).
`k_ryg`: smoking status of the female.

`m_ryg`: smoking status of the male.
`m_zkon0`: sperm concentration of male (mill/ml).

We expect that smoking and caffeine is bad for fertility and that a high sperm concentration leads to better fertility.

Before we fit any models we start by centering the continuous variables. The reason for this is that the baseline function will be easier to interpret. The baseline will here be a couple, where the female intake the average amount of caffeine, she doesn't smoke, the male doesn't smoke, and his sperm concentration is the average. See how this is done below:

```
ttp$k_cof <- ttp$k_cof-mean(ttp$k_cof)
ttp$logm_zkon0 <-log(ttp$m_zkon0+1)- mean(log(ttp$m_zkon0+1),na.rm=TRUE)
ttp$loglogm_zkon0 <- log(log(ttp$m_zkon0+1)+1)-mean(log(log(ttp$m_zkon0+1)+1),na.rm=TRUE)
ttp$m_zkon0 <- ttp$m_zkon0-mean(ttp$m_zkon0,na.rm=TRUE)
```

1) We fit a Cox regression model and do a goodness-of-fit (GOF) analysis.
   To do this we first note that two of the variables are factors, that is `k_ryg` and `m_ryg`, and should be included as such in the model. Therefore the model we fit in R will be:

   ```
   fit1 <- coxph(Surv(ttp, k_gravid)~k_cof+factor(k_ryg)+factor(m_ryg)+m_zkon0,ttp)
   ```

   By taking summary of the fit we get the following output, with the estimated regression coefficients given in the first column and the estimated relative risks given in the second.

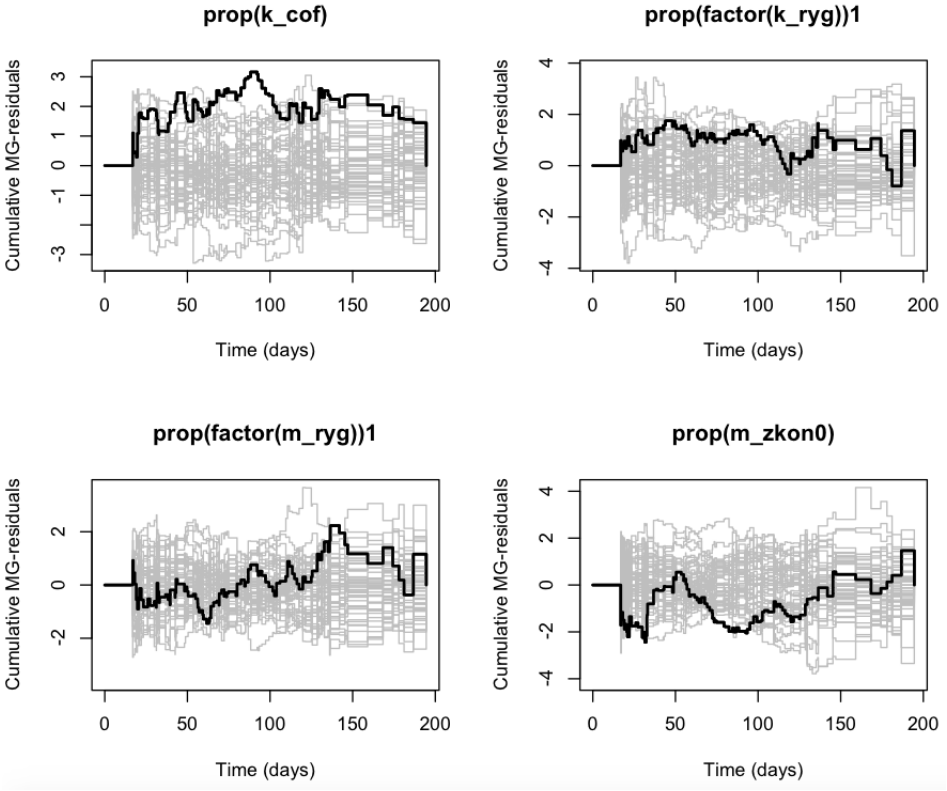
   ```
                        coef  exp(coef)   se(coef)      z Pr(>|z|)
   k_cof          -0.0003589  0.9996412  0.0003308 -1.085 0.277930
   factor(k_ryg)1 -0.1341806  0.8744321  0.1754937 -0.765 0.444516
   factor(m_ryg)1 -0.1099665  0.8958641  0.1682006 -0.654 0.513252
   m_zkon0         0.0039481  1.0039559  0.0010804  3.654 0.000258 ***
   ```

   From the output we see that only one of the variables is significant that is `m_zkon0`, but since we are interested in examinating the effect of all the mentioned variables we will keep them in the model. From the output we can also interpret the effect of the covariates on time to pregnancy. What we see is that when a female intake 1 mg of caffeine more than the average, the chance of getting pregnant decreases with 0.9996, which is like 0.046 %. We also note that when a female smokes the chance of getting pregnant decreases with 0.8744, which is like 12.66 %, and when a male smokes the chance of getting pregnant decreases as well but with 0.8959, which is 11.51%. So it definitely has an impact, though not significant, on time to pregnancy whether the couple smoke or not. Lastly we see that increasing the sperm koncentration of a male increased, significantly, the chance of getting pregnant by 0.4 %.

