# Modelling the MESS Data

*Suzie Brown and Marco Palma*

*26 October 2017*

## Introduction

Th MESS data is from a randomised controlled trial about treatment for epilepsy. The subjects in the study are patients who have had only one seizure or hve not had epilepsy for long. There are two treatments; either immediate or deferred. Patients assigned the deferred treatment are not prescribed anti-epileptic drugs after their first seizure, but the decision is revisited if they have further seizures.

Because many people have only one seizure in their life, we don't necessarily want to put patients on anti-epileptic drugs after just one seizure. However, we also don't want patients going on to have further seizures that could have been prevented or reduced by starting treatment immediately.

We are therefore interested in the level of benefit associated with immediate treatment relative to deferred treatment. Ultimately we could weigh this up against the costs of ongoing medication in order to decide whether a new patient should be given immediate treatment or not. This decision could depend on other data about the patient; for instance the trial recorded demographic details, results of some medical tests, and information about previous seizures.

Our aim is to construct a model considering some of these factors, which can be used to predict the outcome for a specific type of patient under immediate and deferred treatment. The model could then be used with some loss function to make optimal decisions about how to treat new patients.

## Methods

### Aalen's additive model

The proportional hazards assumption underlying Cox model might be in many cases too strict and lead to unrealistic conclusions. Moreover, it is not . . . to consider that the effect of one or more covariates might change over time. In these cases the model proposed by Aalen (1989) provides a flexible way to tackle these issues. Unlike Cox model, in Aalen's model the effect of the covariate on the hazard function at time t is additive. In other words, the hazard function at time t is a linear function of the covariates. Each coefficient is a function allowed to vary over time; in this sense, the model is fully nonparametric. At time t, the value of the coefficient function is to be read as the variation with respect to the baseline hazard function induced by a change of one unit (level in case of categorical variables) in the covariates. Aalen's additive model is widely used because of the easiness in the interpretation of the graphical results, that report the cumulative regression coefficient over time with 95% confidence bands. When the effect of the covariate is constant over time (like in Cox model), the cumulative regression coefficient will look like a straight line; on the contrary, if the effect becomes null only after a certain time point, the slope of the cumulative regression coefficient will decrease towards zero. When the covariate does not play a role in explaining the variability in the hazard function, the value 0 will be included in the confidence bands for all the time interval.

**Parametric survival models**

# Data wrangling

```
mess <- read.csv("MESSdat.csv", sep="\t", header=T)
```

## Change data types

Treatment 0/1 and sex 0/1 have been stored as numeric; we change the data type to factor.

```
mess$trt<-as.factor(mess$trt)
mess$sex<-as.factor(mess$sex)
```

We see that there are many NAs in the columns relating to dates, but this is not shown on the summary because they are formatted as character. We first format the dates as dates.

```
mess$d1sp<-as.Date(mess$d1sp,"%d/%m/%y")
mess$d1cp<-as.Date(mess$d1cp,"%d/%m/%y")
mess$d1ps<-as.Date(mess$d1ps,"%d/%m/%y")
mess$d1myo<-as.Date(mess$d1myo,"%d/%m/%y")
mess$d1ab<-as.Date(mess$d1ab,"%d/%m/%y")
mess$d1aab<-as.Date(mess$d1aab,"%d/%m/%y")
mess$d1tc<-as.Date(mess$d1tc,"%d/%m/%y")
mess$d1oth<-as.Date(mess$d1oth,"%d/%m/%y")
mess$d1seiz<-as.Date(mess$d1seiz,"%d/%m/%y")
mess$rand<-as.Date(mess$rand,"%d/%m/%y")
```

We assume this variable must be an identifying number for the clinic where the subject entered the trial. As such, we format it as a factor.

```
table(mess$centre)
```

```
##
##  101  102  103  105  107  201  202  203  205  206  207  301  302  303  304
##    5   28    9    2    5    6   29    1    2    1   26   20   15    7    7
##  305  307  402  403  501  601  701  702  703  801  802  804 1004 1101 1102
##    2    1    3   25   32    1    3    3    1   16    1    1    5   66   74
## 1103 1105 1106 1201 1204 1206 1301 1302 1303 1401 1403 1404 1503 1505 1601
##    3    5    2   22    2   23  129   11   16   15   32    1    6    1   22
## 1602 1603 1605 1606 1701 1702 3001 3101 3103 3110 3111 3112 3501 3601 3603
##    1    4    5    4   14    1   14    5   28   28    3    1    2   57   10
## 3604 3701 3703 3801 4001 4201 4401 4402 4403 4501 4502 4504 4505 4506 4507
##   32   17    8    5   34    4  261    7    3   24   29   33   17   12    5
## 4508 7001 8001 8003 8004 8005 8006 8007
##   11   12    1    6    6   14    1   19
```

```
mess$centre <- as.factor(mess$centre)
# maybe include in the model, by grouping them together? but it's a bit arbitrary - would be nice to kn
```

We find that the binary results for the various medical tests have been stored as $1/2 = $ yes/no. We convert these to $0/1 = $ no/yes.

```
# as.logical ?
mess$eeg <- as.factor(ifelse(mess$eeg==2,0,mess$eeg))
mess$abeeg <- as.factor(ifelse(mess$abeeg==2,0,mess$abeeg))
```

```
mess$nsab <- as.factor(ifelse(mess$nsab==2,0,mess$nsab))
mess$gparsp <- as.factor(ifelse(mess$gparsp==2,0,mess$gparsp))
mess$gparnsp <- as.factor(ifelse(mess$gparnsp==2,0,mess$gparnsp))
mess$fparsp <- as.factor(ifelse(mess$fparsp==2,0,mess$fparsp))
mess$fparnsp <- as.factor(ifelse(mess$fparnsp==2,0,mess$fparnsp))
mess$swab <- as.factor(ifelse(mess$swab==2,0,mess$swab))
```

## Check censoring indicator

Next we want to see which way round the censoring indicator is used. We see that where the censoring indicator `ind1yr` is 1, the time to one year remission `int1yr` has minimum 365, whereas where `ind1yr` is 0, the minimum is 1. From this we deduce that the indicator is 1 when the variable is observed and 0 if the variable is censored.

```
summary(mess$int1yr[mess$ind1yr==1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   365.0   365.0   365.0   459.5   454.0  2236.0
```

```
summary(mess$int1yr[mess$ind1yr==0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     1.0   229.0   473.2   485.0  3111.0
```

## Missing data

We verify that the same data are missing from `d1seiz` and `period`. These were probably subjects who couldn't remember the date of their first seizure, and/or whose medical records were missing. This suggests they are missing not at random - for instance if the subject can't remember the date it is likely to be a long time ago, i.e. higher values of `period`. However, since they are only 5 out of 1425 observations, we suspect it will not make much difference to treat them as if missing at random.

```
all.equal(which(is.na(mess$d1seiz)),which(is.na(mess$period)))
```
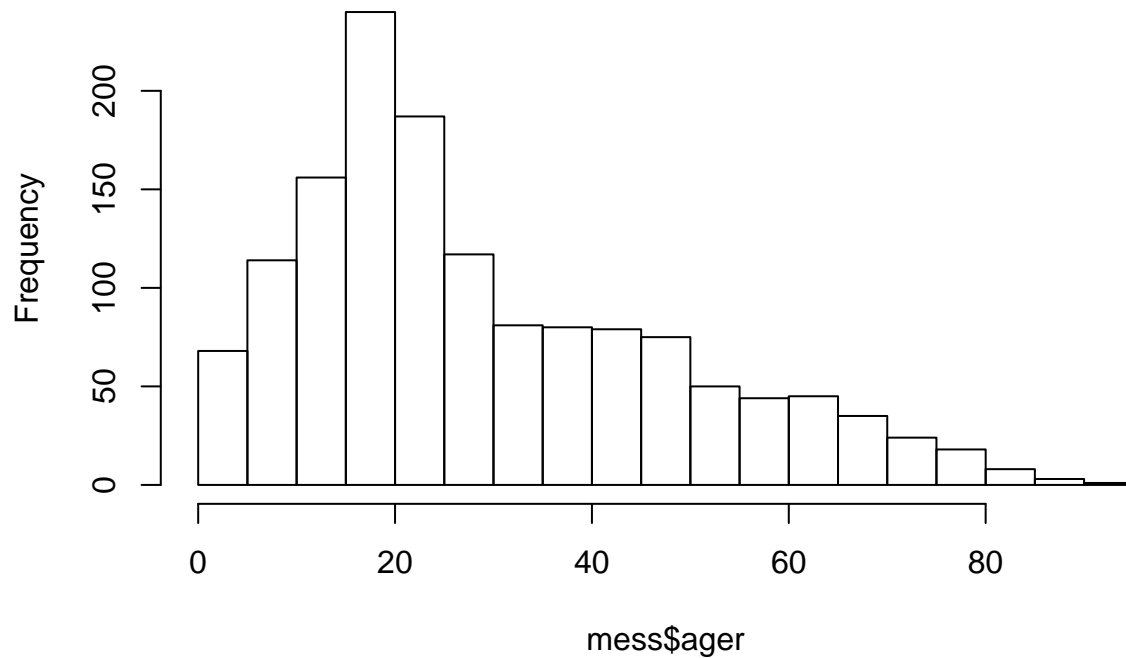
```
## [1] TRUE
```

## Investigate some variables

Plotting the age at randomisation `ager` we see that it is positively skewed. This is plausible since it is a study of people with single seizure or recent diagnosis of epilepsy, and it is more likely that an individual has their first seizure at a younger age.

```
hist(mess$ager,breaks=20)
```

