1 Theoretical Part

1.1 Part A

We let

$$N^g(t) = (N_i^g(t); i = 1, ..., n_g),$$

be a multivariate counting process with filtration F_t^g and F_t^g- intensity given by

$$\lambda_i^g(t) = \alpha_i^g(t, X_i^g(t), Z_i^g(t))Y_i^g(t).$$

The counting processes are related to survival times T_i^g such that the hazard of T_i^g is given by

$$\alpha_i^g(t, X_i^g(t), Z_i^g(t)).$$

We assume that the males have intensities by

$$\lambda_i^m(t) = Y_i^m(t)((X_i^m(t))^T\beta(t) + h(\gamma^T Z_i^m(t))),$$

and that the females have intensity

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

such that a female is like a 4 year younger male.

a)

We wish to write up the hazard for the survival time of a female living 4 years longer, that is $T_i^f + 4$. We observe that

$$P(T + 4 > t) = P(T > t - 4) = S(t-4) = \frac{f(t-4)}{\alpha(t-4)}$$

Since there is a one-one relation between the survival function and the hazard we can find the hazard for $T_i^f + 4$ by inserting $t - 4$ in the hazard we already know. Thus we obtain

$$\alpha_i^f(t-4, X_i^f(t-4), Z_i^f(t-4)) = (X_i^f(t-4)^T\beta(t-8) + h(\gamma^T Z_i^f(t-4))$$

b)

In order to establish an estimating equation of $B(t)$ given γ, we define some new processes. Hence we can find a common intensity for both males and females. We let $n_i, i = 1, .., n_f + n_m$ denote the i'th subject and g_i denote the gender of the i'th subject. We define

$$\tilde{X}_i^{n_i}(t) = \begin{cases} X_i^f(t+4) & \text{for } g_i = f \\ X_i^m(t) & \text{for } g_i = m \end{cases}, \qquad \tilde{Y}_i^{n_i}(t) = \begin{cases} Y_i^f(t+4) & \text{for } g_i = f \\ Y_i^m(t) & \text{for } g_i = m \end{cases},$$

$$\tilde{Z}_i^{n_i}(t) = \begin{cases} Zi^f(t+4) & \text{for } g_i = f \\ Z_i^m(t) & \text{for } g_i = m \end{cases}, \qquad \tilde{\lambda}_i^{n_i}(t) = \begin{cases} \lambda_i^f(t+4) & \text{for } g_i = f \\ \lambda_i^m(t) & \text{for } g_i = m \end{cases}.$$

We can then write the overall intensity as

$$\tilde{\lambda}_i^{n_i}(t) = \tilde{Y}_i^{n_i}(t)\left(\tilde{X}_i^{n_i}(t)^T\beta(t) + h\left(\gamma^T\tilde{Z}_i^{n_i}(t)\right)\right).$$

We use the same approach as on page 72 in the book, hence we consider all sub-models on the form $\beta(t) = \beta_0(t) + \eta b(t)$, where η is an one-dimensional parameter and b is a given p-vector of functions. The estimating equation of $B(t)$ given γ becomes

$$U(\eta, t) = \sum_{i=1}^{n}\int_0^t \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T b(s)\left(dN_i^{n_i}(s) - \tilde{Y}_i^{n_i}(s)\left(\tilde{X}_i^{n_i}(s)^T\beta(s) + h\left(\gamma^T\tilde{Z}_i^{n_i}(s)\right)\right)ds\right) = 0$$

This must be equal to zero for every choice of $b(t)$. For this to hold the increments need to be zero such that

$$\sum_{i=1}^{n} \tilde{Y}_i^{n_i}(t)(\tilde{X}_i^{n_i}(t))^T \left(dN_i^{n_i}(t) - \tilde{Y}_i^{n_i}(t)\left(\tilde{X}_i^{n_i}(t)^T\beta(t) + h\left(\gamma^T\tilde{Z}_i^{n_i}(t)\right)\right)dt\right) = 0 \qquad \Leftrightarrow$$

$$\sum_{i=1}^{n} \tilde{Y}_i^{n_i}(t)(\tilde{X}_i^{n_i}(t))^T \; dN_i^{n_i}(t) - \tilde{Y}_i^{n_i}(\tilde{X}_i^{n_i}(t))^T\tilde{X}_i^{n_i}(t)\beta(t)dt - \tilde{Y}_i^{n_i}(t)(\tilde{X}_i^{n_i}(t))^T h\left(\gamma^T\tilde{Z}_i^{n_i}(t)\right)dt = 0 \qquad \Leftrightarrow$$

$$\sum_{i=1}^{n} \tilde{Y}_i^{n_i}(t)(\tilde{X}_i^{n_i}(t))^T \; dB(t) = \sum_{i=1}^{n} \tilde{Y}_i^{n_i}(t)(\tilde{X}_i^{n_i})^T dN_i^{n_i}(t) - \tilde{Y}_i^{g_i}(\tilde{X}_i^{n_i})^T h\left(\gamma^T\tilde{Z}_i^{n_i}(t)\right)dt \qquad \Leftrightarrow$$

$$d\hat{B}(t) = \left(\sum_{i=1}^{n} \tilde{Y}_i^{n_i}(t)(\tilde{X}_i^{n_i}(t))^T\tilde{X}_i^{n_i}(t)\right)^{-1} \sum_{i=1}^{n} \tilde{Y}_i^{n_i}(t)(\tilde{X}_i^{n_i}(t))^T dN_i^{n_i}(t) - \tilde{Y}_i^{n_i}(t)(\tilde{X}_i^{n_i})^T h\left(\gamma^T\tilde{Z}_i^{n_i}(t)\right)dt \qquad \Leftrightarrow$$

$$\hat{B}(t) = \int_0^t \left(\sum_{i=1}^{n} \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T\tilde{X}_i^{n_i}(s)\right)^{-1} \sum_{i=1}^{n} \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \left(dN_i^{n_i}(s) - h\left(\gamma^T\tilde{Z}_i^{n_i}(s)\right)\right)ds.$$

c)

Now we wish to write up an estimating equation for γ and afterwards profile out $B(t)$. Since γ is time invariant we can use equation (3.19). This yields

$$U(\gamma, t) = \sum_{i=1}^{n}\int_0^t \frac{\partial}{\partial\gamma}\tilde{\lambda}_i^{n_i}(s)(dN_i^{n_i}(s) - \tilde{\lambda}_i^{n_i}(s)\; ds) = 0 \qquad \Leftrightarrow$$

$$\sum_{i=1}^{n}\int_0^t \tilde{Y}_i^{n_i}(s)h'\left(\gamma^T\tilde{Z}_i^{n_i}(s)\right)\tilde{Z}_i^{n_i}(s)\left(dN_i^{n_i}(s) - \tilde{Y}_i^{n_i}(s)\left((\tilde{X}_i^{n_i}(s))^T\beta(s) + h\left(\gamma^T\tilde{Z}_i^{n_i}(s)\right)ds\right)\right) = 0 \qquad \Leftrightarrow$$

$$\sum_{i=1}^{n}\int_0^t \tilde{Y}_i^{n_i}(s)h'\left(\gamma^T\tilde{Z}_i^{n_i}(s)\right)\tilde{Z}_i^{n_i}(s)\left(dN_i^{n_i}(s) - \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T\beta(s) - \tilde{Y}_i^{n_i}(s)h\left(\gamma^T\tilde{Z}_i^{n_i}(s)\right)ds\right) = 0 \qquad \Leftrightarrow$$

$$\sum_{i=1}^{n}\int_0^t \tilde{Y}_i^{n_i}(s)h'\left(\gamma^T\tilde{Z}_i^{n_i}(s)\right)\tilde{Z}_i^{n_i}(s)\left(dN_i^{n_i}(s) - \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T d\hat{B}(s) - \tilde{Y}_i^{n_i}(s)h\left(\gamma^T\tilde{Z}_i^{n_i}(s)\right)ds\right) = 0.$$

And $\hat{\gamma}$ is then the γ that solves this equation.

d1)

We will give some arguments on how to estimate γ and how to establish that $\sqrt{n}(\hat{\gamma} - \gamma)$ is asymptotically normal.
If we know the smooth link function h we can solve the estimating equation for γ from c). We thus obtain an estimate for γ, which does not depend on $B(t)$, since this has been profiled out. To investigate the asymptotic behaviour of $\sqrt{n}(\hat{\gamma}-\gamma)$ we could use the Doob-Meyer decomposition $dN_i^{n_i}(t) = d\Lambda_i^{n_i}(t) + dM_i^{n_i}(t)$. Hopefully we will end up with a martingale plus some other term which converges towards something nice. We can then, under certain regularity conditions, use the Martingale Central Limit Theorem. This gives convergence in distribution towards some Gaussian martingale. The Martingale CLT furthermore provides that the covariance function can be estimated by the quadratic variation process.

d2)

Now we will give some arguments on how to estimate the cumulatives $B(t)$ and that $\sqrt{n}(\hat{B}(t) - B(t))$ is asymptotically normal.