   The Cox regression model can fail in various ways, it's therefore imperative to do a GOF analysis. This includes investigating functional form of the covariates and checking the assumption of the proportional hazards. We start with the latter that can be checked in R by using the `cox.aalen` function, which yields:

   ```
   Test of Proportionality
                        sup|  hat U(t) | p-value H_0
   prop(k_cof)                    3.16        0.026
   prop(factor(k_ryg))1           1.75        0.654
   prop(factor(m_ryg))1           2.23        0.328
   prop(m_zkon0)                  2.44        0.222
   ```

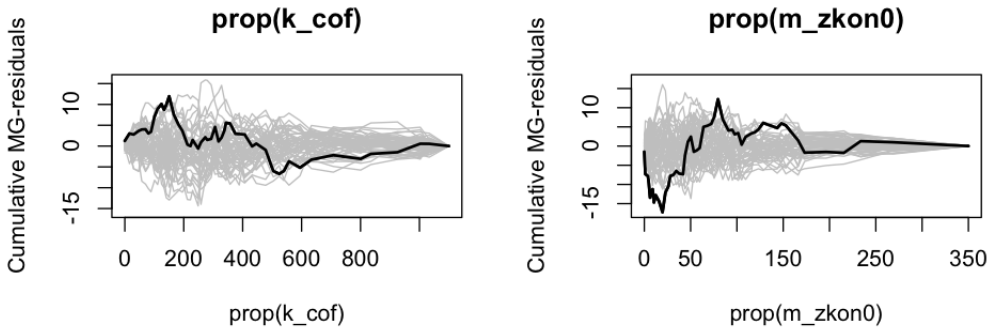
We've computed the weighted version of the supremum test statistics taking the variances of the score processes into account. From the summary output we see that the assumption of proportionality cannot be rejected for three of the variables, but for the last variable `k_cof`, we see that there is a lacking of fit of the Cox model. The same interpretation can be made from figure 1.



Figur 1: Score processes (weighted) with 50 simulated processes under the model.

Now we will use the cumulative residuals to get information about misspecification of the functional form of the continuous covariates. We know that there is a lacking of fit for the model with respect to `k_cof` but figure 2 and the summary statistics shown beneath show that the functional form of the covariate seems to be sensible enough. However for `m_zkon0` the summary statistics and figure 2 suggest that it should not be included in the model on its original scale.

```
Residual versus covariates consistent with model

               sup|  hat B(t) | p-value H_0: B(t)=0
prop(k_cof)              11.944                0.120
prop(m_zkon0)            17.230                0.004
```



Figur 2: Observed cumulative residuals versus continuous covariates with 50 random realisations under the model.

Therefore we try with two different transformations:

1. $\log(\texttt{m\_zkon0} + 1)$,
2. $\log(\log(\texttt{m\_zkon0} + 1) + 1)$.

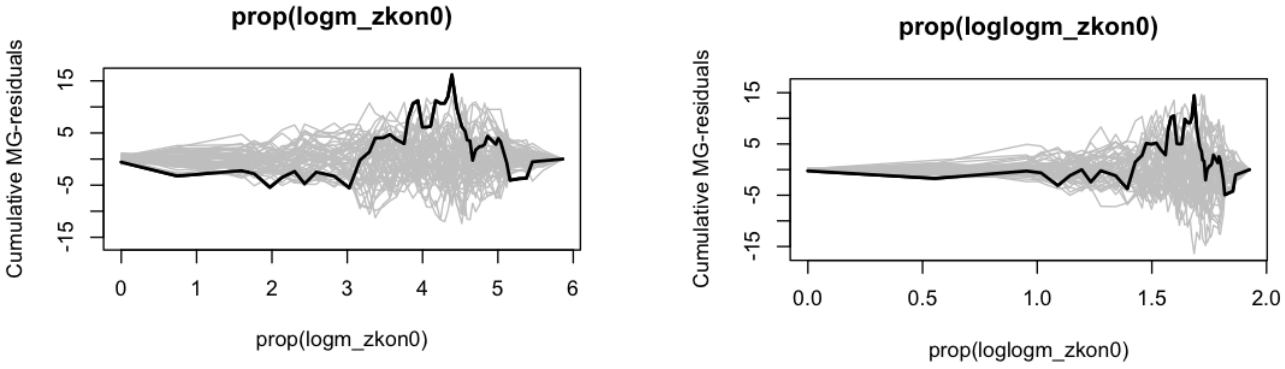
The first transformation doesn't do much for the functional form, see the summary output below.

```
Residual versus covariates consistent with model

                  sup|  hat B(t) | p-value H_0: B(t)=0
prop(k_cof)                11.598                0.096
prop(logm_zkon0)           16.144                0.008
```

But the second transformation has an effect. Now we cannot reject that it should be included on this new scale. See the summary below.

```
Residual versus covariates consistent with model

                     sup|  hat B(t) | p-value H_0: B(t)=0
prop(k_cof)                   11.042                0.202
prop(loglogm_zkon0)           14.421                0.048
```

In figure 3 we compare the two transformations, and we see that the double log-transformation is clearly the better transformation of the two. Therefore we choose to include the loglog-transformation of the covariate `m_zkon0` in the model from now on.