# Histogram of mess$ager



We see that the number of subjects having each treatment is roughly equal, as was the design of the study.

```
table(mess$trt)
```

```
##
##   0   1
## 712 713
```

There is a mystery as to how some patients who have not had an EEG have recorded an abnormal EEG result. There are several possible explanations (different types of recording error), but we will assume the patients with an abnormal EEG have had an EEG. We adjust the `eeg` varaible accordingly. Since there are only five subjects in this category it shoudln't substantially affect any results.

```
t.eeg <- table(mess$eeg,mess$abeeg)
dimnames(t.eeg) <- list(EEG=c(0,1), abEEG=c(0,1))
t.eeg
```

```
##     abEEG
## EEG   0   1
##   0  80   5
##   1 554 786
```

```
mess$eeg[mess$eeg==0 & mess$abeeg==1] <- 1
```

## Split age into bins

We use the age categories referred to in Marson et al 2005.

```
mess$agecat <- cut(mess$ager,c(-1,5,9,19,29,39,49,59,69,100))
```

## Transform some variables

We plotted histograms of each (non-binary) variable we were considering including. The variables `nseiz` and `period` are highly positively-skewed. This is not surprising since it is a study on early epilepsy, so most subjects have only had one seizure and the frist seizure was not long before entering the study.

```
par(mfrow=c(2,2))
hist(mess$intseiz1, breaks=20)
hist(mess$ager, breaks=20)
hist(mess$nseiz, breaks=20) #this one should be transformed
hist(mess$period, breaks=20) #and this one
```

### Histogram of mess$intseiz1

### Histogram of mess$ager

### Histogram of mess$nseiz

### Histogram of mess$period
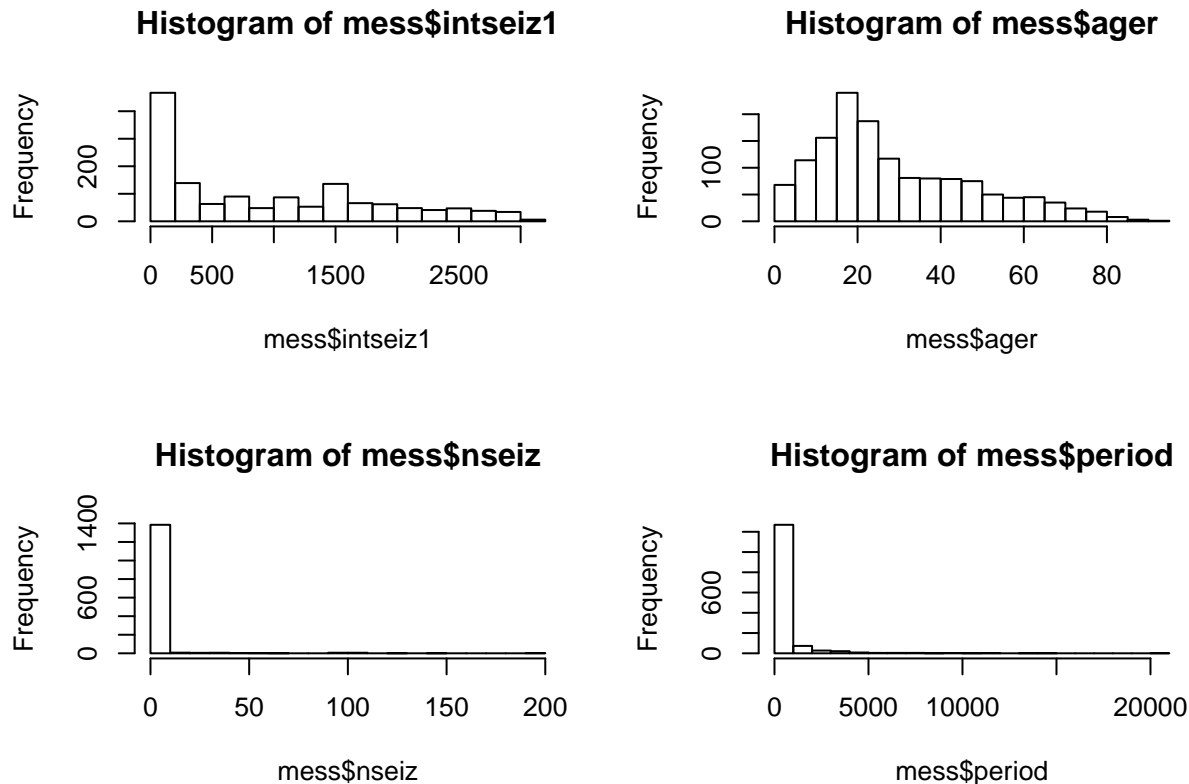
We will try transforming `period` and `nseiz`. For `nseiz`, we simply use the indicator of whether it is 1 or more than 1, since over half of the patients have `nseiz==1`. We take the logarithm of `period` to reduce the skew.

```
#par(mfrow=c(1,2))
mess$period.tr <- log(mess$period + 1)
mess$nseiz.tr <- mess$nseiz > 1
hist(mess$period.tr, breaks = 20)
```

## Histogram of mess$period.tr



```
#hist(mess$nseiz.tr)
```

### Select initial set of covariates

For the purposes of the initial exploratory model fitting, we use a small subset of the covariates. We address the issue of variable selection later on.

We drop all of the date variables because we suspect the absolute time does not matter; relative times are available for many of the events. We also drop the sub-categories for different types of seizures, retaining just the number of tonic-clonic seizures (the most severe type) and the overall number of seizures. We also drop the sub-categories of EEG abnormality, retaining just the overall indicator. We do not include the centre, since there are so many levels it adds too many degrees of freedom, and we have no information about how they might be grouped e.g. geographically.

```
library(dplyr)
covars1 <- transmute(mess, ager, sex, trt, ntc, nseiz.tr, period.tr, eeg, abeeg)
```

## Check balance of randomisation

We check that the randomisation is balanced over sex and age. There is a noticeable deviation from an even split in some categories, but nothing too major.

```
par(mfrow=c(1,2))

t.sex <- table(mess$sex,mess$trt)
colnames(t.sex)<-c('immediate','deferred')
rownames(t.sex)<-c('male','female')
plot(t.sex, main="randomisation by sex", ylab="treatment")
```

```r
mtext("sex",3)
text(0.25,0.75,t.sex[1,1])
text(0.75,0.75,t.sex[2,1])
text(0.25,0.25,t.sex[1,2])
text(0.75,0.25,t.sex[2,2])

t.age <- table(mess$agecat, mess$trt)
colnames(t.age)<-c('immediate','deferred')
plot(t.age, main="randomisation by age", ylab="treatment")
mtext("age",3)
```
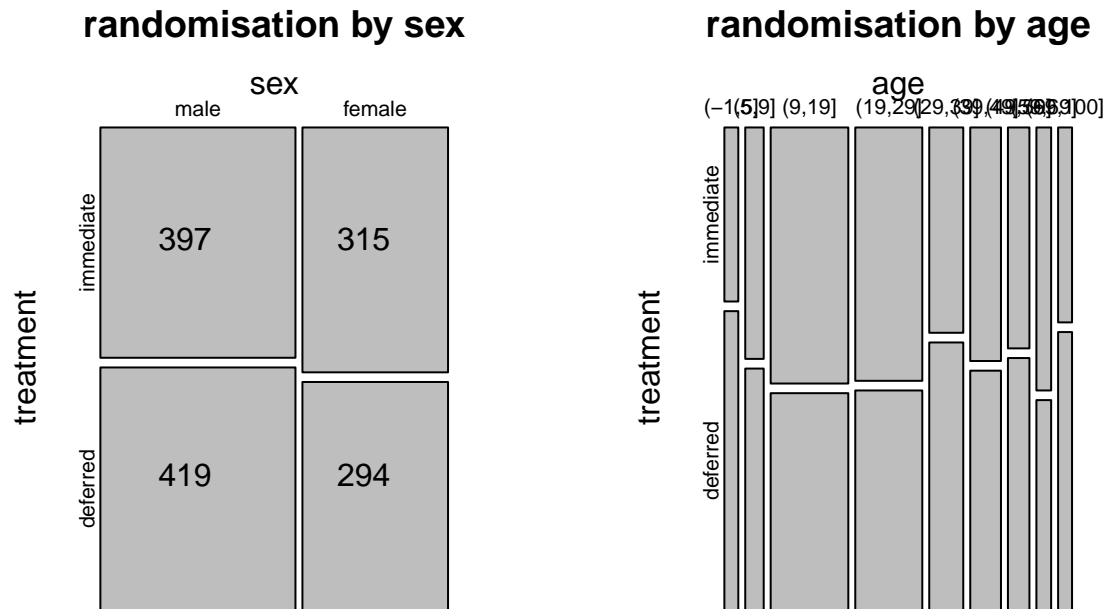


