Survival Analysis: Exam 2021

11 November 2021

1.1 Theoretical part - Marginal Models

1. What is the marginal survival distribution of W_1 given X, $P(W_1 > w_1|X)$, and what is the related hazard function.

First, we calculate the marginal survival distribution:

$$\begin{aligned} P(W_1 > w_1|X) &= E(P(W_1 > w_1|X, \rho)|X) \\ &= E(e^{-\rho\Lambda_1(w_1)\exp(X^T\beta_1)}|X) \end{aligned}$$

The second inequality is obtained by the tower property. We know that ρ is gamma distributed with mean 1 and variance θ, which translates into the following (using shape and scale parameters):

$$\rho \sim \Gamma(1/\theta, \theta),$$

so we have the following Laplace transform:

$$L_\rho(u) = (\theta u + 1)^{-1/\theta},$$

which we can use with $u = \Lambda_1(w_1)\exp(X^T\beta_1)$. We get:

$$P(W_1 > w_1|X) = (1 + \theta\Lambda_1(w_1)\exp(X^T\beta_1))^{-1/\theta}$$

We now wish to calculate the observed related hazard function. We therefore take -log and differentiate the expression for the survival distribution. Note that in the following calculations, the at risk indicator, $Y(t)$ is left out for simplicity, and will be added in the final expression:

$$\begin{aligned} \lambda_1(w_1) &= \frac{d}{dw_1} - \log((1 + \theta\Lambda_1(w_1)\exp(X^T\beta_1))^{-1/\theta}) \\ &= \frac{d}{dw_1} 1/\theta(\log(1 + \theta\Lambda_1(w_1)\exp(X^T\beta_1))) \\ &= 1/\theta(\frac{1}{1 + \theta\Lambda_1(w_1)\exp(X^T\beta_1)}\frac{d}{dw_1}(1 + \theta\Lambda_1(w_1)\exp(X^T\beta_1))) \\ &= \lambda_{01}(w_1)\exp(X^T\beta_1)\frac{1}{1 + \theta\Lambda_1(w_1)\exp(X^T\beta_1)} \end{aligned}$$

Using the at risk indicator, we get:

$$\lambda_1(w_1) = Y(t)\lambda_{01}(w_1)\exp(X^T\beta_1)\frac{1}{1 + \theta\Lambda_1(w_1)\exp(X^T\beta_1)Y(t)}$$

2. What is the observed hazard function of W_1 given X and ρ, ie. what is

$$\mathbf{lim}_{h\to 0}\frac{1}{h}P(W_1 \in [t, t+h]|W_1 > t, C > t, X, \rho).$$

For the j'th individual in cluster i, by the innovation theorem, we get the following observed hazard function (note that $\mathcal{H}_t^i$ is the unobserved cluster i conditional filtration):

$$\begin{aligned}\lambda_{ij}^{\mathcal{F}}(t) &= E(\lambda_{ij}^{\mathcal{H}}|\mathcal{F}_{t-}^i) \\ &= Y_{ij}E(\rho_i|\mathcal{F}_{t-}^i)\lambda_{01}(t)\exp(X_{ij}^T\beta_1) \\ &= \rho_i\lambda_{01}(t)\exp(X_{ij}^T\beta_1),\end{aligned}$$

where the last inequality holds as we condition on ρ such that $E(\rho_i|\mathcal{F}_{t-}^i) = \rho_i$.

2a. What is the observed hazard function of W_1 given only X.

Following the previous exercise, now only given X, we need to calculate the expectation. We get (note that we use Bayes' rule in the 3rd inequality to obtain an expression for $E(\rho_i|\mathcal{F}_{t-}^i)$:

$$\begin{aligned}\lambda_{ij}^{\mathcal{F}}(t) &= E(\lambda_{ij}^{\mathcal{H}}|\mathcal{F}_{t-}^i) \\ &= Y_{ij}E(\rho_i|\mathcal{F}_{t-}^i)\lambda_{01}(t)\exp(X_{ij}^T\beta_1) \\ &= Y_{ij}\frac{\int \rho^{1+\sum_j N_{ij}(t)}\exp(-z\sum_j \Lambda_{ij}^*)p(z)dz}{\int \rho^{\sum_j N_{ij}(t)}\exp(-z\sum_j \Lambda_{ij}^*)p(z)dz}\lambda_{01}(t)\exp(X_{ij}^T\beta_1) \\ &= -Y_{ij}(t)\frac{D^{1+\sum_j N_{ij}(t)}\phi(\sum_j \Lambda_{ij}^*(t))}{D^{\sum_j N_{ij}(t)}\phi(\sum_j \Lambda_{ij}^*(t))}\lambda_{01}(t)\exp(X_{ij}^T\beta_1)\end{aligned}$$

