Normative modeling in Schizophrenia - Analysis of the 308 regions parcellation

Noemi González Lois

2021-06-29

Packages and libraries

We get a list of all the needed packages, then we load them all and also get the number of cores available in PC.

```
library("easypackages")

list.of.packages <- c("viridis", "tidyverse", "MatchIt", "grid", "png",
                      "gridExtra", "parallel", "nlme", "JMbayes", "plyr",
                      "BiocManager", "Biostrings", "lme4", "ggplot2", "dplyr",
                      "Hmisc", "devtools", "longCombat", "neuroCombat",
                      "tinytex", "knitr", "ggExtra", "variancePartition",
                      "Rmisc", "doParallel", "hrbrthemes", "cowplot")

new.packages <- list.of.packages[!(list.of.packages %in%
                                     installed.packages()[,"Package"])]
  if(length(new.packages)>0) { install.packages(new.packages)}

libraries(list.of.packages)
ncores <- detectCores()
rm(list.of.packages, new.packages)
```

Set working directory and load functions

```
setwd("/data_J/Scripts")
source("1_DataPreparation.R")
source("2_RegressionModel.R")
source("3_Statistics.R")
source("4_EDA.R")
```

Data preparation

Options of DataPreparation function:

- parc = “parc35” or “parc308” (whether to use the 34 regions parcellation or the 308 regions one)
- harmonization= “lC” or “nC” (whether to use lonCombat or NeuroCombat harmonization)

- match = T or F (whether to use match-it or not)

```
df_lC_NO_matched <- DataPreparation(parc = "parc308",
                                    harmonization = "lC",
                                    match = F)
```

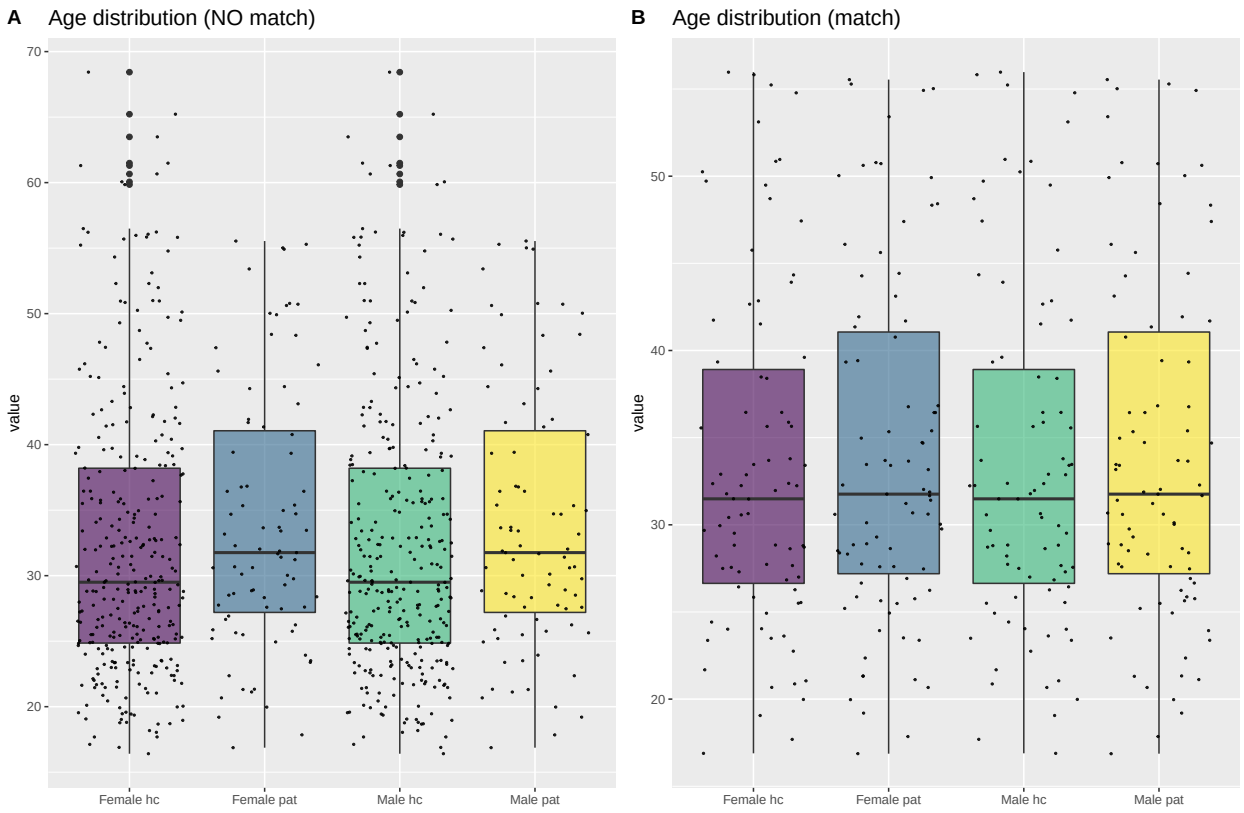
```
## File parc308 already exists, reading file...
## File with longCombat harmonization already exists, reading file...
## Done!
```

```
df_lC_matched <- DataPreparation(parc = "parc308",
                                 harmonization = "lC",
                                 match = T)
```

```
## File parc308 already exists, reading file...
## File with longCombat harmonization already exists, reading file...
## File with MATCH-IT already exists, reading file...
## Done!
```