## Exploratory Kaplan-Meier curves

We plot Kaplan-Meier curves of the first seizure after randomization, grouped by treatment and sex. The first plot shows that the survival probability for those who received a deferred treatment is lower than for those immediately treated, at least for the first 5 years after the randomization. From the second plot it seems that sex does not have a significant effect on the survival probability.

```r
surv1 <- with(mess, Surv(intseiz1, indseiz1 == 1))
mfit <- survfit(surv1 ~ 1, conf.type = "log-log", data = mess)
#options(survfit.print.mean =T)
#plot(mfit, mark.time=T, xlab =factor(c(1,1)) "Days since randomization",ylab = "Survival")
#summary(mfit,censored=T)
par(mfrow=c(1,2))
mfit.trt <- survfit(surv1 ~ trt, conf.type = "log-log", data = mess)
plot(mfit.trt, conf.int = T, mark.time=F, col = 2:3, xlab = "Days since randomization", ylab = "Survival
legend("topright", legend = c("Immediate","Deferred"), col = 2:3, lty=1)
mfit.sex <- survfit(surv1 ~ sex, conf.type = "log-log", data = mess)
plot(mfit.sex, conf.int = T, mark.time=F, col = c("orange","brown"), xlab = "Days since randomization",y
legend("topright", legend = c("Male","Female"), col = c("orange","brown"), lty=1)
```
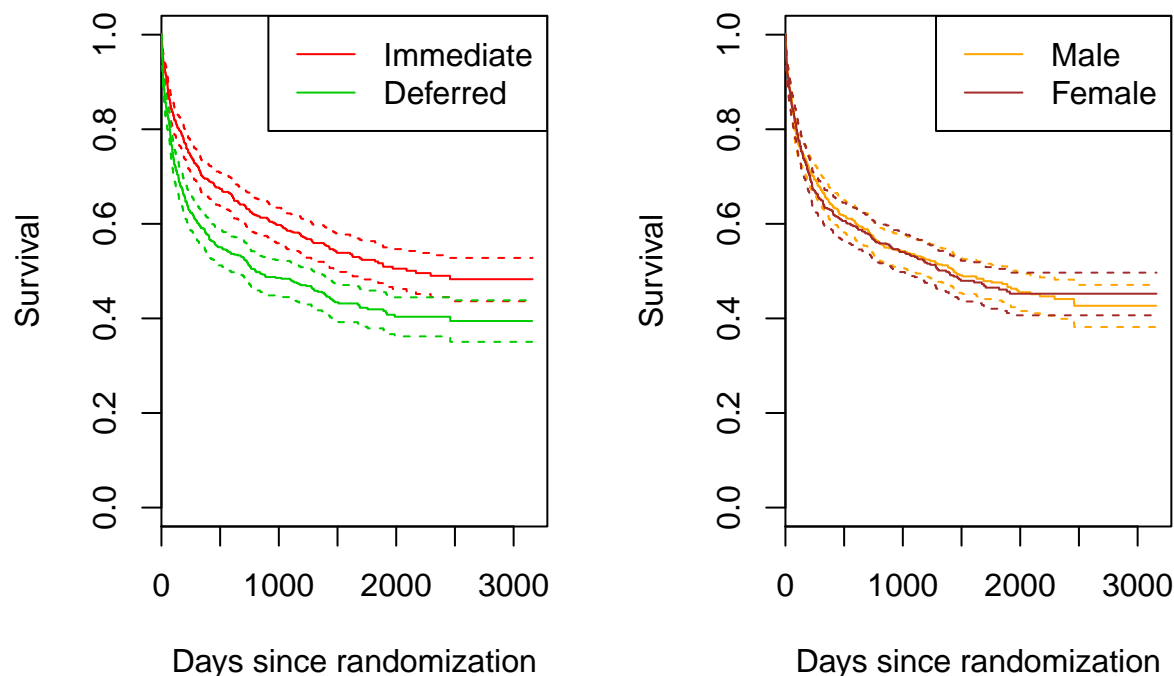
```
#lfit <- survreg(surv1 ~ trt, data=mess)
#pct <- 1:98/100   # The 100th percentile of predicted survival is at +infinity
#ptime <- predict(lfit, newdata=mess, type='quantile',p=pct, se=TRUE)
#matplot(ptime$fit, 1-pct,xlab="Months", ylab="Survival", type='l', lty=c(1,2,2), col=1)
```

## Cox proportional hazards model

The Cox model relies on the strong assumption of *proportional hazards*, that is that a unit increase in a certain covariate has a multiplicative effect on the hazard rate. In particular, the hazard function $\lambda$ has the form

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

where $X$ is the matrix of covariates and $\beta$ is the vector of coefficients to be estimated.

First we fit a naive proportional hazards model with the covariates of interest. We see that the treatment `trt` is significant, as is the number `nseiz` of seizures prior to entering the study. This is to be expected since subjects with a tendency for more frequent seizures are likely to go less time before the first post-randomisation seizure regardless of treatment. The elapsed time between the subject's first seizure and their entry into the trial is slightly significant, which is also unsurprising.

```
mycox <- coxph(Surv(mess$intseiz1,mess$indseiz1,type="right") ~ trt + ager + sex + nseiz + period, data=
summary(mycox)
```

```
## Call:
## coxph(formula = Surv(mess$intseiz1, mess$indseiz1, type = "right") ~
##     trt + ager + sex + nseiz + period, data = mess)
##
##   n= 1420, number of events= 689
##    (5 observations deleted due to missingness)
##
##             coef  exp(coef)   se(coef)       z Pr(>|z|)
## trt1    3.263e-01  1.386e+00  7.678e-02   4.250 2.14e-05 ***
```

```
## ager   -3.223e-03  9.968e-01  2.121e-03 -1.520   0.1286
## sex1    1.290e-02  1.013e+00  7.803e-02  0.165   0.8687
## nseiz   1.535e-02  1.015e+00  1.809e-03  8.481  < 2e-16 ***
## period -6.466e-05  9.999e-01  3.143e-05 -2.057   0.0397 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## trt1       1.3858     0.7216    1.1922     1.611
## ager       0.9968     1.0032    0.9926     1.001
## sex1       1.0130     0.9872    0.8693     1.180
## nseiz      1.0155     0.9848    1.0119     1.019
## period     0.9999     1.0001    0.9999     1.000
##
## Concordance= 0.592  (se = 0.011 )
## Rsquare= 0.042   (max possible= 0.999 )
## Likelihood ratio test= 61.61  on 5 df,   p=5.639e-12
## Wald test            = 92.1  on 5 df,   p=0
## Score (logrank) test = 105.8  on 5 df,   p=0
```

Next we assess whether proportional hazards is a suitable assumption for these data. We see from the residual plots that the treatment seems to have a decreasing effect over time, but the residuals of the other covariates don't seem to vary over time. This is confirmed by the small p-value for `trt` in the test of the proportional hazards assumption. The treatment effect decreasing over time is indicative that a proportional hazards model is not suitable, but we will try a few fixes now.

```
cox.zph(mycox)
```

```
##             rho chisq      p
## trt1   -0.0857 5.038 0.0248
## ager    0.0153 0.172 0.6785
## sex1   -0.0381 1.019 0.3128
## nseiz  -0.0444 0.872 0.3504
## period  0.0487 1.470 0.2254
## GLOBAL      NA 7.799 0.1677
```

```
par(mfrow=c(3,2))
plot(cox.zph(mycox))
```
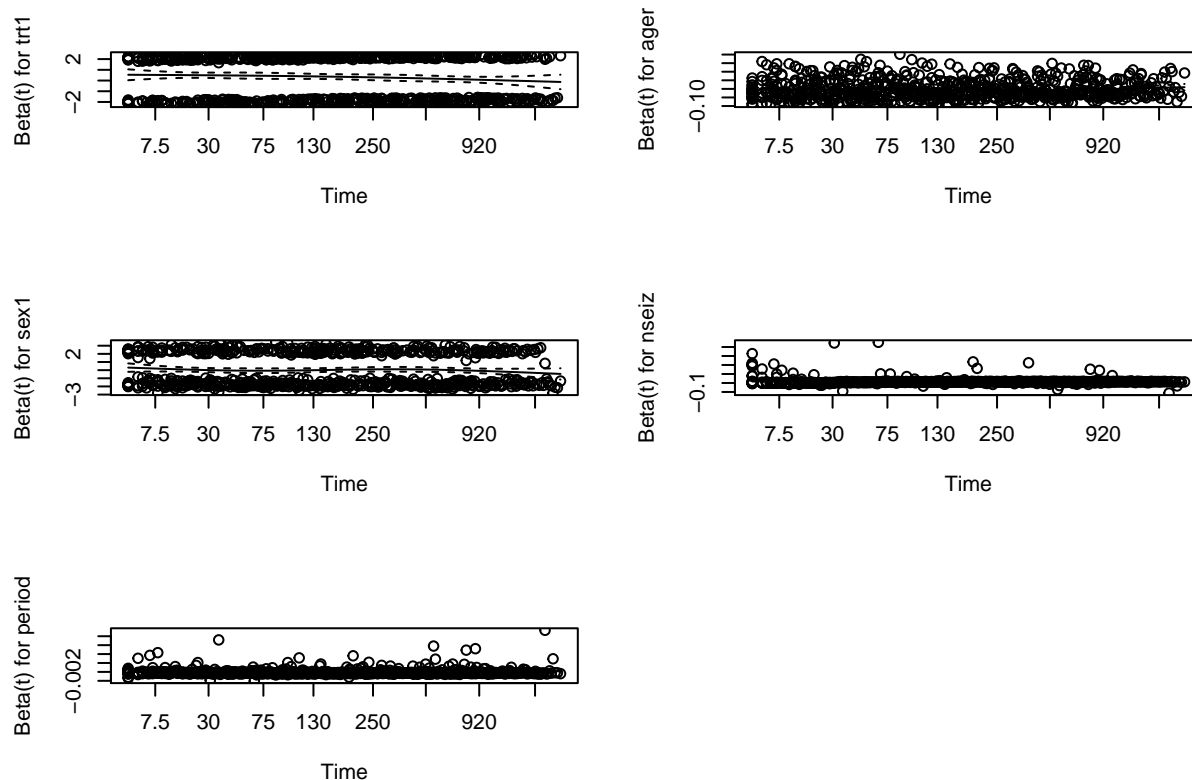
We try adding into the model an interaction term of treatment with time. We see, as expected, that the interaction term is very significant. However the proportional hazards model is still rejected, so even with the interaction term we see that the Cox proportional hazards model is not suitable.

```
mycox2 <- coxph(Surv(mess$intseiz1,mess$indseiz1,type="right") ~ trt + ager + sex + nseiz + period + tr
```

```
## Warning in fitter(X, Y, strats, offset, init, control, weights = weights, :
## Loglik converged before variable 1,2,3,4,5,6,7 ; beta may be infinite.
```

```
summary(mycox2)
```

```
## Call:
## coxph(formula = Surv(mess$intseiz1, mess$indseiz1, type = "right") ~
##     trt + ager + sex + nseiz + period + trt:intseiz1, data = mess)
##
##   n= 1420, number of events= 689
##    (5 observations deleted due to missingness)
##
##                    coef  exp(coef)   se(coef)        z Pr(>|z|)
## trt1          3.345e-02  1.034e+00  1.034e-01    0.323   0.7464
## ager         -2.550e-03  9.975e-01  2.213e-03   -1.152   0.2492
## sex1          4.139e-02  1.042e+00  8.621e-02    0.480   0.6312
## nseiz         4.084e-03  1.004e+00  2.156e-03    1.894   0.0582 .
## period       -1.902e-05  1.000e+00  3.870e-05   -0.492   0.6230
## trt0:intseiz1 -3.120e-01  7.320e-01  1.274e-02  -24.486  <2e-16 ***
## trt1:intseiz1 -3.123e-01  7.318e-01  1.275e-02  -24.493  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
```

```
## trt1                    1.0340      0.9671     0.8443     1.2664
## ager                    0.9975      1.0026     0.9931     1.0018
## sex1                    1.0423      0.9595     0.8802     1.2341
## nseiz                   1.0041      0.9959     0.9999     1.0083
## period                  1.0000      1.0000     0.9999     1.0001
## trt0:intseiz1           0.7320      1.3662     0.7139     0.7505
## trt1:intseiz1           0.7318      1.3665     0.7137     0.7503
##
## Concordance= 1   (se = 0.011 )
## Rsquare= 0.993    (max possible= 0.999 )
## Likelihood ratio test= 7142  on 7 df,    p=0
## Wald test            = 607.2  on 7 df,    p=0
## Score (logrank) test = 1329  on 7 df,    p=0
```

```
cox.zph(mycox2)
```

```
##                     rho  chisq        p
## trt1            -0.0066 0.0272 0.86893
## ager            -0.0124 0.0915 0.76230
## sex1            -0.0204 0.2505 0.61672
## nseiz           -0.0267 0.3813 0.53689
## period          -0.0206 0.2042 0.65139
## trt0:intseiz1    0.4074 7.8526 0.00507
## trt1:intseiz1    0.4073 7.8728 0.00502
## GLOBAL              NA 9.2191 0.23731
```

