Computer Practical 3

Mixed Effect Additive Relational Event Models

Wit, EC., Lomi, A., Lerner, J., Boschi, M. & Lembo, M.

2025-06-24

Remark: Each CP begins by loading data from the previous CP, which is located in the corresponding `_OUTPUT_CP_2_` folder. To ensure this file runs smoothly, the contents of that folder should be copied to the `_INPUT_CP_3_` folder of the current CP.

1.1. Installing libraries

```
if (!require("mgcv", quietly = TRUE)) {
  install.packages("mgcv")
  library("mgcv")
} else {
  if (!require("mgcViz", quietly = TRUE)) {
    install.packages("mgcViz")
    library("mgcViz")
  } else {
    if (!require("ggplot2", quietly = TRUE)) {
      install.packages("ggplot2")
      library("ggplot2")
    } else {
      if (!require("survival", quietly = TRUE)) {
        install.packages("survival")
        library("survival")
      } else {
        if (!require("RColorBrewer", quietly = TRUE)){
          install.packages("RColorBrewer")
        } else {
          if (!require("dplyr", quietly = TRUE)){
          install.packages("dplyr")
          } else {
            library("mgcv")
            library("mgcViz")
            library("ggplot2")
            library("survival")
            library("RColorBrewer")
            library(dplyr)
          }
        }
      }
    }
  }
}
```

1.2. Loading Data

During Computer Practical 1, you computed the necessary statistics to support the inference techniques that will be explored in this second practical. In Computer practical 2, you processed the data to allow for inference using partial likelihood techniques. In particular you saw how, when only one non-event is available per observed event ($m = 1$), we can perform inference using a degenerate logistic regression. This is also the case, when we are to include time-varying, non-linear and random effect. Therefore `dat_gam_1` is the dataset we need.

```
load("_INPUT_CP_3_/dat_gam_1.RData")
head(dat_gam_1)
```

```
##   IS_OBSERVED_ev           SOURCE_ev EVENT_INTERVAL individual.activity_ev
## 1              1          |MY|NP|MB|              2                       0
## 2              1          |MY|ME|MB|              3                       2
## 3              1               |MY|              4                       2
## 4              1 |MY|ME|CL|GE|MC|MB|              5                       6
## 5              1          |MY|MB|ME|              6                       9
## 6              1             |ME|MY|              7                       8
##   dyadic.activity_ev closure_ev female_ev diff.female_ev .row_type_ev
## 1                  0          0         1              2           ev
## 2                  1          2         2              2           ev
## 3                  0          0         0              0           ev
## 4                  4          8         3              9           ev
## 5                  7         32         2              2           ev
## 6                  3         12         1              1           ev
##   IS_OBSERVED_nv           SOURCE_nv individual.activity_nv dyadic.activity_nv
## 1              0          |BM|MT|BL|                       0                  0
## 2              0          |JU|EN|TM|                       0                  0
## 3              0               |FV|                       0                  0
## 4              0 |BZ|GA|FV|FE|TM|BS|                       0                  0
## 5              0          |FV|CN|JV|                       0                  0
## 6              0             |BT|FT|                       0                  0
##   closure_nv female_nv diff.female_nv .row_type_nv y female diff_female
## 1          0         1              2          nv 1      0           0
## 2          0         1              2          nv 1      1           0
## 3          0         1              0          nv 1     -1           0
## 4          0         2              8          nv 1      1           1
## 5          0         1              2          nv 1      1           0
## 6          0         1              1          nv 1      0           0
##   individual_activity dyadic_activity closure
## 1                   0               0       0
## 2                   2               1       2
## 3                   2               0       0
## 4                   6               4       8
## 5                   9               7      32
## 6                   8               3      12
```