We now use Frank's lecture notes p. 8, to get an expression for the h'th derivative of the Laplace transform of the gamma distribution with mean 1, so we get (following Frank's notes and leaving out the details):

$$\begin{aligned}\lambda_{ij}^{\mathcal{F}}(t) &= -Y_{ij}(t)\frac{D^{1+\sum_j N_{ij}(t)}\phi(\sum_j \Lambda_{ij}^*(t))}{D^{\sum_j N_{ij}(t)}\phi(\sum_j \Lambda_{ij}^*(t))}\lambda_{01}(t)\exp(X_{ij}^T\beta_1) \\ &= Y_{ij}(t)\frac{1/\theta + \sum_{j'} N_{ij'}(t)}{1/\theta + \sum_{j'} \Lambda_{ij'}^*(t)}\lambda_{01}(t)\exp(X_{ij}^T\beta_1)\end{aligned}$$

2b. Can we estimate $\Lambda_1()$ and β_1 based on the observed data.

The short answer is "yes". We know that inference for marginal models with an unobserved frailty can be achieved in a two-stage method (Martinussen/Scheike book and Frank's lecture notes):

- 1) Estimate the marginal parameters (ignoring the frailty).
- 2) Plug the marginal estimates into the likelihood for θ and maximize this. This can be done by noting that with a cox marginal model and gamma distributed frailty, the observed likelihood for cluster i os proportional to:

$$\sum_j \int_0^\tau \log(1+\theta \sum_j N_{ij}(t-))dN_{ij}(\tau)H_{ij} - (1/\theta + \sum_j N_{ij}(\tau)) \log(1+\sum_j \exp(\theta H_{ij})-1)),$$

where $H_{ij} = \exp(\beta^T X_{ij})\Lambda_{01}(T_{ij})$. So, replacing β and Λ_{01} by the marginal estimates, we can obtain a pdseudolikelihood, for which consistency and asymptotic normality have been shown.

Another approach is the one-step approach, where an estimator can be obtained by maximizing over both the marginal parameters and frailty parameter in one step.

2.1 Practical part - Colorectal Cancer

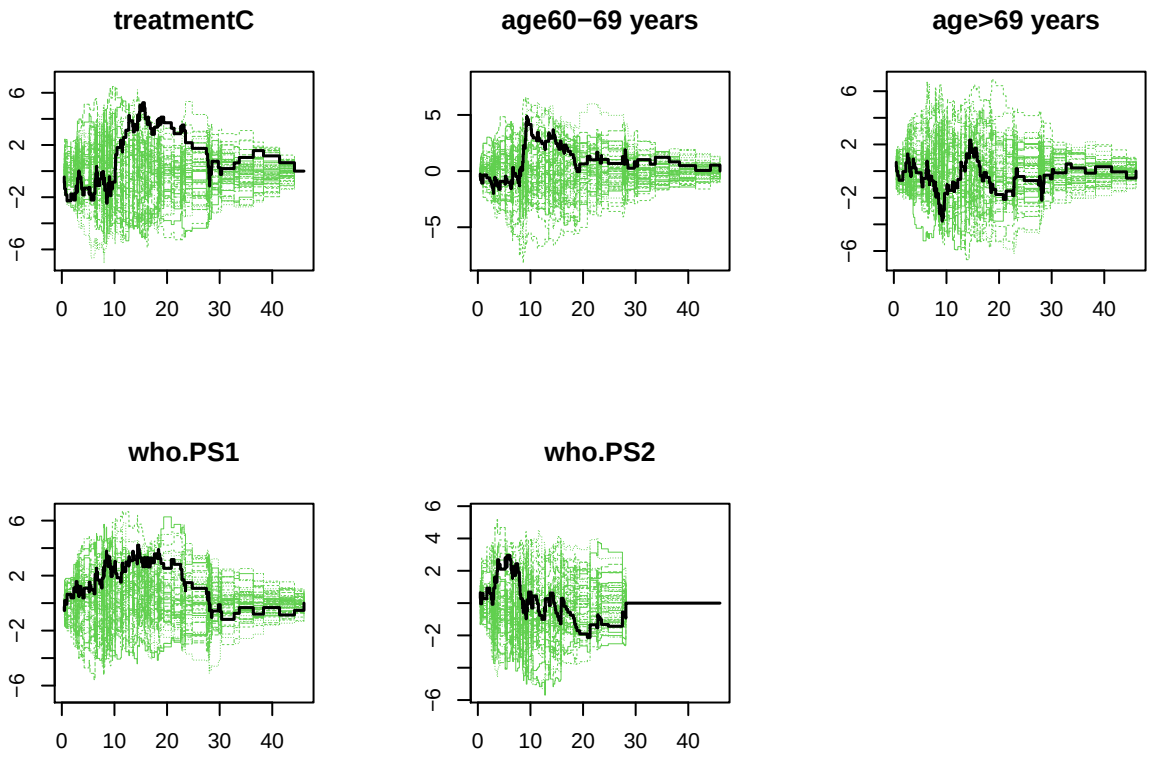
1. Show if covariates are important for death using Cox modelling. Estimate the survival distribution for relevant covariate combinations.