Next we will fit the model using the transformed versions of `period` and `nseiz`. The significant covariates are pretty much the same as in the first model, however the assumption of proportional hazards is not supported in the `trt` and `nseiz` terms.

```
mycox3 <- coxph(Surv(mess$intseiz1,mess$indseiz1,type="right") ~ trt + ager + sex + nseiz.tr + period.t
summary(mycox3)
```

```
## Call:
## coxph(formula = Surv(mess$intseiz1, mess$indseiz1, type = "right") ~
##      trt + ager + sex + nseiz.tr + period.tr, data = mess)
##
##   n= 1420, number of events= 689
##     (5 observations deleted due to missingness)
##
##                    coef exp(coef)  se(coef)       z Pr(>|z|)
## trt1           0.350542  1.419836  0.076872   4.560 5.11e-06 ***
## ager          -0.002727  0.997277  0.002094  -1.302   0.1929
## sex1          -0.040488  0.960321  0.078155  -0.518   0.6044
## nseiz.trTRUE   0.632579  1.882459  0.094127   6.720 1.81e-11 ***
## period.tr     -0.044294  0.956673  0.022166  -1.998   0.0457 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## trt1            1.4198     0.7043    1.2213    1.6507
## ager            0.9973     1.0027    0.9932    1.0014
## sex1            0.9603     1.0413    0.8239    1.1193
## nseiz.trTRUE    1.8825     0.5312    1.5653    2.2638
## period.tr       0.9567     1.0453    0.9160    0.9992
##
```

```
## Concordance= 0.596  (se = 0.011 )
## Rsquare= 0.049   (max possible= 0.999 )
## Likelihood ratio test= 70.87  on 5 df,   p=6.761e-14
## Wald test            = 71.28  on 5 df,   p=5.54e-14
## Score (logrank) test = 72.42  on 5 df,   p=3.209e-14
```

```
cox.zph(mycox3)
```

```
##                    rho  chisq      p
## trt1          -0.08570 5.0596 0.0245
## ager           0.00455 0.0147 0.9036
## sex1          -0.02122 0.3134 0.5756
## nseiz.trTRUE  -0.08135 4.5110 0.0337
## period.tr      0.06225 2.2954 0.1298
## GLOBAL              NA 9.7668 0.0821
```