Exploratory Data Analysis

Show relevant figures and analytics before and after data preparation. First we explore the effects of the match-it:



We show ages than doesn't match between controls and patients (applying floor function to age variable):

Matching ages:

```
EDA_match_ages(df_lC_NO_matched, "NO match-it")
```

```
## Ages that doesn't match in patients vs controls in  NO match-it  dataset are:
##   59 60 62 63 64 65 67 68 69
```

```
EDA_match_ages(df_lC_matched, "match-it")
```

```
## Ages that doesn't match in patients vs controls in  match-it  dataset are:
##   46 56 57 58 60
```

```
## Age in the NO matched dataset is between 16.41 and 69.8
```

```
##
## Age in the matched dataset is between 16.48 and 60.03
```

NO MATCHED DATASET

NO MATCHED	timepoint 1	timepoint 2	timepoint 3
# controls	298	293	109
# patients	169	168	50

Timepoint 1	sex 0	sex 1
# controls	131	167
# patients	38	131

Timepoint 2	sex 0	sex 1
# controls	130	163
# patients	38	130

Timepoint 3	sex 0	sex 1
# controls	50	59
# patients	7	43

MATCHED DATASET

Number of patients vs number of controls per timepoint is not exactly the same:

MATCHED	timepoint 1	timepoint 2	timepoint 3
# controls	169	164	49

MATCHED	timepoint 1	timepoint 2	timepoint 3
# patients	169	164	49

Timepoint 1	sex 0	sex 1
# controls	38	131
# patients	38	131

Timepoint 2	sex 0	sex 1
# controls	37	127
# patients	38	126

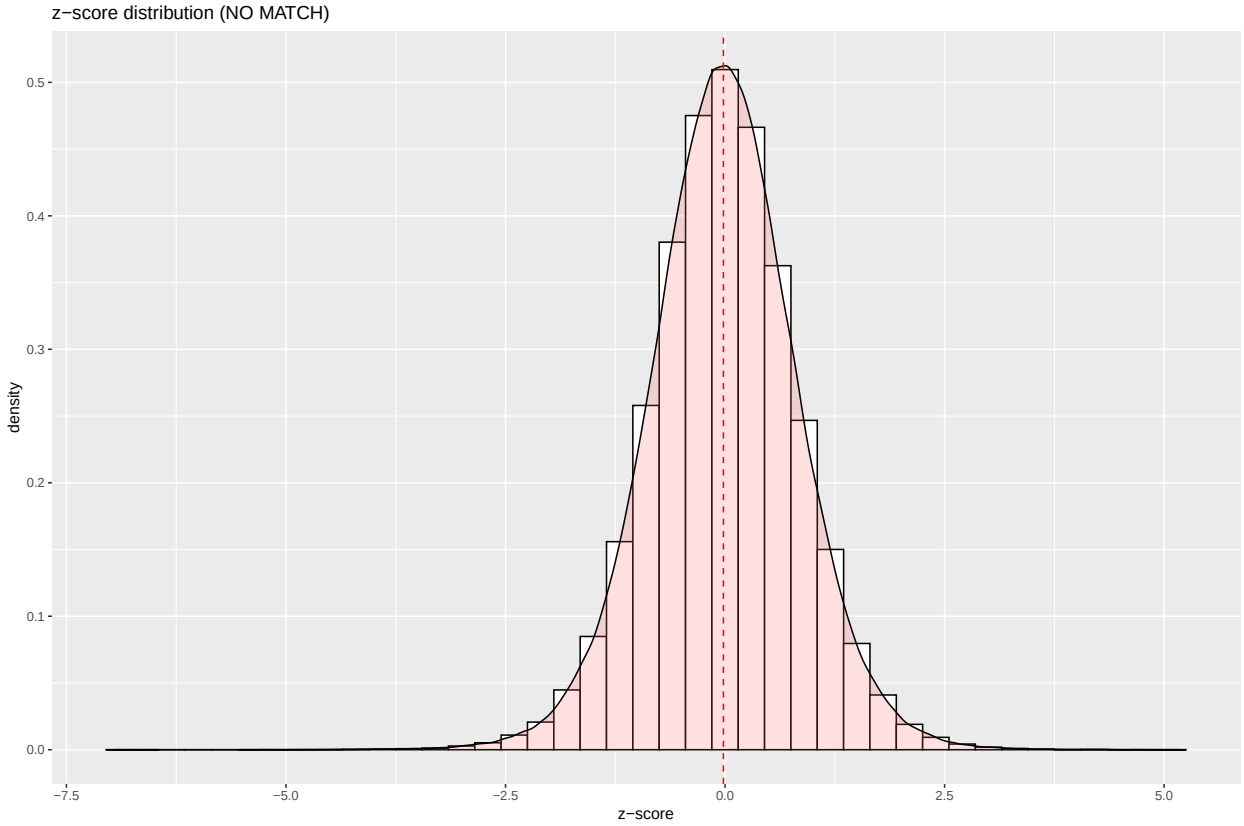
Timepoint 3	sex 0	sex 1
# controls	8	41
# patients	7	42

