Estimate the Gender from the age and intra-cranial volume of the brain

Lets generate some test data with volume, site, age and gender. We will use linear functions of volume by age with fixed variances but different offsets for male and females. We will add a site variable (sites A, B) again with a small offset. This allows us to show the more complex mixed effect models that can also cope with repeated measures (same id, several scans by id) and nesting (random effects as (~(1|site/family)). We will generate data without nesting and a single scan by participant.

```
data = data.frame(age=runif(2000,50,90))
data$id = as.factor(paste("PAT0", seq(1,dim(data)[1],1),1,sep=""))
data$gender = sample(c("M","F"), size=dim(data)[1], replace=TRUE, prob=c(0.5,0.5))
data$gender = as.factor(data$gender)
data$site = sample(c("A","B"), size=dim(data)[1], replace=TRUE, prob=c(0.5,0.5))
data$site = as.factor(data$site)
data$volume[data$gender == "M"] = -400.9 * data$age[data$gender == "M"] + 600000 +
                                  rnorm(length(data$age[data$gender == "M"]), 0, 21000) +
                                  (as.numeric(data$site[data$gender == "M"]) * 100);
data$volume[data$gender == "F"] = -400.9 * data$age[data$gender == "F"] + 550000 +
                                  rnorm(length(data$age[data$gender == "F"]), 0, 18000) +
                                  (as.numeric(data$site[data$gender == "F"]) * 100);
# cubic centimeter from cubic mm
data$volume = data$volume * 0.001
summary(data)
```

```
##       age                id       gender   site         volume
##  Min.   :50.02   PAT010001:   1   F:1037   A: 991   Min.   :461.1
##  1st Qu.:59.85   PAT01001 :   1   M: 963   B:1009   1st Qu.:521.2
##  Median :69.70   PAT010011:   1                     Median :542.6
##  Mean   :69.84   PAT010021:   1                     Mean   :546.4
##  3rd Qu.:79.89   PAT010031:   1                     3rd Qu.:572.8
##  Max.   :89.99   PAT010041:   1                     Max.   :628.9
##                  (Other)  :1994
```

We can show more information about the data:

```
library(tableone)
tab=CreateTableOne(data=data, vars = c("age", "volume"), factorVars = c("gender", "site"), strata=c("ge
print(tab, showAllLevels = TRUE)
```

```
##                      Stratified by gender:site
##                       level F:A            M:A            F:B
##   n                             510            481            527
##   age (mean (SD))            69.49 (11.67)  69.95 (11.71)  70.42 (11.71)
##   volume (mean (SD))        521.76 (18.27) 572.64 (20.93) 522.18 (17.56)
##                      Stratified by gender:site
##                       M:B            p      test
##   n                      482
##   age (mean (SD))     69.48 (11.79)  0.522
##   volume (mean (SD)) 572.72 (21.53) <0.001
```