Finally, we will include an interaction term between `trt` and `nseiz`, since it is plausible that the treatment may have a different effect on people who suffer more severely (indicated by having had more seizures). With this model the proportional hazards assumption seems to be satisfied. However, the interaction term is only slightly significant, and including the interaction reduces the estimated treatment effect.

```
mycox4 <- coxph(Surv(mess$intseiz1,mess$indseiz1,type="right") ~ trt + ager + sex + nseiz.tr + period.t
summary(mycox4)
```

```
## Call:
## coxph(formula = Surv(mess$intseiz1, mess$indseiz1, type = "right") ~
##     trt + ager + sex + nseiz.tr + period.tr + trt:nseiz.tr, data = mess)
##
##   n= 1420, number of events= 689
##    (5 observations deleted due to missingness)
##
##                        coef exp(coef)  se(coef)      z Pr(>|z|)
## trt1               0.218270  1.243923  0.109949  1.985   0.0471 *
## ager              -0.002818  0.997186  0.002092 -1.347   0.1779
## sex1              -0.045141  0.955863  0.078191 -0.577   0.5637
## nseiz.trTRUE       0.493679  1.638332  0.125392  3.937 8.25e-05 ***
## period.tr         -0.044871  0.956120  0.022140 -2.027   0.0427 *
## trt1:nseiz.trTRUE  0.256384  1.292249  0.153413  1.671   0.0947 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## trt1                 1.2439     0.8039    1.0028    1.5431
## ager                 0.9972     1.0028    0.9931    1.0013
## sex1                 0.9559     1.0462    0.8200    1.1142
## nseiz.trTRUE         1.6383     0.6104    1.2814    2.0948
## period.tr            0.9561     1.0459    0.9155    0.9985
## trt1:nseiz.trTRUE    1.2922     0.7738    0.9567    1.7455
##
## Concordance= 0.596  (se = 0.011 )
## Rsquare= 0.051   (max possible= 0.999 )
## Likelihood ratio test= 73.66  on 6 df,   p=7.239e-14
## Wald test            = 77.9  on 6 df,   p=9.659e-15
## Score (logrank) test = 80.83  on 6 df,   p=2.442e-15
```

```
cox.zph(mycox4)
```

```
##                       rho  chisq     p
## trt1              -0.03125 0.6735 0.412
## ager              0.00466  0.0154 0.901
## sex1              -0.02076 0.3002 0.584
## nseiz.trTRUE      -0.03370 0.7521 0.386
## period.tr         0.06073  2.1829 0.140
## trt1:nseiz.trTRUE -0.03316 0.7581 0.384
## GLOBAL                 NA  9.4025 0.152
```
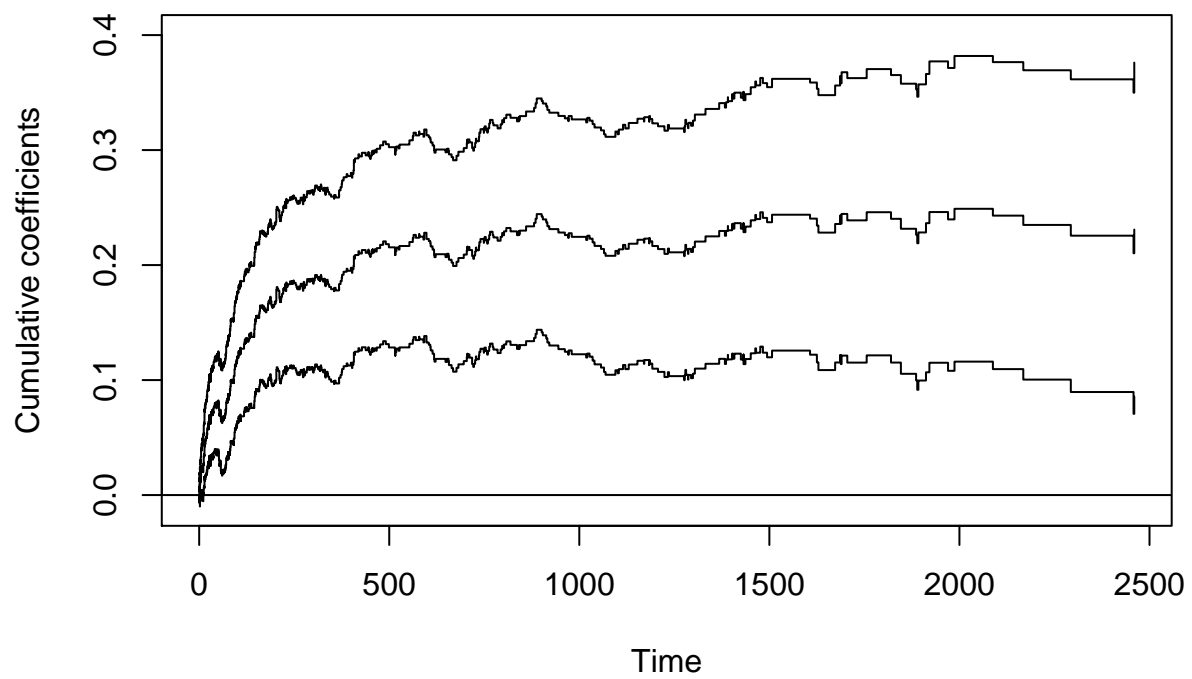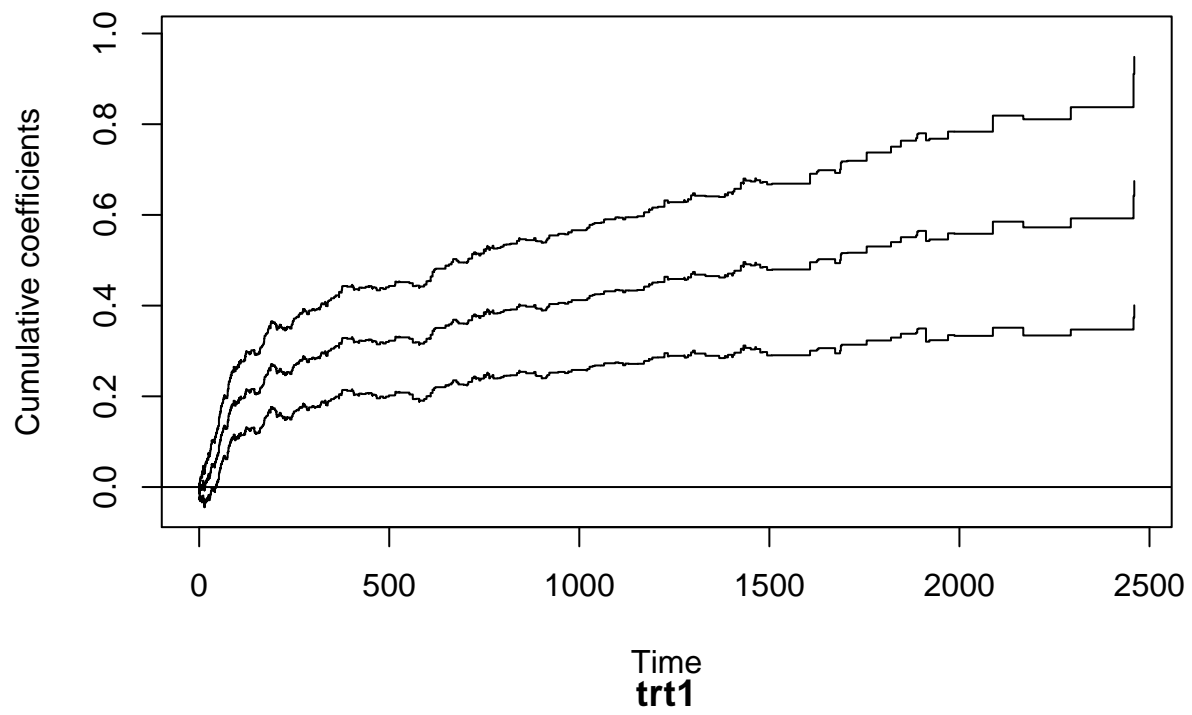
## Aalen's additive model

```
modaalen<-aalen(surv1 ~ trt + const(ager) + const(sex) + const(nseiz) + const(period) + const(abeeg), da
summary(modaalen)
```

```
## Additive Aalen Model
##
## Test for nonparametric terms
##
## Test for non-significant effects
##              Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                           16.50                   0
## trt1                                   4.78                   0
##
## Test for time invariant effects
##                 Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                       0.256                            0
## trt1                              0.164                            0
##                   Cramer von Mises test p-value H_0:constant effect
## (Intercept)                        78.0                            0
## trt1                               26.5                            0
##
## Parametric terms :
##                Coef. SE Robust SE    z P-val lower2.5% upper97.5%
## const(ager)        0  0         0  NaN   NaN        0          0
## const(sex)1        0  0         0  NaN   NaN        0          0
## const(nseiz)       0  0         0  NaN   NaN        0          0
## const(period)      0  0         0  NaN   NaN        0          0
## const(abeeg)1      0  0         0  NaN   NaN        0          0
##
##   Call:
## aalen(formula = surv1 ~ trt + const(ager) + const(sex) + const(nseiz) +
##     const(period) + const(abeeg), data = mess, n.sim = 100)
```

```
#par(mfrow=c(3,3))
plot(modaalen)
```