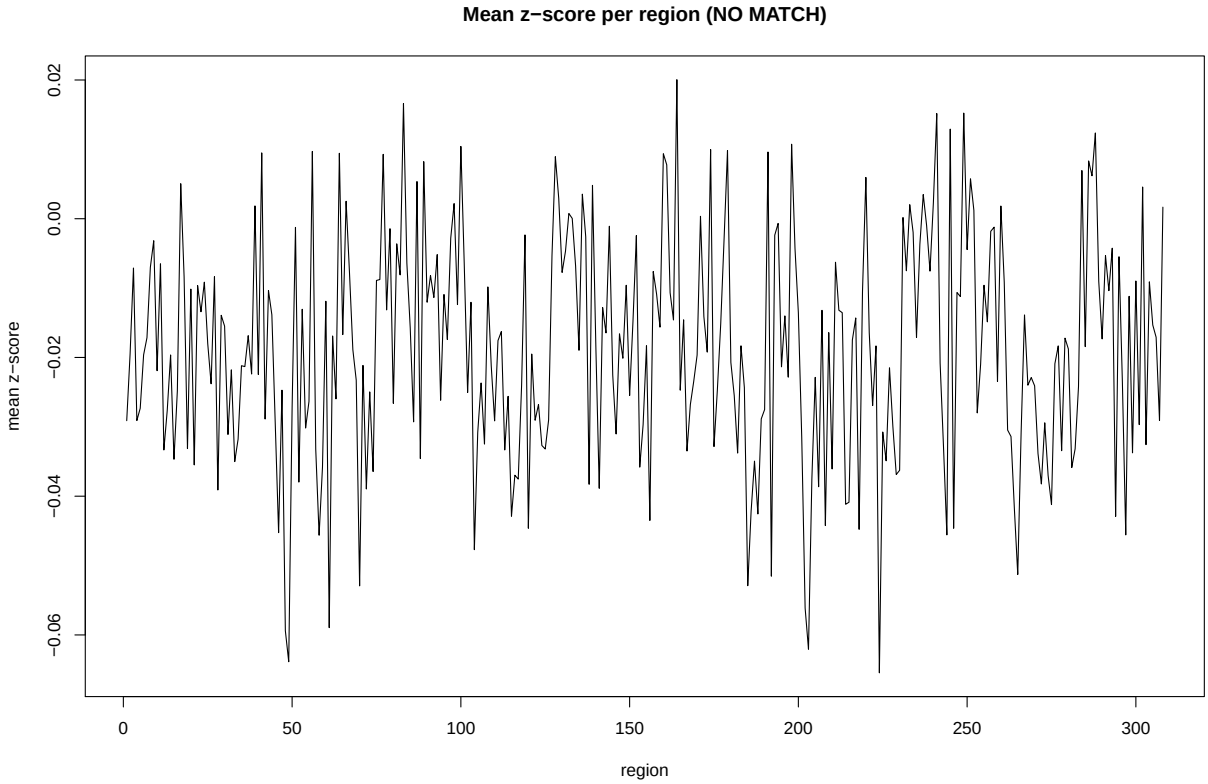
Linear Mixed Effects Model Regresion

Calling the *run_NormativeModel* function with different datasets. We build the regression with the data from healthy controls and make individual predictions over controls as well as patients. This function will return the z-scores (one for each region for each timepoint of each subject)

NO Match-it dataframe (longCombat):

