Description of system performance

AC

Contents

A first attempt

The goal here is to link system performance to characteristics of files and/or speakers on files. See this file for explanation of the fields.

Read in descriptor data

We read in background data (only exists for babytrain), data describing files and speakers. Typically, this does not depend on your system results, so you do not need to change it.

```
# options(warn=2) for debugging, then options(warn=1)

#background information (only available for babytrain)
read.csv("../BabyTrain_ages.csv")->ages
ages[ages$corpus!="corpus",]->ages
ages$age=as.numeric(as.character(ages$age))

allres=dir("../computation/results/",pattern=".csv")

#descriptors per speaker
datsp=NULL
for(j in allres[grep("perSpeaker",allres)])  datsp=rbind(datsp,cbind(j,read.csv(paste0("../computation/

#descriptors per file
datf=NULL
for(j in allres[grep("perSpeaker",allres,invert=T)])  datf=rbind(datf,cbind(j,read.csv(paste0("../compu

#descriptors per speaker
merge(datsp,ages,by.x="file",by.y="basename",all.x=T)->datsp
merge(datf,ages,by.x="file",by.y="basename",all.x=T)->datf

#show dimensions and summary of the 2 datasets
dim(datsp)
```

```
## [1] 8238    10
```

```
summary(datsp)
```

```
##                        file                                            j
##  namibia_aiku_20161111_19980:    9    BabyTrain_train_perSpeaker.csv:4163
##  namibia_eiun_20161113_30780:    9    BabyTrain_dev_perSpeaker.csv  :2071
##  namibia_oegd_20161109_23580:    8    BabyTrain_test_perSpeaker.csv :1254
##  namibia_oekd_20160712_12780:    8    AMI_train_perSpeaker.csv      : 471
##  namibia_oekd_20160712_16380:    8    AMI_dev_perSpeaker.csv        : 104
##  namibia_oekd_20170308_19980:    8    AMI_test_perSpeaker.csv       :  95
##  (Other)                    :8188    (Other)                       :  80
##     speaker        role       tot_ovl_speech      tot_nonovl_speech
##  MOT*   : 925    FEM   :2617   Min.   :   0.0000   Min.   :-297.429
##  C1     : 909    MAL   :1370   1st Qu.:   0.0000   1st Qu.:   0.000
##  MA1    : 563    CHI   :1375   Median :   0.0308   Median :   0.000
##  FA1    : 551    KCHI  :2329   Mean   :  33.5954   Mean   :  39.993
##  C2     : 414    SPEECH: 547   3rd Qu.:   2.9493   3rd Qu.:   7.659
##  UU     : 378                  Max.   :2688.8500   Max.   :2665.000
##  (Other):4498
##    snr            corpus            child.id          age
##  []   :8174    namibia  :3977    fhugo       : 208    Min.   : 1.00
##  NA's:  64     lena_lyon: 907    nath        : 122    1st Qu.:13.00
##                tsay     : 751    marin       : 118    Median :27.00
##                paido    : 649    uebn_20170309: 115   Mean   :25.04
##                tsimane  : 447    ern         : 109    3rd Qu.:33.00
##                (Other)  : 330    (Other)     :6389    Max.   :66.00
##                NA's     :1177    NA's        :1177    NA's   :1177
```

dim(datf)

```
## [1] 2881   16
```

summary(datf)

```
##                      file                        j          key_child_age
##  ES2003a.Mix-Headset:   1    BabyTrain_train.csv:1544    Mode:logical
##  ES2003b.Mix-Headset:   1    BabyTrain_dev.csv  : 736    NA's:2881
##  ES2003c.Mix-Headset:   1    BabyTrain_test.csv : 413
##  ES2003d.Mix-Headset:   1    AMI_train.csv      : 118
##  ES2011a.Mix-Headset:   1    AMI_dev.csv        :  26
##  ES2011b.Mix-Headset:   1    AMI_test.csv       :  24
##  (Other)            :2875    (Other)            :  20
##   clip_length     nb_diff_speakers  nb_children      nb_fem_ad
##  Min.   :   60.0  Min.   :1.000    Min.   :0.000    Min.   :0.0000
##  1st Qu.:   60.0  1st Qu.:1.000    1st Qu.:1.000    1st Qu.:0.0000
##  Median :   60.0  Median :3.000    Median :1.000    Median :1.0000
##  Mean   :  467.6  Mean   :2.859    Mean   :1.286    Mean   :0.9084
##  3rd Qu.:  300.0  3rd Qu.:4.000    3rd Qu.:2.000    3rd Qu.:1.0000
##  Max.   :10723.0  Max.   :9.000    Max.   :4.000    Max.   :4.0000
##
##    nb_mal_ad       nb_uncertain     prop_ovl_speech    prop_nonovl_speech
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.2000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.9200
##  Median :0.0000   Median :0.0000   Median :0.01000   Median :0.9900
##  Mean   :0.4755   Mean   :0.1899   Mean   :0.06742   Mean   :0.9326
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.08000   3rd Qu.:1.0000
##  Max.   :4.0000   Max.   :1.0000   Max.   :0.80000   Max.   :1.0000
##
```

```
##    avg_voc_dur            snr                   corpus          child.id
##  Min.   :   0.42   Min.   :  0.03003   namibia  :1062   fhugo  :  71
##  1st Qu.:  11.87   1st Qu.:  0.71457   paido    : 649   nath   :  48
##  Median :  24.70   Median :  0.84570   lena_lyon: 323   ern    :  36
##  Mean   : 190.21   Mean   : 13.56329   tsay     : 237   flore  :  36
##  3rd Qu.:  61.69   3rd Qu.:  1.96800   tsimane  : 154   leon   :  36
##  Max.   :7659.23   Max.   :104.58705   (Other)  : 122   (Other):2320
##                    NA's   :34          NA's     : 334   NA's   : 334
##       age
##  Min.   : 1.00
##  1st Qu.:15.50
##  Median :30.00
##  Mean   :29.19
##  3rd Qu.:39.00
##  Max.   :66.00
##  NA's   :334
```

Read in system data

HUMAN LOOK HERE Typically you WILL need to change line 50 below, so that you read in your own system results. Please use pyannote.metrics to generate your results. They should be space separated **HUMAN LOOK HERE**

```
file_eval <- read_table("../system_eval/BabyTrain_ConvRNN.txt", comment = "--")
```

```
## Warning: Duplicated column names deduplicated: '%' => '%_1' [10]
```

```
## Parsed with column specification:
## cols(
##   `Detection (collar = 0 ms)` = col_character(),
##   `detection error rate` = col_double(),
##   accuracy = col_double(),
##   precision = col_double(),
##   recall = col_double(),
##   total = col_double(),
##   `false alarm` = col_double(),
##   `%` = col_double(),
##   miss = col_double(),
##   `%_1` = col_double()
## )
```

```
dim(file_eval)
```

```
## [1] 414  10
```

```
summary(file_eval)
```

```
##  Detection (collar = 0 ms) detection error rate    accuracy
##  Length:414                Min.   :   1.87      Min.   :23.23
##  Class :character          1st Qu.:  14.94      1st Qu.:73.11
##  Mode  :character          Median :  38.53      Median :84.25
##                            Mean   :  76.83      Mean   :81.38
##                            3rd Qu.:  72.99      3rd Qu.:92.48
##                            Max.   :2177.42      Max.   :99.33
##     precision         recall           total           false alarm
##  Min.   :  0.00   Min.   :  0.00   Min.   :   0.97   Min.   :   0.00
```

```
##   1st Qu.: 60.83    1st Qu.: 78.56    1st Qu.:    19.81    1st Qu.:    3.59
##   Median : 81.83    Median : 91.59    Median :    37.30    Median :    8.20
##   Mean   : 73.59    Mean   : 84.93    Mean   :  389.02    Mean   :   51.30
##   3rd Qu.: 92.45    3rd Qu.: 97.86    3rd Qu.:    66.45    3rd Qu.:   20.46
##   Max.   :100.00    Max.   :100.00    Max.   :80526.23    Max.   :10619.92
##         %                 miss                %_1
##   Min.   :   0.000  Min.   :    0.000  Min.   :   0.000
##   1st Qu.:   7.303  1st Qu.:    0.565  1st Qu.:   2.143
##   Median :  17.050  Median :    2.410  Median :   8.405
##   Mean   :  61.760  Mean   :   30.841  Mean   :  15.071
##   3rd Qu.:  55.862  3rd Qu.:    8.620  3rd Qu.:  21.435
##   Max.   :2177.420  Max.   : 6384.020  Max.   : 100.000
```

```r
#the first col must be renamed
colnames(file_eval)[1]<-"file"


#you may also want to rename some variables into something that is more readable
colnames(file_eval)[colnames(file_eval)=="%"]<-"fa.pc"
colnames(file_eval)[colnames(file_eval)=="%_1"]<-"miss.pc"
```

Combine descriptor and system data

If all goes well, you won't need to change this section. After this code, the table file_eval has a combination of results and descriptors at the level of files.

```r
merge(file_eval,datf,all.x=T)->file_eval

dim(file_eval) #**human** check that the number of rows (first number) outputted here is the same as th
```

```
## [1] 414  25
```

```r
summary(file_eval)
```

```
##      file            detection error rate    accuracy        precision
##   Length:414       Min.   :    1.87     Min.   :23.23    Min.   :   0.00
##   Class :character  1st Qu.:   14.94     1st Qu.:73.11    1st Qu.:  60.83
##   Mode  :character  Median :   38.53     Median :84.25    Median :  81.83
##                     Mean   :   76.83     Mean   :81.38    Mean   :  73.59
##                     3rd Qu.:   72.99     3rd Qu.:92.48    3rd Qu.:  92.45
##                     Max.   : 2177.42     Max.   :99.33    Max.   : 100.00
##
##      recall          total          false alarm         fa.pc
##   Min.   :  0.00  Min.   :    0.97  Min.   :    0.00  Min.   :   0.000
##   1st Qu.: 78.56  1st Qu.:   19.81  1st Qu.:    3.59  1st Qu.:   7.303
##   Median : 91.59  Median :   37.30  Median :    8.20  Median :  17.050
##   Mean   : 84.93  Mean   :  389.02  Mean   :   51.30  Mean   :  61.760
##   3rd Qu.: 97.86  3rd Qu.:   66.45  3rd Qu.:   20.46  3rd Qu.:  55.862
##   Max.   :100.00  Max.   :80526.23  Max.   :10619.92  Max.   :2177.420
##
##       miss             miss.pc                       j
##   Min.   :   0.000  Min.   :   0.000  BabyTrain_test.csv:413
##   1st Qu.:   0.565  1st Qu.:   2.143  AMI_dev.csv       :  0
##   Median :   2.410  Median :   8.405  AMI_test.csv      :  0
##   Mean   :  30.841  Mean   :  15.071  AMI_train.csv     :  0
```

```
## 3rd Qu.:   8.620   3rd Qu.: 21.435   BabyTrain_dev.csv :   0
## Max.   :6384.020   Max.   :100.000   (Other)           :   0
##                                      NA's              :   1
## key_child_age  clip_length    nb_diff_speakers  nb_children
## Mode:logical   Min.   :  60.0  Min.   :1.000   Min.   :0.000
## NA's:414       1st Qu.:  60.0  1st Qu.:2.000   1st Qu.:1.000
##                Median :  60.0  Median :3.000   Median :1.000
##                Mean   : 326.4  Mean   :3.036   Mean   :1.433
##                3rd Qu.: 201.4  3rd Qu.:4.000   3rd Qu.:2.000
##                Max.   :3420.0  Max.   :8.000   Max.   :3.000
##                NA's   :1       NA's   :1       NA's   :1
##    nb_fem_ad       nb_mal_ad       nb_uncertain     prop_ovl_speech
## Min.   :0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
## 1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
## Median :1.00   Median :0.0000   Median :0.0000   Median :0.02000
## Mean   :1.01   Mean   :0.3729   Mean   :0.2203   Mean   :0.05332
## 3rd Qu.:2.00   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.09000
## Max.   :3.00   Max.   :2.0000   Max.   :1.0000   Max.   :0.70000
## NA's   :1      NA's   :1        NA's   :1        NA's   :1
## prop_nonovl_speech  avg_voc_dur         snr              corpus
## Min.   :0.3000    Min.   :   0.97   Min.   : 0.0977   namibia  :155
## 1st Qu.:0.9100    1st Qu.:  11.21   1st Qu.: 0.6831   paido    : 68
## Median :0.9800    Median :  20.51   Median : 0.7791   lena_lyon: 48
## Mean   :0.9467    Mean   : 105.37   Mean   : 9.8108   tsimane  : 42
## 3rd Qu.:1.0000    3rd Qu.:  34.51   3rd Qu.: 0.9818   tsay     : 40
## Max.   :1.0000    Max.   :1334.05   Max.   :88.2541   (Other)  : 26
## NA's   :1         NA's   :1         NA's   :1         NA's     : 35
##          child.id          age
## nohlan        : 36   Min.   : 3.00
## uebn_20170309: 26    1st Qu.:20.50
## C24           : 17   Median :27.00
## uoga_20170311: 14    Mean   :26.93
## uoga_20170313: 14    3rd Qu.:34.00
## (Other)      :272    Max.   :66.00
## NA's          : 35   NA's   :35
```

Now you are ready to do some inspection. You can turn chunks off by adding ", eval=F" (e.g. {r spl,fig.height=10} below, it would become {r spl,fig.height=10, eval=F})

Example of analysis: Explaining misses and false alarms across files based on file characteristics (Marvin VAD on BabyTrain - 5 class, old architecture)

A scatter plot matrix shows many bivariate plots. In the one below, we focus exclusively on descriptors at the level of the file and only for BabyTrain because that's what I drew results for. We only have false alarms and misses because we are looking at a VAD system. (In particular, this is Marvin's system for week 1.)

```r
library(lattice)
selected=c("fa.pc","miss.pc","prop_ovl_speech","avg_voc_dur","age","snr","nb_diff_speakers")
selnames=gsub(".","\n",gsub("_","\n",selected),fixed=T)
splom(file_eval[c(selected)],pch=".",groups=file_eval$corpus,varnames=selnames,auto.key = list(columns
panel = function(x, y, ...) {
        panel.xyplot(x, y, ...)
        fm <- lm(y ~ x)
        panel.abline(fm,col.line = "gray")
```

```
}
)
```

Focus on the last two rows, which show the correlations between percent misses (penultimate row) or percent false alarms (last row) and the following selected characteristics (from left to right):

- proportion of speech that is overlapping
- average vocalization/utterance/sentence duration
- key child age
- SNR calculated as RMS(x_speech)/RMS(x_sil) where x_speech is an array with all the areas of speech in the gold annotation, and x_sil is un array with all the areas of silence
- number of different speakers

So focusing on the last row, false alarms look unrelated to all of these predictors, although this may be because the scale is too large.

