EDA: GAMs

Virgilio Gonzenbach

1/13/2021

Objectives

1. To visualize the nonlinear relationship between age and (1) Intracranial Volume (ICV) and (2) ROIs (e.g., GM, WM).
2. To visualize effects of applying Combat-GAM to data

Age Descriptives Stats

Full dataset:

```
##        n  mean   sd median  min   max
## age 1185 15.12 3.74  15.33 8.08 23.08
```

Males:

```
##       n  mean   sd median  min   max
## age 555 14.78 3.72  14.67 8.17 22.92
```

Females:

```
##       n  mean   sd median  min   max
## age 630 15.42 3.74  15.75 8.08 23.08
```

ICV GAM

```
gam_icv = mgcv::gam(raw.ICV ~ s(age,  by = sex, bs = 'tp') + sex,
                    data = pnc,
                    family = gaussian,
                    method = 'REML')
```

Model indicates ICV-to-Age relationship in males is approx. linear (EDF $\approx$ 1) and significant. In females, however, the relationship is quadratic (EDF 2).

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## raw.ICV ~ s(age, by = sex, bs = "tp") + sex
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1350084       5033  268.25   <2e-16 ***
## sexMALE       161166       7357   21.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df     F p-value
## s(age):sexFEMALE 2.084  2.617 0.741  0.4166
## s(age):sexMALE   1.309  1.558 4.635  0.0115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.291   Deviance explained = 29.4%
## -REML =  15558  Scale est. = 1.5851e+10  n = 1185
```

Compare with a R-squared of the equivalent linear model 0.29

ICV (Males)

ICV (Females)

Predictors

Estimates

CI

p

Estimates

CI

p

(Intercept)

-0.00

-0.08 – 0.08

1.000

0.00

-0.08 – 0.08

1.000

age

0.12

0.03 – 0.20

0.005

-0.04

-0.12 – 0.04

0.348

Observations

555

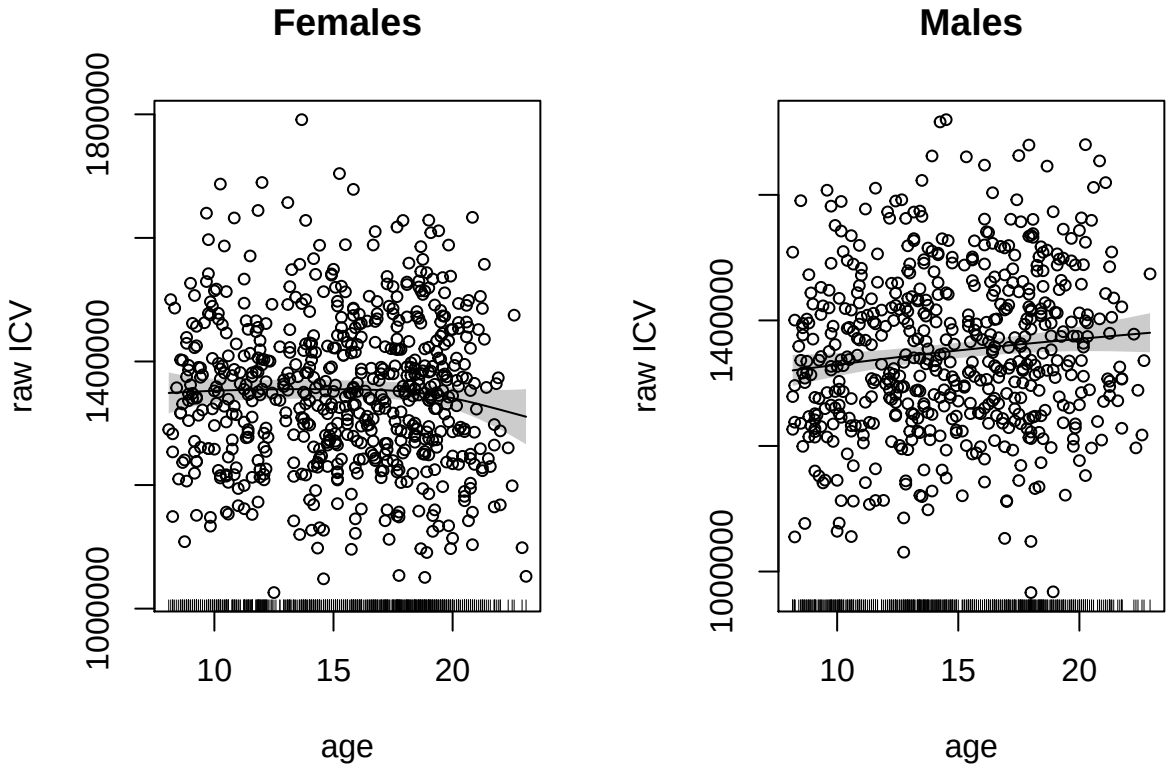
630

R2 / R2 adjusted

0.014 / 0.012

0.001 / -0.000

Plot ICV GAM



WM GAM

Model indicates ICV-to-Age relationship...

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## WM ~ s(age, by = sex, bs = "tp") + sex
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   460364       2003   229.9   <2e-16 ***
## sexMALE        57086       2927    19.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df     F p-value
## s(age):sexFEMALE 2.633  3.305 31.27  <2e-16 ***
## s(age):sexMALE   1.494  1.842 85.81  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.335   Deviance explained = 33.8%
```