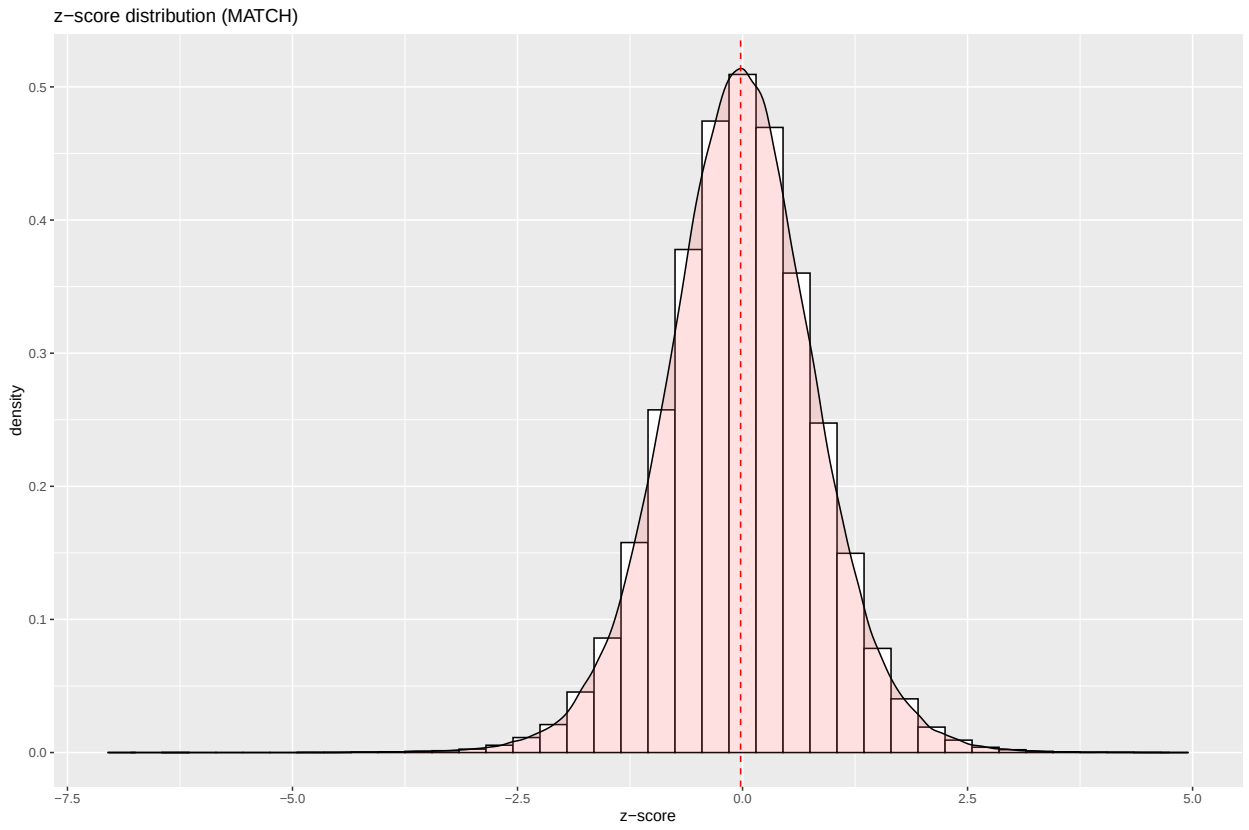
```
Zs_NOmatch <- run_NormativeModel(df_lC_NO_matched,
                                 measure = "thickness",
                                 parc = "parc308",
                                 match = "NOmatch",
                                 harmonization = "lC")
```