One row up, misses does not relate to proportion overlap or number of different speakers, but is anticorrelated with the duration of speech, child age, and SNR.

Example of analysis 2: ConvRNN version

You can also focus on specific outcome and predictor variables and trim their distribution to see them more clearly.

```
cor_color=rainbow(length(levels(factor(file_eval$corpus)))) #get different colors for diff datasets
names(cor_color)<-levels(factor(file_eval$corpus))

file_eval_metrics=c("fa.pc","miss.pc")
predictors=c("total","clip_length","prop_ovl_speech","avg_voc_dur","age","snr","nb_diff_speakers")

for(thismet in file_eval_metrics){
  iqr=IQR(file_eval[,thismet])
  med=median(file_eval[,thismet])
  no_outliers=file_eval[file_eval[,thismet]<med+1.5*iqr,]
  print(paste("removing",dim(file_eval)[1]-dim(no_outliers)[1], "outliers in",thismet))
  for(thispred in predictors){
    plot(no_outliers[,thismet]~no_outliers[,thispred], pch=20,col=alpha(cor_color[no_outliers$corpus],.
    abline(lm(no_outliers[,thismet]~no_outliers[,thispred]))
    if(max(no_outliers[,thismet])>300){
        plot(no_outliers[,thismet]~no_outliers[,thispred], pch=20,col=alpha(cor_color[no_outliers$corpu
    abline(lm(no_outliers[,thismet]~no_outliers[,thispred]))