First, we note that the data given consists of a subset containing 150 patients of the original randomized trial including 410 patients. We assume that this is a randomly chosen subset such that the randomization is still valid (this is a plausible assumption according to help(colorectal)). The time variables are reported in years, so we choose to multiply with 12 to obtain a time scale of months instead.

In this and the following question, we consider the terminal event death as the event of interest. First, we fit a cox model using the covariates treatment, age and who.ps (WHO performance status at baseline), and ignore the number of new lesions for now:

```
colo.sub <- colo[order(id, time1)]
colo.sub <- colo.sub[,.SD[.N],by=id] #Only last event (death)
m.cox <- coxph(Surv(time1, state)~treatment+age+who.PS, data=colo.sub)
```

Before interpreting the model results, we need to chech that the important assumption of proportional hazards is fulfilled, ie. that the effect of the covariates is constant over time. We do this by considering the cumulative martingale residuals, and the cumulative score process test for proportionality. This is done using the gof (goodness of fit) function in R:



We see for all covariates, that the blacl line lies within the green ones, which indicates that the proportional hazards assumption is fulfilled. If we consider the goodness of fit object, we get the cumulative score process test for proportionality:

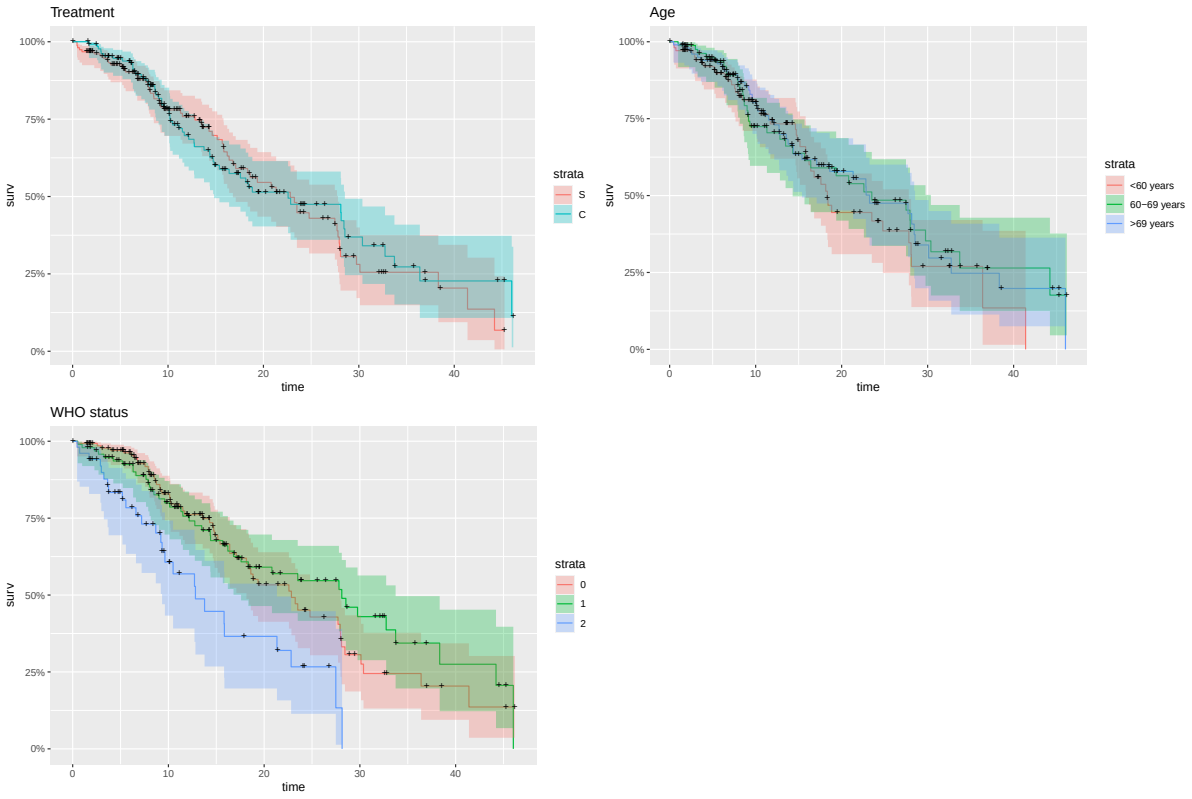
```
## Cumulative score process test for Proportionality:
##                 Sup|U(t)|  pval
## treatmentC       5.240899 0.246
## age60-69 years   4.844747 0.264
## age>69 years     3.763129 0.594
## who.PS1          4.217036 0.423
## who.PS2          2.949119 0.623
```