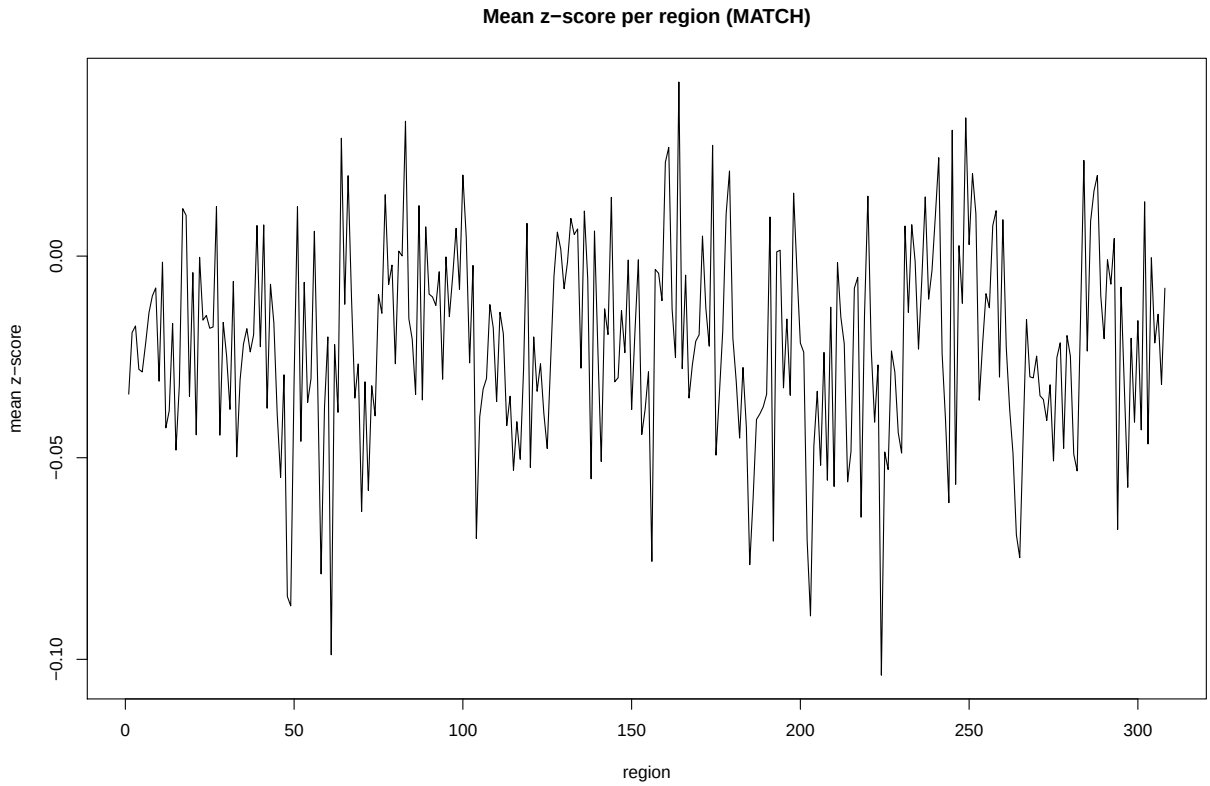

z−score distribution (NO MATCH)
density
0.0
0.1
0.2
0.3
0.4
0.5
−7.5
−5.0
−2.5
0.0
2.5
5.0
z−score


Mean z−score per region (NO MATCH)
mean z−score
0.02
0.00
−0.02
−0.04
−0.06
0
50
100
150
200
250
300
region

Match-it dataframe (longCombat):

```
Zs_match <- run_NormativeModel(df_lC_matched,
                               measure = "thickness",
                               parc = "parc308",
                               match = "match",
                               harmonization = "lC")
```





Match-it dataframe (Age*Diagnosis) BEFORE EXCLUDING DEVIANTS

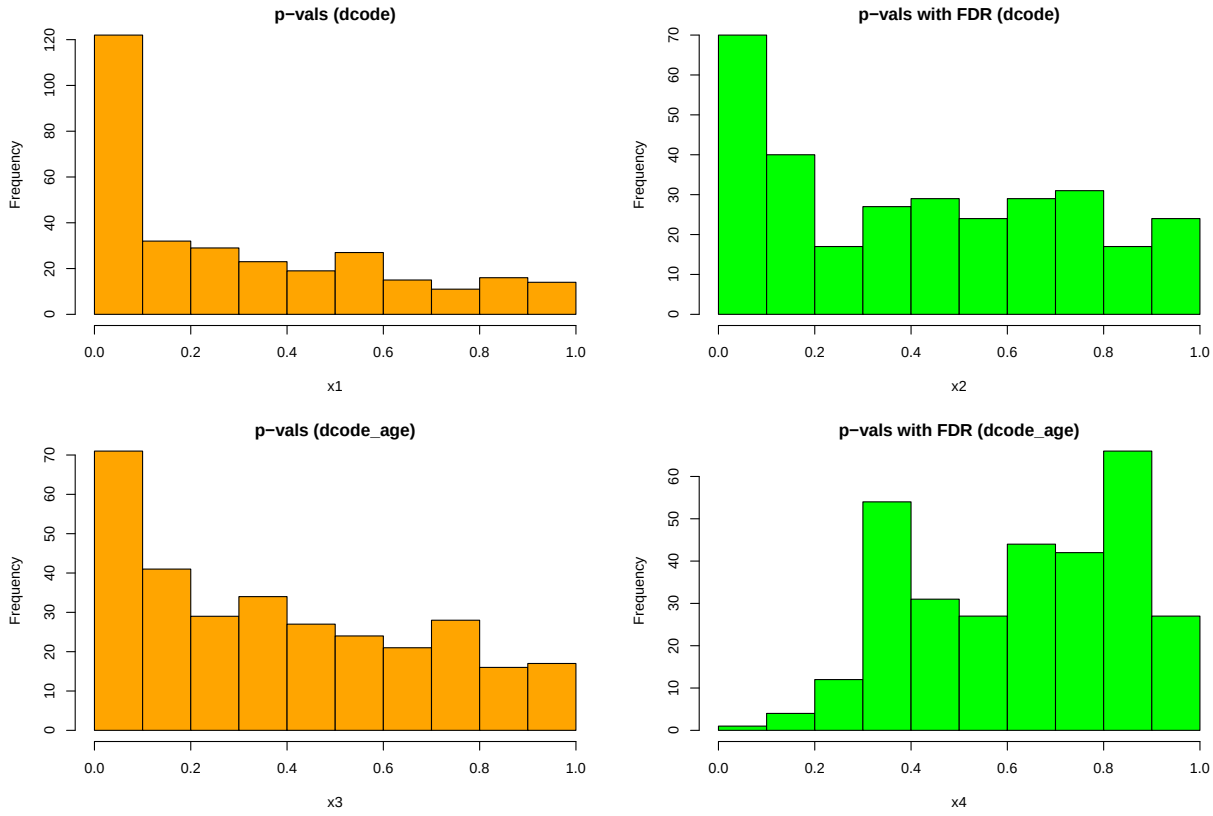
Calling the *run_AgeDiagnosisModel* function. In this case, we build the regression with the data from healthy controls as well as patients. Next we exclude the deviants from the model and run again the function