1. Time-varying effect

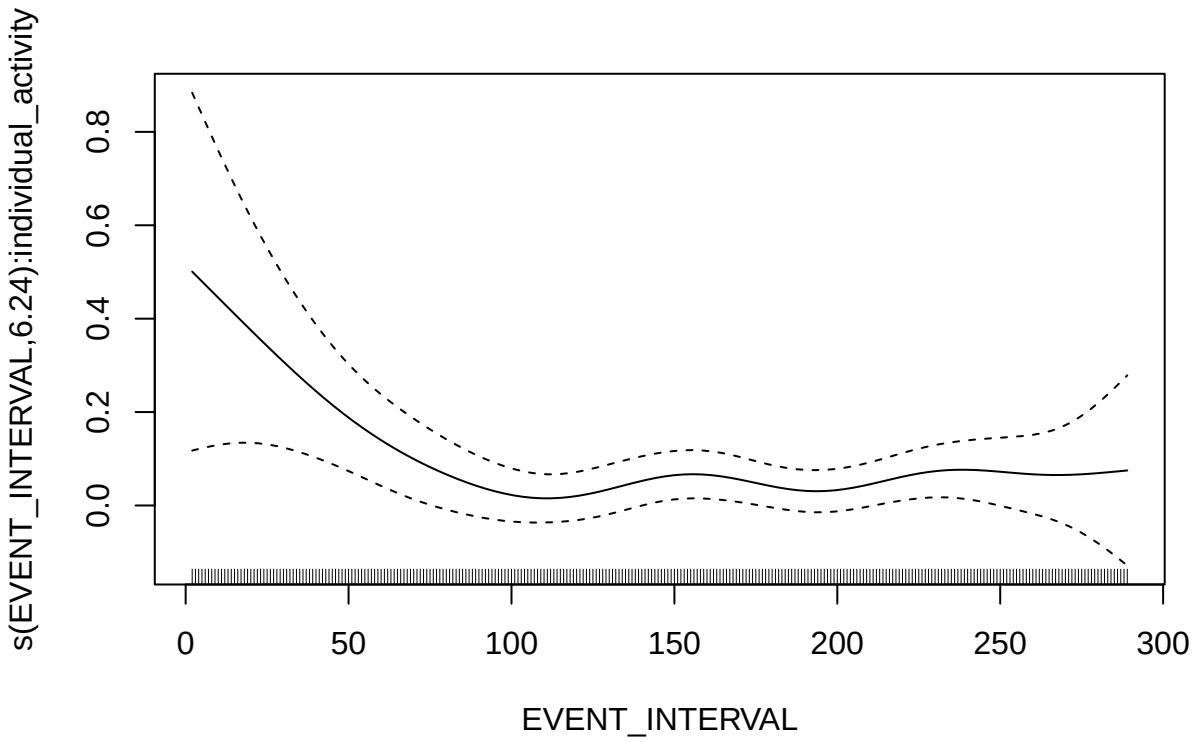
Let's investigate the possibility of **individual activity** having a TVE. Perhaps the impact it has on the characters' co-occurence dynamics depends on the chapter...

```
gam_fit_tve <- gam(y ~
                 diff_female
                #  + female
```

```
                #individual_activity
                + dyadic_activity
                + closure
                + s(EVENT_INTERVAL, by = individual_activity)
                #+ s(EVENT_INTERVAL, by = dyadic_activity)
                #+ s(EVENT_INTERVAL, by = closure)
                 - 1
                , data = dat_gam_1,
                family="binomial")
summary(gam_fit_tve)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## y ~ diff_female + dyadic_activity + closure + s(EVENT_INTERVAL,
##     by = individual_activity) - 1
##
## Parametric coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## diff_female     -0.31340    0.16693  -1.877  0.06046 .
## dyadic_activity  1.57101    0.49827   3.153  0.00162 **
## closure         -0.08281    0.04443  -1.864  0.06234 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                          edf Ref.df Chi.sq  p-value
## s(EVENT_INTERVAL):individual_activity 6.242  7.317  28.49 0.000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   -Inf   Deviance explained = -Inf%
## UBRE = -0.48664  Scale est. = 1         n = 288
```

```
plot(gam_fit_tve)
```



MODEL INTERPRETATION

Individual activity appears to have a positive effect on characters' co-occurrence that decreases until around chapter 50, after which it stabilizes at a value slightly above zero (orange horizontal line at 0.123) through to the end. However, the wider confidence intervals at the beginning and end of the timeline reduce the reliability of these estimates in those regions, suggesting that the true effect might differ from the apparent trend. Dyadic activity maintains a positive effect around 1.5: prior co-occurence in pairs increases the rate of a hyperevent. Closure, on the other hand, has a negative effect: hyperevents with actors that have appeared with a common third-party in the past, are less likely. Gender homophily also ha a negative effect, favouring co-appearence with other actors of the same gender.