If we have found an estimate of γ, perhaps by the method described in d), we can obtain an estimate of $\hat{B}$. This is done by inserting the estimate $\hat{\gamma}$ in the estimating equation from b).
Now we look at the asymptotic behaviour of $\sqrt{n}(\hat{B}(t) - B(t))$. Again we will use the Doob-Meyer decomposition.

$$\sqrt{n}(\hat{B}(t) - B(t)) =$$

$$\sqrt{n}\left(\int_0^t \left(\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \tilde{X}_i^{n_i}(s)\right)^{-1} \sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \left(dN_i^{n_i}(s) - h\left(\gamma^T \tilde{Z}_i^{n_i}(s)\right) ds\right) - B(t)\right) =$$

$$\sqrt{n}\left(\int_0^t \left(\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \tilde{X}_i^{n_i}(s)\right)^{-1} \sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \left(d\Lambda_i^{n_i}(s) + dM_i^{n_i}(s) - h\left(\gamma^T \tilde{Z}_i^{n_i}(s)\right) ds\right) - B(t)\right) =$$

$$\sqrt{n}\left(\int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \; d\Lambda_i^{n_i}(s)}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \tilde{X}_i^{n_i}(s)} + \int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \; dM_i^{n_i}(s)}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \tilde{X}_i^{n_i}(s)}\right)$$

$$- \sqrt{n}\left(\int_0^t \frac{\sum_{i=1}^n h\left(\gamma^T \tilde{Z}_i^{n_i}(s)\right) \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \; ds}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \tilde{X}_i^{n_i}(s)} - B(t)\right)$$

We look at the first term

$$\int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \; d\Lambda_i^{n_i}(s)}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \tilde{X}_i^{n_i}(s)} = \int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \tilde{Y}_i^{n_i}(s) \left((\tilde{X}_i^{n_i}(s))^T \beta(s) + h\left(\gamma^T \tilde{Z}_i^{n_i}(s)\right)\right) ds}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^{\otimes 2}} =$$

$$\int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)\tilde{X}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T \beta(s) \; ds}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^{\otimes 2}} + \int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T h\left(\gamma^T \tilde{Z}_i^{n_i}(s)\right) \; ds}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^{\otimes 2}} =$$

$$\int_0^t \beta(s) \; ds + \int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T h\left(\gamma^T \tilde{Z}_i^{n_i}(s)\right) \; ds}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^{\otimes 2}} = B(t) + \int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T h\left(\gamma^T \tilde{Z}_i^{n_i}(s)\right) \; ds}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^{\otimes 2}}$$

Hence

$$\sqrt{n}(\hat{B}(t) - B(t)) = \sqrt{n} \int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T dM_i^{n_i}(s)}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^{\otimes 2}}$$

Since $\tilde{Y}_i^{n_i}$ and $\tilde{X}_i^{n_i}(t)$ both are predictable the term above is a martingale. Under certain regularity conditions the Martingale CLT provides that

$$\sqrt{n}(\hat{B}(t) - B(t)) \overset{\mathcal{D}}{\to} U,$$

where $U \sim \mathcal{N}(0, V(t))$ and $V(t)$ can be estimated by the quadratic variation process

$$\left[M^{(n)}(s)\right] = \left[\sqrt{n} \int_0^t \frac{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^T dM_i^{n_i}(s)}{\sum_{i=1}^n \tilde{Y}_i^{n_i}(s)(\tilde{X}_i^{n_i}(s))^{\otimes 2}}\right]$$

d3)

In order to answer the question of, for which t we can estimate $B(t)$ on the data, we first need to clearify on which scale we look at t. You can either see t as the subject's ages ($t \in [30, 90]$), or you can see t as time after enrollment in the study / follow-up time ($t \in [0, 60]$). Here we will consider t as the time after each individuals enrollment in the study. We also have to remember that a female is estimated as a 4 year younger male. The t's for which we can estimate $B(t)$ from the data depends on the subjects gender. For the males we can estimate their hazards for $t \in [0, 60]$ but for the females this is only possible for $t \in [0, 56]$. In order to estimate the hazard for the females at time $t = 57, 58, 59, 60$ we would need them to be observed until age 94, which is not possible in this study. Therefore it is possible to estimate $B(t)$ for both males and females for $t \in [0, 56]$.

e)

We wish to find the predicted survival probability after 10 years, of a male subject with covariates (X_0, Z_0) that enters the study at 30 years of age. We let T_0 be the subjects survival time after enrollment in the study and we assume the covariates X_0 and Z_0 to be time invariant. We let $\alpha(t, X, Z)$ denote the hazard of this male. We find

$$P\left(T_0 > 10 | X = X_0, Z = Z_0, T_0 > 0\right) = \exp\left(-\int_0^{10} \alpha(t|X = X_0, Z = Z_0)\ dt\right)$$

$$= \exp\left(-\int_0^{10} X_0^T \hat{\beta}(t) + h\left(\gamma^T Z_0\right)\ dt\right) = \exp\left(-\int_0^{10} X_0^T d\hat{B}(t)\right) \exp\left(-\int_0^{10} h\left(\gamma^T Z_0\right) dt\right)$$

$$= \exp\left(-X_0^T \int_0^{10} d\hat{B}(t)\right) \exp\left(-h\left(\gamma^T Z_0\right) \int_0^{10} 1\ dt\right) = \exp\left(-X_0^T \hat{B}(10)\right) \exp\left(-10 \cdot h\left(\gamma^T Z_0\right)\right)$$

In the previous problems d1) and d2) we argued how to estimate the covariance function of $\hat{B}$ and $\hat{\gamma}$. By theorem 5.3.1 and the Delta Method it can beshown that $\sqrt{n}(\hat{S}_0 - S_0)$ converges towards a zero mean Gaussian process U with some variance function Q. We could do a Taylor expansion and then we would end up with an expression depending on the true parameter γ. We could then insert the estimate of γ, since we know the asymptotic behaviour of this and hopefully we would end up with some martingale decomposition. We could then find an optional variance process estimator of its variance, denoted $\hat{Q}(t)$. The variance function Q can then be estimated by

$$\tilde{Q}(t) = \hat{S}_0^2 \hat{Q}(t).$$

An alternative way of finding standard errors for the estimate is to use resampling. This is probably easier to implement in practice.

f)

If females in the study are recruited at different ages we are looking at a case of left truncation. Now we need to take into account that the females are at-risk at a different amount of time. We can include this in our analysis if we change the at-risk process to the following

$$Y_i = \mathbb{1}(T_i > t, V_i < t)$$

Here V_i is the time you are recruited to the study, a truncation time. From this formulation you are only at risk if you have been recruited to the study, such that the individuals are at risk from their (individual) truncation time and onwards.

1.2 Part B

Now we consider the hazard

$$\lambda(t) = \beta(t) + X\alpha(t + A),$$

for $t \in [0, \tau]$ given stochastic variables X and A. X is a binary covariate where $P(X = 1) = P(X = 0) = \frac{1}{2}$, and A is a stochastic variable with density $f(a)$. We assume that X and A are independent.

a)

We wish to find the hazard function given only X, that is when X is observed and A is not observed. First, we note that the distribution of A is

$$F(A) = \int_0^a f(s)ds = \int_0^a dF(s).$$

We start by finding the Survival function given the covariate X.

$$S(t|X) = P(T > t|X = x) = E(P(T > t|X = x, A = a)|X = x) = \int_0^\infty P(T > t|X = x, A = a)dF(A|X)$$

$$= \int_0^\infty P(T > t|X = x, A = a)dF(A) = \int_0^\infty \exp\left(-\int_0^t \beta(s) + x\alpha(s+a)ds\right) dF(a)$$

$$= \int_0^\infty \exp\left(-\int_0^t \beta(s)ds\right) \cdot \exp\left(-\int_0^t x\alpha(s+a)ds\right) dF(a)$$

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

We then find the hazard in the following way

$$\lambda(t|X) = \frac{-\partial}{\partial t}\log(S(t|X)) = \frac{\int_0^\infty \frac{-\partial}{\partial t} \exp\left(-\int_0^t \beta(s) + x\alpha(s+a)ds\right) dF(a)}{\int_0^\infty \exp\left(-\int_0^t \beta(s) + x\alpha(s+a)ds\right) dF(a)}$$

$$= \frac{\int_0^\infty \exp\left(-\int_0^t \beta(s) + x\alpha(s+a)ds\right) (\beta(t) + x\alpha(t+a)) f(a)da}{\int_0^\infty \exp\left(-\int_0^t \beta(s) + x\alpha(s+a)ds\right) f(a)da}$$

$$= \frac{\int_0^\infty \exp\left(-\int_0^t \beta(s) + x\alpha(s+a)ds\right) \beta(t)f(a)da + \int_0^\infty \exp\left(-\int_0^t \beta(s) + x\alpha(s+a)ds\right) x\alpha(t+a)f(a)da}{\int_0^\infty \exp\left(-\int_0^t \beta(s) + x\alpha(s+a)ds\right) f(a)da}$$

$$= \beta(t) + \frac{\exp\left(-\int_0^t \beta(s)ds\right) \cdot \int_0^\infty \exp\left(-x\int_0^t \alpha(s+a)ds\right) x\alpha(t+a)f(a)da}{\exp\left(-\int_0^t \beta(s)ds\right) \int_0^\infty \exp\left(-x\int_0^t \alpha(s+a)ds\right) f(a)da}$$

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

b)

We now wish to find the hazard function when both X and A are not observed.
Again we start by writing up the Survival function in the situation where none of the variables are observed.

