Global Covariates (Student)

Bianchi, Boschi, Filippi-Mazzola, Juozaitiene, Wit

2024-06-21

Exercise: Email communication in Manufacturing Company

We the history of internal email communication among employees of a mid-sized manufacturing company located in Poland (Michalski et al., 2014). The study contains information about 57,791 email communications observed between 159 employees, where each record consists of a sender and recipient, as well as the timestamp of the sent email. The analysed communication history covers the period of nine months, starting from 2010-01-01 to 2010-09-30. Each timestamp records the day since 2010-01-01. The original data was pre-processed, ensuring no simultaneity of events. Although it is understood that repetition, reciprocity and general heterogeneity play an important role, the specific aim of this tutorial is to evaluate the role of the *time of day* and *time of the week*, which are global covariates.

Exercise 1.1. Load email data

Download the data and load the data in memory.

```
load("00-Data/mnf-raw.RData")
n<-nrow(mnf.raw)
```

Exercise 1.2. Visualize time of day and day of week

Consider the data and make a histogram of the time of day and day of the week. Draw some conclusions whether both terms are important to be included for an analysis of the email interactions in this context.

```
#Use the function hist with an appropriate number of breaks to visualize when interactions
#tend to take place
```

As can be seen from both graphs, information about the time of day and day of the week clearly affect the intensity of the interactions. For this reason, it is important that both terms are included in the relational event model.

Exercise 1.3. Creating time shifts

For all possible events (s, r) sample time shifts,

$$\tau_{12}, \tau_{13}, \ldots, \tau_{159,158} \overset{\text{iid}}{\sim} \text{Exp}(\mu)$$

Lembo et al. (2024) show that large shifts, i.e., small μ, are most informative about the global covariates. However, if too large shifts are applied, it may be that there are some events that do not correspond to any non-events in the shifted process. These events are effectively lost.

Select a shift distribution $\text{Exp}(\mu)$ such that not more than 1% of the events do not have associated non-events.

```
# Simulate shift with as a mean roughly the event time horizon
p<-159
# Choose mean shift m
```

```
m <- 1 # YOUR CHOICE (don't choose 1! it is bad)
tau<-rexp(p^2,1/m)
dgnl<-(0:(p-1))*p+(1:p)
tau[dgnl]<-NA
tms.shift<-mnf.raw$tms + tau[mnf.raw$s+(mnf.raw$r-1)*p]
T<-max(mnf.raw$tms)

# How many events do not have a matching non-event
sum(sapply(tms.shift, function(tm,tau){sum((tm>tau)&(tm<(tau+T)),na.rm = TRUE)==0},tau=tau))
```

```
## [1] 0
```

Exercise 1.4. Sample non-events

On the shifted process, sample for each (shifted) events a non-event at the same shifted time.

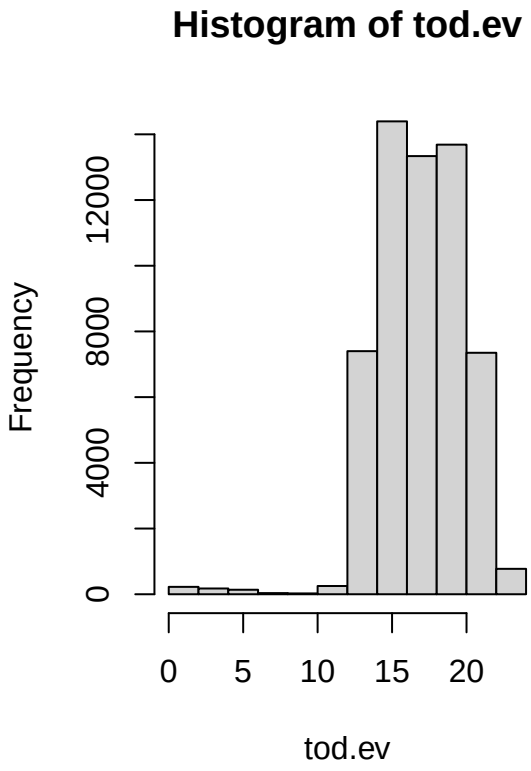
```
resample <- function(x, ...) x[sample.int(length(x), ...)]
event.ids<-mnf.raw$s+(mnf.raw$r-1)*p
non.event.ids<-apply(cbind(tms.shift,event.ids),1,
function(x,tau){resample(setdiff((1:length(tau))[(x[1]>tau)&(x[1]<(tau+T))],c(NA,x[2])))[1]},
tau=tau)
```

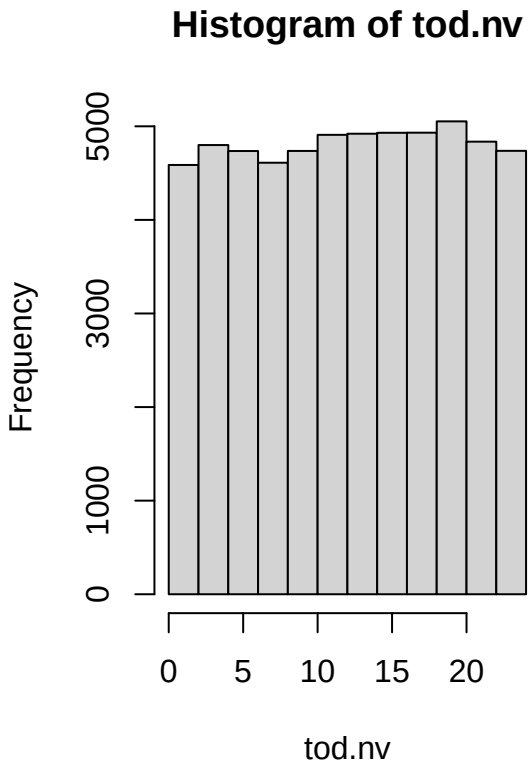
Exercise 1.5. Create covariates

First we create the Time of Day covariate. Then you are asked to create the Day of the Week covariate, as well as the baseline hazard.