2. Non-linear effect

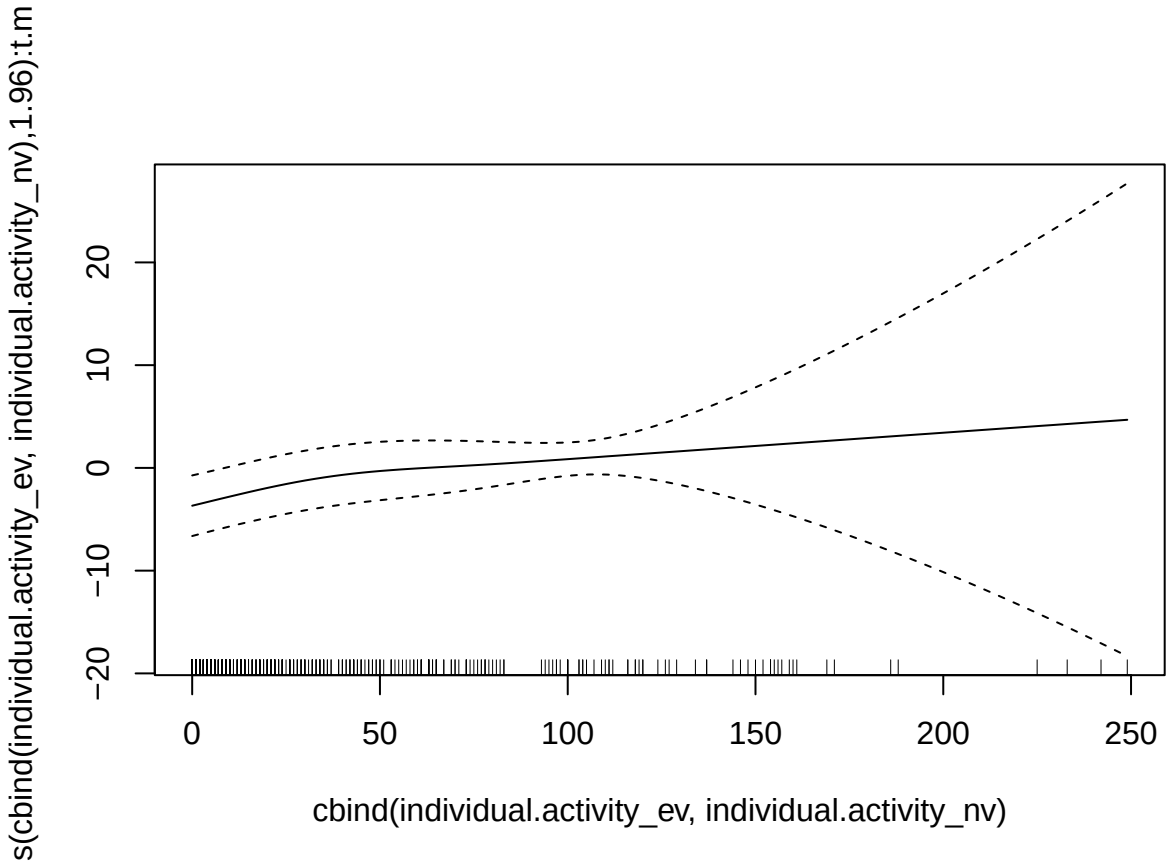
Let's try with a non-linear effect for individual activity instead...

```
t.mat <- cbind(rep(1, nrow(dat_gam_1)),rep(-1, nrow(dat_gam_1)))
gam_fit_nle <- gam(y ~
                    diff_female
                   #  + female
                  #individual_activity
                   + dyadic_activity
                   +closure
                   +  s(cbind(individual.activity_ev,individual.activity_nv),
                        by = t.mat)
                  #+ s(cbind(dyadic.activity_ev, dyadic.activity_nv), by = t.mat)
                  #+ s(cbind(closure_ev,closure_nv),by = t.mat)
                    - 1
                   , data = dat_gam_1,
                   family="binomial")
summary(gam_fit_nle)
```

```
##
```

```
## Family: binomial
## Link function: logit
##
## Formula:
## y ~ diff_female + dyadic_activity + closure + s(cbind(individual.activity_ev,
##     individual.activity_nv), by = t.mat) - 1
##
## Parametric coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## diff_female     -0.36039    0.14804  -2.434  0.01491 *
## dyadic_activity  1.60886    0.45615   3.527  0.00042 ***
## closure         -0.09160    0.03948  -2.320  0.02033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                                                    edf Ref.df
## s(cbind(individual.activity_ev, individual.activity_nv)):t.mat 1.962  2.398
##                                                                  Chi.sq  p-value
## s(cbind(individual.activity_ev, individual.activity_nv)):t.mat   28.8 1.58e-06
##
## s(cbind(individual.activity_ev, individual.activity_nv)):t.mat ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   -Inf   Deviance explained = -Inf%
## UBRE = -0.46442  Scale est. = 1         n = 288
```

```
plot(gam_fit_nle)
```