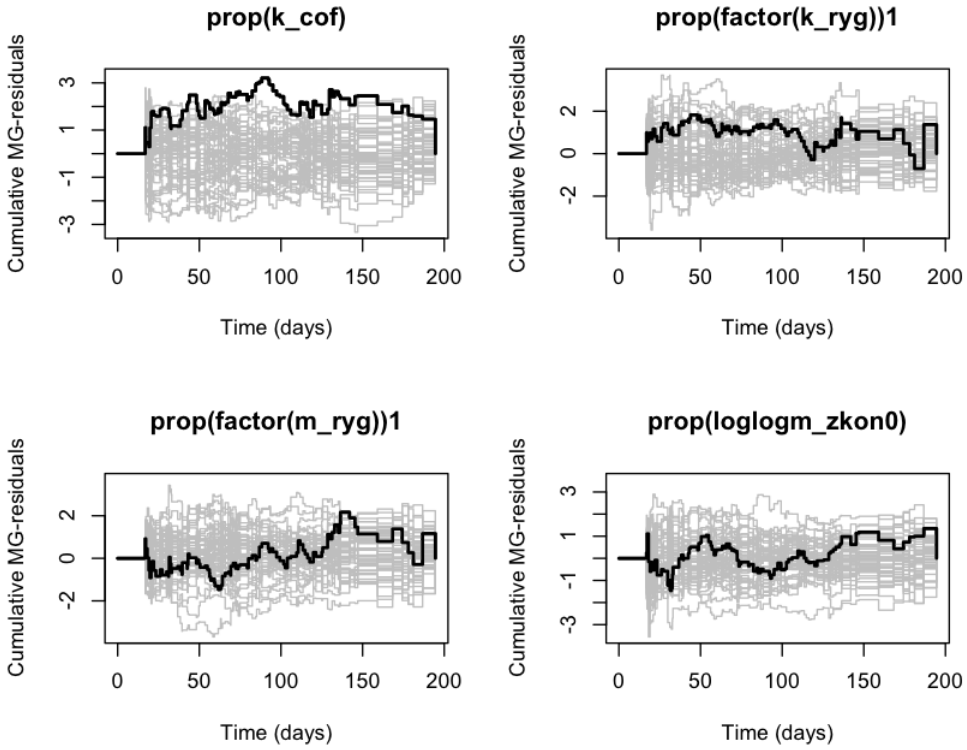


Figur 3: Observed cumulative residuals versus the log-transformation and loglog-transformation of `m_zkon0` respectively.

Since we've come up with a new model, we choose to check the proportional hazards assumption again, which yields the summary statistics below

```
Test of Proportionality
                      sup|  hat U(t) | p-value H_0
prop(k_cof)                       3.21        0.030
prop(factor(k_ryg))1              1.83        0.638
prop(factor(m_ryg))1              2.17        0.396
prop(loglogm_zkon0)               1.46        0.860
```

From the output and figure 4 we see that the result from before hasn't changed, there is still a lacking of fit of the Cox model for `k_cof`.



Figur 4: Score processes (weighted) with 50 simulated processes under the new model with a loglog-transformation of `m_zkon0`.

Lastly we want to check for interesting interactions. After having considered all possible interactions we deemed only to interactions interesting, namely the interaction between female and male smokers, and the interaction between male smokers and sperm concentration. We can imagine that if one partner smokes the other will be more likely to smoke as well. Also since smoking is very bad for ones health, we can imagine that it has an effect on the male sperm concentration. From the output below we conclude that non of the interactions are significant in this analysis.

```
                                    coef  exp(coef)   se(coef)      z Pr(>|z|)
k_cof                         -0.0003655  0.9996346  0.0003374 -1.083    0.279
factor(k_ryg)1                -0.1339222  0.8746581  0.2348509 -0.570    0.569
factor(m_ryg)1                -0.0674016  0.9348197  0.2144299 -0.314    0.753
m_zkon0                        0.0049987  1.0050112  0.0012505  3.997  6.4e-05 ***
factor(k_ryg)1:factor(m_ryg)1 -0.0346553  0.9659383  0.3573730 -0.097    0.923
```

```
factor(m_ryg)1:m_zkon0        -0.0036716  0.9963351  0.0024831 -1.479    0.139
```