$$\begin{aligned}
S(t) = P(T > t) &= \int_0^\infty \int_0^\infty P(T > t|X, A) dF(A) dX(P) = \int_0^\infty S(t|X) dX(P) \\
&= S(t|X = 0) \cdot P(X = 0) + S(t|X = 1) \cdot P(X = 1) = \frac{1}{2} S(t|X = 0) + \frac{1}{2} S(t|X = 1) \\
&= \frac{1}{2} \cdot \int_0^\infty \exp\left(-\int_0^t \beta(s) ds\right) f(a) da + \frac{1}{2} \cdot \int_0^\infty \exp\left(-\int_0^t \beta(s) ds\right) \cdot \exp\left(-\int_0^t \alpha(s + a) ds\right) f(a) da \\
&= \frac{1}{2} \exp\left(-B(t)\right) \int_0^\infty f(a) da + \frac{1}{2} \exp\left(-B(t)\right) \int_0^\infty \exp\left(-\int_0^t \alpha(s + a) ds\right) f(a) da \\
&= \frac{1}{2} \exp(-B(t)) + \frac{1}{2} \exp(-B(t)) \int_0^\infty \exp\left(-\int_0^t \alpha(s + a) ds\right) f(a) da.
\end{aligned}$$

Hence the hazard becomes

$$\begin{aligned}
\lambda(t) = \frac{-\partial}{\partial t} \log(S(t)) &= \frac{\frac{-\partial}{\partial} \frac{1}{2} \exp(-B(t)) + \int_0^\infty \frac{-\partial}{\partial t} \frac{1}{2} \exp\left(-\int_0^t \beta(s) + \alpha(s + a) ds\right) f(a) da}{\frac{1}{2} \exp(-B(t)) + \frac{1}{2} \int_0^\infty \exp\left(-\int_0^t \beta(s) ds + \alpha(s + a) ds\right) f(a) da} \\
&= \frac{\frac{1}{2} \exp\left(-\int_0^t \beta(s) ds\right) \beta(t) + \frac{1}{2} \int_0^\infty \exp\left(-\int_0^t \beta(s) + \alpha(s + a) ds)(\beta(t) + \alpha(t + a)\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} \\
&= \frac{\exp\left(-\int_0^t \beta(s) ds\right) \cdot \beta(t) + \exp\left(-\int_0^t \beta(s) ds\right) \int_0^\infty \exp\left(-\int_0^t \alpha(s + a) ds\right) (\beta(t) + \alpha(t + a)) f(a) da}{\exp\left(-\int_0^t \beta(s) ds\right) + \exp\left(-\int_0^t \beta(s) ds\right) \int_0^\infty \exp\left(-\int_0^t \alpha(s + a) ds\right) f(a) da} \\
&= \frac{\beta(t) + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) (\beta(t) + \alpha(t + a)) f(a)\ da}{1 + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a) ds\right) f(a)\ da} \\
&= \frac{\beta(t) + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\right) ds\ \beta(t) f(a) da + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\right) ds\ \alpha(t + a) f(a) da}{1 + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) f(a)\ da} \\
&= \frac{\beta(t) + \beta(t) \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) f(a)\ da + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) \alpha(t + a) f(a)\ da}{1 + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) f(a)\ da} \\
&= \frac{\beta(t) \left(1 + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) f(a)\ da + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) \alpha(t + a) f(a)\ da\right)}{1 + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) f(a)\ da} \\
&= \beta(t) + \frac{\int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) \alpha(t + a) f(a)\ da}{1 + \int_0^\infty \exp\left(-\int_0^t \alpha(s + a)\ ds\right) f(a)\ da}
\end{aligned}$$

2 Practical Part

We consider a data-set on 423 first pregnancy planners. We wish to study risk-factors that may affect the ”time to pregnancy”. We follow 423 couples for 6 months until conception was achieved. We consider the following risk factors

`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)

First we note that there is 113 observations with missing values in the dataset. Furthermore we note that there is a problem with observation 96, hence this is removed from the data. We make the analysis based on the 309 observations, which does not have missing values. We read in data.

```
ttp_data <- read.xport("ttp.xpt")
names(ttp_data) <- tolower(names(ttp_data))
ttp_data <- ttp_data[-96,]
```

1)

Before we begin our goodness-of-fit analysis we wish to investigate whether or not to include interactions in our model. We have four explanatory variables, which means that we have four-way, three-way and two-way interactions. In order to actually interpret the model we choose to exclude three-way and four-way interactions. We then have 6 two-way interactions left. This will still give a quite large model and we don’t want to include them all. In order to decide which interactions to include we try to use our intuition. It is sensible to imagine that smoking status of the male and female could have some effect on each other, hence we include the interaction between `k_ryg` and `m_ryg`. Furthermore you could imagine that the sperm concentration of the male and the smoking status of the male could interact, hence we also include this interaction between `m_zkon0` and `m_ryg`. At last we find it possible that the females caffeine intake and smoking status could have an effect upon each other, hence we include the interaction betweem `k_cof` and `k_ryg`. This is the three interaction terms we consider in our model.

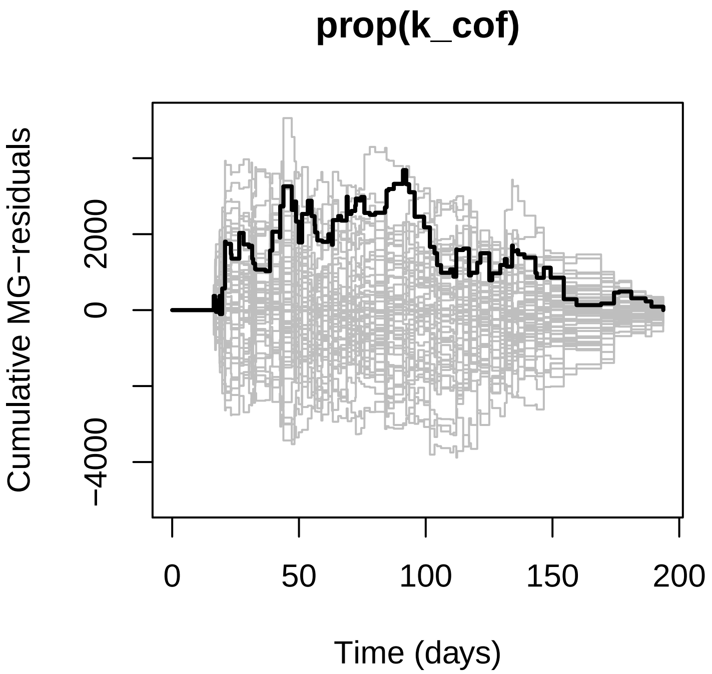
We fit a model including all interactions and remove these one at a time and find that all of them are redundant, hence they are excluded from the model. We see that some of the fixed effects also are non-significant but since these are the risk factors we will investigate, we are not interested in excluding these. We thus end up with an additive model with four fixed effects, two categorical variables and two continuous covariates.

Now we wish to do a goodness-of-fit analysis of the chosen model and start by checking the proportional hazards assumption. If the assumption is satisfied the hazard ratio will not depend on time. We fit a cox regression model using the `cox.aalen` function for each of the four variables, one at a time.

First we do it for the variable `k_cof`.

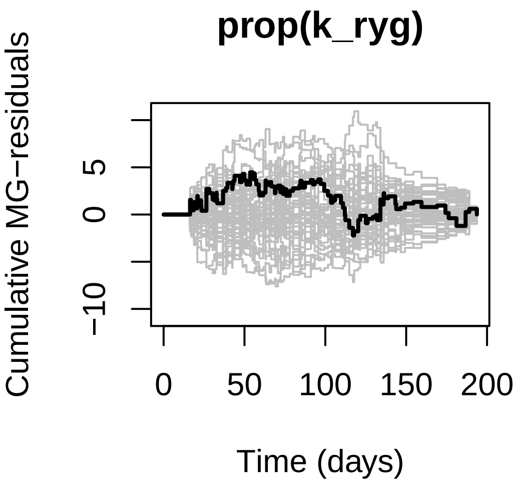
```
cox_reg_kcof <- cox.aalen(Surv(ttp,k_gravid)~prop(k_cof), data=ttp_data, weighted.test=0)
summary(cox_reg_kcof)

## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
```

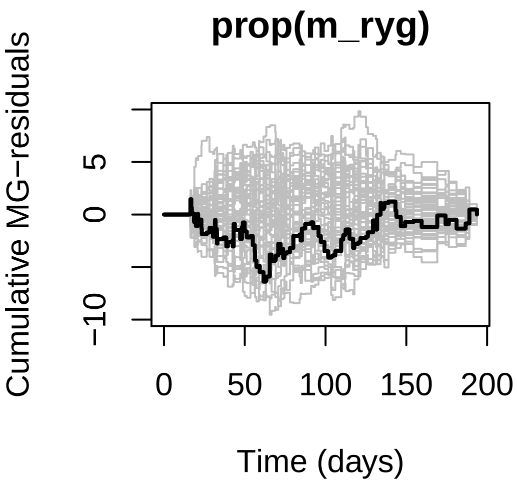
```
##             Coef. SE Robust SE D2log(L)^-1      z P-val
## prop(k_cof)     0  0           0         0 -0.143 0.887
## Test for Proportionality
##             sup|  hat U(t) | p-value H_0
## prop(k_cof)             3690        0.156

plot(cox_reg_kcof, score=T, xlab="Time (days)")
```

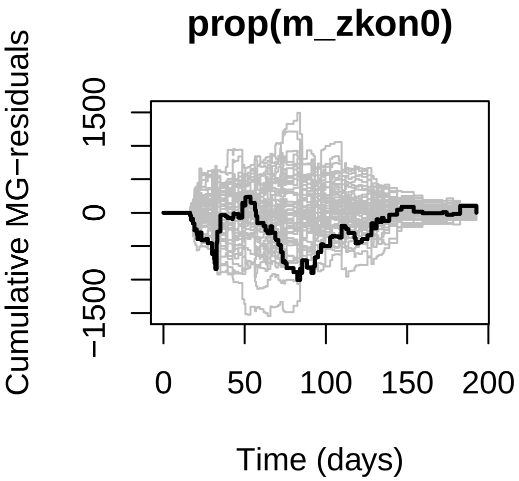


In the summary we see that the p-value under `Test for proportionality` is high (> 0.05) for the unweighted test. It seems like the proportional hazards assumption is satisfied for the covariate `k_cof`. To further analyze the hypothesis we have illustrated the score process (black) to 50 random processes simulated under the proportional hazards assumption (gray). If the assumption is fulfilled, the black curve should look like the simulated gray ones, which it does in this case.