## (Intercept)



## trt1
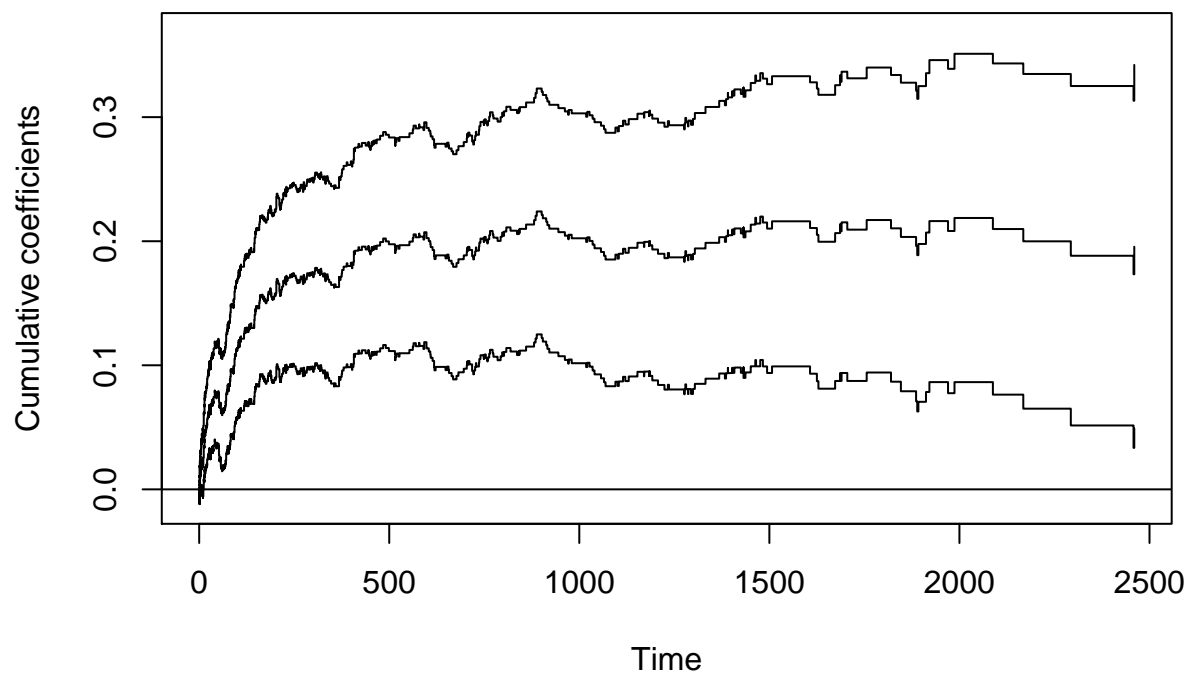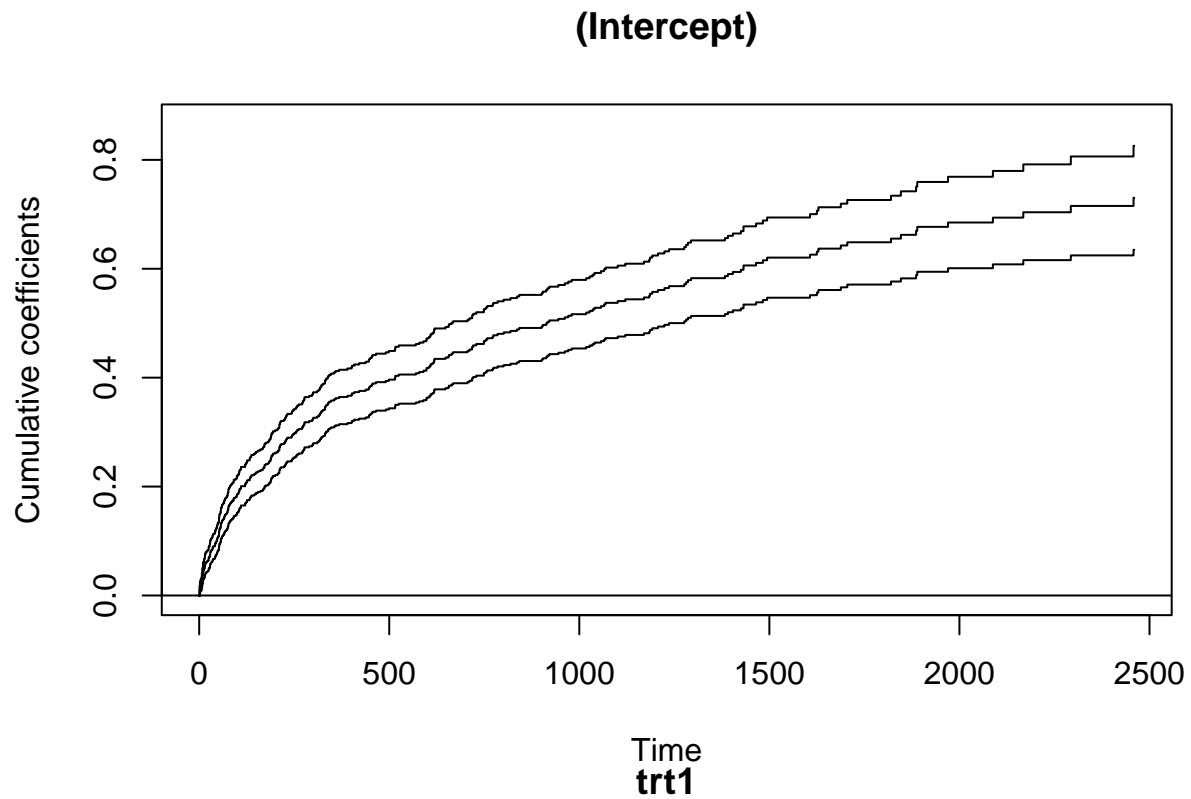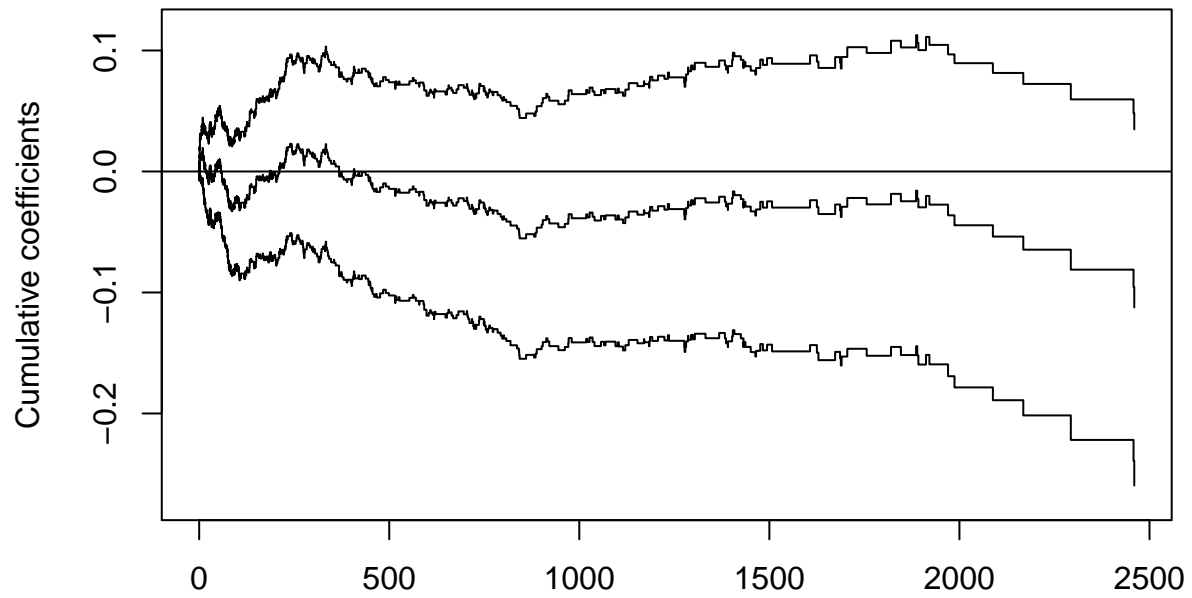


```
modaalen2<-aalen(surv1 ~ trt + const(ager) + sex + nseiz.tr + period.tr + abeeg , data = mess , n.sim=1(
#modaalen3<-aalen(surv1 ~ . - ager + const(ager), data = covars , n.sim=100)
x11()
plot(modaalen2)
```
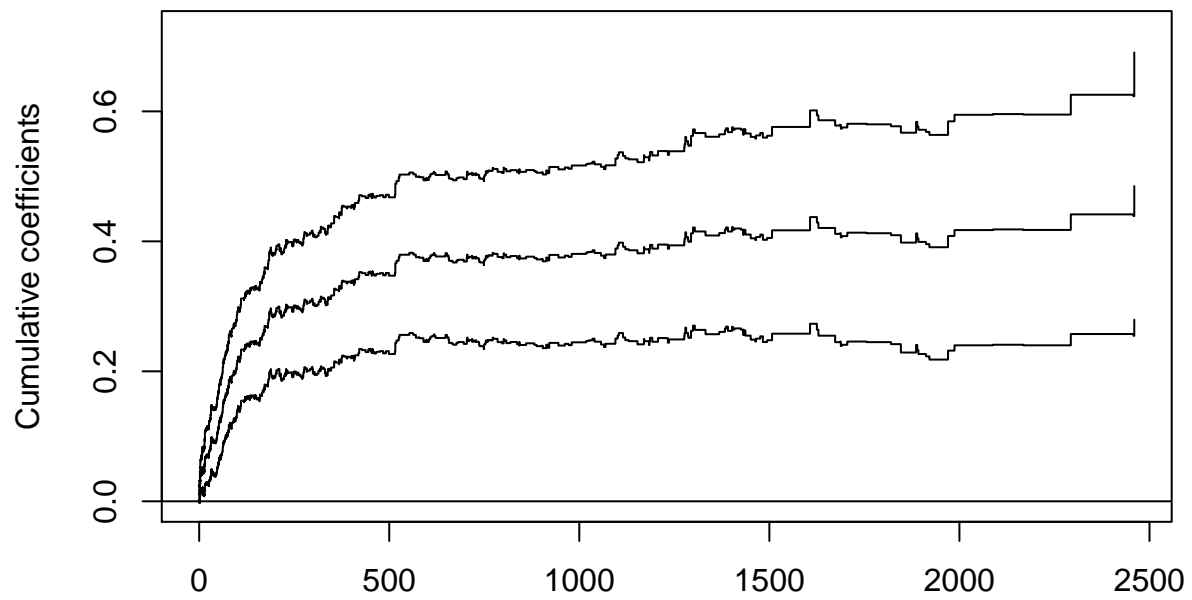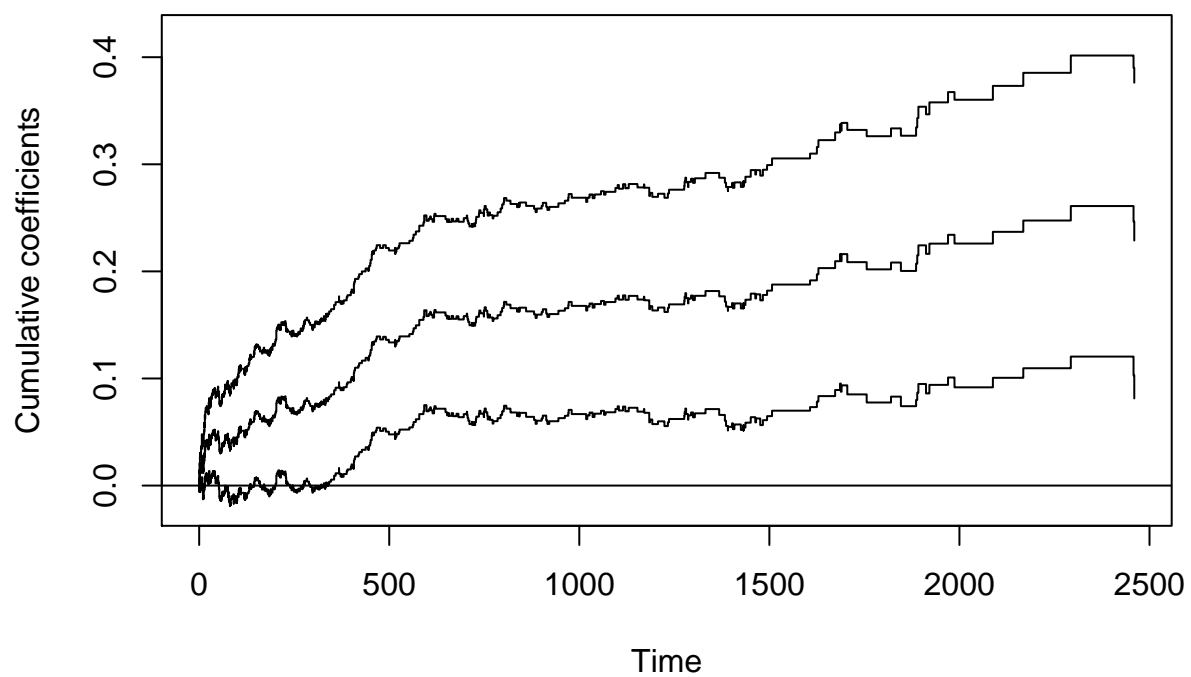
14