This function will return the p-values (one for each region of each subject)

```
p_val <- run_AgeDiagnosisModel(df_lC_matched,
                               measure = "thickness",
                               Z = NULL,
                               exclude_deviants = F)

p_val_FDR <- Apply_FDR_Correction(p_val)
```

```
## Without FDR correction:
## Variable dcode has statistical significance for  94 / 308 regions
## Variable dcode_age has statistical significance for  41 / 308 regions

##      reg       scode          age        euler      dcode    dcode_age
## [1,]   0 0.925853805 9.886278e-08 0.0306395581 0.05195468 0.6919928081
## [2,]   0 0.862469074 1.118862e-09 0.0005925169 0.17111506 0.0833292698
## [3,]   0 0.034451738 2.227107e-02 0.2599459951 0.26266032 0.0002843601
## [4,]   0 0.009293535 3.050449e-12 0.0037449896 0.14422297 0.5690237690

## With FDR correction:
## Variable dcode has statistical significance for  42 / 308 regions
## Variable dcode_age has statistical significance for  0 / 308 regions
```

```
##      reg     scode          age       euler     dcode  dcode_age
## [1,]   0 0.9528341 2.435979e-07 0.068883094 0.1651340 0.86288982
## [2,]   0 0.9288128 3.961028e-09 0.003578338 0.3561043 0.38628909
## [3,]   0 0.1310017 2.618700e-02 0.384920031 0.4544909 0.08758292
## [4,]   0 0.0556858 2.135314e-11 0.012141650 0.3299198 0.80333969
```



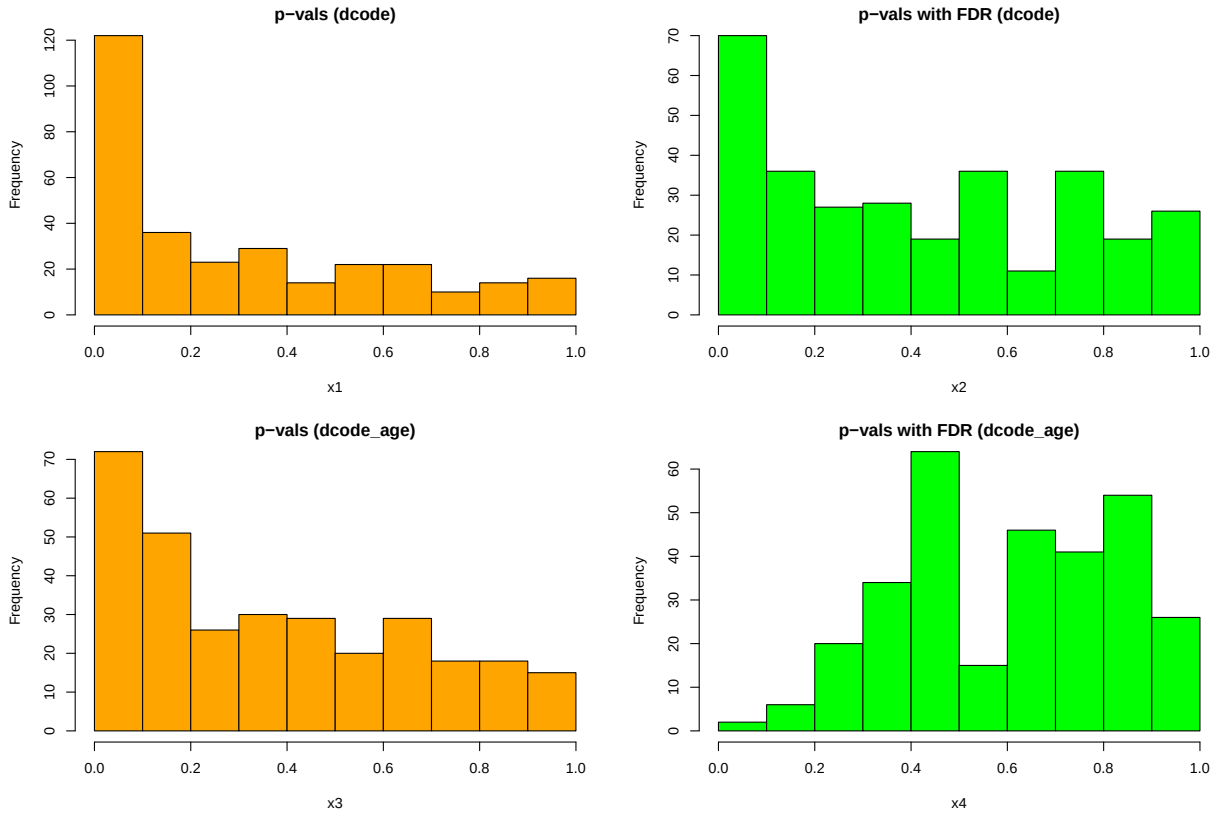
Match-it dataframe (Age*Diagnosis) AFTER EXCLUDING DEVIANTS

```
p_val <- run_AgeDiagnosisModel(df_lC_matched,
                               measure = "thickness",
                               Z = Zs_match,
                               exclude_deviants = T)

p_val_FDR <- Apply_FDR_Correction(p_val)
```

```
## Without FDR correction:
## Variable dcode has statistical significance for  95 / 308 regions
## Variable dcode_age has statistical significance for 43 / 308 regions
```

```
##      reg     scode          age        euler      dcode    dcode_age
## [1,]   0 0.9739481 6.452295e-08 0.0149329013 0.04325705 0.8090443529
## [2,]   0 0.8662380 2.413578e-10 0.0001261839 0.18265030 0.0416145704
## [3,]   0 0.0312670 3.554556e-02 0.2929800901 0.33246511 0.0002741094
## [4,]   0 0.0117006 2.160050e-12 0.0053548943 0.14432086 0.5483439815
```

```
## With FDR correction:
## Variable dcode has statistical significance for  43 / 308 regions
## Variable dcode_age has statistical significance for  0 / 308 regions

##      reg      scode          age       euler     dcode dcode_age
## [1,]   0 0.98031374 1.633728e-07 0.035932294 0.1510536 0.8893680
## [2,]   0 0.93287173 1.009676e-09 0.001038716 0.3750419 0.3230328
## [3,]   0 0.11755282 4.131333e-02 0.417767906 0.5389434 0.0844257
## [4,]   0 0.06825413 1.446294e-11 0.016012694 0.3268443 0.7532983
```