We now do the same for the rest of the variables. The 3 plots are shown below.





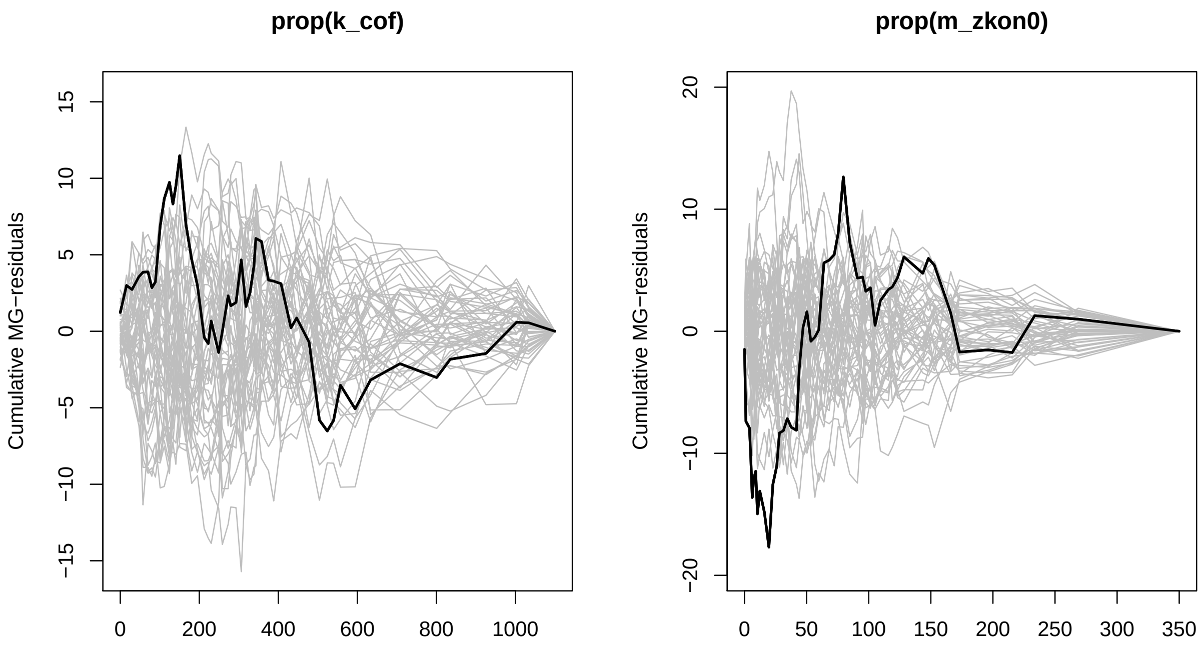


Both the plots and the summaries agree upon the hazard assumption being fulfilled for all 4 variables. We could alse compute a weighted version of the supremum test statistic, where the variance of the score process is taken into account. This test leads to the same conclusion as the unweighted test statistic. So we can conclude that the proportional hazards assumption is fulfilled.

We will now use the cumulative martingale residuals to investigate if the functional form of the continuous covariates are misspecified and thus should be changed. We do this by fitting the Cox model with the four fixed effects. The categorical variables will here be ignored.

```
cox_reg_func_1 <- cox.aalen(Surv(ttp,k_gravid)~prop(k_cof)+prop(k_ryg)+prop(m_ryg)+
                            prop(m_zkon0), data=ttp_data, residuals=1)
resids_func_1 <- cum.residuals(cox_reg_func_1, ttp_data, cum.resid=1)
summary(resids_func_1)

## 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
##            11.479               0.162
##            17.688               0.004

par(mfrow=c(1,2))
plot(resids_func_1, score=2, main=c("prop(k_cof)", "prop(m_zkon0)"))
```

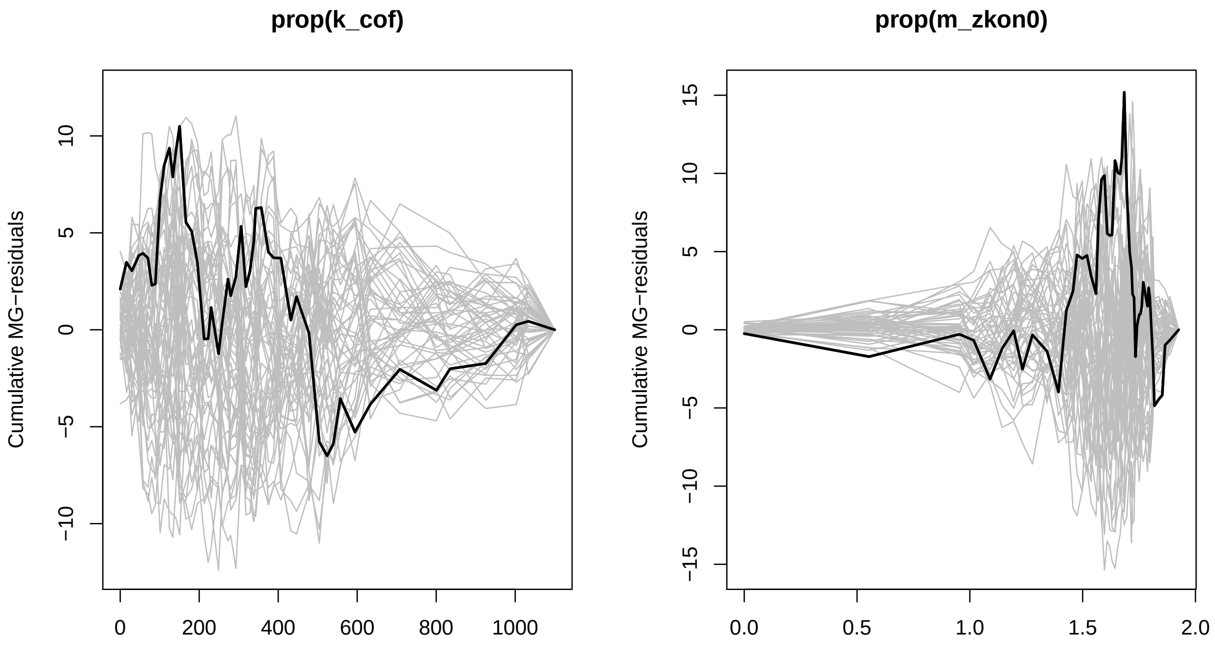


We see that the p-value from the summary suggests that `k_cof` can be included on its original scale. The plot supports this conclusion, since the observed black curve does not look extreme compared to the simulated ones. We find a p-value of < 0.05 with respect to `m_zkon0`, suggesting that this covariate should not be included on its original scale. From the plot we see that the observed curve is quite different from the simulated ones, in the beggining of the time period. Again supporting the conclusion that this covariate should be transformed. Thus the functional representation of `m_zkon0` (male sperm concentration) does not seem to be sensible.

We try to include `m_zkon0` in the model on log-log scale. We add 1 in each log, to avoid taking the logarithm of zero. Since the functional form of `k_cof` seems to be sensible enough we do not transform this.

```
ttp_data$loglogm_zkon0 <- log(log(ttp_data$m_zkon0+1)+1)
cox_reg_func_2 <- cox.aalen(Surv(ttp,k_gravid)~prop(k_cof)+prop(k_ryg)+prop(m_ryg)+
                              prop(loglogm_zkon0), data=ttp_data, weighted.test=1, residuals=1)
resids_func_2 <- cum.residuals(cox_reg_func_2, ttp_data, cum.resid=1)
summary(resids_func_2)

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

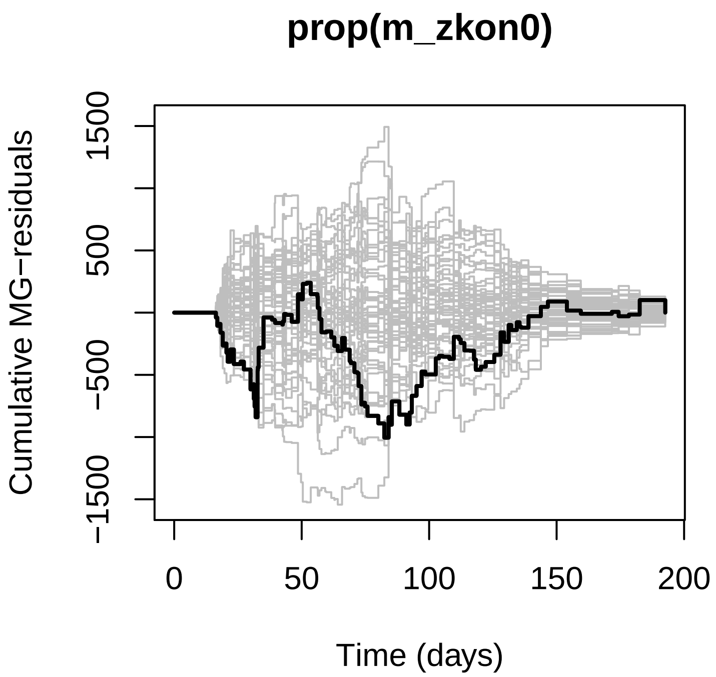
```
##
## Residual versus covariates consistent with model
##
##   sup|  hat B(t) | p-value H_0: B(t)=0
##            10.497                0.204
##            15.192                0.036

par(mfrow=c(1,2))
plot(resids_func_2, score=2, main=c("prop(k_cof)", "prop(m_zkon0)"))
```



Both the summary and the plot suggest that the functional form of the covariate m_zkon0 still are not acceptable, but we are not interested in making some wild transformation. We therefore keep the log-log transformation. Since we changed the functional form of the variable m_zkon0, we will re-check the proportional hazards assumption for this variable.