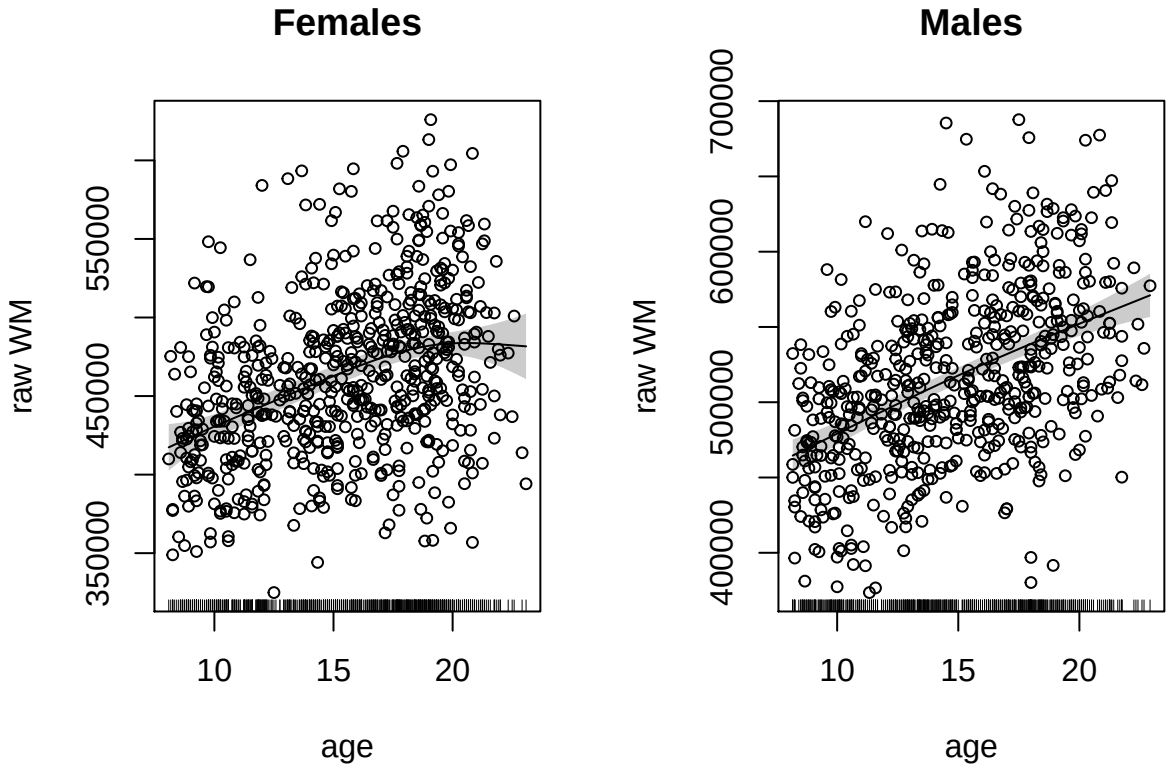
```
## -REML =  14470  Scale est. = 2.5087e+09  n = 1185
```