    }
  }
}
```
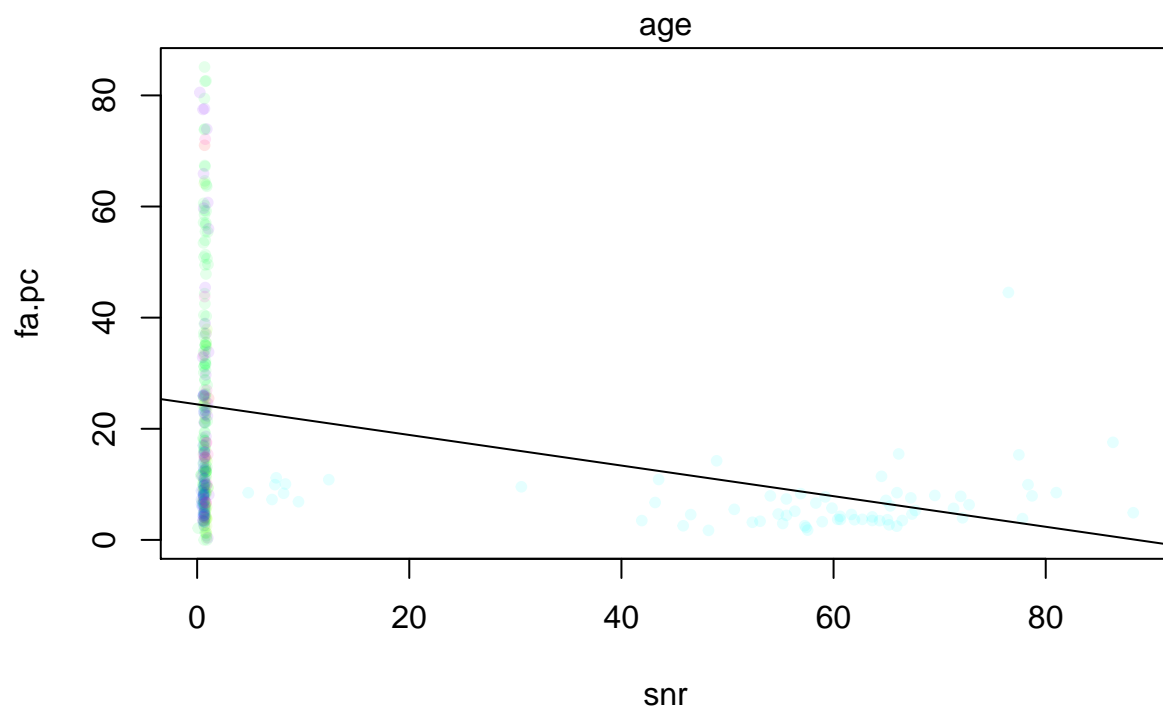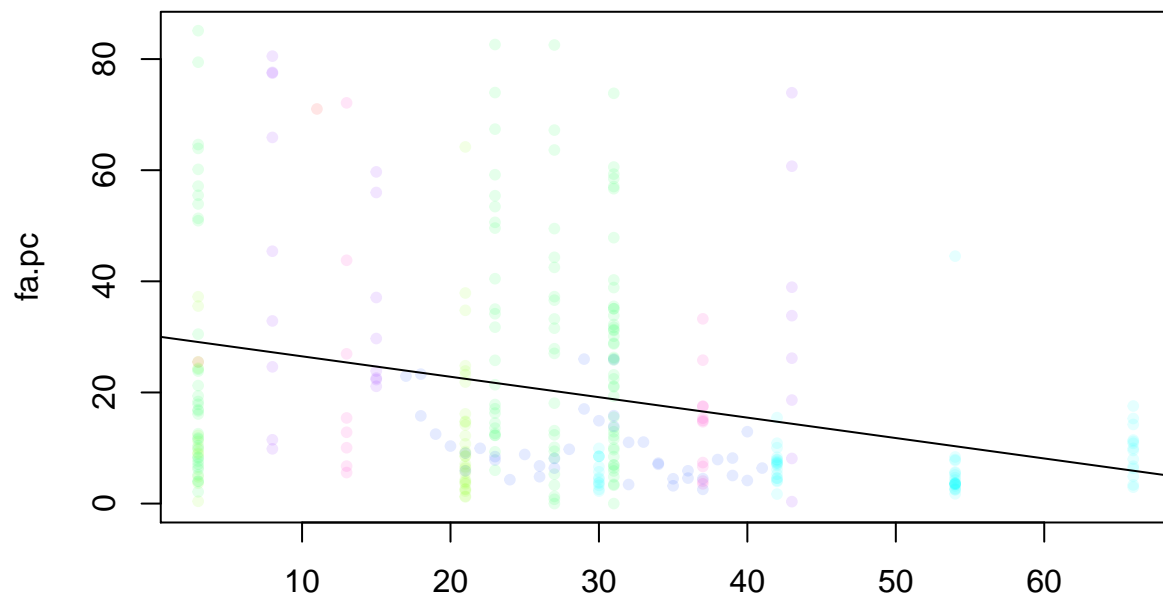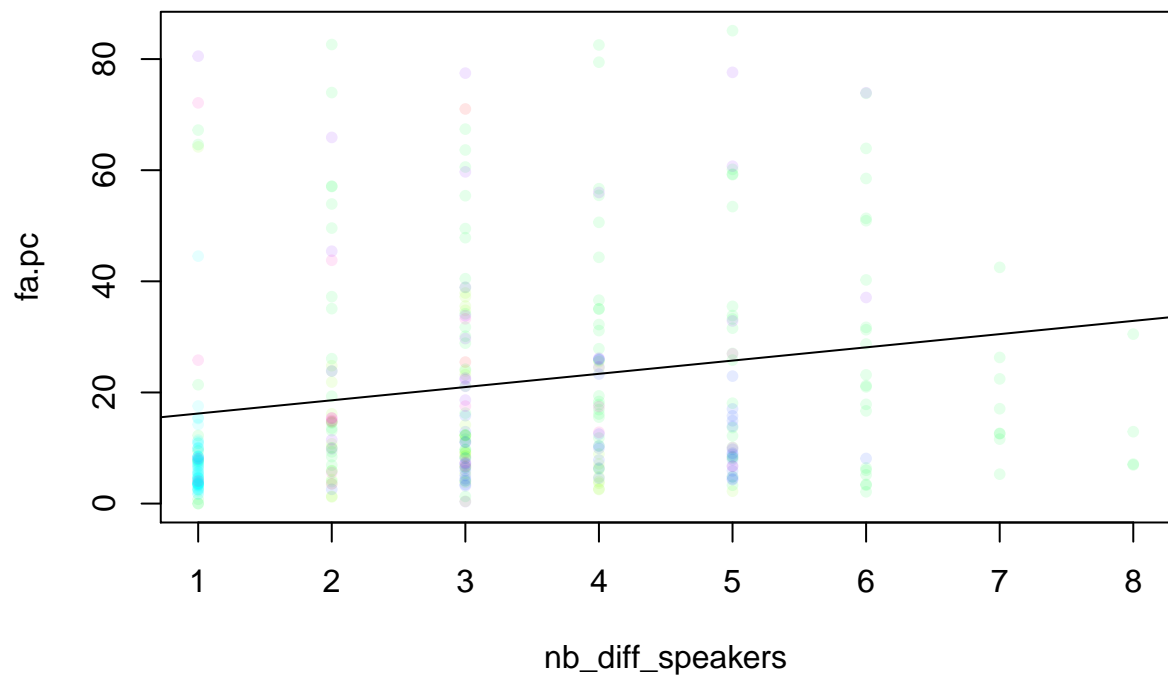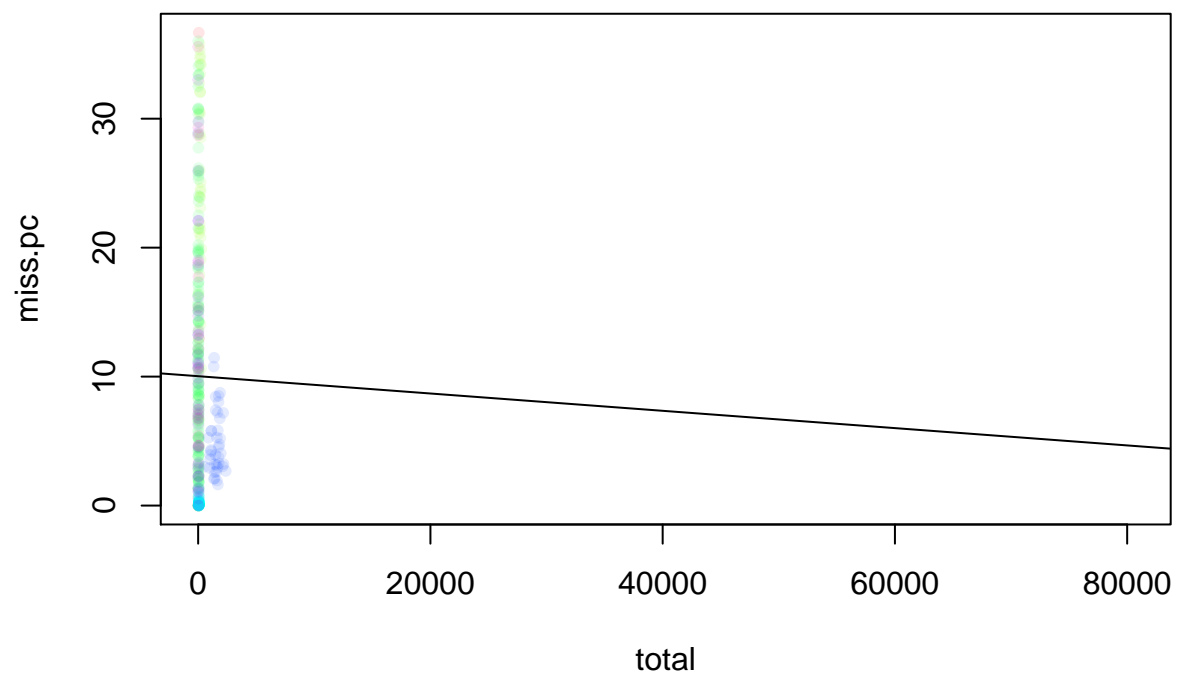
```
## [1] "removing 67 outliers in fa.pc"
```
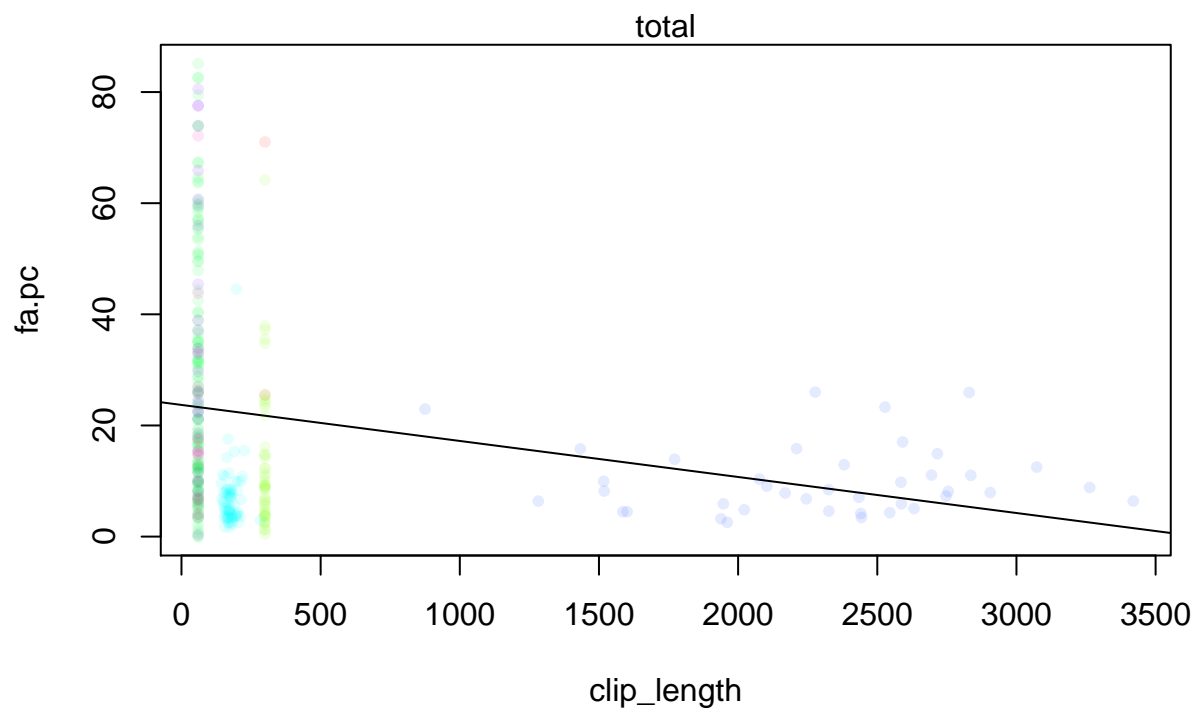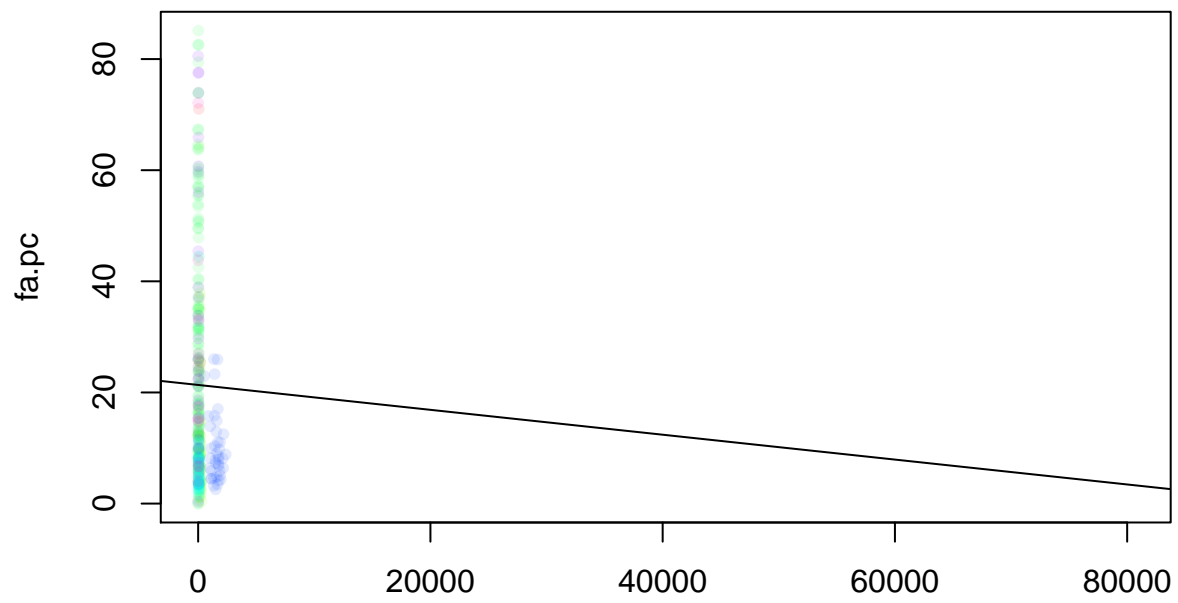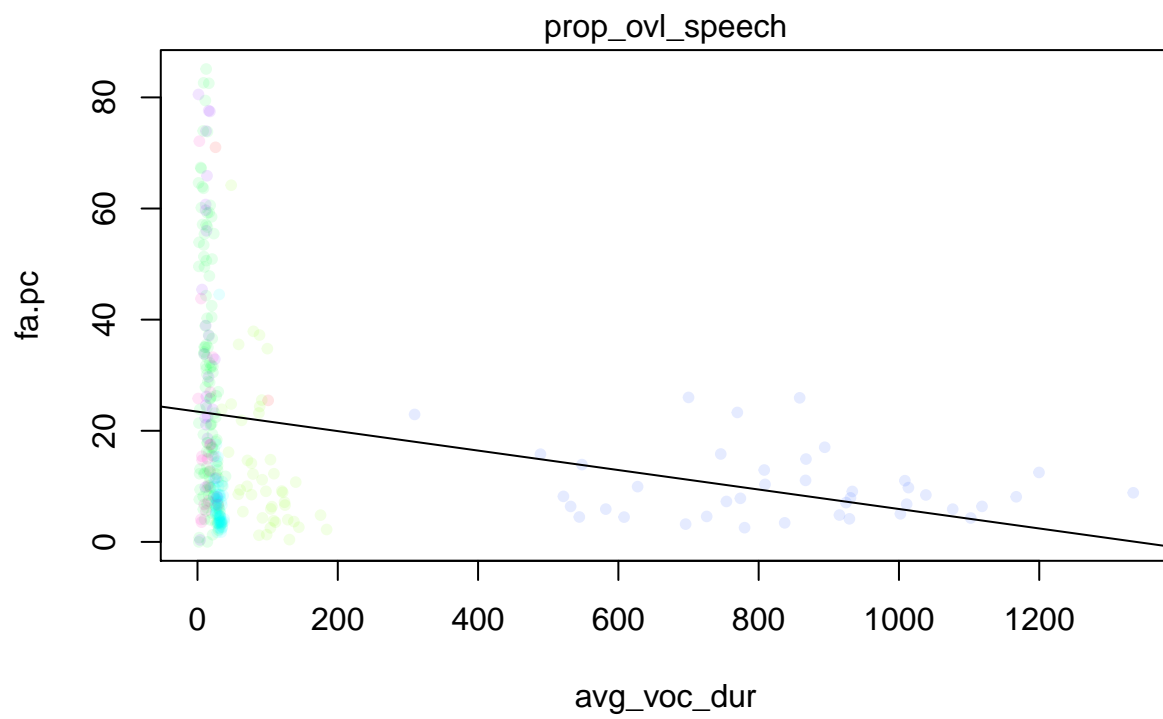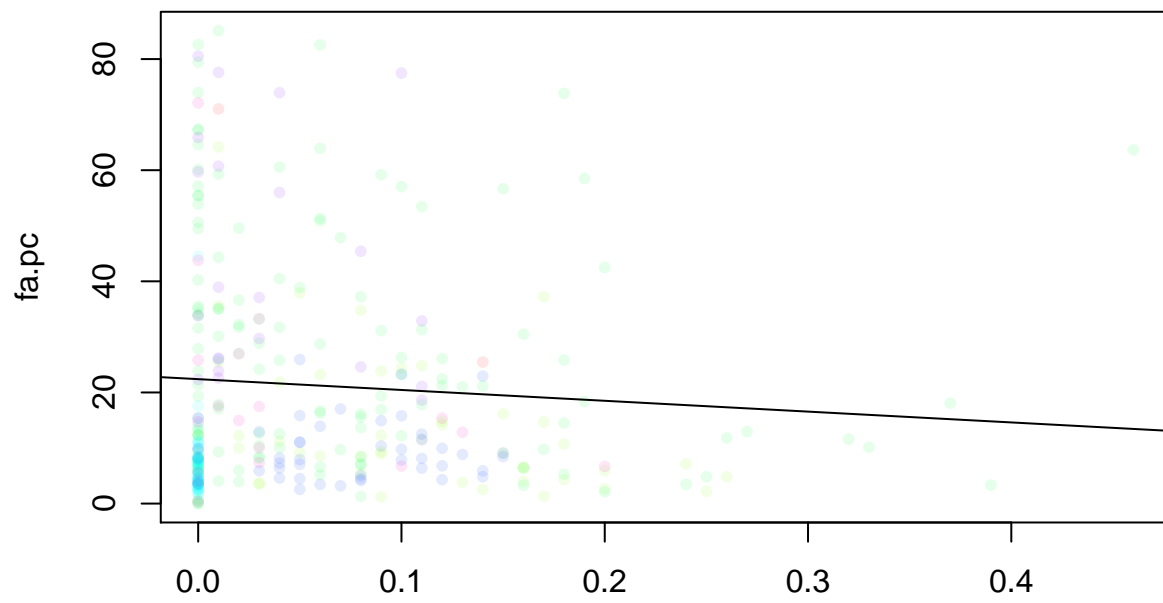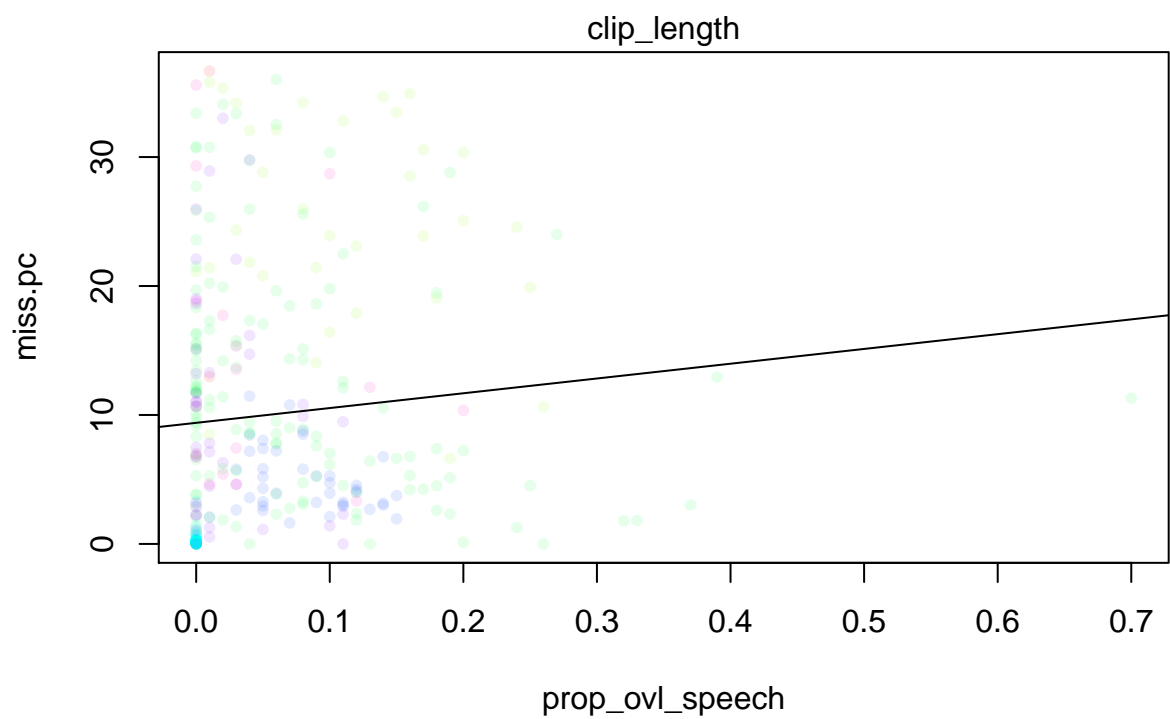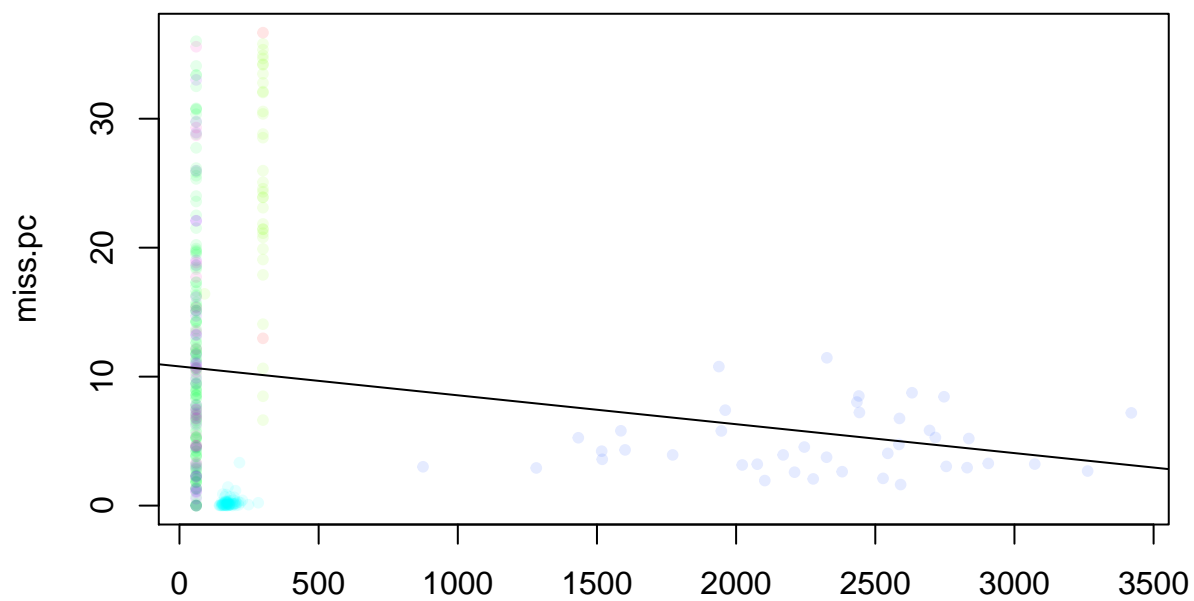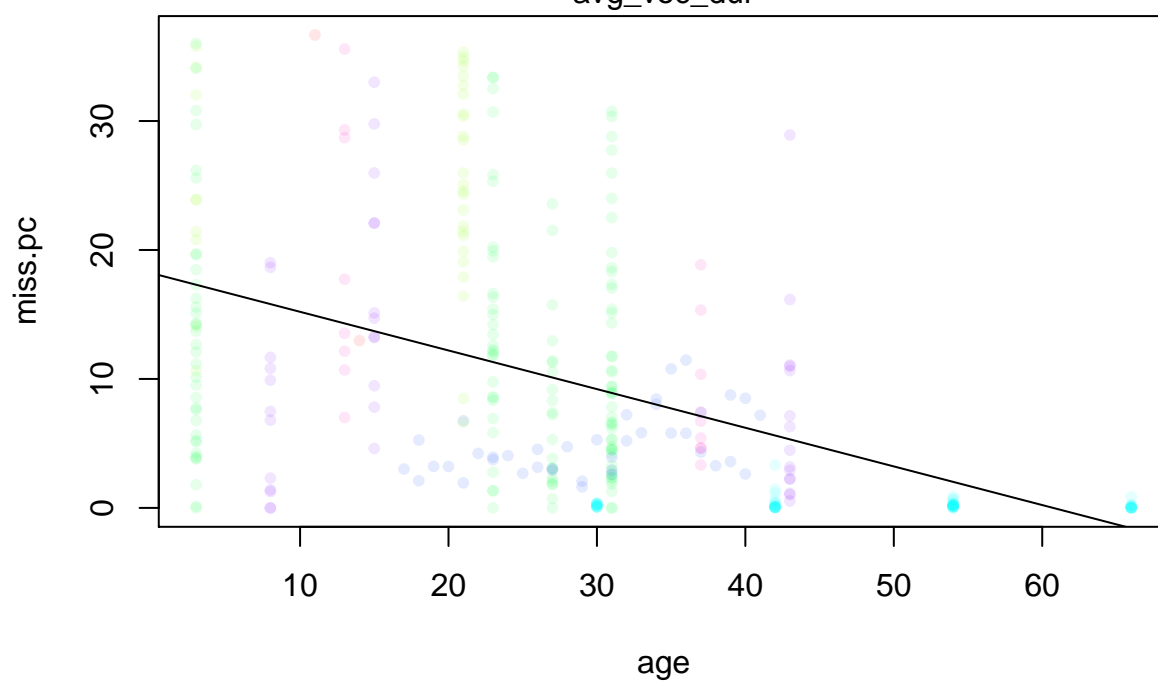
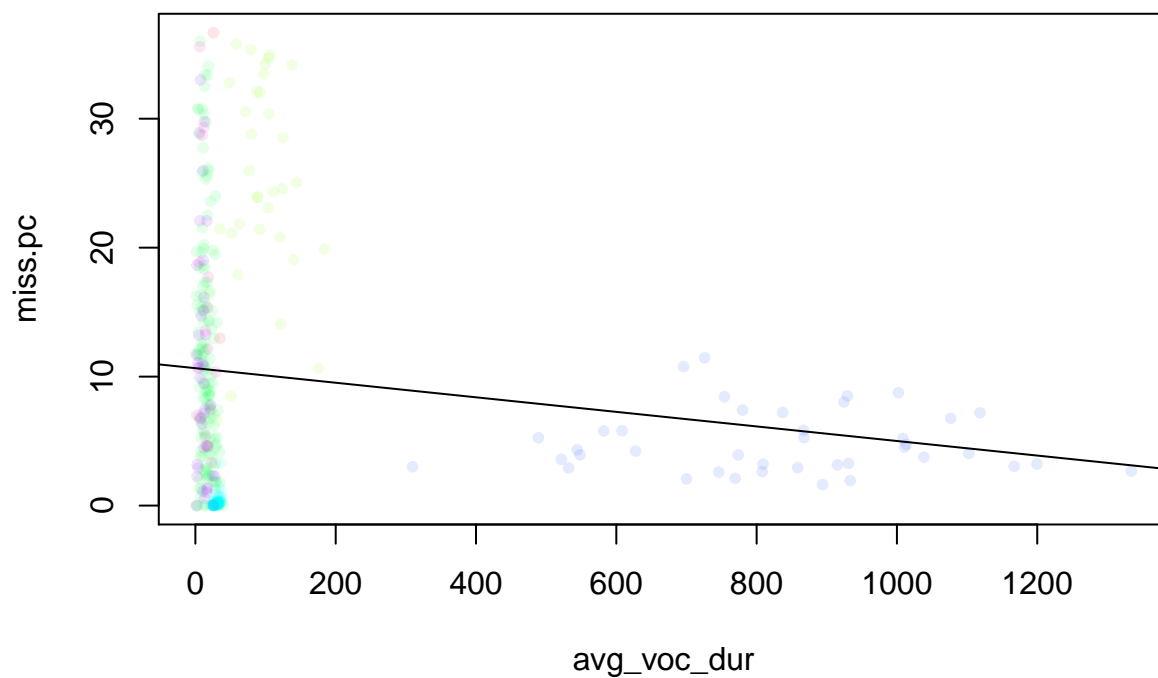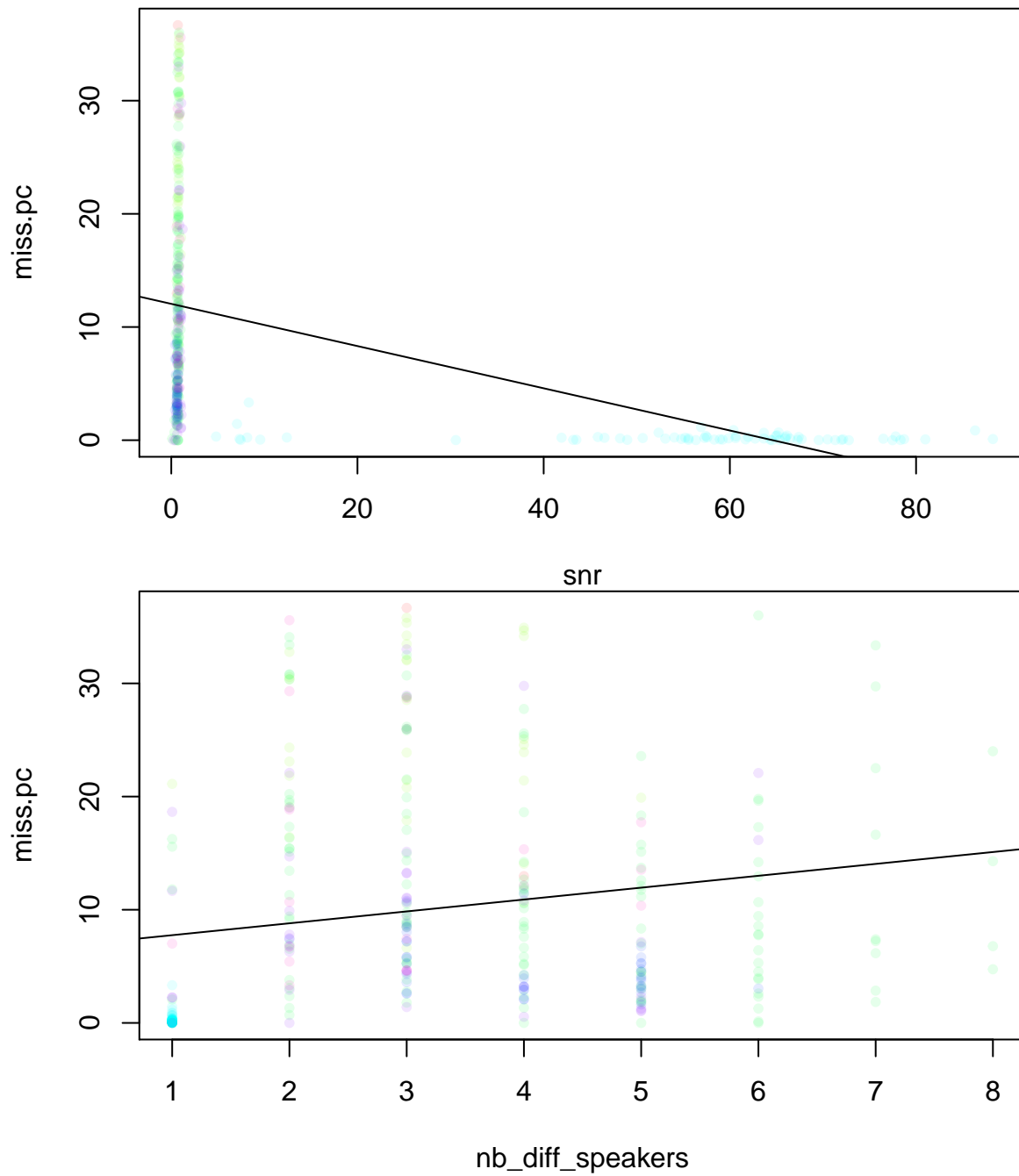## [1] "removing 46 outliers in miss.pc"