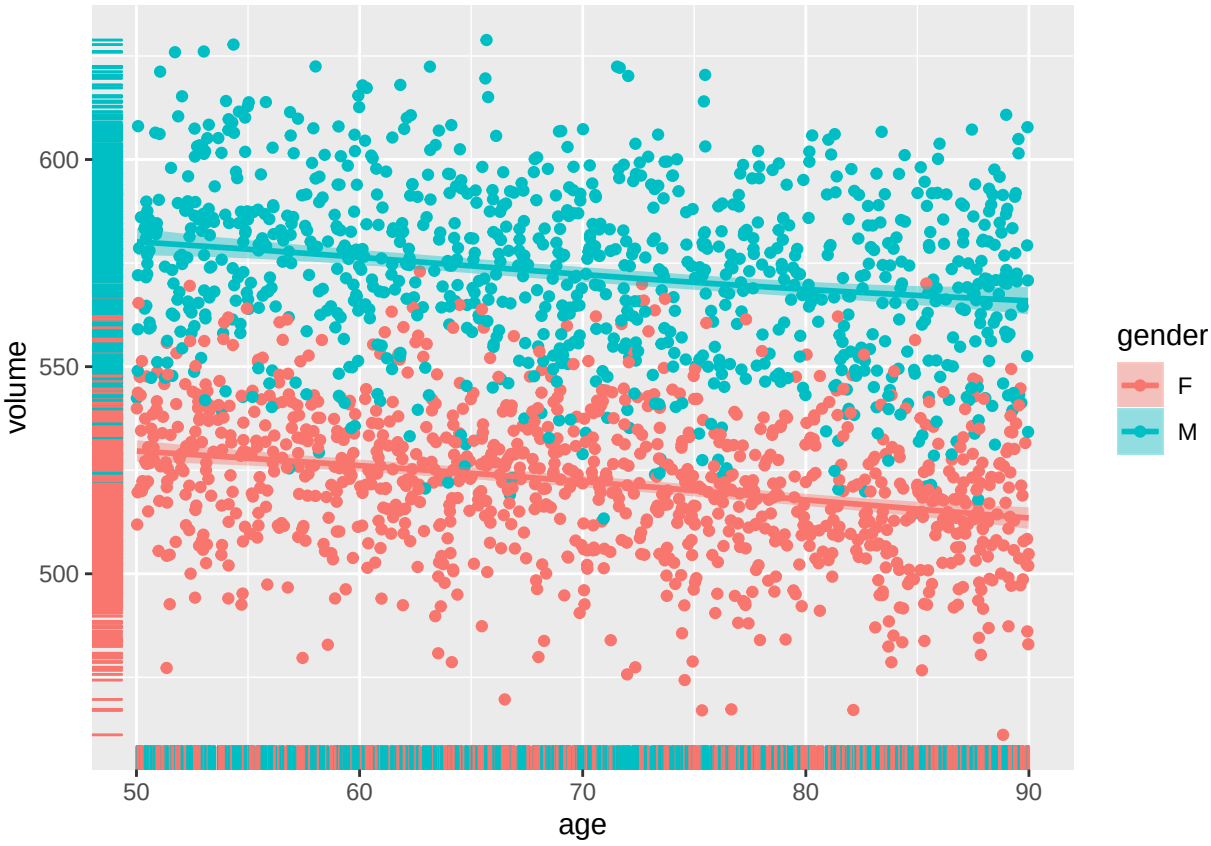
```
# summary(tab)
```

We can plot these data as a scatter plots. Shows the difference between male and female over age and the

randomness of the values - just toy data.

```
library("ggplot2")
ggplot(data=data, aes(x=age, y=volume, color=gender)) + geom_point(aes(color=gender)) + geom_smooth(se="
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Now we can fit a model similar to the implicit model used above for the plotting.

```
library('gamm4')
```

```
## Loading required package: Matrix
## Loading required package: lme4
## Loading required package: mgcv
## Loading required package: nlme
##
## Attaching package: 'nlme'
## The following object is masked from 'package:lme4':
##
##     lmList
## This is mgcv 1.8-28. For overview type 'help("mgcv-package")'.
## This is gamm4 0.2-5
```

```
formula_full = volume ~ s(age,bs="ts",k=3) + gender
formula_part = volume ~ gender
model_full = gamm4(formula_full, data=data, random=~(1|site))
```

```
model_part = gamm4(formula_part, data=data, random=~(1|site))
#anova(model_full$gam)
#summary(model_full$gam)
#summary(model_full$mer)
```

To compare the model with and without age - to see if age really helps - we compute the AIC values of both models (smaller is better). Compute R-squared value (proportion of variance of volume explained by age).

```
if (!('gamm4'     %in% installed.packages()[,"Package"])) install.packages('gamm4')
if (!('rjson'     %in% installed.packages()[,"Package"])) install.packages('rjson')
if (!('stargazer' %in% installed.packages()[,"Package"])) install.packages('stargazer')
if (!('knitr'     %in% installed.packages()[,"Package"])) install.packages('knitr')
if (!('MuMIn'     %in% installed.packages()[,"Package"])) install.packages('MuMIn')
if (!('R.matlab'  %in% installed.packages()[,"Package"])) install.packages('R.matlab')
if (!('tableone'  %in% installed.packages()[,"Package"])) install.packages('tableone')

if(!"rjson" %in% .packages())     library(rjson)
if(!"stargazer" %in% .packages()) library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
if(!"knitr" %in% .packages())     library(knitr)
if(!"MuMIn" %in% .packages())     library(MuMIn)
```

```
## Registered S3 method overwritten by 'MuMIn':
##   method         from
##   predict.merMod lme4
```