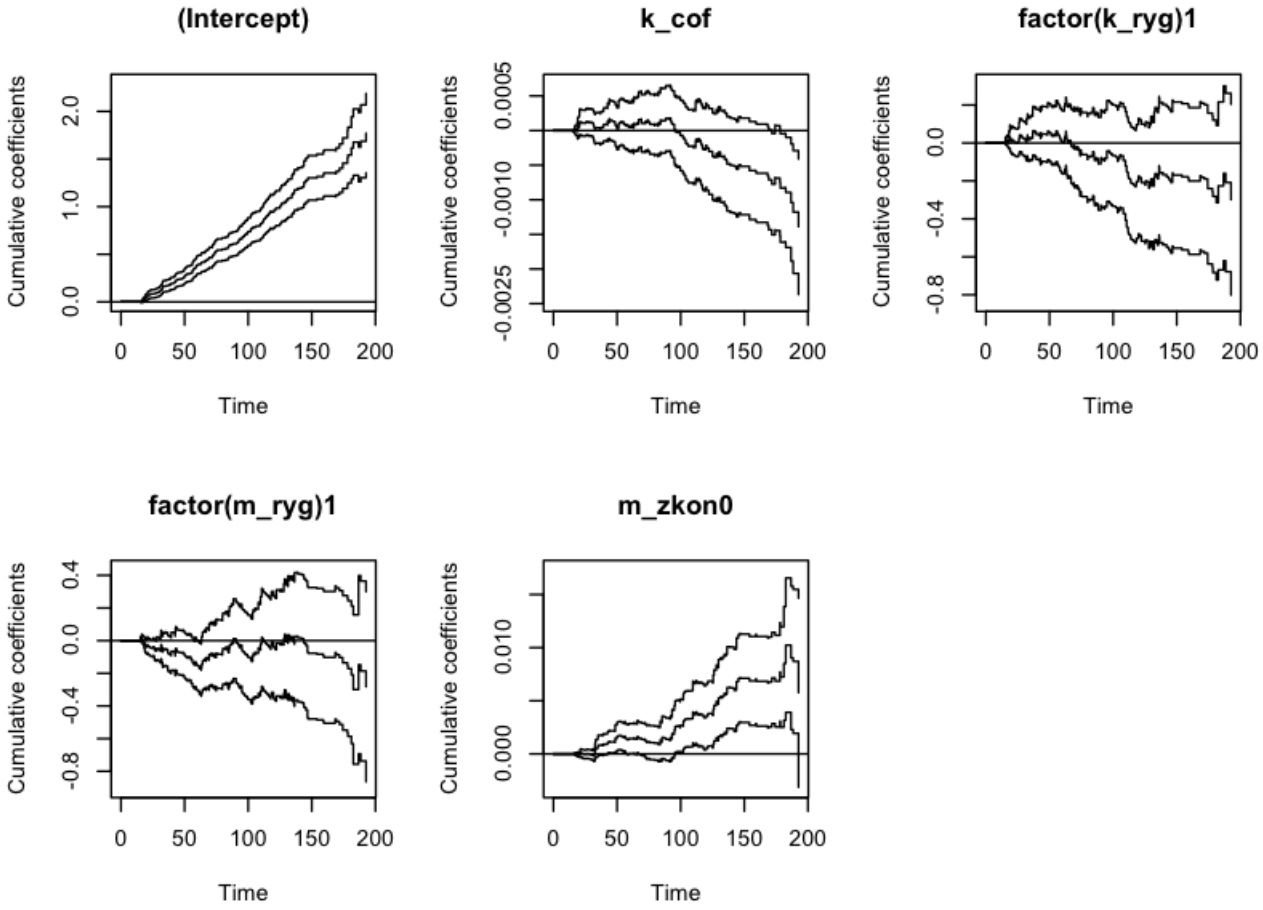
2) Now we fit an additive hazards model in R:

```
fit.add <- aalen(Surv(ttp,k_gravid)~k_cof+factor(k_ryg)+factor(m_ryg)+m_zkon0,ttp)
```

It gives an error, and we find that it's because of observation number 96, so we remove this observation from our dataset, and fit the model again with the new dataset. The summary test statistics is given below:

```
Test for non-significant effects
               Supremum-test of significance p-value H_0: B(t)=0
(Intercept)                            11.40               0.000
k_cof                                   2.57               0.153
factor(k_ryg)1                          1.53               0.766
factor(m_ryg)1                          2.22               0.373
m_zkon0                                 3.98               0.001

Test for time invariant effects
                    Kolmogorov-Smirnov test p-value H_0:constant effect
(Intercept)                        0.218000                       0.261
k_cof                              0.000836                       0.249
factor(k_ryg)1                     0.158000                       0.742
factor(m_ryg)1                     0.228000                       0.660
m_zkon0                            0.004760                       0.282
                      Cramer von Mises test p-value H_0:constant effect
(Intercept)                        4.79e+00                       0.071
k_cof                              5.12e-05                       0.181
factor(k_ryg)1                     1.05e+00                       0.638
factor(m_ryg)1                     2.32e+00                       0.561
m_zkon0                            3.84e-04                       0.558
```

First we see, using a supremum test, that only one covariate effect, `m_zkon0` is significant. Figure 5 depicts the estimated cumulative regression coefficient with 95% pointwise confidence intervals. It appears from these that the effect of at least `k_ryg` is constant with time. Looking at the test for time invariant effects in the summary we see that actually all covariate effects are constant with time.



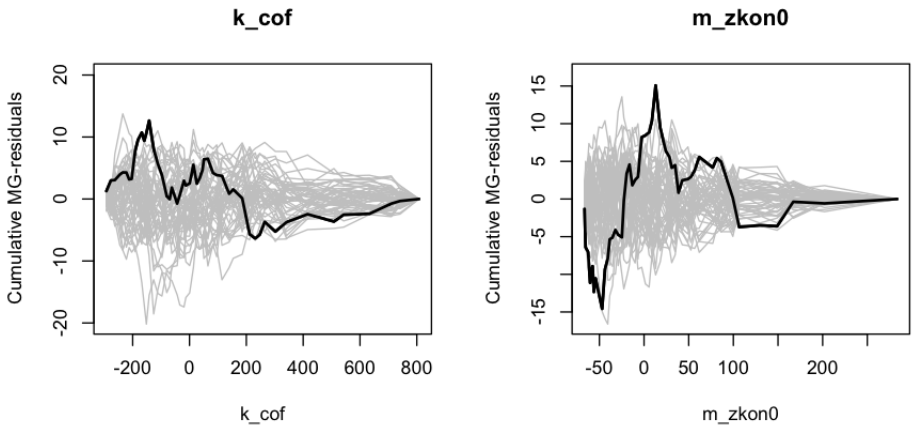
Figur 5: Estimated cumulative regression functions with 95% poitnwise confidence intervals based on Aalen's additive model.

Before doing a number of succesive test to try to simplify the model, we turn to the GOF analysis of the model we specified above. It's important to check if the model provides an adequate fit to the data, even though the additive hazards model is quite flexible.