Messages I take away from this:

For FA rate

- no strong rel with proportion of overlapping speech
- higher FA for files with shorter voc duration
- lower FA for files from older children
- lower FA for files with higher SNR
- higher FA when higher number of different speakers

For miss rate

- slight higher miss for files with higher prop overlapping speech

- lower miss for files with longer voc dur
- strong trend for fewer misses for files from older children
- strong grend for fewer misses for files with higher SNR
- higher miss when more speakers

ConvRNN version continued, now excluding Tsay and Paido

```
# subanalyses without paido and tsay

print("**removing paido and tsay**")
```

```
## [1] "**removing paido and tsay**"
```

```
npnt=file_eval[!(file_eval$corpus %in% c("tsay","paido")),]

for(thismet in file_eval_metrics){
  iqr=IQR(npnt[,thismet])
  med=median(npnt[,thismet])
  no_outliers=npnt[npnt[,thismet]<med+1.5*iqr,]
  print(paste("removing",dim(npnt)[1]-dim(no_outliers)[1], "outliers in",thismet))
  for(thispred in predictors){
    plot(no_outliers[,thismet]~no_outliers[,thispred], pch=20,col=alpha(cor_color[no_outliers$corpus],.
    abline(lm(no_outliers[,thismet]~no_outliers[,thispred]))
    if(max(no_outliers[,thismet])>300){
        plot(no_outliers[,thismet]~no_outliers[,thispred], pch=20,col=alpha(cor_color[no_outliers$corpu
    abline(lm(no_outliers[,thismet]~no_outliers[,thispred]))
    }
  }
}
```
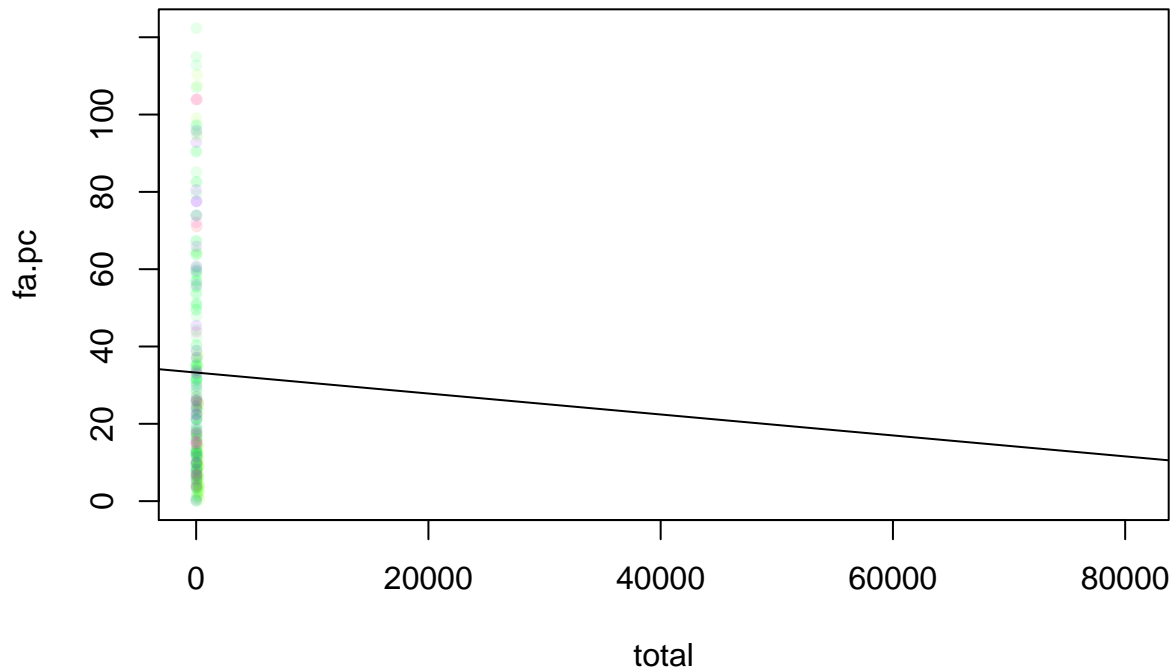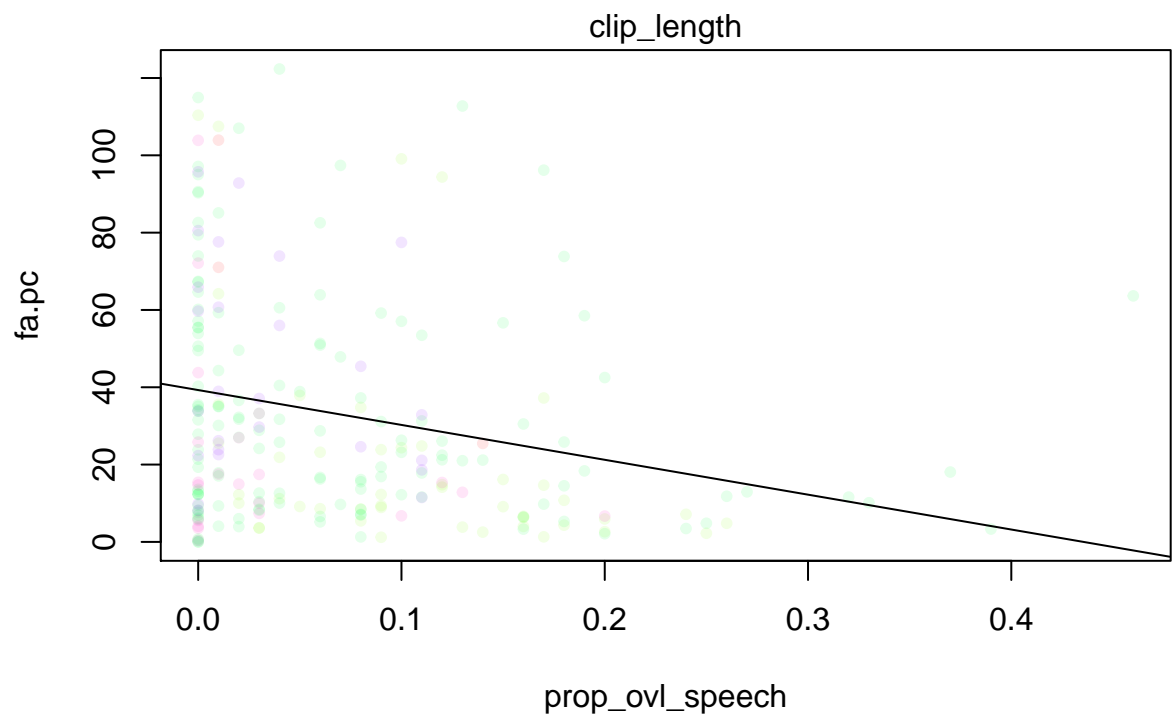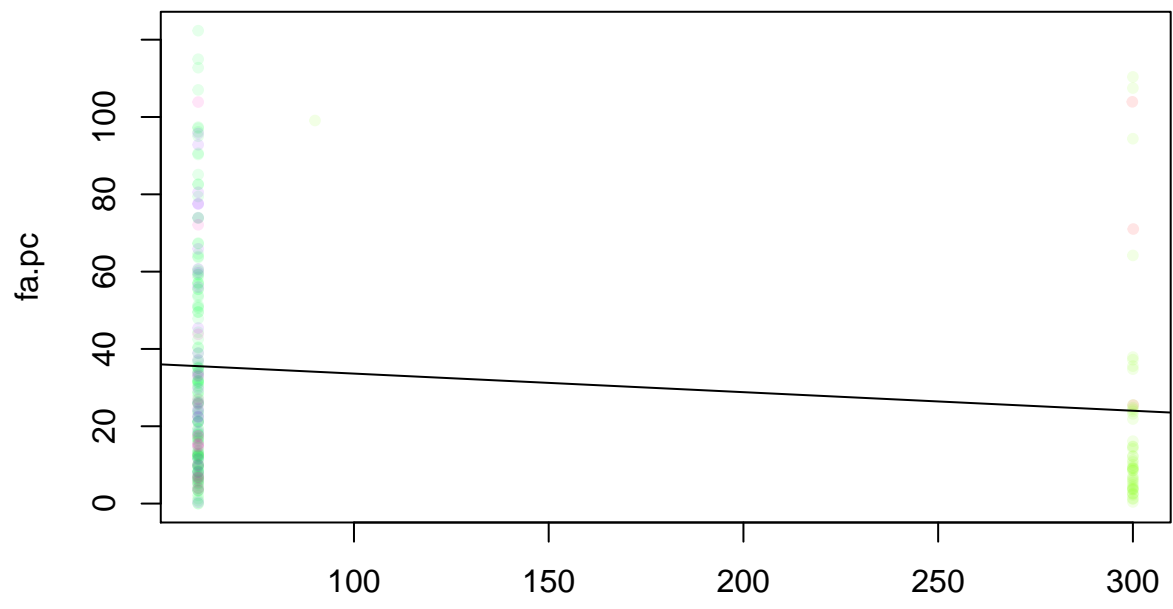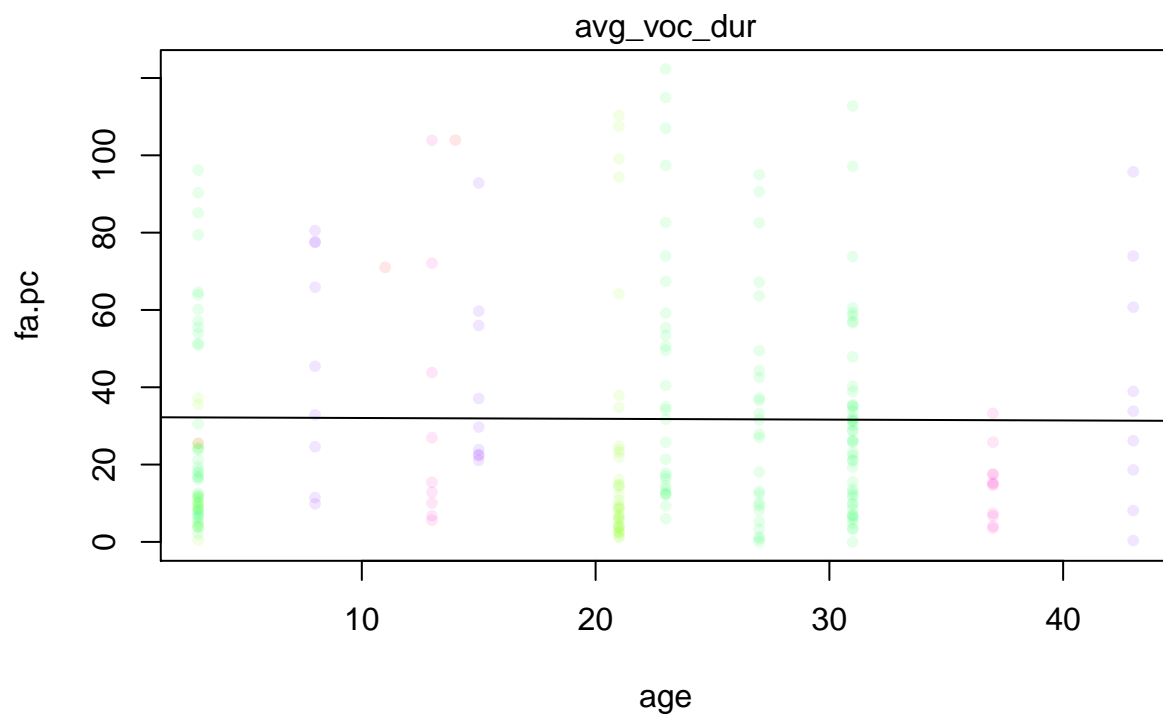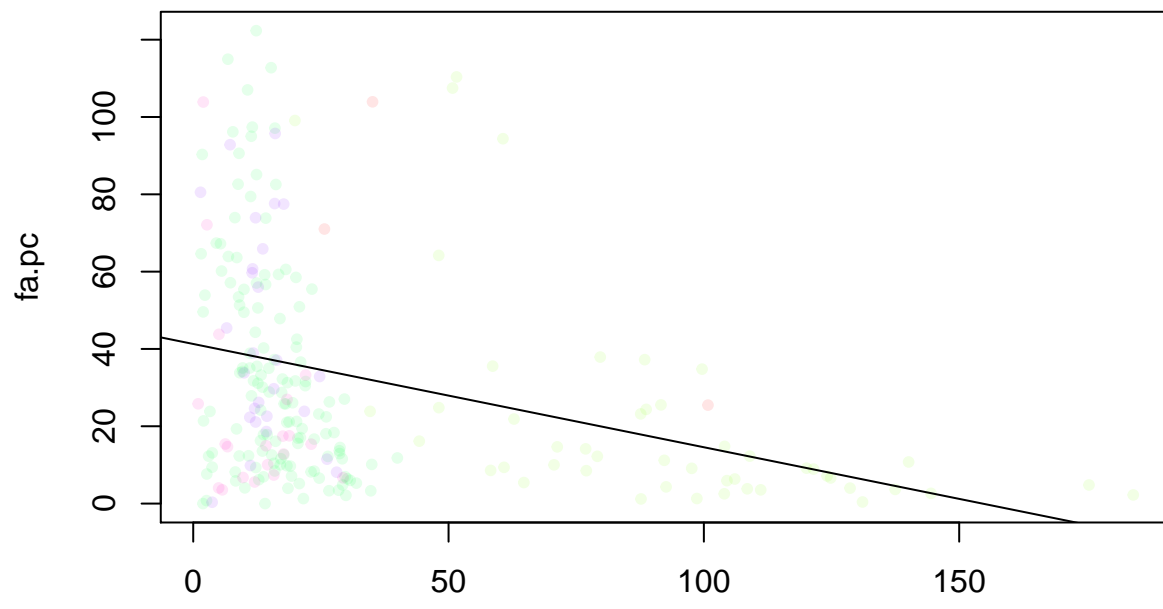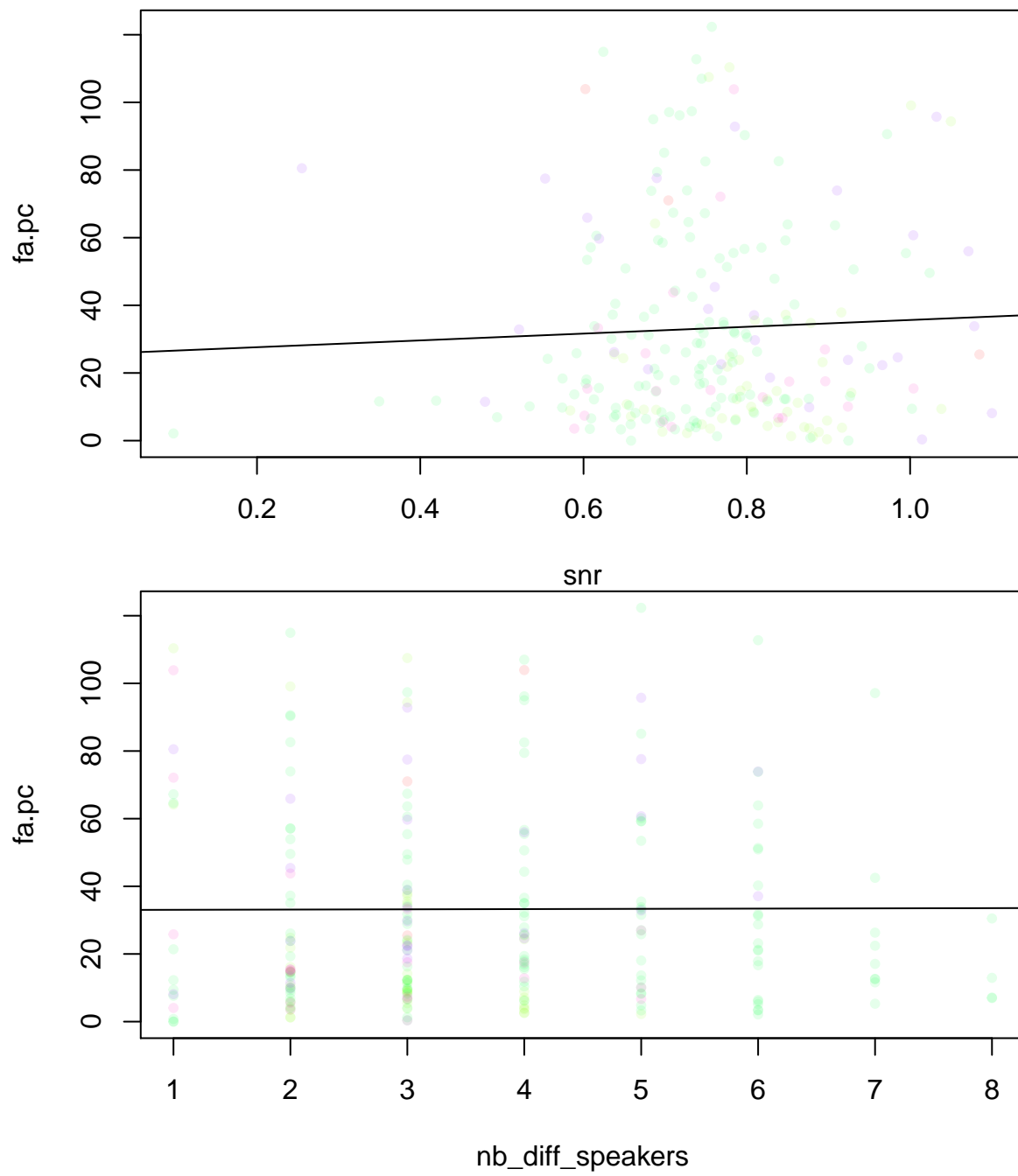
```
## [1] "removing 45 outliers in fa.pc"
```