```
if(!"R.matlab" %in% .packages())  library(R.matlab)
```

```
## R.matlab v3.6.2 (2018-09-26) successfully loaded. See ?R.matlab for help.
```

```
##
## Attaching package: 'R.matlab'
```

```
## The following objects are masked from 'package:base':
##
##     getOption, isOpen
```

```
if(!"tableone" %in% .packages())  library(tableone)

model_full_aic = AIC(model_full$mer)
model_part_aic = AIC(model_part$mer)
rsq = round(as.numeric(r.squaredLR(model_full$mer,model_part$mer)),5)
```

The above AIC values compare the model with and without the independent variable of interest (x-axis). If the AIC score of two model changes by at least -30 (in our model its -119) between model B (full model without independent variable of interest) and model A (full model), the model A can be accepted as a better model describing the data (information loss as measured by Akaiken Information Criterion). This is of interest because limited numbers of data points allow for a limited number of variables in the model to be tested, but increasing the number of variables will likely yield a better model fit (overfitting). The AIC value of two models can be compared even if the models have different numbers of variables.

The R^2 value is the variance explained in the model by adding the independent variable of interest. The

larger that value the more variance is explained - increases the model fit. The value is in percent, so in this case we end up with 5.859% of the variance in volume explained by age. Assuming that we know the gender and site the new participant comes from.

We can use this model to predict the volume given values for site, age and gender. If we don't know if the participant is male or female we can test both of these hypothesis:

```
vol_if_male = predict(model_full$gam,data.frame(age=65,site="A", gender="M"))
vol_if_female = predict(model_full$gam,data.frame(age=65,site="A", gender="F"))
```

The expected volume, given all the data we have, of a participant that is male at age 65 for site A would therefore be 575. If the same participant is female the expected volume would be 524.

We could hope to get a gender estimate given a volume if we compare an observed volume (for age = 65 and site A) with the two volumes we got above. If our observed volume is closer to vol_if_male we could guess that the gender of the participant is male, if the volume is instead closer to vol_if_female we could guess that the gender is female. This would work if the variances of the volumes at that age and site for both males and females are the same. That is probably a good assumption - if we have the same number of participants at each age interval for both genders.

A better way of implementing the model to decide if someone looks like the female or male participants in the data is to use gender as the dependent variable (the variable on the right that we want to predict).

```
formula_gender_full = gender ~ volume + s(age,bs="ts",k=3)
formula_gender_part = gender ~ s(age,bs="ts",k=3)
```

In this case we now use a categorical variable as the outcome variable and we need to change our model to using logistic regression. This is done by adding the 'family = binomial' option to gamm4. This will generate some errors because of scaling issues between the predictor variables. Probably that is caused by the units of the volumes. We can change the units from cubic milimeter to cubic meters before estimating the model (see change in generated data above).

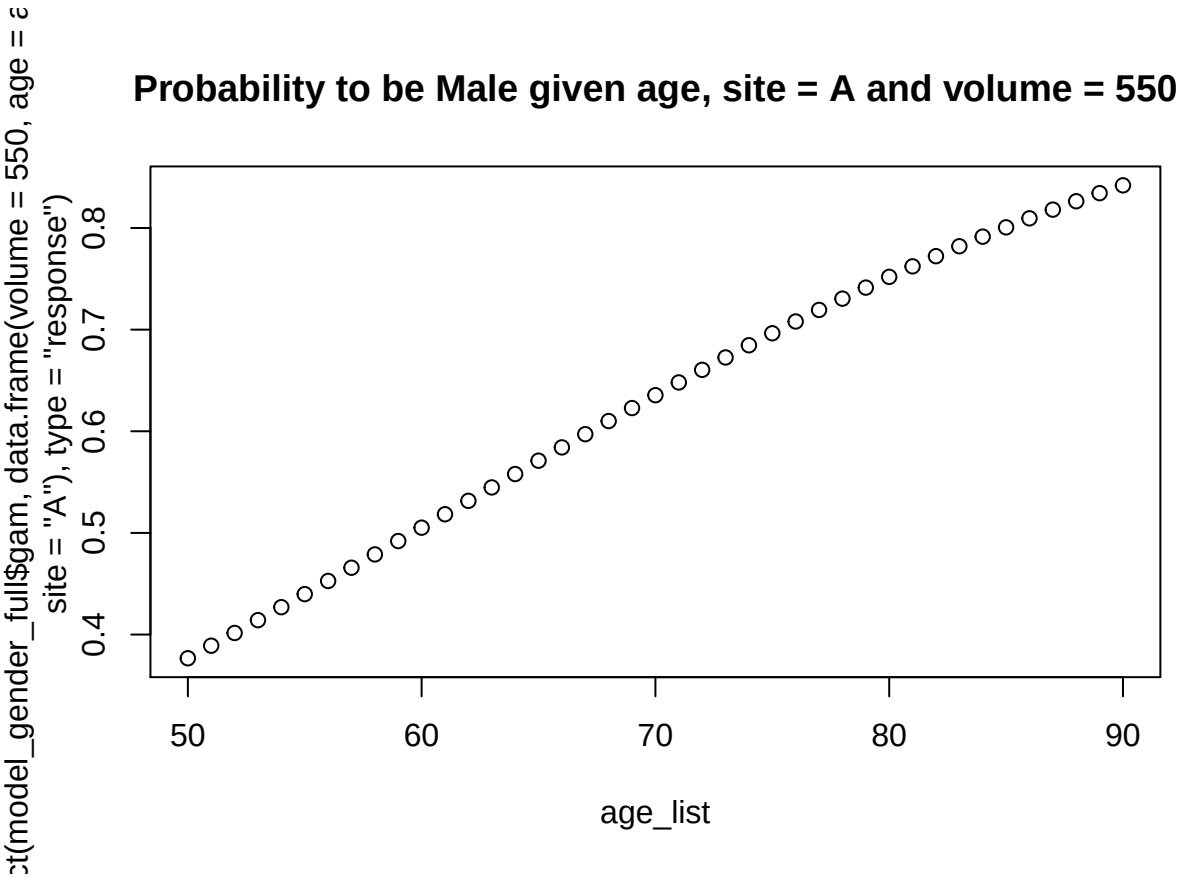
As a site note: The model can fail to converge with a "pwrssUpdate Error" if the noise added to the test data is not gaussian but for example uniformely distributed (adding noise by ifrun instead of rnorm).