First we check the continuous variables in the model, and we get the following output:

```
Residual versus covariates consistent with model

 sup|  hat B(t) | p-value H_0: B(t)=0
          12.638                0.082
          15.076                0.008
```

From this output we see that it's only `k_cof` that leads to a performance that is consistent with the model. Figure 6 shows the cumulative residuals with 50 resampled processes under the model. `m_zkon0` shows observed cumulative processes that does not behave as they should under the model, which is also reflected in the insignificant supremum test statistic.

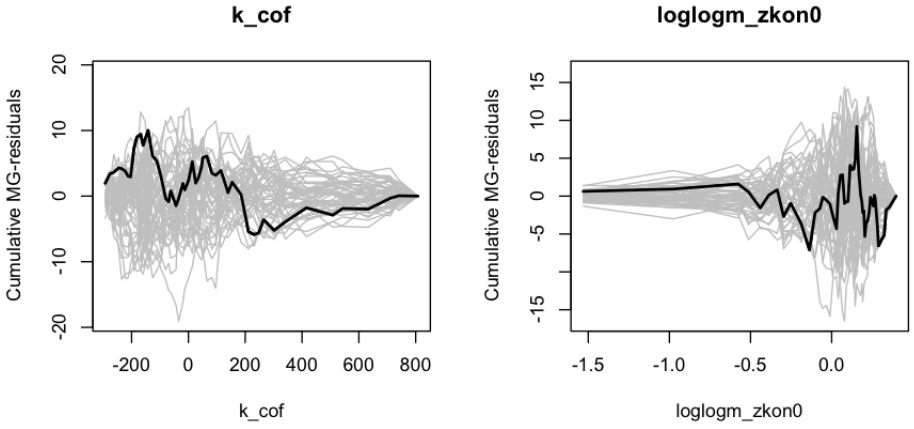


Figur 6: Observed cumulative residuals versus continuous covariates and 50 random realisations under the model.

Therefore we choose to transform the covariate `m_zkon0` by $\log(\log(\texttt{m\_zkon0} + 1) + 1)$, fit the model again, and do the same analysis as above, which yields:

```
Residual versus covariates consistent with model

 sup|  hat B(t) | p-value H_0: B(t)=0
          10.011                0.278
           9.206                0.458
```

Now we see that both covariates lead to a performance that is consistent with the model, which is also reflected in the plots in figure 7.



Figur 7: Observed cumulative residuals versus continuous covariates and 50 random realisations under the new model with the loglog-transformation of `m_zkon0`.

Now we've examined the continuous variables, so we move on to the factors in the model, `k_ryg` og `m_ryg`. We compute the cumulative residuals for this model for `m_ryg`:

```
Test for cumulative MG-residuals

Grouped Residuals consistent with model

                sup|  hat B(t) | p-value H_0: B(t)=0
```

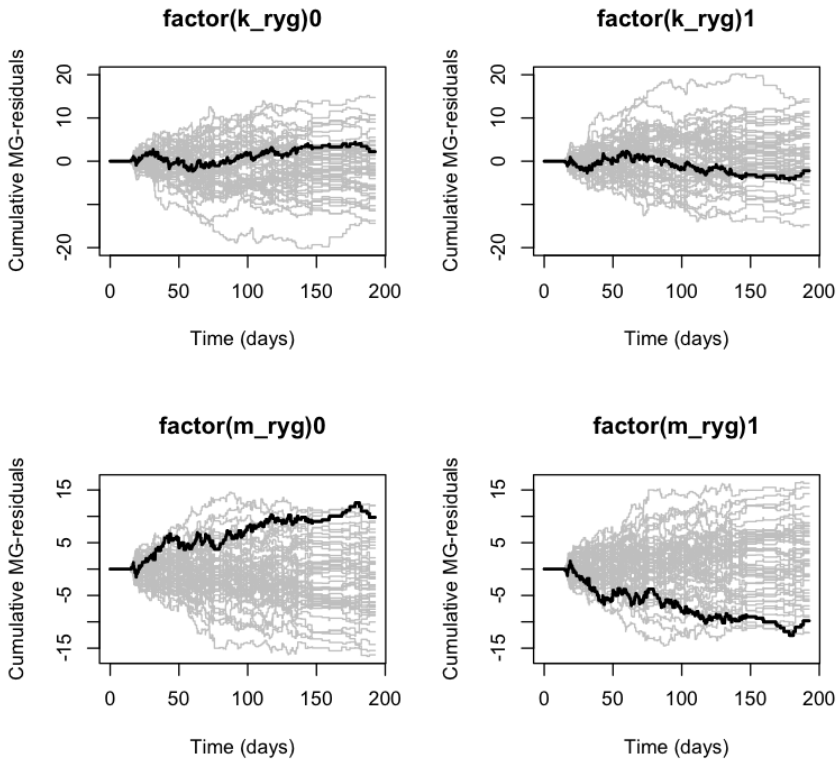
```
factor(m_ryg)0           12.594                0.118
factor(m_ryg)1           12.594                0.118

               int ( B(t) )^2 dt p-value H_0: B(t)=0
factor(m_ryg)0          10988.08                 0.14
factor(m_ryg)1          10988.08                 0.14
```

Looking at the p-values above we deduce that there should be no problem with the fit, but looking at the plots in figure 8 we would not quite have been convinced, had it not been for the insignificant p-values.