MODEL INTERPRETATION

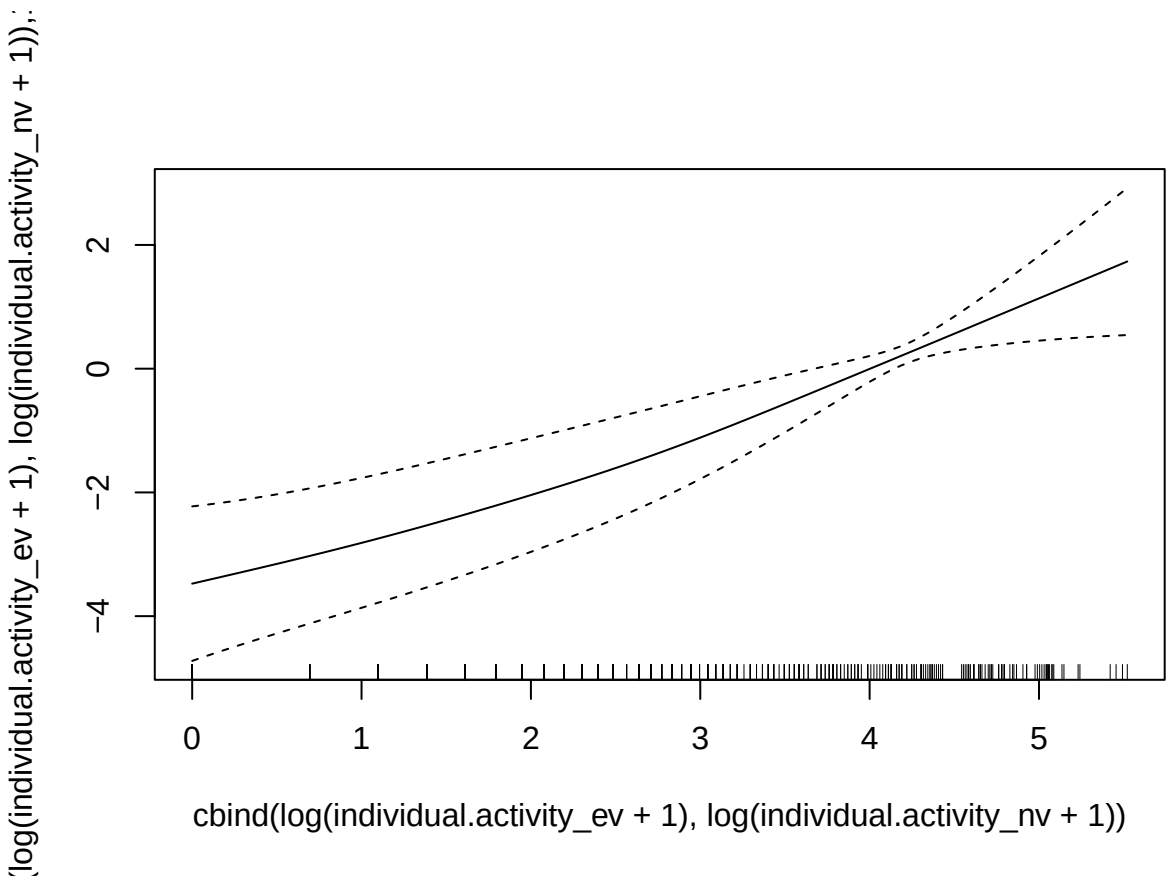
The result suggests an increasing trend but for value of the covariate greater than 50, the confidence interval width suggest poor reliability. Indeed, 85% of the observed values of individual activity are between 0 and 100. Perhaps a variance stabilizing transformation, e.g. log, of the covariate could help. We therefore proceed to include in the model a smooth term of log(individual activity + 1). Dyadic activity maintains a positive effect around 1.6, while closure appears to have a negative effect as before. Similarly for gender homophily.

```
t.mat <- cbind(rep(1, nrow(dat_gam_1)),rep(-1, nrow(dat_gam_1)))
gam_fit_nle_log <- gam(y ~
                 diff_female
                #  + female
               #individual_activity
                + dyadic_activity
                +closure
                +  s(cbind(log(individual.activity_ev+1),
                           log(individual.activity_nv+1)), by = t.mat)
               #+ s(cbind(dyadic.activity_ev, dyadic.activity_nv), by = t.mat)
               #+ s(cbind(closure_ev,closure_nv),by = t.mat)
                 - 1
                , data = dat_gam_1,
                family="binomial")
summary(gam_fit_nle)
```