All the covariates have a non-significant p-value (> 0.05), which again indicates that the proportional hazards assumption can be assumed. We therefore continue with the model. Below, the coefficients from the model are shown:

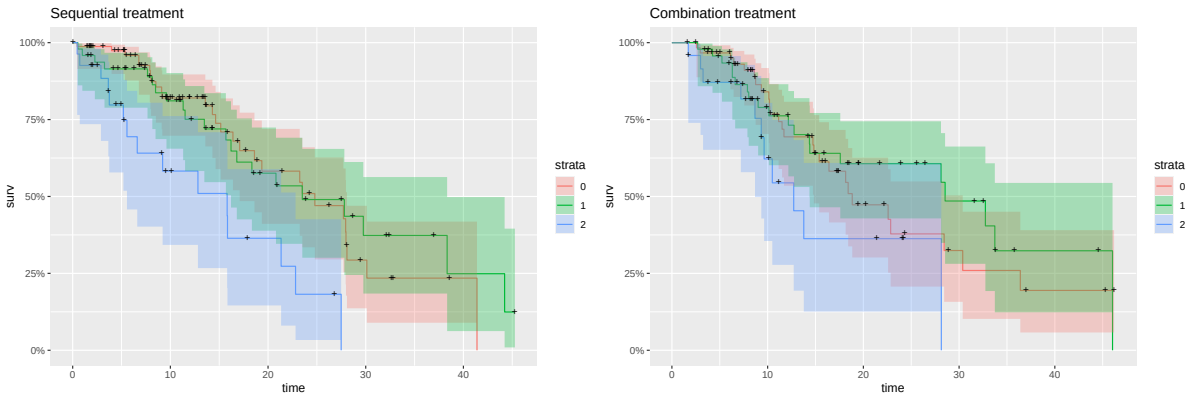
```
summary(m.cox)$coef
```

```
##                       coef exp(coef)  se(coef)          z     Pr(>|z|)
## treatmentC     -0.11670324 0.8898492 0.1906453 -0.6121484 0.5404395821
## age60-69 years -0.18286815 0.8328780 0.2347132 -0.7791132 0.4359130173
## age>69 years   -0.22058062 0.8020530 0.2184044 -1.0099641 0.3125125108
## who.PS1        -0.07389633 0.9287680 0.2083767 -0.3546286 0.7228678711
## who.PS2         0.87175235 2.3910972 0.2449886  3.5583388 0.0003732077
```

Not taking new lesions into account and only considering the effect of the three covariates on death, we see from the summary that treatment and age do not seem to affect death significantly (assuming a confidence level of 5 %). However, it seems that the hazard of dying for who status 2 is significantly higher as compared to status 0 (between 1.44 and 3.82 times higher). This is also seen, when considering the survival functions below, where it seems that the group with who status 2 in general has a lower survival probability than the other two groups. Also, the curves for the different groups of treatment and age are overlapping, which indicate no effect on death of these covariates.