```
cox_reg_mzkon0_2 <- cox.aalen(Surv(ttp,k_gravid)~prop(loglogm_zkon0), data=ttp_data,
                              weighted.test=1)
summary(cox_reg_mzkon0)

## 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(m_zkon0) 0.004 0.001     0.001       0.001 4.44     0
## Test for Proportionality
##               sup|  hat U(t) | p-value H_0
## prop(m_zkon0)              1010        0.122
```

```
plot(cox_reg_mzkon0,score=T, xlab="Time (days)")
```



When we use the transformed variable the proportional hazards assumption is still fulfilled. The model we end up with assumes that the intensity is of the form

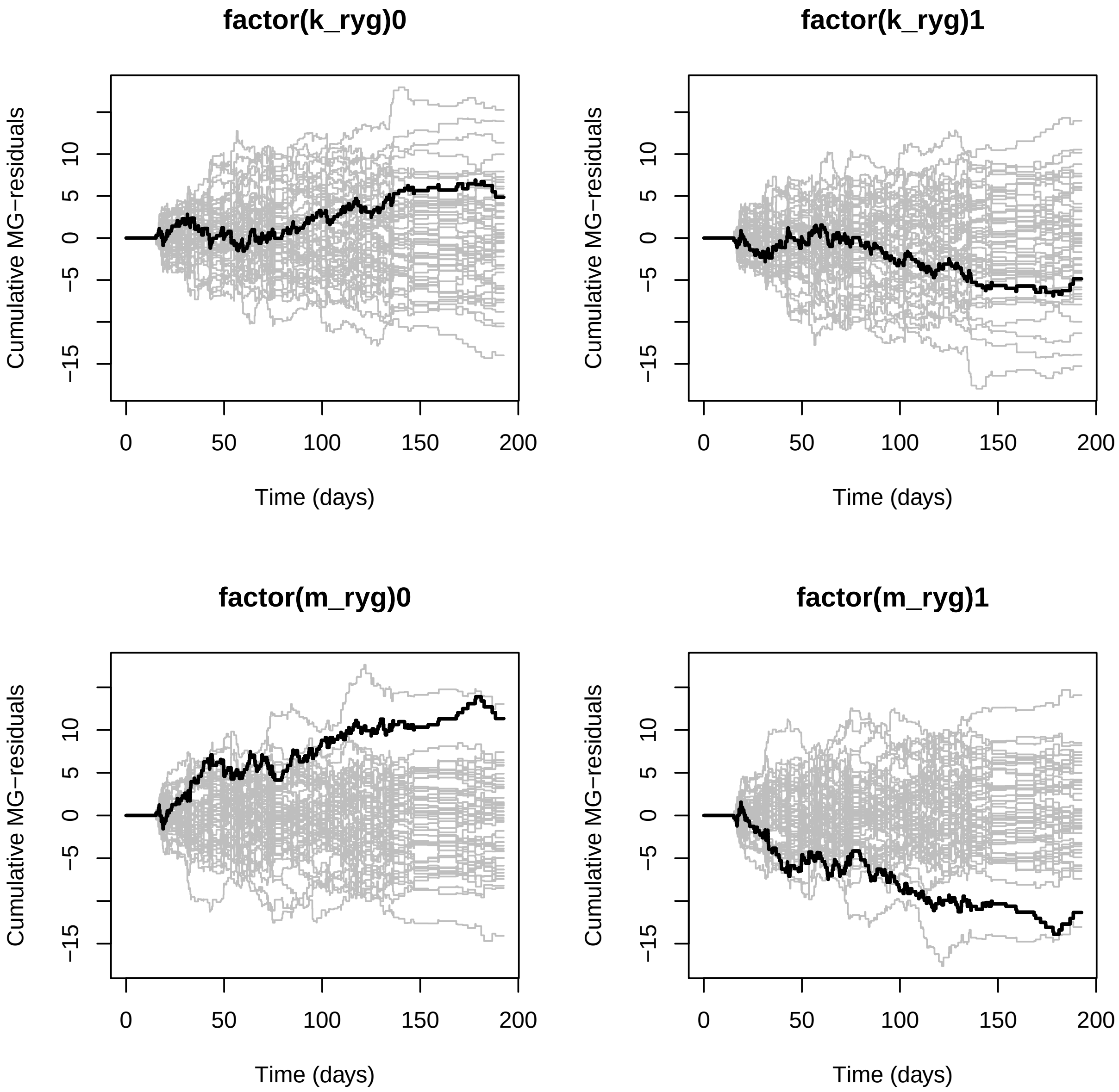
$$\lambda(t) = Y(t)\lambda_0(t)\exp(X^T\beta),$$

where X is a 4-dimensional vector containing the covariates and $Y(t)$ is the at risk indicator.

2)

Now we wish to fit an additive hazards model. We still ignore every possible interaction and focus on the fixed effects. So we consider a model with the four risk factors included in an additive way. We need to check that the model provides an adequate fit to the data. To evaluate the fit of the model we consider various martingale residuals and investigate if they behave consistently with what would be expected under the model. We start by evaluating the fit of the categorical variables.

```
par(mfrow=c(2,2))
aalen_1 <- aalen(Surv(ttp,k_gravid)~k_cof+k_ryg+m_ryg+m_zkon0, data=ttp_data, residuals=1)
X_k <- model.matrix(~-1+factor(k_ryg),ttp_data)
resids_k <- cum.residuals(aalen_1,ttp_data,X_k, n.sim=1000)
X_m <- model.matrix(~-1+factor(m_ryg),ttp_data)
resids_m <- cum.residuals(aalen_1,ttp_data,X_m, n.sim=1000)

plot(resids_k, score=1, xlab="Time (days)")
plot(resids_m, score=1, xlab="Time (days)")
```

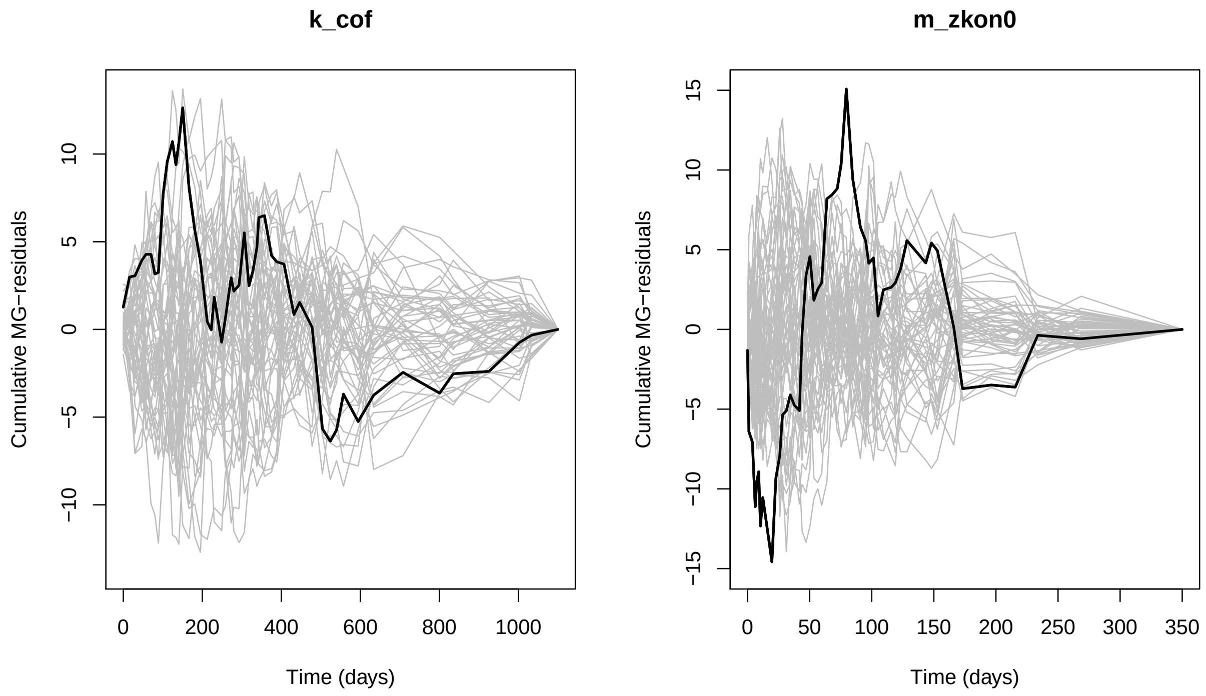