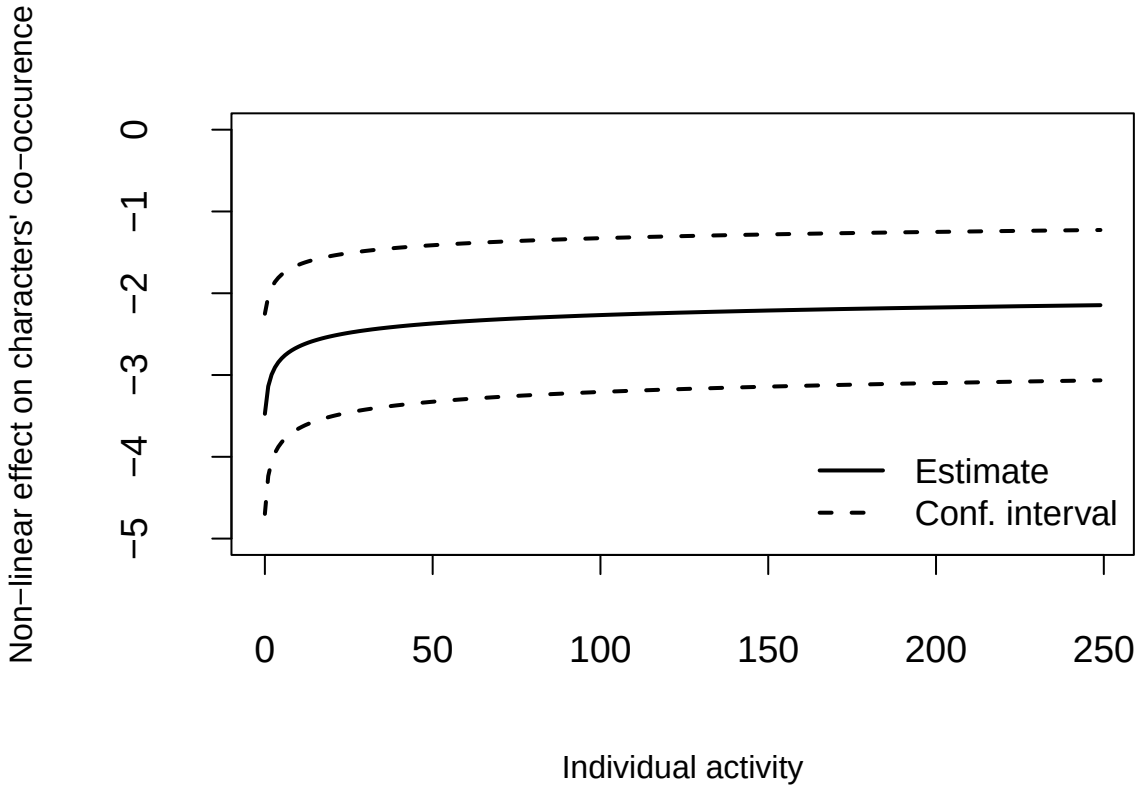
```
##
## Family: binomial
## Link function: logit
##
```

```
## Formula:
## y ~ diff_female + dyadic_activity + closure + s(cbind(individual.activity_ev,
##     individual.activity_nv), by = t.mat) - 1
##
## Parametric coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## diff_female     -0.36039    0.14804  -2.434  0.01491 *
## dyadic_activity  1.60886    0.45615   3.527  0.00042 ***
## closure         -0.09160    0.03948  -2.320  0.02033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                                                    edf Ref.df
## s(cbind(individual.activity_ev, individual.activity_nv)):t.mat 1.962  2.398
##                                                                  Chi.sq  p-value
## s(cbind(individual.activity_ev, individual.activity_nv)):t.mat   28.8 1.58e-06
##
## s(cbind(individual.activity_ev, individual.activity_nv)):t.mat ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   -Inf   Deviance explained = -Inf%
## UBRE = -0.46442  Scale est. = 1         n = 288
```

```
plot(gam_fit_nle_log)
```



```
ind_act<- sort(dat_gam_1$individual.activity_ev)
pred_nle<-predict(gam_fit_nle_log,
                  newdata = list(individual.activity_ev = log(ind_act + 1),
                                 individual.activity_nv = rep(0,length(ind_act)),
                                 t.mat = cbind(rep(1, length(ind_act)),
                                               rep(0, length(ind_act))),
                                 closure =rep(0,length(ind_act)),
                                 dyadic_activity=rep(0, length(ind_act)),
                                 diff_female = rep(0,length(ind_act))
                                 ), se.fit =T,
                  type = "terms")

par(mar=c(5,6,4,1)+1, mgp=c(5,2,0))
plot(ind_act, pred_nle$fit[,4], t='l',xlab ="Individual activity",
     ylab = "Non-linear effect on characters' co-occurence", ylim = c(-5,0)
     ,cex.lab=1, cex.axis=1.2, cex.main=3,col = "black",lwd = 2)
lines(ind_act,pred_nle$fit[,4] + 1.96*pred_nle$se.fit[,4],
      lty = 2, col = "black", lwd = 2)
lines(ind_act,pred_nle$fit[,4] - 1.96*pred_nle$se.fit[,4],
      lty = 2, col = "black", lwd = 2)
legend("bottomright",c("Estimate", "Conf. interval"),
       lty = c(1,2),lwd = c(2,2), col = c("black","black"), cex=1.1, bty = "n")
```