Compare with a linear model (of each sex separately):

```
##
## Call:
## lm(formula = WM ~ scale(age), data = filter(pnc, sex == "MALE"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158301  -36238   -3335   32575  172528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   515011       2187  235.46   <2e-16 ***
## scale(age)     27198       2189   12.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 51530 on 553 degrees of freedom
## Multiple R-squared:  0.2182, Adjusted R-squared:  0.2168
## F-statistic: 154.3 on 1 and 553 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = WM ~ scale(age), data = filter(pnc, sex == "FEMALE"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -133748  -31709   -1484   28810  144580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   461942       1953  236.51   <2e-16 ***
## scale(age)     19784       1955   10.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49020 on 628 degrees of freedom
## Multiple R-squared:  0.1402, Adjusted R-squared:  0.1389
## F-statistic: 102.4 on 1 and 628 DF,  p-value: < 2.2e-16
```

Plot WM GAM



GM GAM

Model indicates...

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## GM ~ s(age, by = sex, bs = "tp") + sex
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   731329       2887  253.30   <2e-16 ***
## sexMALE        93369       4221   22.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df      F p-value
## s(age):sexFEMALE 1.017  1.033 107.71  <2e-16 ***
## s(age):sexMALE   1.017  1.034  37.37  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.366   Deviance explained = 36.8%
```

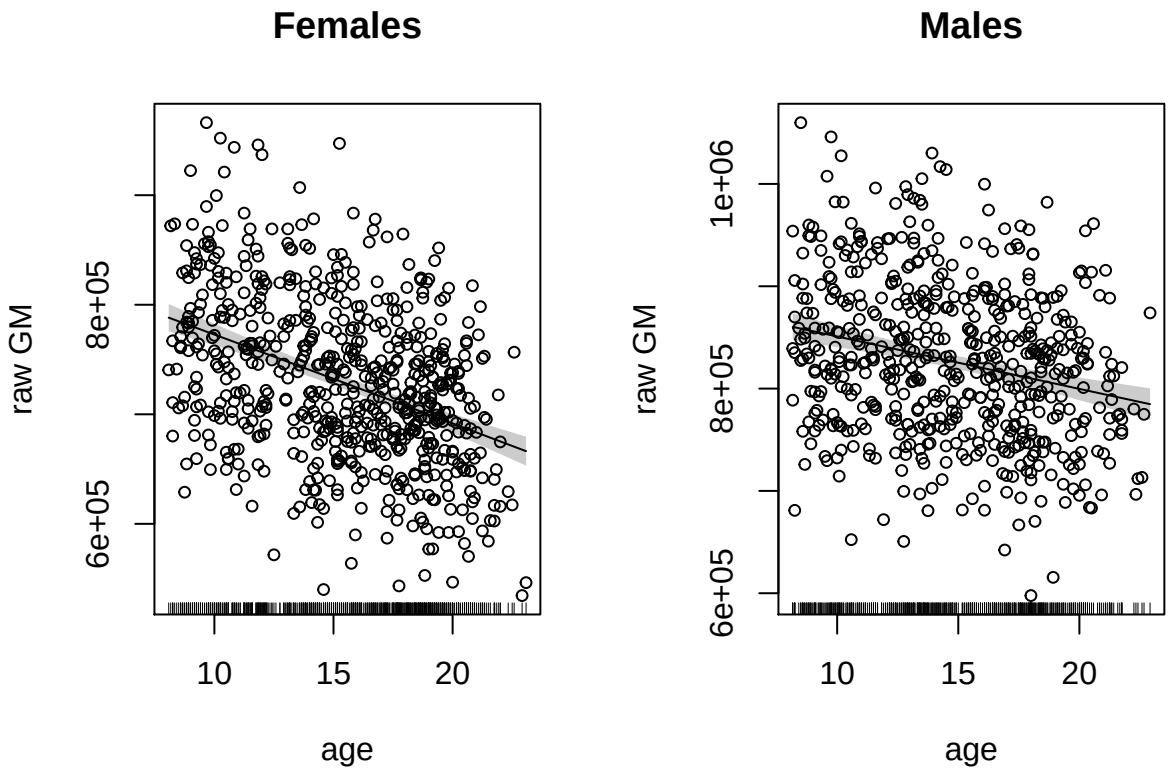
```
## -REML =  14901  Scale est. = 5.2188e+09  n = 1185
```

Compare with a linear model (of each sex separately):