We compute the cumulative residuals for this model for `k_ryg`

```
Test for cumulative MG-residuals

Grouped Residuals consistent with model

               sup|  hat B(t) | p-value H_0: B(t)=0
factor(k_ryg)0            4.381                0.896
factor(k_ryg)1            4.381                0.896

               int ( B(t) )^2 dt p-value H_0: B(t)=0
factor(k_ryg)0           943.262                0.812
factor(k_ryg)1           943.262                0.812
```

Looking at the p-values above and at the plots in figure 8 we are convinced that there is absolutely no problem with fit.



Figur 8: Observed cumulative residuals versus the factors and 50 random realisations under the model.

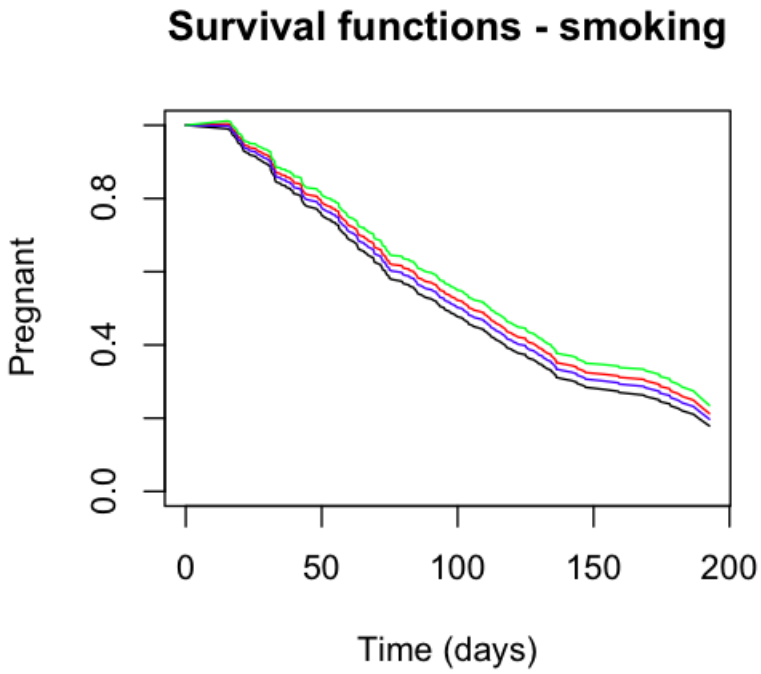
By the GOF analysis we've now established that the model fits the data well, so we move on to simplifing the model by a number of succesive tests for time invariance. We start with the following model:

```
fit.add1 <- aalen(Surv(ttp,k_gravid)~k_cof+factor(k_ryg)+factor(m_ryg)+loglogm_zkon0,
 substtp)
```

For this model we find that `k_ryg` does not seem to have a time varying effect, so we make it constant in the model. After doing this we see that `loglogm_zkon0` is also time invariant, so we also make this constant in the model and so on. In the end we find that all the covariates are time invariant. So the final model is:

```
fit.final <- aalen(Surv(ttp,k_gravid)~const(k_cof)+const(factor(k_ryg))
+const(factor(m_ryg))+const(loglogm_zkon0),substtp)
```

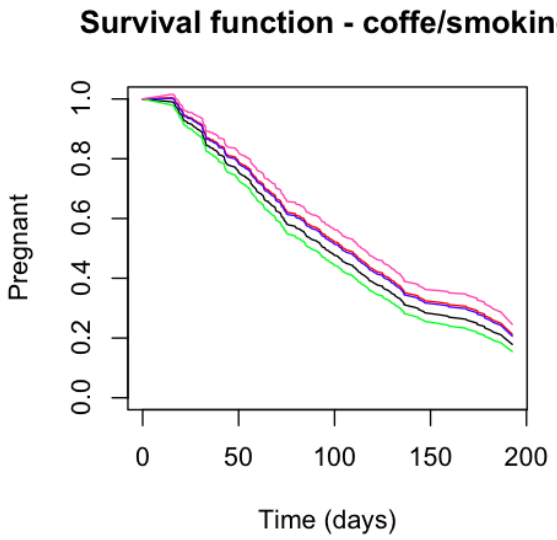
Finally we compute estimates of the survival function for various subgroups. Figure 9 shows estimates of the survival function for our baseline (black), which is a non-smoking couple, where the female's intake of coffeine is average, and the male's sperm concentration is average. The plot in figure 9 also shows survival functions for when the female smokes (red), the male smokes (blue) and both male and female in the relationship smoke (green). What we see is that smoking has an impact, though non-significant, of how long it takes for a couple to get pregnant. It has a bigger impact that the female smokes compared to if the male smokes, but the impact is largest if both smoke.



Figur 9: Survival function estimates for a baseline couple (black), a couple where the female smokes (red), a couple where the male smokes (blue), and a couple where both male and female smoke (green).