The first plot displays the estimated smooth non-linear effect (NLE) as a function of the log-transformed covariate. It is important to note that the values on the y-axis do not directly represent the magnitude or direction (positive/negative) of the effect. Due to identifiability constraints—specifically, identifiability up to an additive constant—NLEs are estimated to be centered around zero. Therefore, only the shape of the effect (e.g., increasing or decreasing trends) can be meaningfully interpreted.

The second plot shows the same estimated effect but with the covariate on its original scale. From this, we observe that hyper-edges with greater individual activity are more likely to occur. This increasing effect is most pronounced for lower values of the covariate (up to around 5), after which it tends to plateau.

Findings for dyadic activity, closure and gender homophily remain consistent with those observed in previous models.

3. Random effect

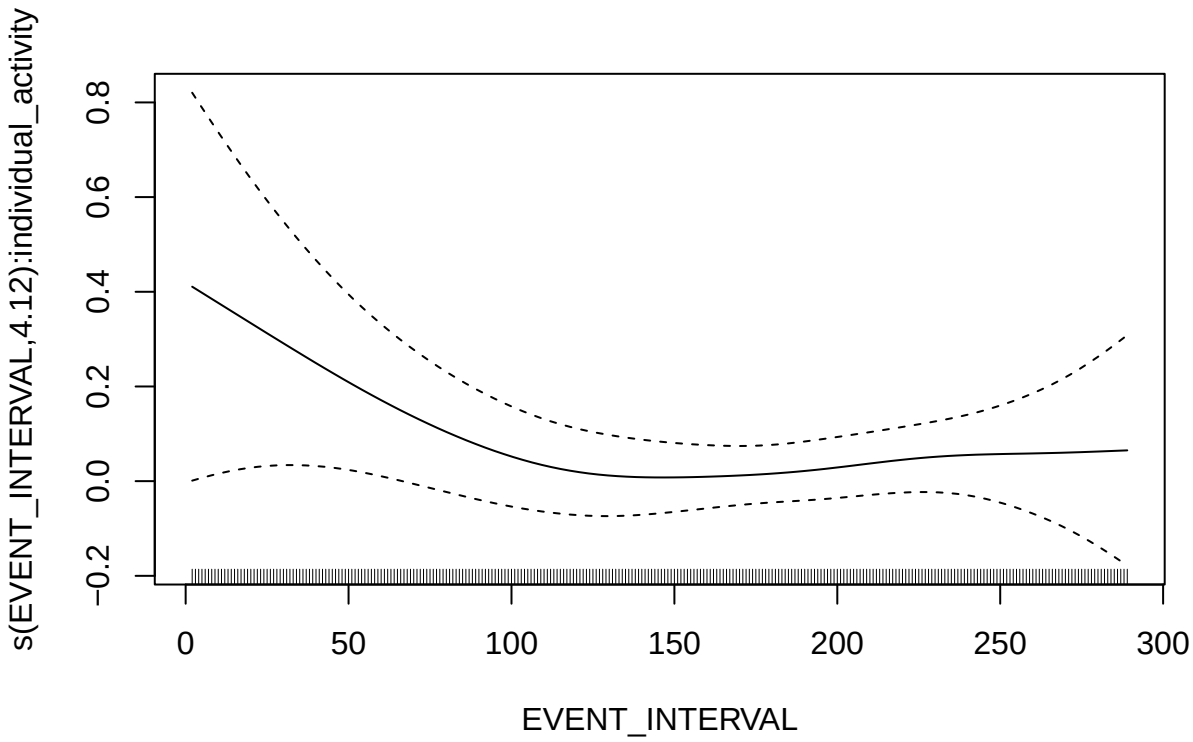
Not all group of characters (aka hyperedges) are the same. Therefore, we fit a model that includes random effects, specifically by introducing a random intercept for each hyper-edge. This allows us to account for unobserved heterogeneity at the hyper-edge level that is not explained by the observed covariates. In this specification, dyadic activity and closure are kept as linear effects, while individual activity is included as a time-varying effect.

```
source_factor <- factor(c(dat_gam_1$SOURCE_ev,dat_gam_1$SOURCE_nv))