Analysis of z-scores and computation of global scores

The following tables represent, **in terms of samples**, the number of deviations ($|Z| > 1.96$). The number of samples refers to the total number of regions, calculated for each timepoint ($timepoint = 1, 2, 3$) and differentiated between controls and subjects.

NO MATCHED	timepoint 1	timepoint 2	timepoint 3
# total samples	143.528	141.988	48.664
# samples dev (%)	3.421 (2.38%)	3.136 (2.21%)	1.122 (2.31%)
# controls samples	91.784	90.244	33.572
# controls dev (%)	2.042 (**2.22%**)	1.696 (**1.88%**)	680 (**2.03%**)
# patients samples	51.744	51.744	15.092
# patients dev (%)	1.379 (2.67%)	1.440 (2.78%)	442 (2.93%)

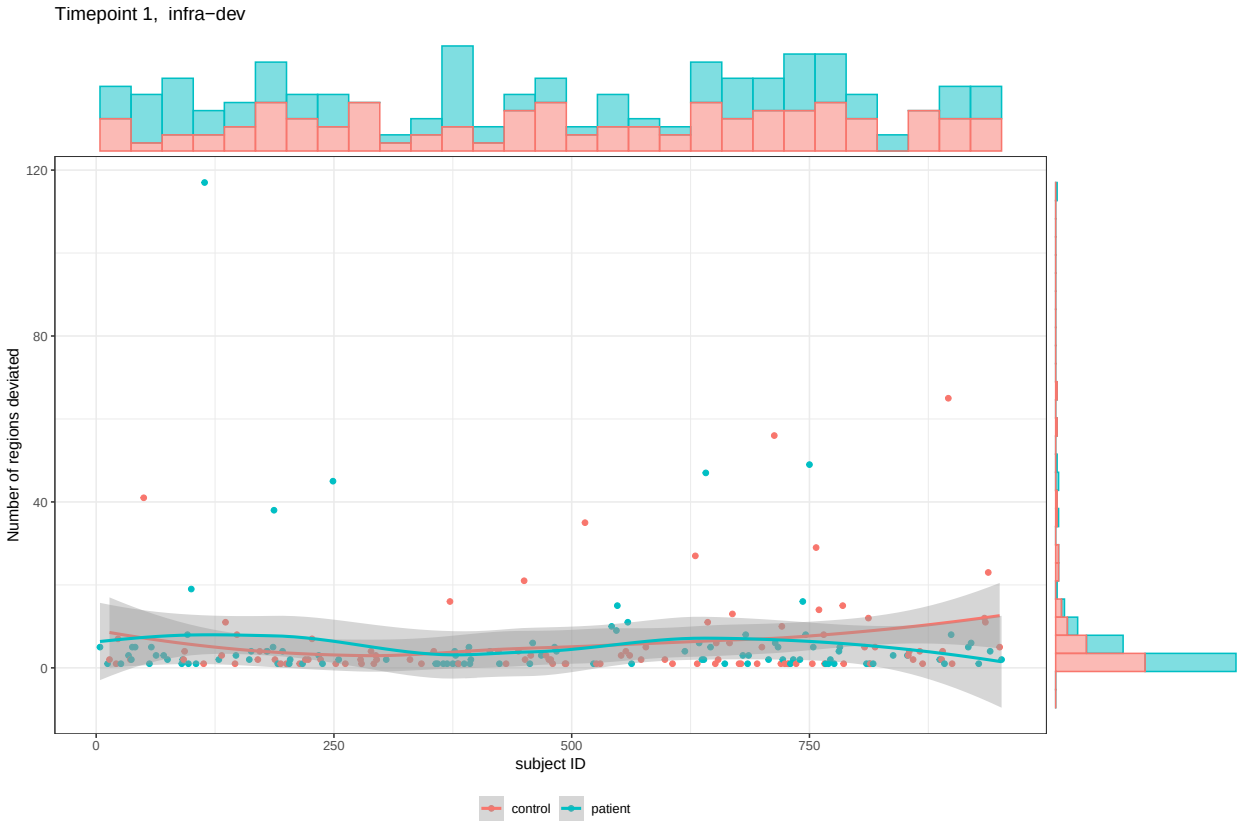
MATCHED	timepoint 1	timepoint 2	timepoint 3
# total samples	103.796	101.024	30.184
# samples dev (%)	2.518 (2.42%)	2.318 (2.27%)	760 (2.33%)
# controls samples	52.052	50.512	15.092
# controls dev (%)	1.111 (**2.13%**)	910 (**1.8%**)	288 (**1.91%**)
# patients samples	51.744	50.512	15.092
# patients dev (%)	1.363 (2.63%)	1.389 (2.75%)	440 (2.92%)