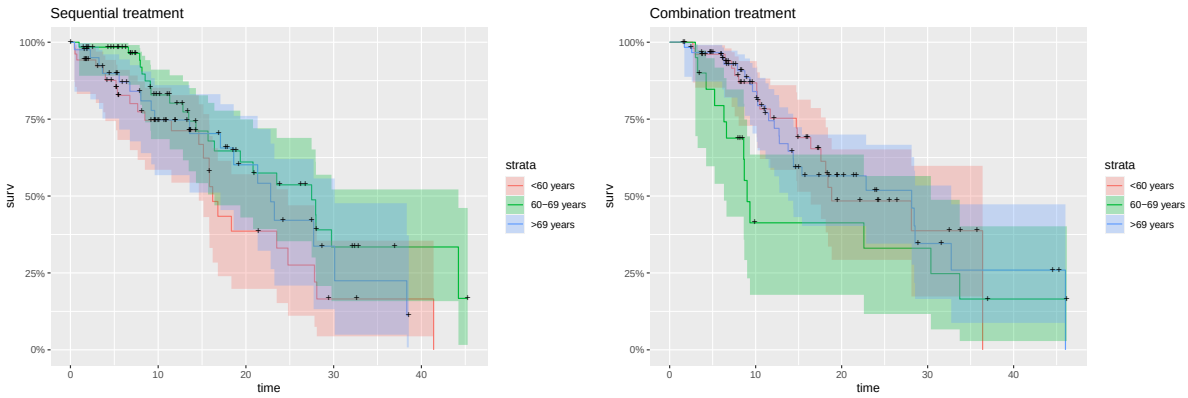
We now consider the survival function for chosen combinations of covariates. First, we take a look at treatment and the who status:



When stratifying by treatment, we see that the difference in survival on the who groups is mostly driven by the sequential treatment group, where the curve for strata 2 is lower at all

times. The survival curve for who status 2 for combination treatment is also lower than the other two at some time points, but at other time points, the curves are crossing. A general picture when stratifying by treatment is that who group 0 and 1 are more similar as compared to group 2, which the cox model and the overall summary also indicated.

We now consider a combination of treatment and age:



For the sequential treatment, the survival curves are overlapping, however, the combination treatment seems to be working more poorly for people aged 60-69, especially during the first 10 months. Note however, that there is only 15 subjects in this subgroup.

2. Is death and the number new lesions related?

To answer this question, we construct a variable, stating the number of new lesions per person (id), as a cumulated sum per time point. This is included in the cox model as a factor. Note that in order not to condition on the end of the time interval, when we are at the beginning of the interval, we shift the sum of lesions to the next time interval. This is done in the following way:

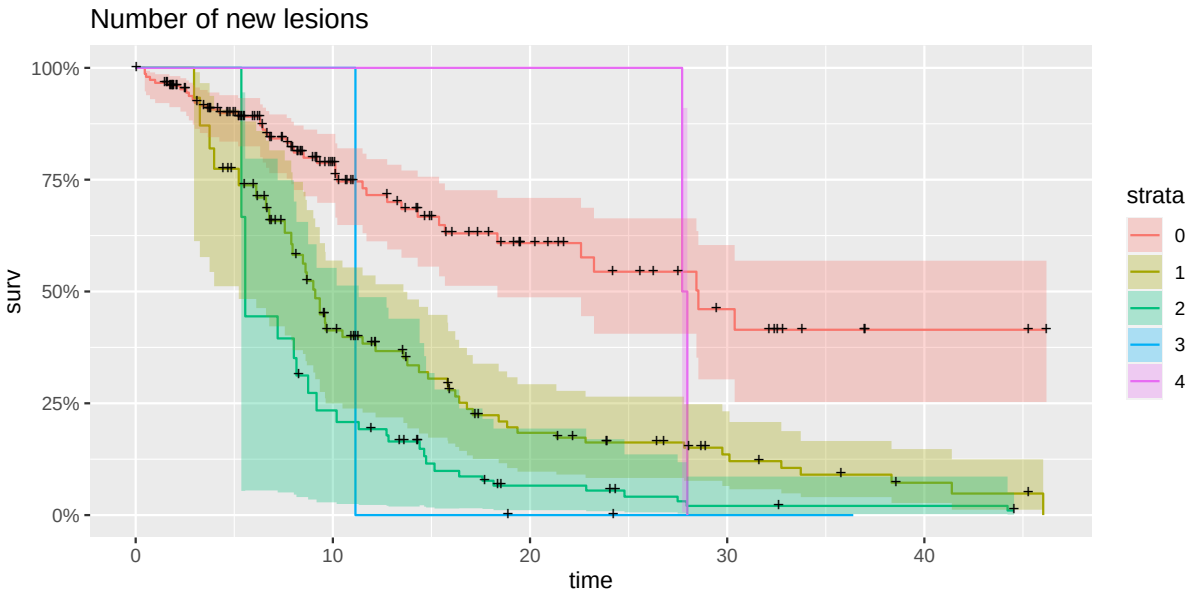
```
colo <- colo[order(id, time1)]
colo[, num.lesions:=head(c(0,cumsum(new.lesions)),n=-1), by=id] #Calculate number of new les
```