```
summary(resids_k)

## 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            6.904               0.572
## factor(k_ryg)1            6.904               0.572
##
##                int ( B(t) )^2 dt p-value H_0: B(t)=0
## factor(k_ryg)0          2617.565                0.55
## factor(k_ryg)1          2617.565                0.55
##
```

```
## Residual versus covariates consistent with model
##
##  sup|  hat B(t) | p-value H_0: B(t)=0
##            12.638               0.067
##            15.076               0.011

summary(resids_m)

## 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           13.906               0.077
## factor(m_ryg)1           13.906               0.077
##
##                 int ( B(t) )^2 dt p-value H_0: B(t)=0
## factor(m_ryg)0          13601.46               0.094
## factor(m_ryg)1          13601.46               0.094
##
## Residual versus covariates consistent with model
##
##  sup|  hat B(t) | p-value H_0: B(t)=0
##            12.638               0.078
##            15.076               0.008
```

The above plots show the cumulative test processes with 50 random simulations under the null for the two categorical variables `m_ryg` and `k_ryg`. We see that the observed process and the simulated ones fluctuate around 0, suggesting a good model fit with respect to the categorical variables.
We now want to look at the continuous variables. We compute the cumulative residuals versus the covariates.

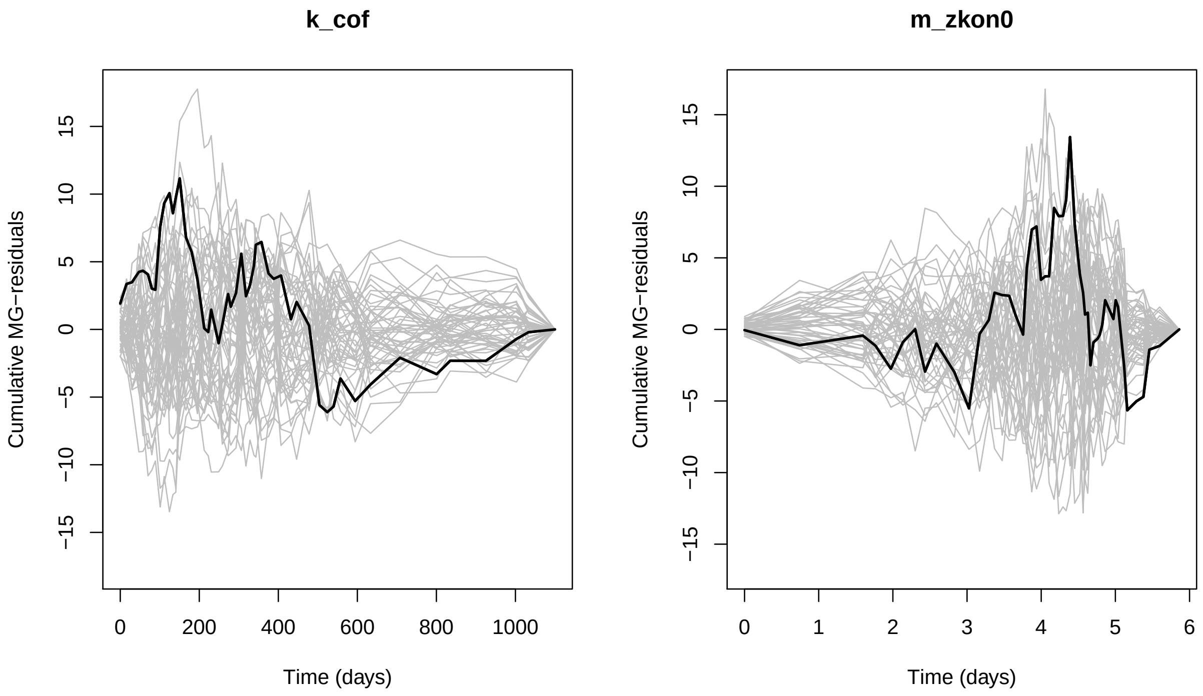
```
par(mfrow=c(1,2))
resids_aalen_1 <- cum.residuals(aalen_1, ttp_data, cum.resid=1)
plot(resids_aalen_1, score=2, main=c("k_cof", "m_zkon0"), xlab="Time (days)")
```



```
summary(resids_aalen_1)

## 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
##            12.638                0.068
##            15.076                0.012
```

Both the summary and the plot show that the performance of the covariate `m_konz0` is not consistent with the one expected under the model. This means that the variable are not linear on the scale where it is included. We therefore wish to transform this continuous covariate. The plot indicates that there is no problem with the covariate `k_cof`, so we choose to keep it on its original scale. We transform `m_zkon0` on log scale and add a +1 to avoid taking log of zero.

```
par(mfrow=c(1,2))
ttp_data$logm_zkon0 <- log(ttp_data$m_zkon0+1)
aalen_2<- aalen(Surv(ttp,k_gravid)~k_cof+k_ryg+m_ryg+logm_zkon0, data=ttp_data, residuals=1)
resids_aalen_2 <- cum.residuals(aalen_2, ttp_data, cum.resid=1)
plot(resids_aalen_2, score=2, main=c("k_cof", "m_zkon0"), xlab="Time (days)")
```



```
summary(resids_aalen_2)

## 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
##            11.157               0.152
##            13.436               0.032
```

Both the plot and the p-value suggest a small problem even after the transformation. But the deviation are not dramatic and the p-value are in some of the simulations non-significant. We therefore choose to use this transformation. Thus `k_cof` enters the model linearly on log scale.

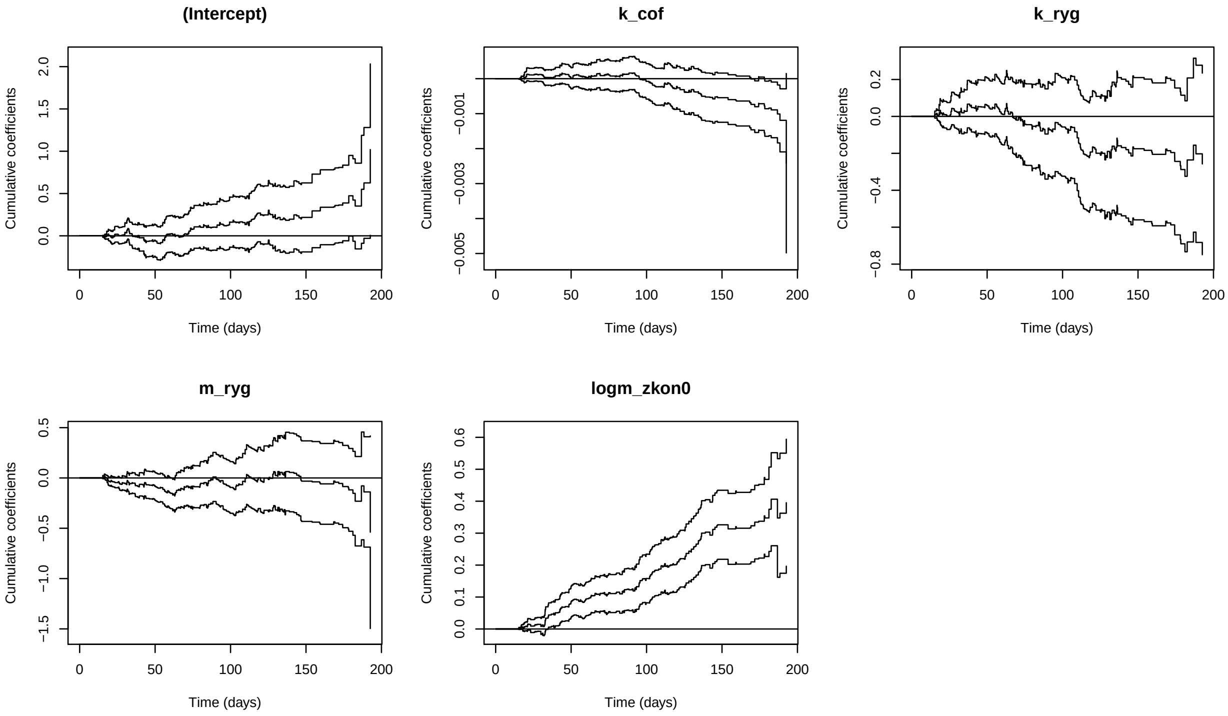
We have established Aalens additive hazards model with all components having timevarying effects. We will now simplify the model by a number of successive tests. First we investigate if the covariates are time invariant. Hence we test the hypothesis

$$H_0 : \beta_j(t) \equiv \gamma,$$

where γ is some constant not depending on time.