dim(source_factor) <- c(nrow(dat_gam_1),2)

gam_fit_re <- gam(y ~ s(source_factor, by = t.mat, bs = "re") +
                    diff_female
                    #  + female
                    # + individual_activity
                    + dyadic_activity
                    #+ triadic_activity
                    + closure
                    #+ s(cbind(log(individual.activity_ev+1),
                    #+ log(individual.activity_nv+1)), by = t.mat)
                    + s(EVENT_INTERVAL, by = individual_activity)
                    - 1
                , data = dat_gam_1,
                family="binomial", method = "REML")

summary(gam_fit_re)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## y ~ s(source_factor, by = t.mat, bs = "re") + diff_female + dyadic_activity +
##     closure + s(EVENT_INTERVAL, by = individual_activity) - 1
##
## Parametric coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## diff_female     -0.35384    0.32530  -1.088   0.2767
## dyadic_activity  1.61021    0.72468   2.222   0.0263 *
## closure         -0.07883    0.07398  -1.066   0.2866
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                          edf  Ref.df Chi.sq p-value
## s(source_factor):t.mat                46.508 220.000 59.539  0.0408 *
## s(EVENT_INTERVAL):individual_activity  4.118   4.707  6.757  0.2058
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   -Inf   Deviance explained = -Inf%
## -REML = 73.561  Scale est. = 1         n = 288
```

```
plot(gam_fit_re,select = 2)
```