We fit a cox model using this as a covariate, with a cluster(id), as we have repeated measurements per id. We could also have modelled the repeated measurements per subject with a frailty model, adding frailty(id), but we choose use the cluster(id), where a robust estimate of the standard errors is used. Thereby, possible dependence between the subjects is modelled. As mentioned, this could also have been done with a frailty model, where the dependence between subjects is modelled usually as gamma or log-normal random effects. Furthermore, we now include both time1 as before, as well as time0, which is the start of the time interval (0 or the previous recurrence time). This is done as we now have a time-dependent covariate (number of previous lesions).

```
m.cox.1 <- coxph(Surv(time0, time1, state) ~ treatment + age + who.PS +
                   factor(num.lesions) + cluster(id), ties = "breslow", data = colo)
summary(m.cox.1)$coef[,c(1:3, 6)]
```

```
##                              coef  exp(coef)  se(coef)     Pr(>|z|)
## treatmentC                0.13164623  1.1407047 0.2012575 4.893092e-01
## age60-69 years           -0.07748919  0.9254370 0.2395318 7.375575e-01
## age>69 years             -0.12656520  0.8811167 0.2234081 5.322327e-01
## who.PS1                   0.06843643  1.0708326 0.2174377 7.345618e-01
## who.PS2                   0.80740339  2.2420786 0.2549318 6.131510e-04
## factor(num.lesions)1      1.04165371  2.8338996 0.2427865 2.034942e-05
## factor(num.lesions)2      1.74143643  5.7055331 0.2941080 3.137937e-10
## factor(num.lesions)3      1.64951228  5.2044409 0.4161579 2.765895e-08
## factor(num.lesions)4      2.81231787 16.6484626 0.8138383 1.842198e-06
```

From the summary, we can see that the hazard of dying (as compared to no new lesions) is increasing significantly with the number of new lesions. The same picture is seen in the survival curve below. Note that in the four lesion group, there are only two subjects that both die after around 27 months, which gives this sudden fall in the survival curve. So, it could seem that subjects with four lesions have a longer lifespan, but after a certain time threshold, they die. All the 9 subjects in the three lesion group also die eventually. From the figure, we can also see that having one and two lesions affects the hazard of death similarly over time.



As we have so few subjects in the four and three lesion group, we try to model the number of lesions as continous instead. We fit a cox model using this continous covariate, again with a time0 and cluster(id), as we have repeated measurements per id:

```
m.cox.2 <- coxph(Surv(time0, time1, state) ~ treatment + age + who.PS +
                   num.lesions + cluster(id), data = colo)
summary(m.cox.2)$coef[,c(1:3, 6)]
```

```
##                       coef exp(coef)  se(coef)     Pr(>|z|)
## treatmentC      0.08024854 1.0835563 0.2001348 6.762475e-01
```

```
## age60-69 years -0.05947281 0.9422612 0.2360205 7.972544e-01
## age>69 years   -0.11083703 0.8950846 0.2219450 5.931337e-01
## who.PS1         0.10627110 1.1121233 0.2147533 6.025588e-01
## who.PS2         0.89508456 2.4475427 0.2469690 9.699862e-05
## num.lesions     0.66554204 1.9455448 0.1040912 4.929986e-13
```

We can see that the hazard of dying increases significantly with number of lesions which we also expected from the graphical interpretation.