```
par(mfrow=c(2,3))
aalen_3<- aalen(Surv(ttp,k_gravid)~k_cof+k_ryg+m_ryg+logm_zkon0, data=ttp_data)
summary(aalen_3)

## Additive Aalen Model
```

```
##
## Test for nonparametric terms
##
## Test for non-significant effects
##             Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                          1.96               0.514
## k_cof                                2.52               0.192
## k_ryg                                1.47               0.827
## m_ryg                                2.20               0.357
## logm_zkon0                           6.16               0.000
##
## Test for time invariant effects
##                   Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                       0.61500                       0.219
## k_cof                             0.00148                       0.228
## k_ryg                             0.15400                       0.749
## m_ryg                             0.44500                       0.346
## logm_zkon0                        0.06800                       0.781
##                     Cramer von Mises test p-value H_0:constant effect
## (Intercept)                      2.87e+01                       0.145
## k_cof                            2.15e-04                       0.111
## k_ryg                            8.12e-01                       0.738
## m_ryg                            1.40e+01                       0.258
## logm_zkon0                       2.15e-01                       0.692
##
##
##
##   Call:
## aalen(formula = Surv(ttp, k_gravid) ~ k_cof + k_ryg + m_ryg +
##     logm_zkon0, data = ttp_data)

plot(aalen_3, xlab="Time (days)")
```



We see from the summary that both the Kolmogorov-Smirnov and the Cramer von Mises test for time invariant effects have p-values above 0.05 for all 4 variables. Thus the hypothesis of the variables being time invariant are accepted. We can therefore assume constant covariates. This is consistent with the plot where the cumulative estimates are approximately straight lines, expect in the end of the time period where there seems to be a small problem.
The reduced model with all effects being constant has intensity of the form

$$\lambda_i(t) = Y_i(t)\left(X_i^T\beta\right),$$

where X is a 4-dimensional vector containing our covariates and Y_i is the at risk indicator. The model is fitted and the resulting output is shown below.

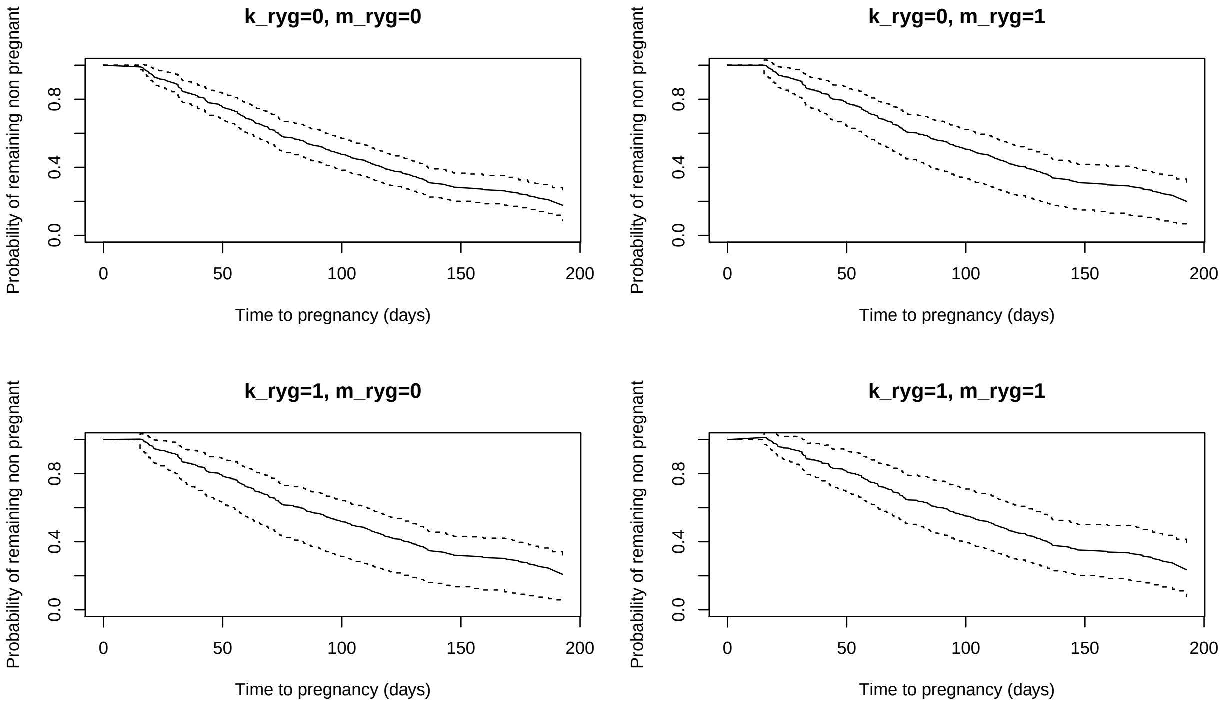
```
aalen_4<- aalen(Surv(ttp,k_gravid)~const(k_cof)+const(k_ryg)+const(m_ryg)+
                  const(logm_zkon0), data=ttp_data)
summary(aalen_4)

## Additive Aalen Model
##
## Test for nonparametric terms
##
## Test for non-significant effects
##             Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                          3.92                   0
##
## Test for time invariant effects
##                   Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                         0.184                       0.221
##                     Cramer von Mises test p-value H_0:constant effect
## (Intercept)                          3.86                       0.062
##
## Parametric terms :
```

```
##                     Coef.    SE Robust SE      z P-val
## const(k_cof)        0.000 0.000     0.000 -1.230 0.220
## const(k_ryg)       -0.001 0.001     0.001 -0.718 0.473
## const(m_ryg)       -0.001 0.001     0.001 -0.571 0.568
## const(logm_zkon0)  0.002 0.000     0.000  5.830 0.000
##
##   Call:
## aalen(formula = Surv(ttp, k_gravid) ~ const(k_cof) + const(k_ryg) +
##     const(m_ryg) + const(logm_zkon0), data = ttp_data)
```

We notice that some of the variables have non significant effects, but since we only included the risk factors we will investigate, we will not reduce the model any further.
In order to estimate the survival function for various subgroups, we choose to look at the average of the continuous variables. We will then investigate how the survival function changes for various combinations of the categorical variables.

```
fit <- aalen(Surv(ttp,k_gravid)~const(k_cof)+const(k_ryg)+const(m_ryg)+const(logm_zkon0),
           data=ttp_data, silent=0,resample.iid=1)

x0<-1 # Kun intercept er tidsafhængigt
z0<-c(mean(ttp_data$k_cof, na.rm=T),0,0,mean(ttp_data$logm_zkon0, na.rm=T))
delta<-matrix(0,length(fit$cum[,1]),309);
for (i in 1:309) {delta[,i]<-x0%*%t(fit$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))
se<-apply(delta^2,1,sum)^.5

mpt<-c()
for (i in 1:309) {g<-rnorm(309); pt<-abs(delta %*% g)/c(se);mpt<-c(mpt,max(pt[-1])); }
Cband0<-percen(mpt,0.95);

z1 <- c(mean(ttp_data$k_cof, na.rm=T),0,1,mean(ttp_data$logm_zkon0, na.rm=T))
delta<-matrix(0,length(fit$cum[,1]),309);
for (i in 1:309) {delta[,i]<-x0%*%t(fit$B.iid[[i]])+fit$cum[,1]*sum(z1*fit$gamma.iid[i,]);}
S1<-exp(- x0 %*% t(fit$cum[,-1])- fit$cum[,1]*sum(z1*fit$gamma))
se1<-apply(delta^2,1,sum)^.5
mpt<-c()
for (i in 1:309) {g<-rnorm(309); pt<-abs(delta %*% g)/c(se1);mpt<-c(mpt,max(pt[-1])); }
Cband1<-percen(mpt,0.95);

z2 <- c(mean(ttp_data$k_cof, na.rm=T),1,0,mean(ttp_data$logm_zkon0, na.rm=T))
delta<-matrix(0,length(fit$cum[,1]),309);
for (i in 1:309) {delta[,i]<-x0%*%t(fit$B.iid[[i]])+fit$cum[,1]*sum(z2*fit$gamma.iid[i,]);}
S2<-exp(- x0 %*% t(fit$cum[,-1])- fit$cum[,1]*sum(z2*fit$gamma))
se2<-apply(delta^2,1,sum)^.5
mpt<-c()
for (i in 1:309) {g<-rnorm(309); pt<-abs(delta %*% g)/c(se2);mpt<-c(mpt,max(pt[-1])); }
Cband2<-percen(mpt,0.95);

z3 <- c(mean(ttp_data$k_cof, na.rm=T),1,1,mean(ttp_data$logm_zkon0, na.rm=T))
delta<-matrix(0,length(fit$cum[,1]),309);
for (i in 1:309) {delta[,i]<-x0%*%t(fit$B.iid[[i]])+fit$cum[,1]*sum(z3*fit$gamma.iid[i,]);}
S3<-exp(- x0 %*% t(fit$cum[,-1])- fit$cum[,1]*sum(z3*fit$gamma))
se3<-apply(delta^2,1,sum)^.5
mpt<-c()
```

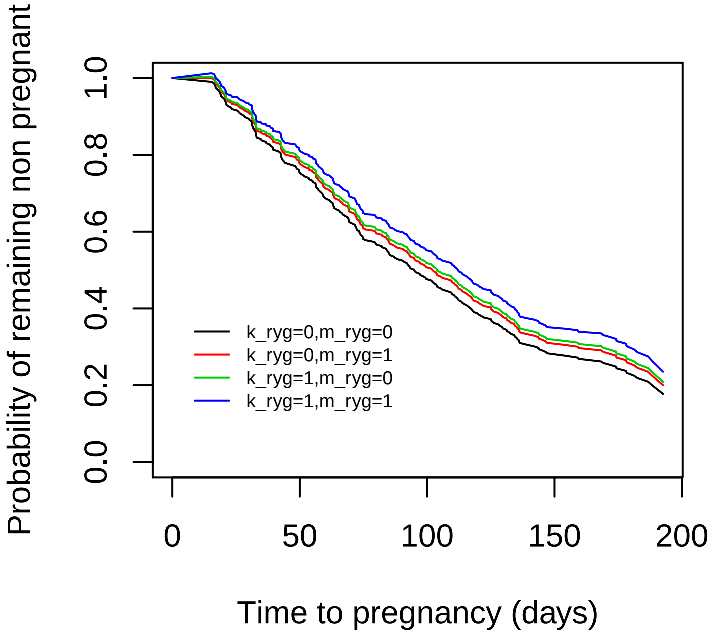
```
for (i in 1:309) {g<-rnorm(309); pt<-abs(delta %*% g)/c(se3);mpt<-c(mpt,max(pt[-1])); }
Cband3<-percen(mpt,0.95);