```
##
## Call:
## lm(formula = GM ~ scale(age), data = filter(pnc, sex == "MALE"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -212220  -58071   -3814   51337  201258
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   826424       3205 257.862  < 2e-16 ***
## scale(age)    -19038       3208  -5.935 5.19e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75500 on 553 degrees of freedom
## Multiple R-squared:  0.05988,    Adjusted R-squared:  0.05818
## F-statistic: 35.22 on 1 and 553 DF,  p-value: 5.191e-09
```

```
##
## Call:
## lm(formula = GM ~ scale(age), data = filter(pnc, sex == "FEMALE"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -195649  -48711   -1265   47744  216991
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   728923       2759  264.22   <2e-16 ***
## scale(age)    -30351       2761  -10.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69240 on 628 degrees of freedom
## Multiple R-squared:  0.1614, Adjusted R-squared:   0.16
## F-statistic: 120.8 on 1 and 628 DF,  p-value: < 2.2e-16
```

Plot GM GAM



CSF GAM

Model indicates...

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## VN ~ s(age, by = sex, bs = "tp") + sex
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12861.8      249.1  51.626  < 2e-16 ***
## sexMALE       1885.1      364.2   5.176 2.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df     F  p-value
## s(age):sexFEMALE 1.004  1.007 18.16 2.13e-05 ***
## s(age):sexMALE   1.004  1.008 50.02  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0676   Deviance explained =    7%
```

```
## -REML =  12008  Scale est. = 3.8858e+07  n = 1185
```

Compare with a linear model (of each sex separately):

```
##
## Call:
## lm(formula = VN ~ scale(age), data = filter(pnc, sex == "MALE"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10847  -3899  -1171   2271  49080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14576.6      269.6   54.08  < 2e-16 ***
## scale(age)    1880.5      269.8    6.97 9.06e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6350 on 553 degrees of freedom
## Multiple R-squared:  0.08075,    Adjusted R-squared:  0.07909
## F-statistic: 48.58 on 1 and 553 DF,  p-value: 9.061e-12
```

```
##
## Call:
## lm(formula = VN ~ scale(age), data = filter(pnc, sex == "FEMALE"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9282  -3850  -1331   1533  46747
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12946.1      244.2  53.018  < 2e-16 ***
## scale(age)    1063.3      244.4   4.351 1.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6129 on 628 degrees of freedom
## Multiple R-squared:  0.02926,    Adjusted R-squared:  0.02772
## F-statistic: 18.93 on 1 and 628 DF,  p-value: 1.581e-05
```

Plot CSF GAM

```
# Males
gam_vn_M = mgcv::gam(VN ~ s(age,
                            k=10,
                            bs = 'tp'),
                     data = filter(pnc, sex == "MALE"),
                     family = gaussian,
                     method = 'REML')
```

```r
# Females
gam_vn_F = mgcv::gam(VN ~ s(age,
                            k=10,
                            bs = 'tp'),
                     data = filter(pnc, sex == "FEMALE"),
                     family = gaussian,
                     method = 'REML')

par(mfrow= c(1, 2))
plot(gam_vn_F, residuals = TRUE, pch = 1, ylab = "raw CSF", cex = 0.75, shade = TRUE, seWithMean = TRUE
plot(gam_vn_M, residuals = TRUE, pch = 1, ylab = "raw CSF", cex = 0.75, shade = TRUE, seWithMean = TRUE
```