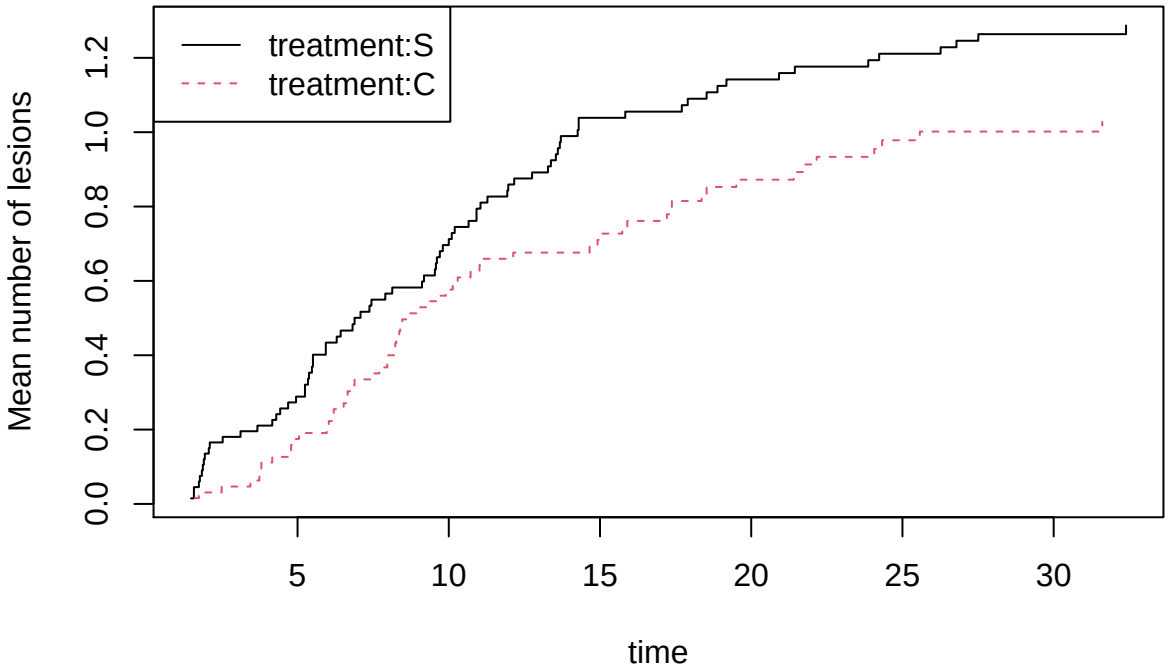
3. Estimate the mean number of new lessions as a function of time, i.e., the marginal mean of the recurrent events ($\mu(t)$ in the previous exercise).

We now wish to take the re-occurence of new lesions into account, and estimate the mean number of new lesions as a function of time. Following Per's slides, we assume that there is no "gap" times, ie. a periods where a new lesion was not possible. It is also a plausible assumption that a subject is at risk for getting a lesion again immediately after experiencing a lesion. As we saw in the previous exercise, a large number of the subjects experience the terminal event death, so we have presence of a non-negligible mortality rate that we need to take into account in the modelling.

When mortality plays a role, the Nelson-Aalen estimator for $\mu(t)$ will be upwards biased, as events can only happen as long as the subject is still alive. We will use a simple estimator for $\mu(t)$ that accounts for mortality, which is given by the "Gosh-Lin" estimator. To do this in R, we use the phreg function from the mets package. First, we estimate the mean number of new lesions non-parametrically in the following way (inlcuding the same covariates as in the previous exercises, stratifying by treatment as using the cluster(id) option as argued earlier):

```
library(mets)
survobj <- phreg(Surv(time0, time1, state == 1) ~ strata(treatment) + age +
                   who.PS + cluster(id), data = colo, km = TRUE)
recevobj <- phreg(Surv(time0, time1, new.lesions == 1) ~ strata(treatment) +
                  age + who.PS + cluster(id), data = colo, km = TRUE)

obj <- recmarg(recevobj, survobj)
bplot(obj, ylab = "Mean number of lesions")
```



This shows, that over time, the mean number of lesions is larger for sequential than combination treatment. The effect seems roughly constant over time (one might argue that the effect is more pronounced after around 12 months, but the effect is not noticeable from the plot). We now estimate the treatment effect on the mean function in a Ghosh-Lin regression model using the recreg function in the mets package. Note that to do this, we need to construct a censoring variable, and we also construct the variable "status", which contains information about both death and lesions, such that:

- status=0 means alive and no lesion
- status=1 means dead
- status=2 means alive and new lesion.

```
colo[,status:=state]
colo[new.lesions==1,status:=2]
colo$cens <-ifelse(colo$state==0,1,0) #Censoring variable
m.goshlin <- recreg(EventCens(time0, time1, status, cens) ~ treatment + age + who.PS +
                    cluster(id), data = colo, cause = 2, death.code = 1,
                    cens.code = 1, cens.model ~ 1)
summary(m.goshlin)$coef
```