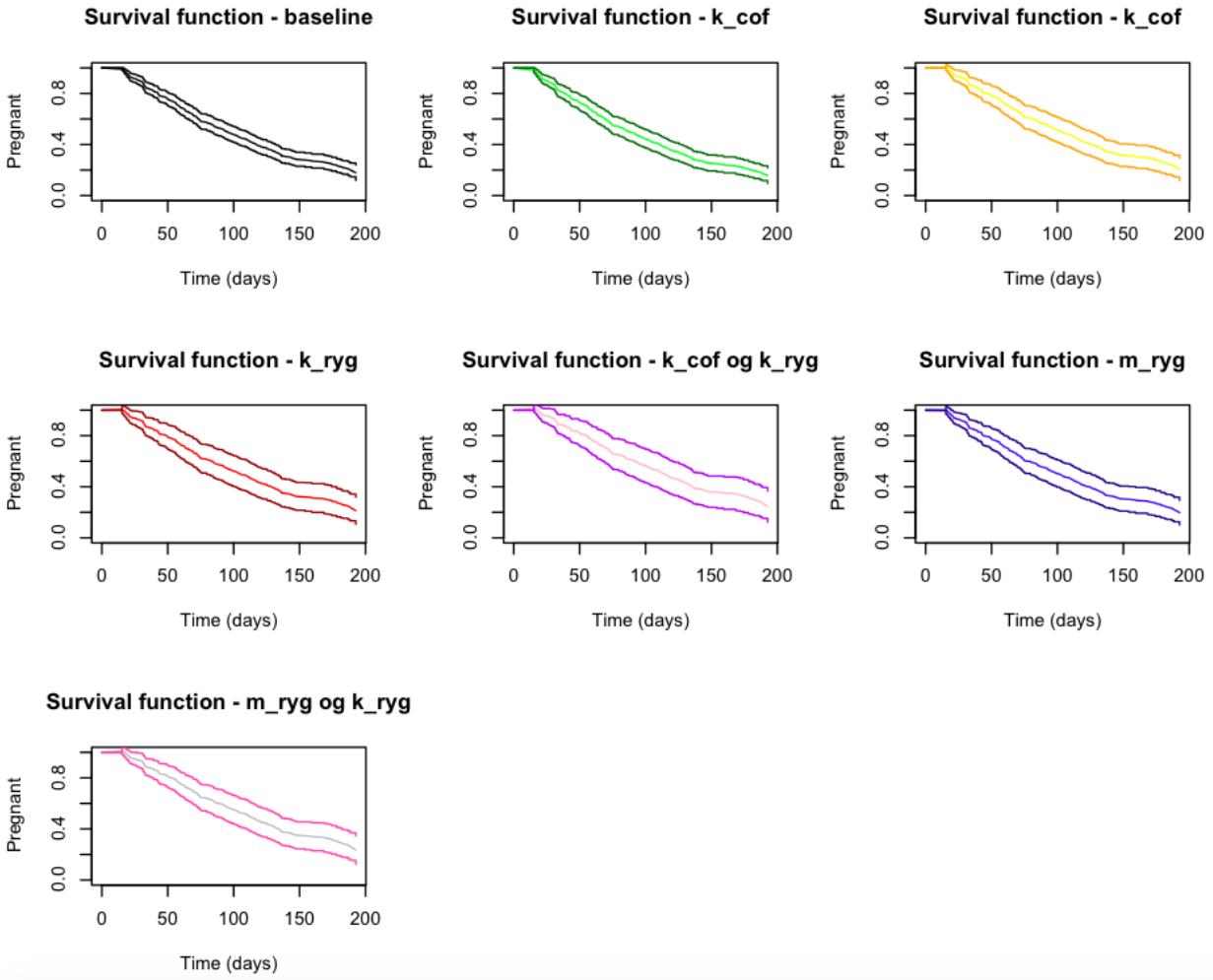
The next figure we look at is figure 10 it depicts the estimates for a survival function for a baseline couple (black), a couple where the female intake twice as much coffeine as the average (blue) - the average intake is approximately 292 mg per day, a couple where the female smoke (red), and a couple where the female both smoke and intake twice as much coffeine as the average (pink), and lastly a couple where the female doesn't smoke and doesn't intake any coffeine (green). The plot shows that doubling the intake of caffeine for a female

almost has the same effect on the chance of getting pregnant as if she smokes, and if the female both smoke and intake a lot of caffeine the chance of getting pregnant worsen even more.