# (Intercept)



Time

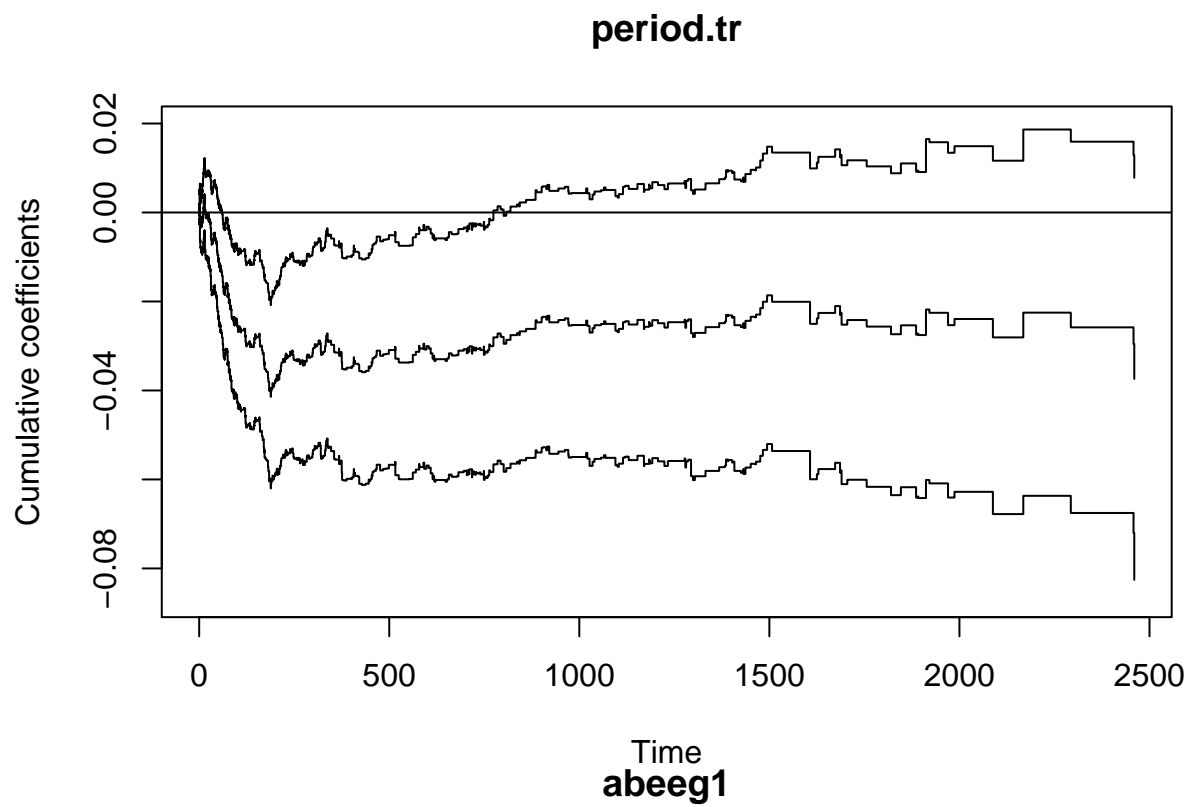# trt1



Time

# sex1



# nseiz.trTRUE

**period.tr**



**abeeg1**
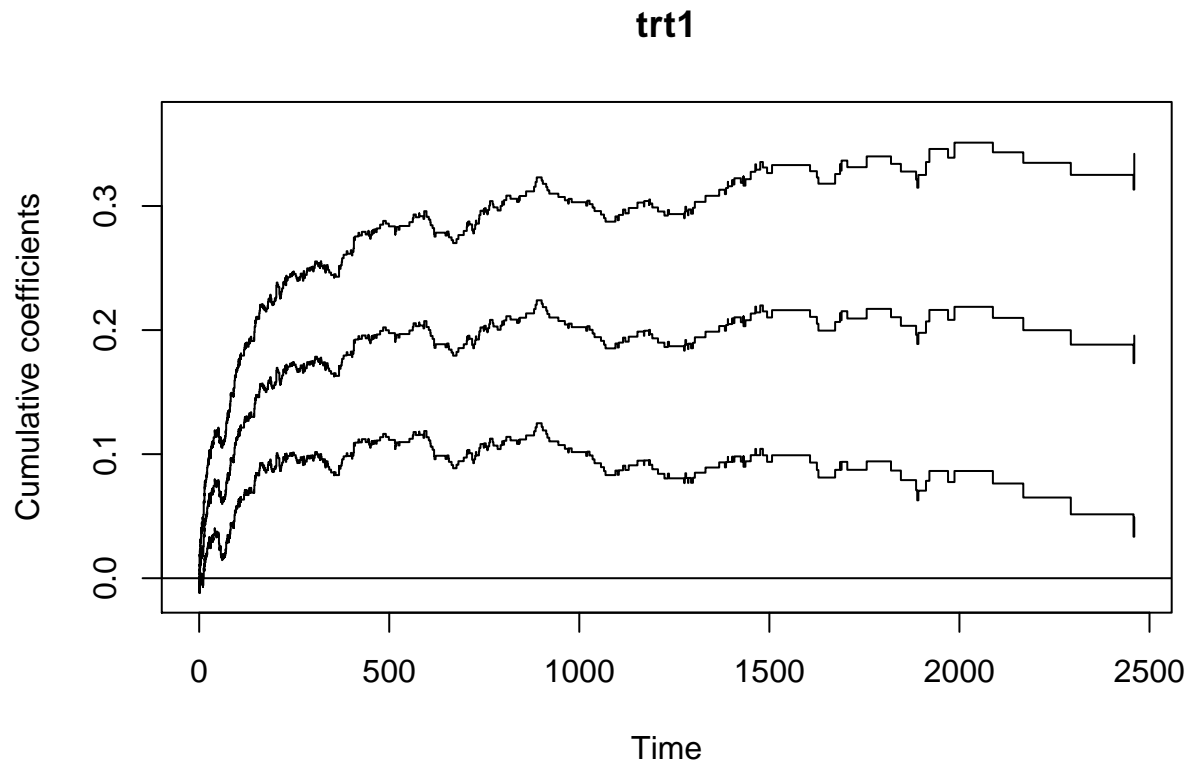


```r
dev.off()
```

```
## pdf
##   2
```

## trt1



## Parametric survival models

```r
modpar1 <- survreg(surv1 ~ trt + ager + sex + nseiz + period + abeeg, mess, dist="exponential")
#summary(modpar1)
modpar2 <- survreg(surv1 ~ trt + ager + sex + nseiz + period + abeeg, mess, dist="weibull")
#summary(modpar2)
modpar3 <- survreg(surv1 ~ trt + ager + sex + nseiz + period + abeeg, mess, dist="lognormal")
#summary(modpar3)
modpar4 <- survreg(surv1 ~ trt + ager + sex + nseiz + period + abeeg, mess, dist="loglogistic")
#summary(modpar4)

# modpar1 <- flexsurvreg(surv1 ~ trt + ager + sex + nseiz + period, data = mess, dist="weibull")
# summary(modpar1)
# modpar2 <- flexsurvreg(surv1 ~ trt + ager + sex + nseiz + period, data = mess, dist="lnorm")
# summary(modpar2)
# modpar3 <- flexsurvreg(surv1 ~ trt + ager + sex + nseiz + period, data = mess, dist="llogis")
# summary(modpar3)
# modpar5 <- flexsurvreg(surv1 ~ trt + ager + sex + nseiz + period, data = mess, dist="gengamma")
# summary(modpar5)

parmodels<-matrix(NA,nrow=4,ncol=4)
parmodels[1,] <- c(summary(modpar1)$n, modpar1$loglik[1],summary(modpar1)$df, AIC(modpar1))
parmodels[2,] <- c(summary(modpar2)$n, modpar2$loglik[1],summary(modpar2)$df, AIC(modpar2))
parmodels[3,] <- c(summary(modpar3)$n, modpar3$loglik[1],summary(modpar3)$df, AIC(modpar3))
parmodels[4,] <- c(summary(modpar4)$n, modpar4$loglik[1],summary(modpar4)$df, AIC(modpar4))
```

```r
colnames(parmodels) <- c("N.obs","Loglik","df","AIC")
rownames(parmodels) <- c("exponential","weibull","lognormal","loglogistic")
parmodels
```

```
##              N.obs    Loglik df      AIC
## exponential   1420 -5900.726  7 11696.03
## weibull       1420 -5548.895  8 11035.36
## lognormal     1420 -5511.851  8 10948.83
## loglogistic   1420 -5527.916  8 10983.18
```

Among the accelerated failure time (AFT) models, lognormal model is preferred.

```r
summary(modpar3)
```

```
##
## Call:
## survreg(formula = surv1 ~ trt + ager + sex + nseiz + period +
##     abeeg, data = mess, dist = "lognormal")
##                 Value Std. Error       z        p
## (Intercept)  7.844287    2.37e-01 33.1448 6.73e-241
## trt1        -0.827606    1.75e-01 -4.7407  2.13e-06
## ager         0.004877    4.70e-03  1.0372  3.00e-01
## sex1        -0.017270    1.78e-01 -0.0972  9.23e-01
## nseiz       -0.044123    6.13e-03 -7.1975  6.13e-13
## period       0.000208    7.42e-05  2.8052  5.03e-03
## abeeg1      -0.573969    1.79e-01 -3.2130  1.31e-03
## Log(scale)   1.066196    2.97e-02 35.9255 1.22e-282
##
## Scale= 2.9
##
## Log Normal distribution
## Loglik(model)= -5466.4   Loglik(intercept only)= -5511.9
##   Chisq= 90.87 on 6 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 3
## n=1420 (5 observations deleted due to missingness)
```

```r
# pct <- 1:98/100   # The 100th percentile of predicted survival is at +infinity
# ptime <- predict(modpar2, newdata=data.frame(trt=0,ager=24,sex=0,nseiz=1,period= 438.4), type='quanti
# matplot(cbind(ptime$fit, ptime$fit + 2*ptime$se.fit, ptime$fit - 2*ptime$se.fit)/30.5, 1-pct,xlab="Mo
```