par(mfrow=c(2,2))
plot(fit$cum[,1],S0,type="l",ylim=c(0,1),xlab="Time to pregnancy (days)",
     ylab="Probability of remaining non pregnant", col=1, main="k_ryg=0, m_ryg=0")
lines(fit$cum[,1],S0-Cband0*S0*se,lty=2,type="s");
lines(fit$cum[,1],S0+Cband0*S0*se,lty=2,type="s")

plot(fit$cum[,1],S1,type="l",ylim=c(0,1),xlab="Time to pregnancy (days)",
     ylab="Probability of remaining non pregnant", col=1, main="k_ryg=0, m_ryg=1")
lines(fit$cum[,1],S0-Cband1*S1*se1,lty=2,type="s");
lines(fit$cum[,1],S0+Cband1*S1*se1,lty=2,type="s")

plot(fit$cum[,1],S2,type="l",ylim=c(0,1),xlab="Time to pregnancy (days)",
     ylab="Probability of remaining non pregnant", col=1, main="k_ryg=1, m_ryg=0")
lines(fit$cum[,1],S0-Cband2*S2*se2,lty=2,type="s");
lines(fit$cum[,1],S0+Cband2*S2*se2,lty=2,type="s")

plot(fit$cum[,1],S3,type="l",ylim=c(0,1),xlab="Time to pregnancy (days)",
     ylab="Probability of remaining non pregnant", col=1, main="k_ryg=1, m_ryg=1")
lines(fit$cum[,1],S3-Cband3*S3*se3,lty=2,type="s");
lines(fit$cum[,1],S3+Cband3*S3*se3,lty=2,type="s")
```



Here we see the estimated survival function for all combinations of `k_ryg` and `m_ryg` with 95%-confidence bands. We note that the confidence band for $k\_ryg = 0, m\_ryg = 0$ are more narrow than the others. This is the case since there are 186 observations where none of the subjects smoke, hence the parameter can be estimated more exact.

Now we plot the four estimated survival functions in one plot, without confidence bands, in order to compare the different combinations.

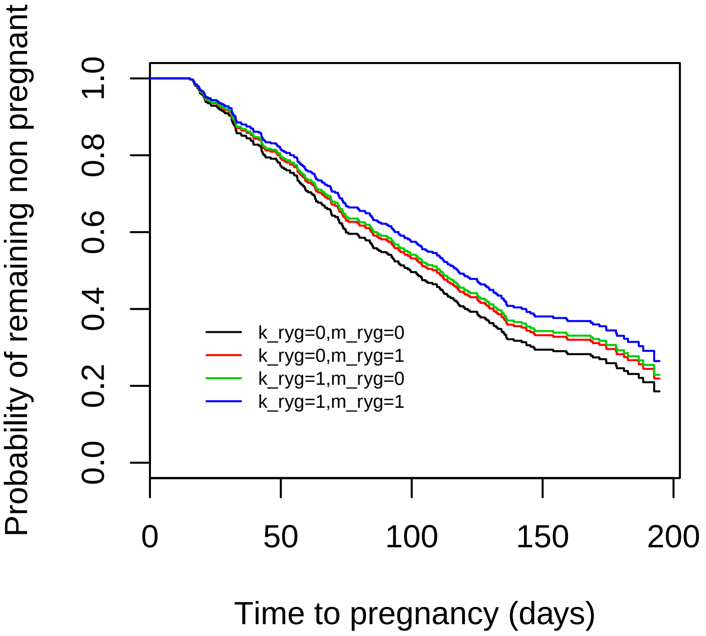
```
par(mfrow=c(1,1))
plot(fit$cum[,1],S0,type="l",ylim=c(0,1),xlab="Time to pregnancy (days)",
     ylab="Probability of remaining non pregnant", col=1)
lines(fit$cum[,1],S1,type="l", col=2)
lines(fit$cum[,1],S2,type="l", col=3)
lines(fit$cum[,1],S3,type="l", col=4)
legend(2,0.4,col=c(1:4),c("k_ryg=0,m_ryg=0","k_ryg=0,m_ryg=1","k_ryg=1,m_ryg=0",
                          "k_ryg=1,m_ryg=1"), lty=1, cex=0.6, bty="n")
```



Here we see that smoking extend the time of conceiving such that it will take longer to get pregnant if both the male and female smoke. Furthermore we see that it has a greater impact on the time to pregnancy when the female smokes compared to when the male smokes. At last we see that the time to pregnancy is shortest for couples where none of the subjects smoke.

In order to compare the findings with the results from question 1) we plot the survival curves predicted by the Cox-regression model.

```
cox_ph <- coxph(Surv(ttp,k_gravid)~k_ryg+m_ryg+k_cof+loglogm_zkon0, data=ttp_data)

plot(survfit(cox_ph, newdata=data.frame(k_ryg=0, m_ryg=0,
                                         k_cof=mean(ttp_data$k_cof,na.rm=T),
                                         loglogm_zkon0=mean(ttp_data$loglogm_zkon0, na.rm=T)),
             se.fit=F), col=1, mark.time=F, xlab="Time to pregnancy (days)",
     ylab="Probability of remaining non pregnant")

legend(15,0.4,col=c(1:4),c("k_ryg=0,m_ryg=0","k_ryg=0,m_ryg=1","k_ryg=1,m_ryg=0",
                           "k_ryg=1,m_ryg=1"), lty=1, cex=0.6, bty="n")
lines(survfit(cox_ph, newdata=data.frame(k_ryg=0, m_ryg=1,
                                          k_cof=mean(ttp_data$k_cof,na.rm=T),
                                          loglogm_zkon0=mean(ttp_data$loglogm_zkon0, na.rm=T)),
               se.fit=F), col=2, mark.time=F)
lines(survfit(cox_ph, newdata=data.frame(k_ryg=1, m_ryg=0,
                                          k_cof=mean(ttp_data$k_cof,na.rm=T),
                                          loglogm_zkon0=mean(ttp_data$loglogm_zkon0, na.rm=T)),
               se.fit=F), col=3, mark.time=F)
lines(survfit(cox_ph, newdata=data.frame(k_ryg=1, m_ryg=1,
                                          k_cof=mean(ttp_data$k_cof,na.rm=T),
                                          loglogm_zkon0=mean(ttp_data$loglogm_zkon0, na.rm=T)),
               se.fit=F), col=4, mark.time=F)
```



We note that the various groups are predicted in the same order. This means, that the models suggest the same effect of the various combinations of smoking status. Furthermore the lines are quite similar, there are no clear sign of different predictions.

3)

To summarize our findings we start by listing estimates in a table.

Variable	Estimate	Variable	Estimate
k_cof2	-0.00043588	k_cof	$-2.57837 \cdot 10^{-6}$
k_ryg	-0.13284994	k_ryg	$-8.413131 \cdot 10^{-4}$
m_ryg	-0.10341833	m_ryg	$-6.237052 \cdot 10^{-4}$
logm_zkon0	1.33473469	logm_zkon0	$1.787128 \cdot 10^{-3}$

Table 1: Estimates from Cox model (left) and Aalen model (right)

Like we concluded in the last part of question 2) the models are quite similar. From table 1 we see that they agree on the sign of each estimate. However we must remember that we used different transformations in the two models, and the covariates are included in different ways. If we want the estimates on their original scale we thus have to transform them back. We are not interested in this, we are only interested in their effect. In question 2) we found that smoking had a negative impact on the time for pregnancy, such that it prolongs the time to succesfully conceiving, in both models. This coincide with what we expected from the beginning, naimly that smoking is bad for fertility. Furthermore our estimates show that caffiene also has a negative influence on your chance of conceiving. A high sperm concentration leads, in contrast, to better fertility. It is worth noting that eventhough there is an effect of smoking and the females caffeine intake these effects are non significant.

There are pros and cons for both models. The Cox model are very flexible but its validity relies on the assumption of proportional hazards and finding suitable functional forms of the variables. The effect of the covariates act multiplicatively on some unknown baseline hazard/intensity, for which we have no assumptions. Furhtermore the regression coefficients are assumed constant. In our case we didn't manage to find a suitable functional form of `m_zkon0`, hence the model does not fit our data well. If the assumptions in the Cox model are not fulfilled, the Aalen's additive model is a nice alternative. Here the covariates act in an additive way on some unknown baseline. We allow the unknown risk factors to be functions of time, but in our model the effects are found to be constant. We saw that a log transformation of `m_zkon0` was acceptable. All in all we conclude that the Aalen's additive model is a more suitable model for this data.