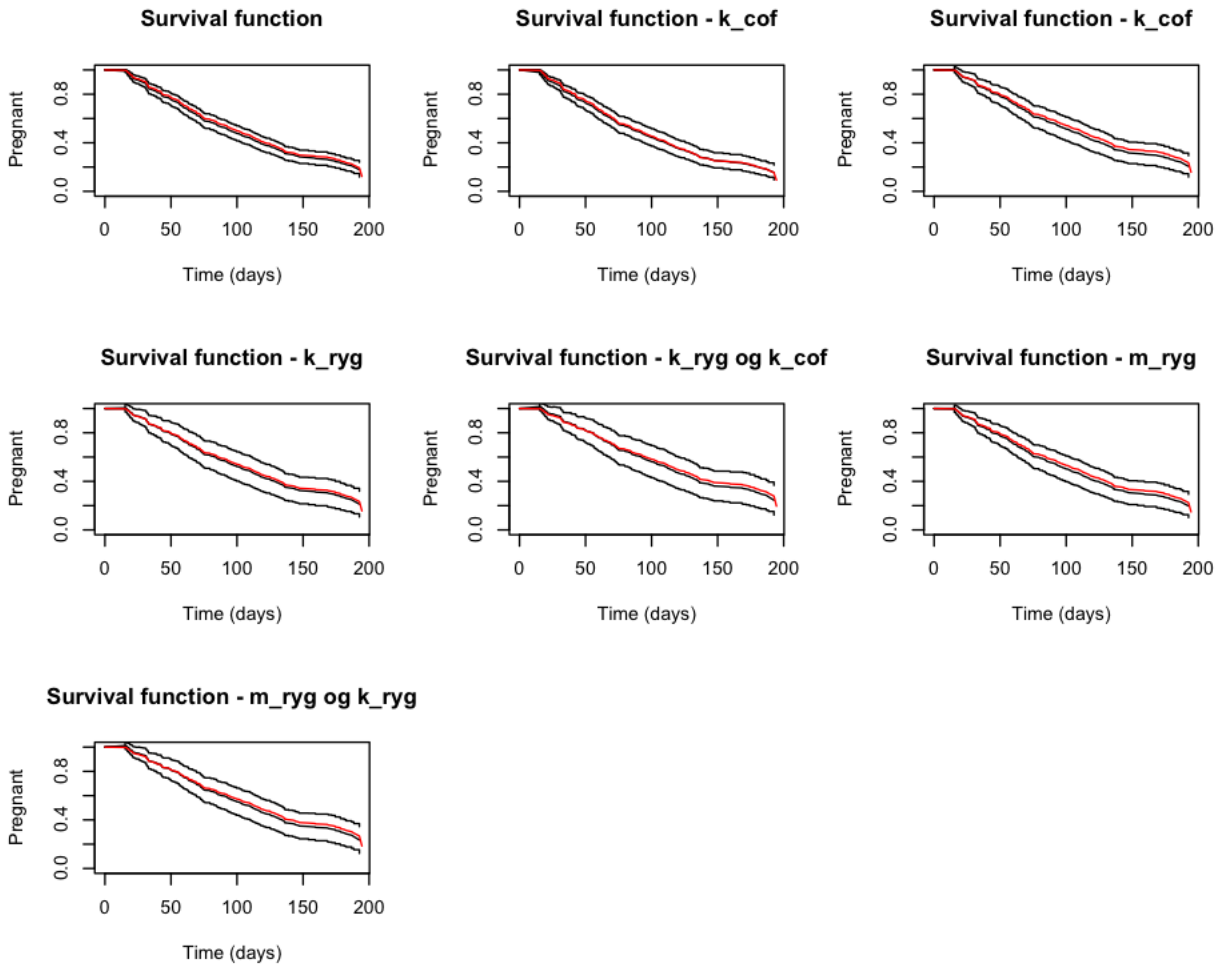
Figur 10: Survival function estimates for a baseline couple (black), a couple where the female smokes (red), a couple where the female intake twice as much coffeine as the average (blue), and a couple where the female smoke and intake twice as much coffeine as the average (pink), and a couple where the female doesn't drink any coffee and doesn't smoke.

In the plots in the above figures we haven't plotted confidence bands, to make it easier for us to compare the groups. But of course we need to see the confidence bands for each survival function estimate to make sure nothing weird is happening, so those are shown in the figure below, figure 11. We have used the headline to try to indicate, what survival function estimate that is drawn, for example the first plot shows the base line survival function estimates with confidence bands, the second shows survival function estimate for a female that doesn't intake any coffeine, the third shows survival function estimate for a couple where the female intake twice as much coffeine as the average, and so on.



Figur 11: In the first row we have survival function estimates for a baseline couple (black), a couple where the female doesn't intake any coffeine at all (green), and a couple where the female intake twice as much coffeine as the average (yellow). The second row shows survival function estimates for a couple where the female smokes (red), a couple where the female both intake twice as much coffeine as the average and smokes (purple), and a couple where the male smokes (blue). The third and last row shows survival function estimates for a couple where both male and female smoke (pink/grey).

We will now compare the results in figure 9 and 10 to the results of the Cox model in 1). To do that we draw the survival function estimates for the same groups as in the two fgures, but for the Cox regression model. See figure 12. The figure depicts survival function estimates and confidence bands for each of the groups we drew survival functions for in figure 9 and 10 (black lines), and survival function estimate for the same groups in the Cox model. What we see is that all the survival function estimates for the cox model is almost exactly on top of the survival function estimates for the additive hazards model.



Figur 12: Comparison of survival function estimates for groups in the Cox model in 1) and the additiv hazards model in 2). The first row shows three plots, the first plot shows survival function estimates for the baseline in the additiv hazards model with confidence bands (black), and survival function estimates for the baseline in the Cox regression model (red). The second plot shows survival function estimates for a couple where the female doesn't intake any coffein at all (black), and survival function estimates for the same couple in the Cox model (red). The third plot shows survival function estimates for a couple where the female intake twice as much coffein as the average (black), and survival function estimates for the same couple in the Cox model (red). The second row shows three plots where the first shows survival function estimates for a couple where the female smokes (black), and survival function estimates for the same couple in the Cox model (red), the second plot shows survival function estimates for a couple where the female both smokes and intake twice as much coffeine as the average (black), and survival function estimates for the same couple in the Cox model (red), the last plot in the second row shows survival function estimates for a couple where the male smokes (black), and survival function estimates for the same couple in the Cox model (red). The third row contains only one plot, that shows survival function estimates for a couple where both the female and male smoke (black), and survival function estimates for the same couple in the Cox model (red).

3) We now summarize the findings based on the above models. Both models deemed only the effect of `m_zkon0` significant, but all variables was kept in both models, since we specifically were interested in their effect on the response. In both models we loglog-transformed the `m_zkon0` variable, since the original scale didn't gave a good fit, and the transformation made the fit substantially better. Furthermore we recall that in the Cox regression model the proportional hazards assumption didn't hold for one of the variables, so we concluded a lacking of fit of the Cox model. Then we drew the survival function estimates for various

groups, both in the Cox regression model and in the additive hazards model, where we found that the estimates was almost exactly on top of each other, so the two models agree on the estimates. In the end we prefer the additive hazards model because the GOF analysis indicates that the model fits the data very well, which is not the case for the Cox regression model.