Now let us consider the transformed variables.

```r
modpar1.tr <- survreg(surv1 ~ trt + scale(ager,scale = F) + sex + nseiz.tr + period.tr + abeeg, mess, d
modpar2.tr <- survreg(surv1 ~ trt + scale(ager,scale = F) + sex + nseiz.tr + period.tr + abeeg, mess, d
modpar3.tr <- survreg(surv1 ~ trt + scale(ager,scale = F) + sex + nseiz.tr + period.tr + abeeg, mess, d
modpar4.tr <- survreg(surv1 ~ trt + scale(ager,scale = F) + sex + nseiz.tr + period.tr + abeeg, mess, d

parmodels.tr<-matrix(NA,nrow=4,ncol=4)
parmodels.tr[1,] <- c(summary(modpar1.tr)$n, modpar1.tr$loglik[1],summary(modpar1.tr)$df, AIC(modpar1.t
parmodels.tr[2,] <- c(summary(modpar2.tr)$n, modpar2.tr$loglik[1],summary(modpar2.tr)$df, AIC(modpar2.t
parmodels.tr[3,] <- c(summary(modpar3.tr)$n, modpar3.tr$loglik[1],summary(modpar3.tr)$df, AIC(modpar3.t
parmodels.tr[4,] <- c(summary(modpar4.tr)$n, modpar4.tr$loglik[1],summary(modpar4.tr)$df, AIC(modpar4.t
colnames(parmodels.tr) <- c("N.obs","Loglik","df","AIC")
rownames(parmodels.tr) <- c("exponential","weibull","lognormal","loglogistic")
parmodels.tr #the difference to be interpreted in terms of chisq(1) = 3.84
```

```
##                N.obs    Loglik df      AIC
## exponential  1420 -5900.726  7 11692.63
## weibull      1420 -5548.895  8 11024.34
## lognormal    1420 -5511.851  8 10950.74
## loglogistic  1420 -5527.916  8 10979.20
modpar.logn <- survreg(surv1 ~ . + trt:nseiz.tr , covars1, dist="lognormal") #using covars before it is
```

```
covars <- transmute(mess, ager, sex, trt, nsp, ncp, nps, nmyo, nab, naab, ntc, noth, nseiz.tr, period.t
###DFNEW NOT WORKING!
#dfNew <- data.frame(sex=factor(c(0,0)), trt=factor(c(0,1)),ager=84,nseiz.tr=factor(c(1,1)),period.tr =
#percs <- (1:99)/100
#FLogN <- predict(modpar3.tr, newdata=dfNew, type="quantile", p=percs, se=TRUE)
#matplot(cbind(FLogN$fit[1, ],
#              FLogN$fit[1, ] - 2*FLogN$se.fit[1, ],
#              FLogN$fit[1, ] + 2*FLogN$se.fit[1, ])/365.25, 1-percs,
#       type="l", main=expression(paste("LogNormal-Fit ", hat(S)(t), " with SE")),xlim = c(0,10),
#       xlab="Time (in years)", ylab="Survival", lty=c(1, 2, 2), lwd=2, col="blue")
#matlines(cbind(FLogN$fit[2, ],
#              FLogN$fit[2, ] - 2*FLogN$se.fit[2, ],
#              FLogN$fit[2, ] + 2*FLogN$se.fit[2, ])/365.25, 1-percs, col="red", lwd=2)
#legend(x="topright", lwd=2, lty=c(1, 2, 1, 2), col=c("blue", "blue", "red", "red"),
#       legend=c("Trt 1 (immediate)", "+- 2*SE", "Trt 0 (deferred)", "+- 2*SE"))
```

## Survival Lasso

```
#
# XX <- model.matrix(Surv(mess$intseiz1, mess$indseiz1, type="right") ~ ager + sex + trt + nsp + ncp + 
#
# YY=Surv(mess$intseiz1, mess$indseiz1, type="right")
# YY <- YY[complete.cases(covars),] #remove rows with missing values (there are only 5/1425)
#
# mynet <- glmnet(x=XX, y=YY, family="cox")
# mycvnet <- cv.glmnet(x=XX, y=YY, family="cox")
#
# plot_glmnet(mynet, s=mycvnet$lambda.min)
```

## Random survival forests

We would like to ensure that the set of covariates we include in the model is relatively stable. Up to now we have only selected covariates insofar as the fitted log-normal model has some significant and some insignificant covariates. We now compare the set of significant covariates in the log-normal model to those obtained using a non-parametric variable selection technique, namely random survival forests.

Again we omit the centre variable since it destabilises the model, having a large number of levels many of which correspond to only one observation. If we had some additonal information say about suitably groupings of centres it would be sensible to include it. For instance, in Marson et al 2005 the authors distinguish between UK and non-UK centres. However we do not have any information linking the number of a centre to its location. However we now inlcude the counts of every type of seizure, and the results of all tests, as possible covariates.

We still omit date variables. we found that the error rate didn't really decrease by creating more than 100-200 trees, so we use `ntree = 200`. There is not a universal threshold on the `vimp` above which variables should be considered selected, but in the literature `vimp=0.02` has been suggested as a rule-of-thumb.

```
rsf.surv1<-rfsrc(Surv(intseiz1, indseiz1, type="right") ~ ager + sex + trt + nsp + ncp + nps + nmyo + na
```

```
#print(rsf.surv1)
#plot(rsf.surv1)
gg_dta <- gg_vimp(rsf.surv1)
plot(gg_dta)
```
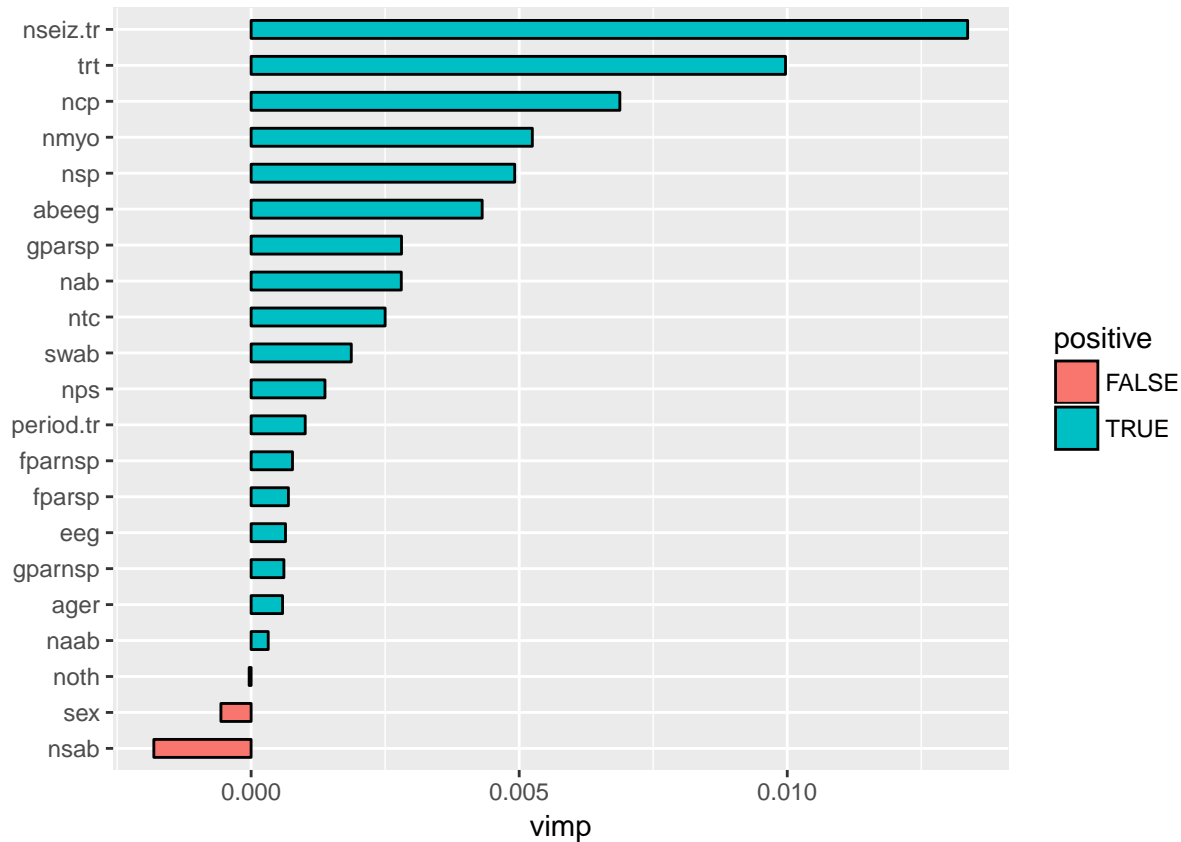


As we found in our exploratory analysis, `nseiz` and `trt` are the most important covariates, with `abeeg` also significant. We also see the counts of several particular types of seizure are important, which weren't included in our initial analysis.

The six most important variables in random forests were all significant in the log-normal model, which is encouraging. On the whole, random forests has selected roughly the same variables as the log-normal model, with only a few discrepancies. For instance, `nps` (one of the seizure counts) was significant in the log-normal model (p=0.009) but was certainly not selected by random forests.

For the purposes of our model, we will include only the covariates which the parametric and non-parametric approaches have both selected, namely `nseiz`, `trt`, `nsp`, `ncp`, `nmyo`, and `abeeg`. This captures the most important covariates of each approach.

Aalen, Odd O. 1989. "A Linear Regression Model for the Analysis of Life Times." *Statistics in Medicine* 8 (8). Wiley Online Library: 907–25.