[1] "removing 26 outliers in miss.pc"
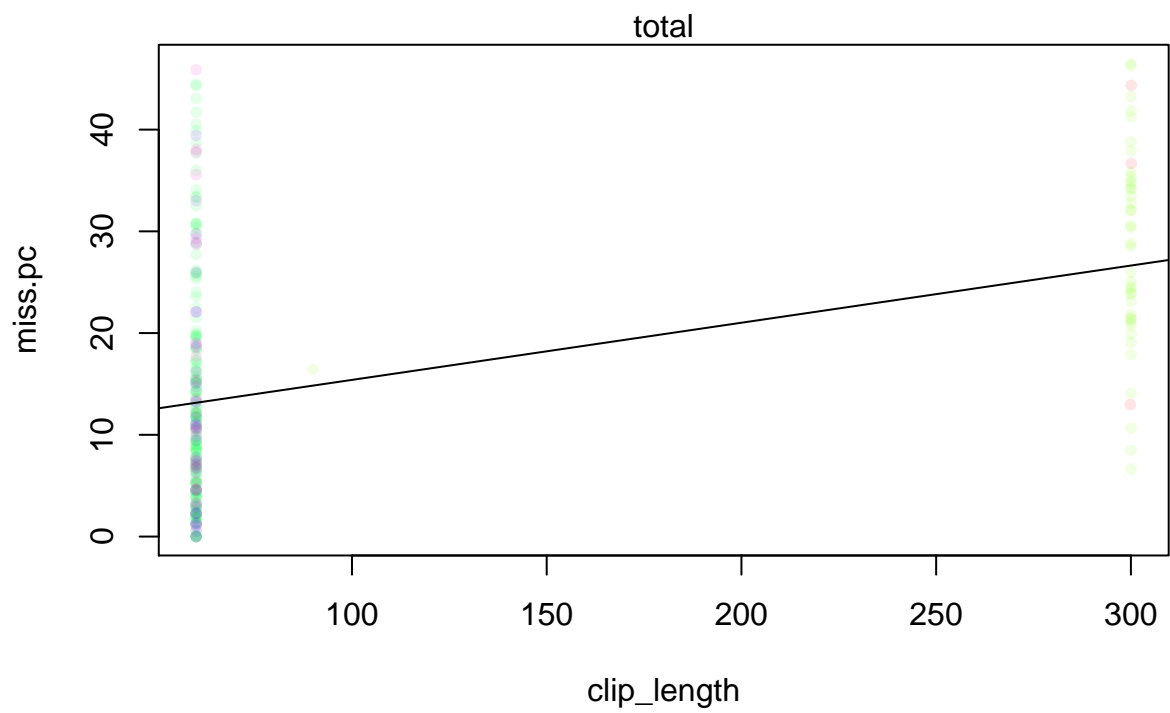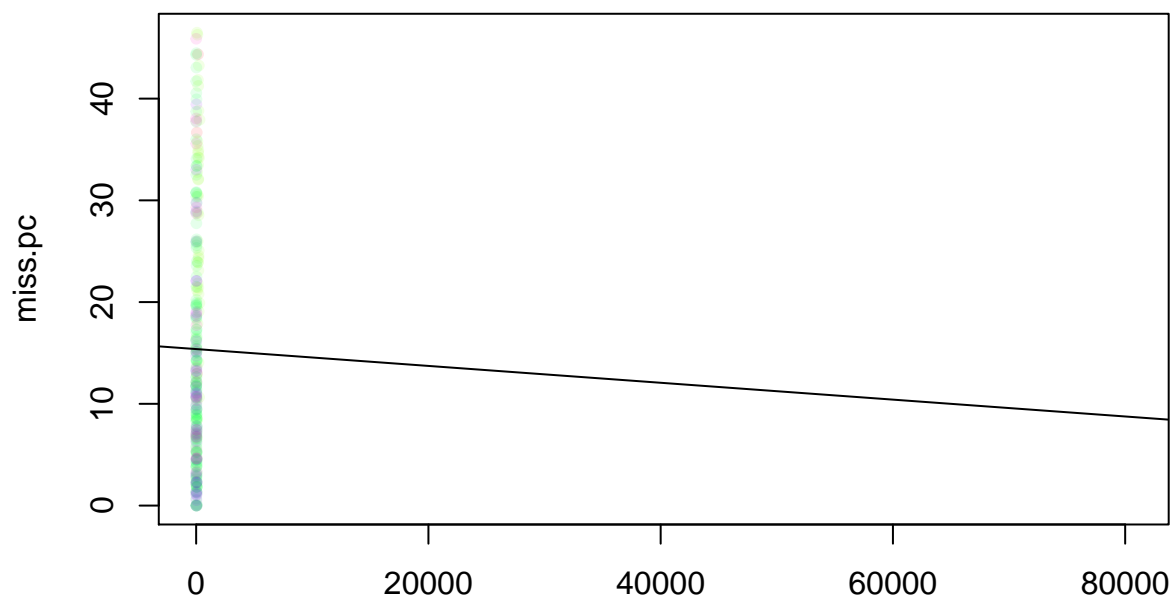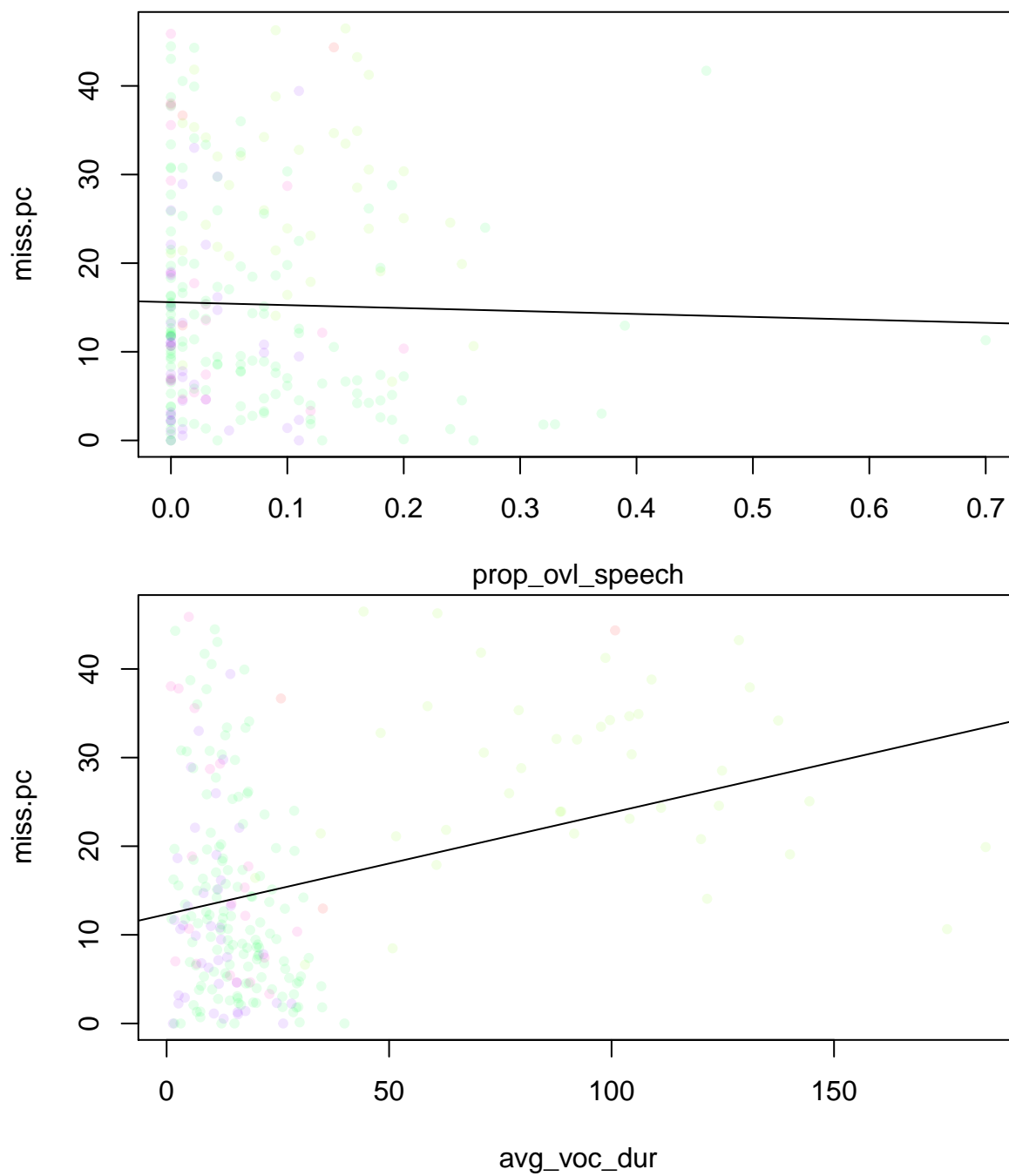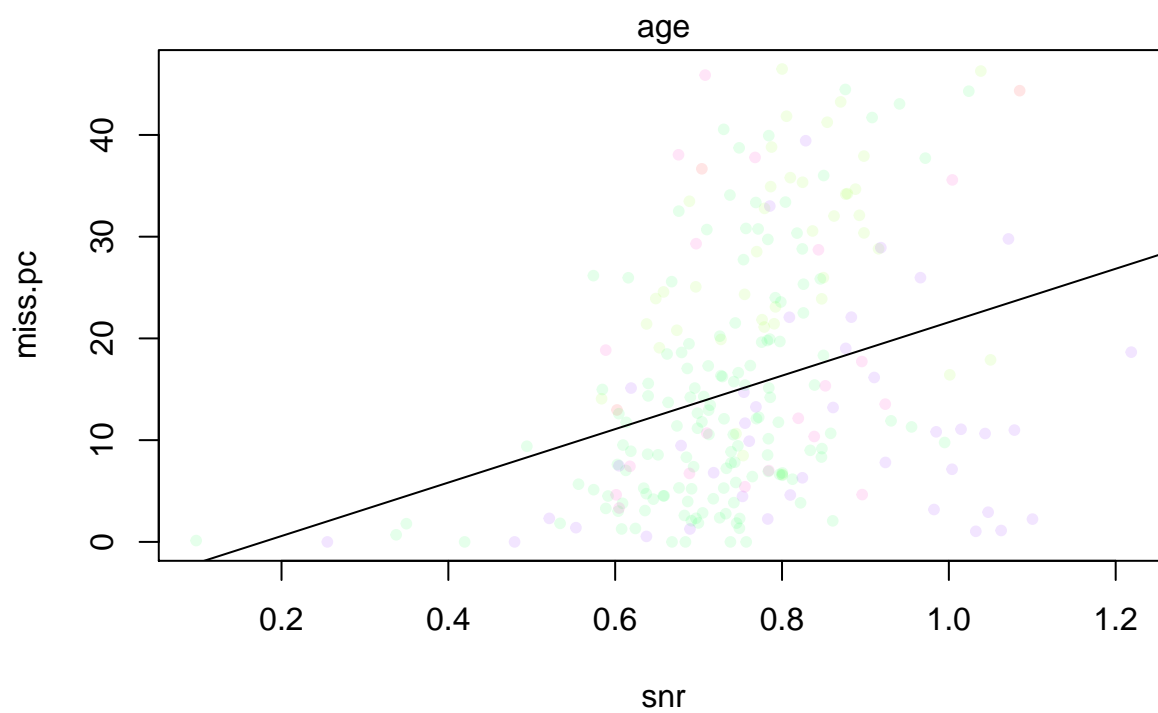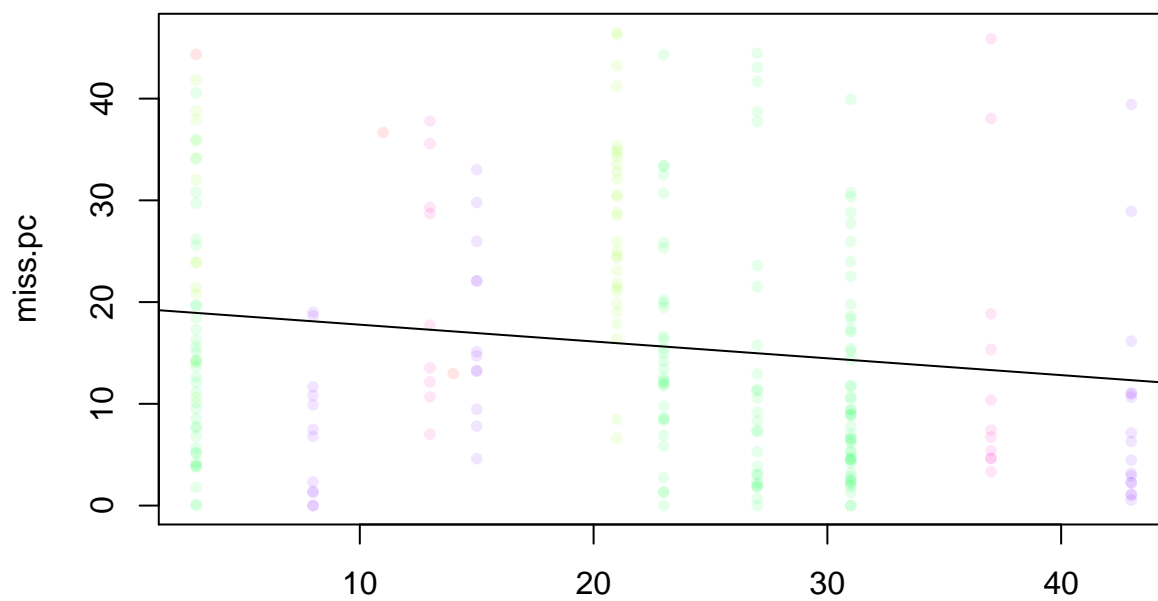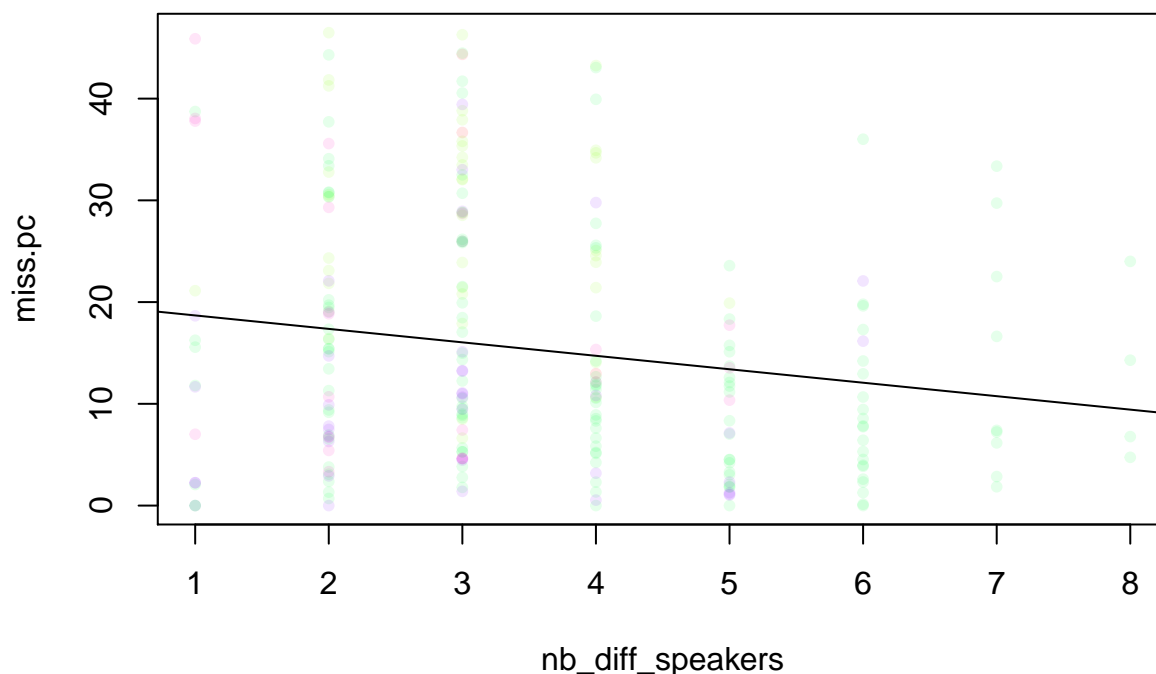
Messages that go away or remain when we excluded paido and tsay:

For FA rate

- no strong rel with proportion of overlapping speech –> NO, the opposite
- higher FA for files with shorter voc duration –> –> NO, the opposite
- lower FA for files from older children –> NO, stable
- lower FA for files with higher SNR –> NO, stable
- higher FA when higher number of different speakers –> NO, stable

For miss rate

- slight higher miss for files with higher prop overlapping speech –> NO, stable M for overlap
- lower miss for files with longer voc dur –> NO, the opposite
- strong trend for fewer misses for files from older children –> YES (but not that strong)
- strong grend for fewer misses for files with higher SNR –> NO, the opposite
- higher miss when more speakers –> NO, the opposite (but weak)

ConvRNN version continued, checking whether subcorpora differences can be explained away via these other variables

We see that many effects are different when paido and tsay are removed. This suggest that some of the apparent correlations are driven by subcorpus differences. So in this section we check whether subcorpus adds any explanatory power once clip diffs are already captured by the other methods

```r
for(thismet in file_eval_metrics){
  iqr=IQR(file_eval[,thismet])
  med=median(file_eval[,thismet])
  no_outliers=file_eval[file_eval[,thismet]<med+1.5*iqr,]
  #print(paste("removing",dim(file_eval)[1]-dim(no_outliers)[1], "outliers in",thismet))
  for(thispred in predictors){
    print(paste("Regressions with and without corpus for",thismet,"and",thispred))
   basemodel=lm(file_eval[,thismet]~file_eval[,thispred],subset=c(!is.na(file_eval[,"corpus"])))
    withcor=lm(file_eval[,thismet]~file_eval[,thispred]+file_eval[,"corpus"])
    print(summary(basemodel))
```

```
    print(summary(withcor))
    print(anova(basemodel,withcor))
  }
}
```

```
## [1] "Regressions with and without corpus for fa.pc and total"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##      "corpus"])))
##
## Residuals:
##     Min     1Q  Median     3Q     Max
##  -62.83  -51.97  -37.06    3.08 2114.58
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             62.90844    8.42474   7.467 5.74e-13 ***
## file_eval[, thispred]  -0.03857    0.01609  -2.397    0.017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 150.1 on 377 degrees of freedom
## Multiple R-squared:  0.01502,    Adjusted R-squared:  0.0124
## F-statistic: 5.747 on 1 and 377 DF,  p-value: 0.017
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##      "corpus"])
##
## Residuals:
##     Min     1Q  Median     3Q     Max
## -140.44  -50.08  -13.09    2.21 2099.28
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  71.63157   84.76934   0.845    0.399
## file_eval[, thispred]        -0.05026    0.05948  -0.845    0.399
## file_eval[, "corpus"]lena_lyon -38.70863   87.26576  -0.444    0.658
## file_eval[, "corpus"]namibia   6.59656   85.49251   0.077    0.939
## file_eval[, "corpus"]paido   -60.32466   86.44943  -0.698    0.486
## file_eval[, "corpus"]tsay     16.24137  123.16776   0.132    0.895
## file_eval[, "corpus"]tsimane  69.53053   87.66359   0.793    0.428
## file_eval[, "corpus"]vanuatu -47.64290   90.80654  -0.525    0.600
## file_eval[, "corpus"]war2    -17.41268  119.66465  -0.146    0.884
##
## Residual standard error: 146.5 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.07983,    Adjusted R-squared:  0.05993
## F-statistic: 4.012 on 8 and 370 DF,  p-value: 0.0001388
##
## Analysis of Variance Table
##
```

```
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    377 8499375
## 2    370 7940098  7    559277 3.7231 0.0006567 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for fa.pc and clip_length"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##    Min     1Q  Median     3Q     Max
##  -62.28  -51.96  -36.50    1.50 2115.14
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           63.82383    8.64195   7.385 9.85e-13 ***
## file_eval[, thispred] -0.02581    0.01111  -2.324   0.0207 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 150.2 on 377 degrees of freedom
## Multiple R-squared:  0.01412,    Adjusted R-squared:  0.0115
## F-statistic: 5.399 on 1 and 377 DF,  p-value: 0.02068
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min      1Q  Median     3Q     Max
## -139.83  -50.73   -8.67    0.18 2100.50
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   67.057771  85.609521   0.783    0.434
## file_eval[, thispred]         -0.000837   0.042417  -0.020    0.984
## file_eval[, "corpus"]lena_lyon -41.984512  87.264048  -0.481    0.631
## file_eval[, "corpus"]namibia    9.908253  86.077919   0.115    0.908
## file_eval[, "corpus"]paido    -58.598296  86.657993  -0.676    0.499
## file_eval[, "corpus"]tsay     -55.181046 121.770796  -0.453    0.651
## file_eval[, "corpus"]tsimane   73.176018  88.219174   0.829    0.407
## file_eval[, "corpus"]vanuatu  -44.102553  91.355061  -0.483    0.630
## file_eval[, "corpus"]war2     -20.476683 119.725013  -0.171    0.864
##
## Residual standard error: 146.6 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.07805,    Adjusted R-squared:  0.05812
## F-statistic: 3.916 on 8 and 370 DF,  p-value: 0.0001862
##
```

```
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    377 8507119
## 2    370 7955414  7    551706 3.6656 0.0007665 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for fa.pc and prop_ovl_speech"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -61.34  -49.44  -34.63   -5.89 2116.08
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               61.339      9.438   6.499 2.56e-10 ***
## file_eval[, thispred]   -118.010     96.775  -1.219    0.223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 151 on 377 degrees of freedom
## Multiple R-squared:  0.003929,   Adjusted R-squared:  0.001287
## F-statistic: 1.487 on 1 and 377 DF,  p-value: 0.2234
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -143.39  -48.90   -8.42    1.63 2090.77
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     74.47      84.62   0.880    0.379
## file_eval[, thispred]         -143.77     103.43  -1.390    0.165
## file_eval[, "corpus"]lena_lyon  -34.49      87.20  -0.396    0.693
## file_eval[, "corpus"]namibia     12.18      85.26   0.143    0.886
## file_eval[, "corpus"]paido      -66.16      86.46  -0.765    0.445
## file_eval[, "corpus"]tsay       -52.83      87.60  -0.603    0.547
## file_eval[, "corpus"]tsimane     69.27      87.45   0.792    0.429
## file_eval[, "corpus"]vanuatu    -46.39      90.57  -0.512    0.609
## file_eval[, "corpus"]war2       -12.33     119.56  -0.103    0.918
##
## Residual standard error: 146.3 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.08284,    Adjusted R-squared:  0.06301
```

```
## F-statistic: 4.178 on 8 and 370 DF,  p-value: 8.39e-05
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1    377 8595042
## 2    370 7914097  7    680945 4.5479 6.93e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for fa.pc and avg_voc_dur"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -62.75  -52.04  -37.01    2.91 2114.66
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             62.88887    8.42568   7.464 5.87e-13 ***
## file_eval[, thispred] -0.07121    0.02979  -2.390   0.0173 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 150.2 on 377 degrees of freedom
## Multiple R-squared:  0.01493,    Adjusted R-squared:  0.01232
## F-statistic: 5.714 on 1 and 377 DF,  p-value: 0.01732
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -140.42  -50.02  -12.67    1.88 2099.40
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   71.39035   84.76794   0.842    0.400
## file_eval[, thispred]         -0.08508    0.10412  -0.817    0.414
## file_eval[, "corpus"]lena_lyon -38.84871   87.26948  -0.445    0.656
## file_eval[, "corpus"]namibia    6.77875   85.49395   0.079    0.937
## file_eval[, "corpus"]paido    -60.49536   86.46237  -0.700    0.485
## file_eval[, "corpus"]tsay       9.51574  119.52569   0.080    0.937
## file_eval[, "corpus"]tsimane   69.69540   87.66667   0.795    0.427
## file_eval[, "corpus"]vanuatu  -47.43199   90.80706  -0.522    0.602
## file_eval[, "corpus"]war2     -18.35329  119.64542  -0.153    0.878
##
## Residual standard error: 146.5 on 370 degrees of freedom
```

```
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.07971,    Adjusted R-squared:  0.05982
## F-statistic: 4.006 on 8 and 370 DF,  p-value: 0.0001415
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    377 8500107
## 2    370 7941092  7    559016 3.7209 0.0006606 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for fa.pc and age"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -64.22  -46.47  -37.71  -10.57 2122.67
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              65.8663    15.6037   4.221 3.05e-05 ***
## file_eval[, thispred]    -0.4115     0.5025  -0.819    0.413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 151.2 on 377 degrees of freedom
## Multiple R-squared:  0.001776,    Adjusted R-squared:  -0.0008719
## F-statistic: 0.6707 on 1 and 377 DF,  p-value: 0.4133
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -159.08  -48.72  -13.27    5.50 2094.94
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     57.2680    84.6274   0.677    0.499
## file_eval[, thispred]            1.0220     0.6717   1.521    0.129
## file_eval[, "corpus"]lena_lyon -49.3051    87.1253  -0.566    0.572
## file_eval[, "corpus"]namibia    -2.3812    85.6022  -0.028    0.978
## file_eval[, "corpus"]paido     -98.3748    90.1316  -1.091    0.276
## file_eval[, "corpus"]tsay      -77.5588    88.5551  -0.876    0.382
## file_eval[, "corpus"]tsimane    58.2173    87.9234   0.662    0.508
## file_eval[, "corpus"]vanuatu   -61.1393    91.2098  -0.670    0.503
## file_eval[, "corpus"]war2      -23.2020   119.3658  -0.194    0.846
```

```
##
## Residual standard error: 146.2 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.08379,    Adjusted R-squared:  0.06398
## F-statistic: 4.229 on 8 and 370 DF,  p-value: 7.159e-05
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    377 8613620
## 2    370 7905961  7    707659 4.7312 4.182e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for fa.pc and snr"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -63.45  -52.57  -33.62   -2.54 2113.98
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             64.0248     8.4986   7.534  3.7e-13 ***
## file_eval[, thispred]   -0.8708     0.3381  -2.576   0.0104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 150 on 377 degrees of freedom
## Multiple R-squared:  0.01729,    Adjusted R-squared:  0.01468
## F-statistic: 6.633 on 1 and 377 DF,  p-value: 0.01039
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -139.85  -50.73   -8.68    0.55 2100.51
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     66.74786   84.66050   0.788    0.431
## file_eval[, thispred]            0.07378    0.87804   0.084    0.933
## file_eval[, "corpus"]lena_lyon -41.98177   87.26306  -0.481    0.631
## file_eval[, "corpus"]namibia    10.11408   85.47297   0.118    0.906
## file_eval[, "corpus"]paido     -62.55180   99.05313  -0.631    0.528
## file_eval[, "corpus"]tsay      -56.83324   87.77501  -0.647    0.518
## file_eval[, "corpus"]tsimane    73.37324   87.62896   0.837    0.403
```

```
## file_eval[, "corpus"]vanuatu    -43.89921    90.78517  -0.484     0.629
## file_eval[, "corpus"]war2        -20.45336  119.72425  -0.171     0.864
##
## Residual standard error: 146.6 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.07807,    Adjusted R-squared:  0.05814
## F-statistic: 3.917 on 8 and 370 DF,  p-value: 0.0001857
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df      RSS Df Sum of Sq      F   Pr(>F)
## 1    377 8479741
## 2    370 7955270  7    524471 3.4847 0.001245 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for fa.pc and nb_diff_speakers"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min     1Q Median     3Q    Max
##  -73.06 -52.28 -29.30  -1.33 2104.36
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             82.016     15.637   5.245 2.61e-07 ***
## file_eval[, thispred]   -8.960      4.472  -2.004   0.0458 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 150.5 on 377 degrees of freedom
## Multiple R-squared:  0.01054,    Adjusted R-squared:  0.007911
## F-statistic: 4.014 on 1 and 377 DF,  p-value: 0.04584
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -179.32  -39.59   -4.95   13.01 2030.40
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   149.508     84.231   1.775   0.0767 .
## file_eval[, thispred]         -24.810      5.363  -4.626 5.17e-06 ***
## file_eval[, "corpus"]lena_lyon -52.835     84.877  -0.622   0.5340
## file_eval[, "corpus"]namibia    22.328     83.146   0.269   0.7884
## file_eval[, "corpus"]paido    -116.388     85.034  -1.369   0.1719
```