```
model_gender_full = gamm4(formula_gender_full, data=data, random=~(1|site), family = binomial)
model_gender_part = gamm4(formula_gender_part, data=data, random=~(1|site), family = binomial)
model_gender_full_aic = AIC(model_full$mer)
model_gender_part_aic = AIC(model_part$mer)
rsq_gender = round(as.numeric(r.squaredLR(model_gender_full$mer,model_gender_part$mer)),5)
```

The difference in AIC if we have volume to estimate gender is -119. As this value is smaller than -30 this indicates that volume is a useful measure to know if one has to estimate gender. The additional variance in gender explained by volume is 61%. This is the variance up and above the variance already explained by site, and a smooth of age.

Ok, we can predict the gender now given a volume, age and site value. We need to specify the type of the value returned by predict as "response" to get a probability for assignment. If we pick a volume of 550 (in the middle between male and female groups) and change the age from 50 to 90 we get an estimated probability to belong to the male group of first below 0.5 and at an age of about 60 the probability is above 0.5 (more likely a male).

```
age_list = seq(50,90,1)
plot(age_list,predict(model_gender_full$gam, data.frame(volume=550,age=age_list,site="A"), type="respons
     main="Probability to be Male given age, site = A and volume = 550")
```



If we want to use this model outside R we could use a lookup table that lists for a given age at a yearly resolution, site and volume the probability that its a male (1- that probability is the probability that its a female). Here we create a table that contains for 100 steps of volume and 100 steps of age (based on the observed min and max ages and volume) for all sites the probability value of being male.

```
lookup_table = data.frame(male_prob=NA,site=NA,volume=NA,age=NA)
for (site in levels(data$site)) {
  for (volume in seq(min(data$volume), max(data$volume), length.out=100)) {
    for (age in seq(min(data$age), max(data$age), length.out=100)) {
      lookup_table = rbind(lookup_table, data.frame(male_prob =
                     predict(model_gender_full$gam,
                             data.frame(age=age, volume=volume, site=site), type="response"),
                     site=site,volume=volume,age=age))
    }
  }
}
```

We can export this table by writing it to a spreadsheet file and use it outside of R to predic the probability of being male or female.

```
write.csv(lookup_table, file="ProbMaleGivenAgeSiteVolume.csv")
```

Done.