Given that we will be fitting both effects non-linearly, we will store covariate separately for both event and non-event.

```
day.ev<-floor(mnf.raw$tms)
tod.ev<-(mnf.raw$tms-day.ev)*24
tms.nv<-mnf.raw$tms+tau[event.ids]-tau[non.event.ids]
day.nv<-floor(tms.nv)
tod.nv<-(tms.nv-day.nv)*24
par(mfrow=c(1,2))
hist(tod.ev)
hist(tod.nv)
```






```
dat<-data.frame(tod.ev=tod.ev,tod.nv=tod.nv)
```

Do the same thing for Day of the Week. Store the day of the week as a continuous value between 0 and 7. The baseline hazard is also a global covariate that merely records the time of the event (and non-event).

```
# create the covariates Day of Week and a general time covariate for the Baseline Hazard.
```

Exercise 1.6. Fit degenerate additive logistic regression model

We show how we can fit the global covariate Time of Day in a non-linear way in the Relational Event Model. Afterwards, you are asked to include also day of week and the baseline hazard. Given that we fit a non-linear effect, consider the material from session 3.

```
require(mgcv)
```

```
## Loading required package: mgcv
```

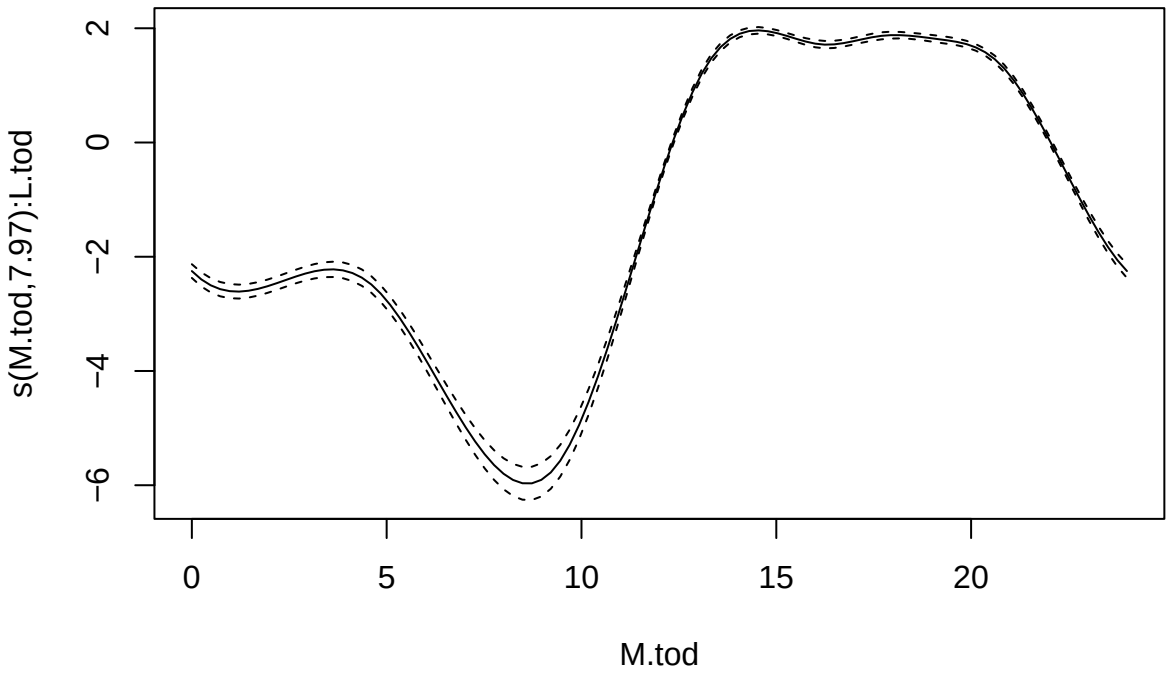
```
## Loading required package: nlme
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

```
library(mgcv)
one<-rep(1,n)
dat<-cbind(dat,one)
L.tod<-cbind(one,-one)
M.tod<-cbind(tod.ev,tod.nv)
# Note that we fit a CYCLIC spline by the option bs = "cc"
mnf1.rem<-gam(one~-1+s(M.tod,by=L.tod,bs="cc"),family = binomial)
summary(mnf1.rem)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## one ~ -1 + s(M.tod, by = L.tod, bs = "cc")
##
## Approximate significance of smooth terms:
##                   edf Ref.df Chi.sq p-value
## s(M.tod):L.tod 7.973      8   8818  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   -Inf   Deviance explained = -Inf%
## UBRE = -0.35266  Scale est. = 1         n = 57791
```

```
plot(mnf1.rem)
```



Add additional covariates to the model.

```
# Extend the previous model with additional covariates, such as the Day of the Week and the Baseline
# hazard. It is also possible to include random sender/receiver effects and perhaps repetition and
# reciprocity variables.
```