```
## file_eval[, "corpus"]tsay        -39.686      85.423  -0.465   0.6425
## file_eval[, "corpus"]tsimane      62.744      85.232   0.736   0.4621
## file_eval[, "corpus"]vanuatu     -60.855      88.345  -0.689   0.4914
## file_eval[, "corpus"]war2        -53.557     116.626  -0.459   0.6463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 142.6 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.1285, Adjusted R-squared:  0.1096
## F-statistic: 6.817 on 8 and 370 DF,  p-value: 2.319e-08
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    377 8538034
## 2    370 7520483  7   1017551 7.1518 4.952e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for miss.pc and total"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min     1Q Median     3Q     Max
## -16.435 -13.921  -4.795   5.785  83.580
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           16.444215   1.011360   16.26  < 2e-16 ***
## file_eval[, thispred] -0.006567   0.001931   -3.40 0.000746 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.02 on 377 degrees of freedom
## Multiple R-squared:  0.02975,    Adjusted R-squared:  0.02717
## F-statistic: 11.56 on 1 and 377 DF,  p-value: 0.0007463
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min     1Q Median     3Q     Max
## -25.704  -9.757  -0.403   1.672  85.712
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             31.719410   9.013120   3.519 0.000487 ***
```

```
## file_eval[, thispred]                -0.004056   0.006324  -0.641 0.521630
## file_eval[, "corpus"]lena_lyon    0.847428   9.278552   0.091 0.927278
## file_eval[, "corpus"]namibia    -12.805623   9.090011  -1.409 0.159747
## file_eval[, "corpus"]paido      -31.209422   9.191755  -3.395 0.000760 ***
## file_eval[, "corpus"]tsay       -20.551954  13.095841  -1.569 0.117421
## file_eval[, "corpus"]tsimane    -17.425673   9.320852  -1.870 0.062338 .
## file_eval[, "corpus"]vanuatu    -13.972949   9.655027  -1.447 0.148682
## file_eval[, "corpus"]war2       -29.969378  12.723372  -2.355 0.019021 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.58 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.289,  Adjusted R-squared:  0.2736
## F-statistic: 18.79 on 8 and 370 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    377 122486
## 2    370  89763  7     32722 19.269 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for miss.pc and clip_length"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min     1Q Median     3Q    Max
## -16.375 -14.070  -4.697   6.076 83.625
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           16.646045   1.037006  16.052  < 2e-16 ***
## file_eval[, thispred] -0.004526   0.001333  -3.395 0.000759 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.03 on 377 degrees of freedom
## Multiple R-squared:  0.02967,    Adjusted R-squared:  0.02709
## F-statistic: 11.53 on 1 and 377 DF,  p-value: 0.0007592
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min     1Q Median     3Q    Max
## -25.286  -9.776  -0.338   1.487 85.785
```

```
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.113e+01  9.098e+00   3.422 0.000692 ***
## file_eval[, thispred]          6.631e-04  4.508e-03   0.147 0.883135
## file_eval[, "corpus"]lena_lyon 5.862e-01  9.274e+00   0.063 0.949632
## file_eval[, "corpus"]namibia  -1.236e+01  9.148e+00  -1.351 0.177393
## file_eval[, "corpus"]paido    -3.098e+01  9.210e+00  -3.364 0.000849 ***
## file_eval[, "corpus"]tsay     -2.777e+01  1.294e+01  -2.146 0.032539 *
## file_eval[, "corpus"]tsimane  -1.696e+01  9.376e+00  -1.808 0.071342 .
## file_eval[, "corpus"]vanuatu  -1.351e+01  9.709e+00  -1.392 0.164859
## file_eval[, "corpus"]war2     -3.022e+01  1.272e+01  -2.375 0.018071 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 15.58 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.2882, Adjusted R-squared:  0.2728
## F-statistic: 18.73 on 8 and 370 DF,  p-value: < 2.2e-16
## 
## Analysis of Variance Table
## 
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    377 122496
## 2    370  89858  7     32638 19.199 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for miss.pc and prop_ovl_speech"
## 
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.124 -12.970  -6.616   6.630  84.957
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            15.0435     1.1438  13.152   <2e-16 ***
## file_eval[, thispred]   0.3097    11.7284   0.026    0.979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 18.3 on 377 degrees of freedom
## Multiple R-squared:  1.849e-06, Adjusted R-squared:  -0.002651
## F-statistic: 0.0006973 on 1 and 377 DF,  p-value: 0.9789
## 
## 
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
```

```
##
## Residuals:
##     Min     1Q Median     3Q    Max
## -27.717  -9.572  -0.398   1.462  84.673
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      33.725      8.812   3.827 0.000152 ***
## file_eval[, thispred]           -44.899     10.772  -4.168 3.83e-05 ***
## file_eval[, "corpus"]lena_lyon    2.922      9.082   0.322 0.747837
## file_eval[, "corpus"]namibia    -11.875      8.880  -1.337 0.181928
## file_eval[, "corpus"]paido      -33.457      9.004  -3.716 0.000234 ***
## file_eval[, "corpus"]tsay       -25.197      9.122  -2.762 0.006028 **
## file_eval[, "corpus"]tsimane    -18.398      9.107  -2.020 0.044091 *
## file_eval[, "corpus"]vanuatu    -14.449      9.432  -1.532 0.126387
## file_eval[, "corpus"]war2       -27.672     12.451  -2.223 0.026853 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.23 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.3201, Adjusted R-squared:  0.3054
## F-statistic: 21.77 on 8 and 370 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    377 126241
## 2    370  85833  7     40408 24.884 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for miss.pc and avg_voc_dur"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min     1Q Median     3Q    Max
## -16.410 -13.899  -4.775   5.832  83.617
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           16.427199   1.011821  16.235  < 2e-16 ***
## file_eval[, thispred] -0.012003   0.003577  -3.355 0.000873 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.03 on 377 degrees of freedom
## Multiple R-squared:  0.029,  Adjusted R-squared:  0.02642
## F-statistic: 11.26 on 1 and 377 DF,  p-value: 0.0008733
##
```

```
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min     1Q  Median     3Q     Max
## -25.774  -9.762  -0.409   1.842  85.710
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     31.774033   9.010666   3.526 0.000474 ***
## file_eval[, thispred]           -0.008242   0.011068  -0.745 0.456943
## file_eval[, "corpus"]lena_lyon   0.886750   9.276576   0.096 0.923898
## file_eval[, "corpus"]namibia   -12.844752   9.087840  -1.413 0.158379
## file_eval[, "corpus"]paido     -31.255502   9.190781  -3.401 0.000745 ***
## file_eval[, "corpus"]tsay      -20.022087  12.705348  -1.576 0.115908
## file_eval[, "corpus"]tsimane   -17.471874   9.318796  -1.875 0.061593 .
## file_eval[, "corpus"]vanuatu   -14.012991   9.652614  -1.452 0.147424
## file_eval[, "corpus"]war2      -30.010970  12.718074  -2.360 0.018808 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.57 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.2892, Adjusted R-squared:  0.2739
## F-statistic: 18.82 on 8 and 370 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    377 122581
## 2    370  89729  7     32852 19.352 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for miss.pc and age"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min     1Q  Median     3Q     Max
## -24.231 -10.367  -4.550   4.321  86.497
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           25.38061    1.78688  14.204  < 2e-16 ***
## file_eval[, thispred] -0.38316    0.05755  -6.658 9.81e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 17.31 on 377 degrees of freedom
## Multiple R-squared:  0.1052, Adjusted R-squared:  0.1028
## F-statistic: 44.33 on 1 and 377 DF,  p-value: 9.814e-11
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.255  -9.705  -0.513   1.508  85.684
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    31.388606   9.022350   3.479 0.000563 ***
## file_eval[, thispred]          -0.006279   0.071615  -0.088 0.930178
## file_eval[, "corpus"]lena_lyon  0.628334   9.288660   0.068 0.946105
## file_eval[, "corpus"]namibia  -12.445388   9.126279  -1.364 0.173495
## file_eval[, "corpus"]paido    -30.816900   9.609170  -3.207 0.001458 **
## file_eval[, "corpus"]tsay      -26.323491   9.441094  -2.788 0.005574 **
## file_eval[, "corpus"]tsimane  -17.022097   9.373743  -1.816 0.070190 .
## file_eval[, "corpus"]vanuatu  -13.565091   9.724117  -1.395 0.163854
## file_eval[, "corpus"]war2     -30.199922  12.725900  -2.373 0.018149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.58 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.2882, Adjusted R-squared:  0.2728
## F-statistic: 18.72 on 8 and 370 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    377 112958
## 2    370  89861  7     23097 13.586 1.194e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for miss.pc and snr"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.965 -12.388  -3.650   4.503  82.303
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.03606    0.97138  18.567  < 2e-16 ***
```

```
## file_eval[, thispred] -0.28033     0.03865  -7.254 2.32e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.14 on 377 degrees of freedom
## Multiple R-squared:  0.1225, Adjusted R-squared:  0.1201
## F-statistic: 52.62 on 1 and 377 DF,  p-value: 2.325e-12
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.283  -9.776  -0.559   1.490  85.784
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   31.32788    8.99795   3.482 0.000558 ***
## file_eval[, thispred]          0.00266    0.09332   0.029 0.977274
## file_eval[, "corpus"]lena_lyon  0.58330    9.27456   0.063 0.949886
## file_eval[, "corpus"]namibia  -12.52195    9.08430  -1.378 0.168908
## file_eval[, "corpus"]paido    -31.20811   10.52764  -2.964 0.003229 **
## file_eval[, "corpus"]tsay     -26.45028    9.32897  -2.835 0.004830 **
## file_eval[, "corpus"]tsimane  -17.11537    9.31344  -1.838 0.066907 .
## file_eval[, "corpus"]vanuatu  -13.67091    9.64889  -1.417 0.157372
## file_eval[, "corpus"]war2     -30.21583   12.72462  -2.375 0.018078 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.58 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.2882, Adjusted R-squared:  0.2728
## F-statistic: 18.72 on 8 and 370 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    377 110780
## 2    370  89863  7     20917 12.303 3.62e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] "Regressions with and without corpus for miss.pc and nb_diff_speakers"
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred], subset = c(!is.na(file_eval[,
##     "corpus"])))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.555 -12.476  -6.090   6.818  83.445
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          17.2877      1.8969   9.114   <2e-16 ***
## file_eval[, thispred] -0.7327     0.5425  -1.351    0.178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.25 on 377 degrees of freedom
## Multiple R-squared:  0.004815,   Adjusted R-squared:  0.002176
## F-statistic: 1.824 on 1 and 377 DF,  p-value: 0.1776
##
##
## Call:
## lm(formula = file_eval[, thismet] ~ file_eval[, thispred] + file_eval[,
##     "corpus"])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.257  -8.355  -0.248   3.751  78.068
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   44.8350     8.5927   5.218 3.02e-07 ***
## file_eval[, thispred]         -4.0515     0.5471  -7.405 8.95e-13 ***
## file_eval[, "corpus"]lena_lyon -1.1892    8.6587  -0.137  0.89084
## file_eval[, "corpus"]namibia  -10.5269    8.4821  -1.241  0.21537
## file_eval[, "corpus"]paido    -40.5154    8.6746  -4.671 4.21e-06 ***
## file_eval[, "corpus"]tsay     -23.6485    8.7144  -2.714  0.00696 **
## file_eval[, "corpus"]tsimane  -18.8516    8.6948  -2.168  0.03079 *
## file_eval[, "corpus"]vanuatu  -16.4395    9.0125  -1.824  0.06895 .
## file_eval[, "corpus"]war2     -35.6187   11.8975  -2.994  0.00294 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.54 on 370 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:   0.38,  Adjusted R-squared:  0.3666
## F-statistic: 28.35 on 8 and 370 DF,  p-value: < 2.2e-16
##
## Analysis of Variance Table
##
## Model 1: file_eval[, thismet] ~ file_eval[, thispred]
## Model 2: file_eval[, thismet] ~ file_eval[, thispred] + file_eval[, "corpus"]
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    377 125633
## 2    370  78265  7     47368 31.991 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```