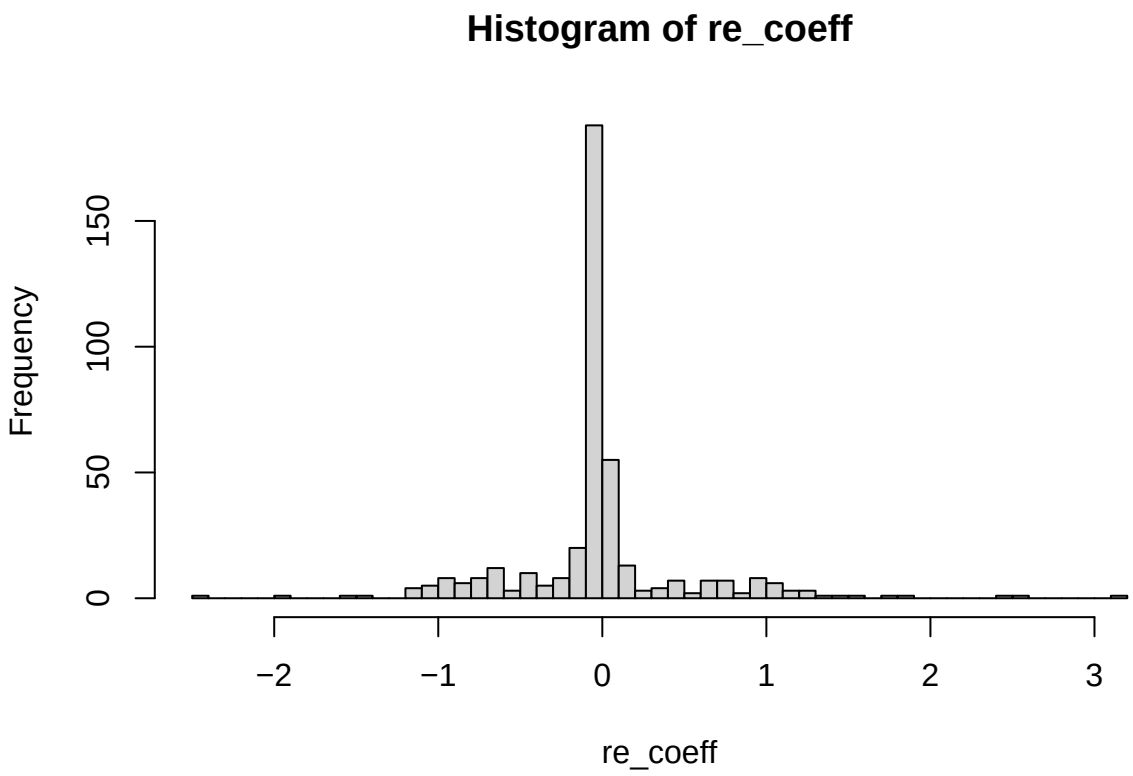

```
gam.vcomp(gam_fit_re)
```

```
##
## Standard deviations and 0.95 confidence intervals:
##
##                                             std.dev        lower       upper
## s(source_factor):t.mat                 2.3222329065 1.0825750071 4.981424508
## s(EVENT_INTERVAL):individual_activity 0.0004827313 0.0001621378 0.001437231
##
## Rank: 2/2
```

```
re_coeff<-coefficients(gam_fit_re)[4:412]
names(re_coeff)<-levels(source_factor)
print(sort(re_coeff, decreasing = TRUE)[1:5])
```

```
##          |MA|          |CO| |GU|BB|QU|MO|          |EN|          |GA|
##      3.103574      2.568576      2.413258      1.842885      1.785698
```

```
hist(re_coeff, breaks = 50)
```



MODEL INTERPRETATION

Gender homophily, closure and individual activity appear non-significant. Conclusion on the effect of dyadic activity remains as before. When examining the estimated random effects — ordered from highest to lowest — we observe that the largest positive effects are associated with small hyper-edges ,most of them of of size 1, indicating that certain characters are intrinsically more likely to appear. Interestingly, the character with the highest frequency, `|JV|` (Jean Valjean), appear only as 23th highest value in terms of random intercept value. In contrast, the second most frequent character, `|MA|` (Marius), has a higher random effect. This suggests that, after accounting for the covariates included in the model, Marius has a higher underlying propensity to appear in a chapter than Jean Valjean. The estimate of the variance of the random effect, ≈ 2.3, suggests moderate variability across groups of character. However, as the histogram shows, most of these intercepts take value close to 0 with only a few of them being largely positive and largely negative.

SAVING MATERIALS FOR CP4!:

```
save(dat_gam_1,
     gam_fit_tve,
     gam_fit_nle,
     gam_fit_nle_log,
     gam_fit_re, file="_OUTPUT_CP_3_/nonlinear_models.RData")
```