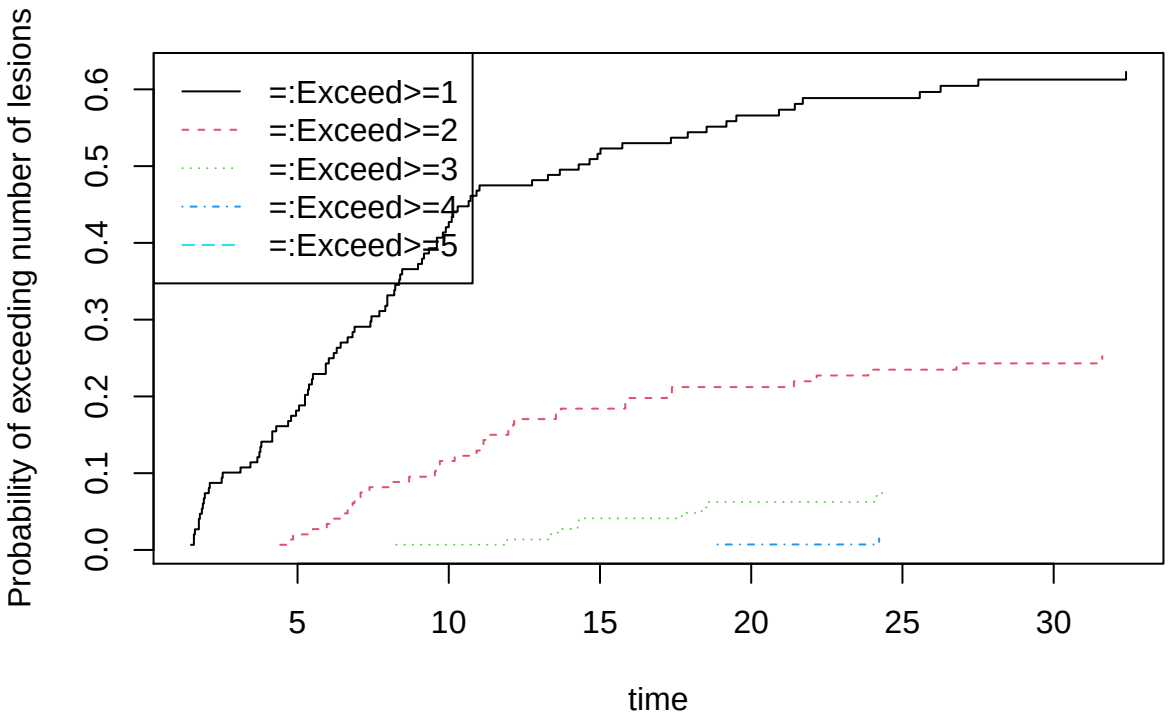
```
##                    Estimate      S.E.   dU^-1/2   P-value
## treatmentC     -0.25053314 0.1592898 0.1758116 0.1157620
## age60-69 years -0.12319697 0.1902138 0.2126833 0.5171944
## age>69 years   -0.13647908 0.1863036 0.2025636 0.4638252
## who.PS1        -0.23099135 0.1792848 0.1929288 0.1976052
## who.PS2        -0.04935763 0.2019346 0.2369963 0.8069026
```

The model is fitted using the same covariates as in the previous exercises. As expected from the plot above, the mean number of lesions is lower over time for combination treatment, with an estimated mean ratio of $\exp(-0.25053314) = 0.78$. This means that, constant over time, we have 22% less lesions at $(1 - 0.78 = 0.22)$ for combination as compared to sequential treatment. Note, however, that the treatment effect is non-significant, with a p-value of $0.11 > 0.05$.

4. Estimate the probability of a patient having more than one new lesion before dying as a function of time.

In the previous exercise, we considered the mean number of new lesions, however, now we wish to investigate another summary measure, namely the probability of a patient having more than one new lesion. For this purpose, we use the function prob.exceedRecurrent from the mets package (https://cran.r-project.org/web/packages/mets/vignettes/recurrent-events.html):

```
prob.obj <- count.history(colo, status="new.lesions") #Set up data with the count.history fc
prob.exceed <- prob.exceedRecurrent(prob.obj, 1, status = "new.lesions",
                                    death="state", start="time0", stop="time1", id="id")
bplot(prob.exceed, ylab="Probability of exceeding number of lesions")
```



So, the probability we are interested in is getting more than one lesion (> 1), which is the red curve (two or above). So, not including any covariates, the probability of getting more than one lesion is increasing with time (as expected), with a steeper increase in the first months. For example, after 10 months, the probability of getting more than one lesion is around 10%, and after 25 months, it is approximately 23%.

- Check prop hazards assumption and validate the model like in question 1.
- Remember to check the assumptions and validate the regression model.

5. Estimate the probability of a patient having more than two new lessions before dying as a function of time.

- Same procedure as in question 4.