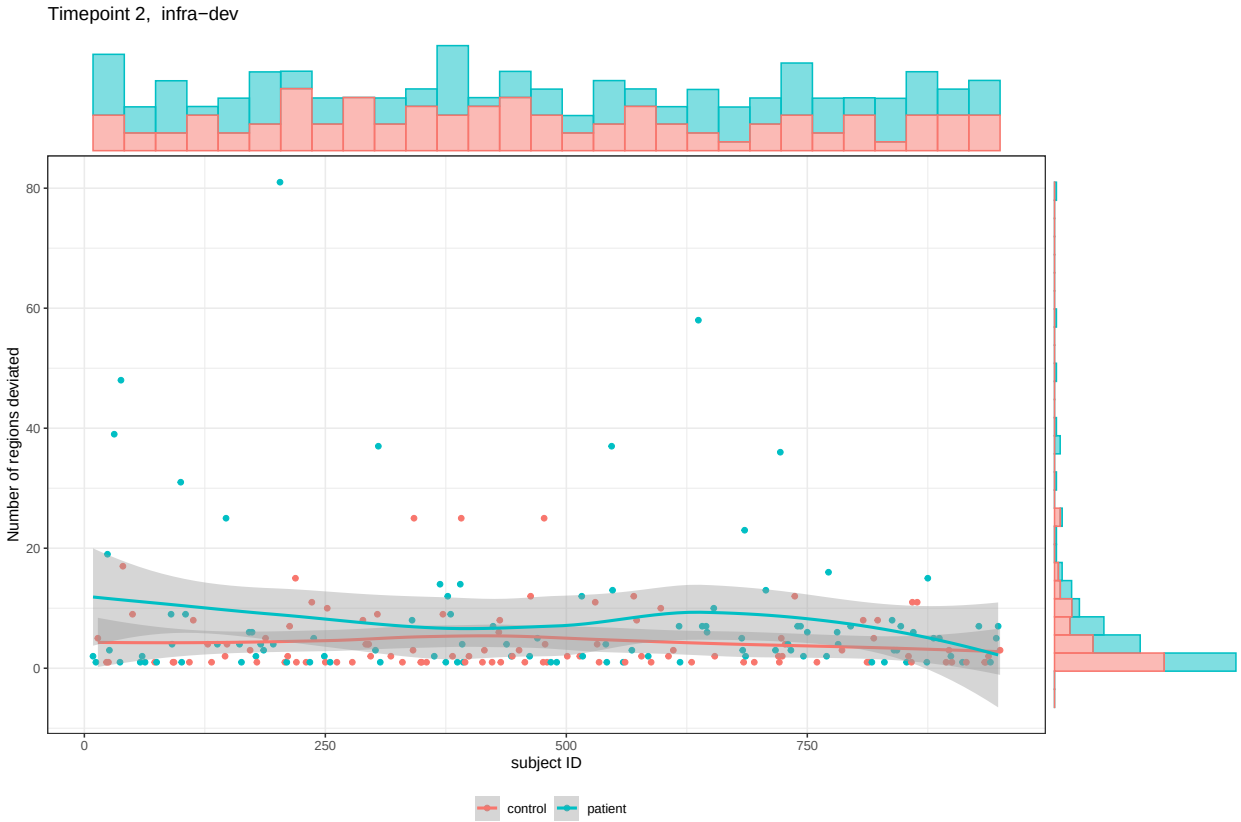
In the following, we chose to work with the matched dataset and the Zs derived from its lme model, providing a better statistical support for the analysis.

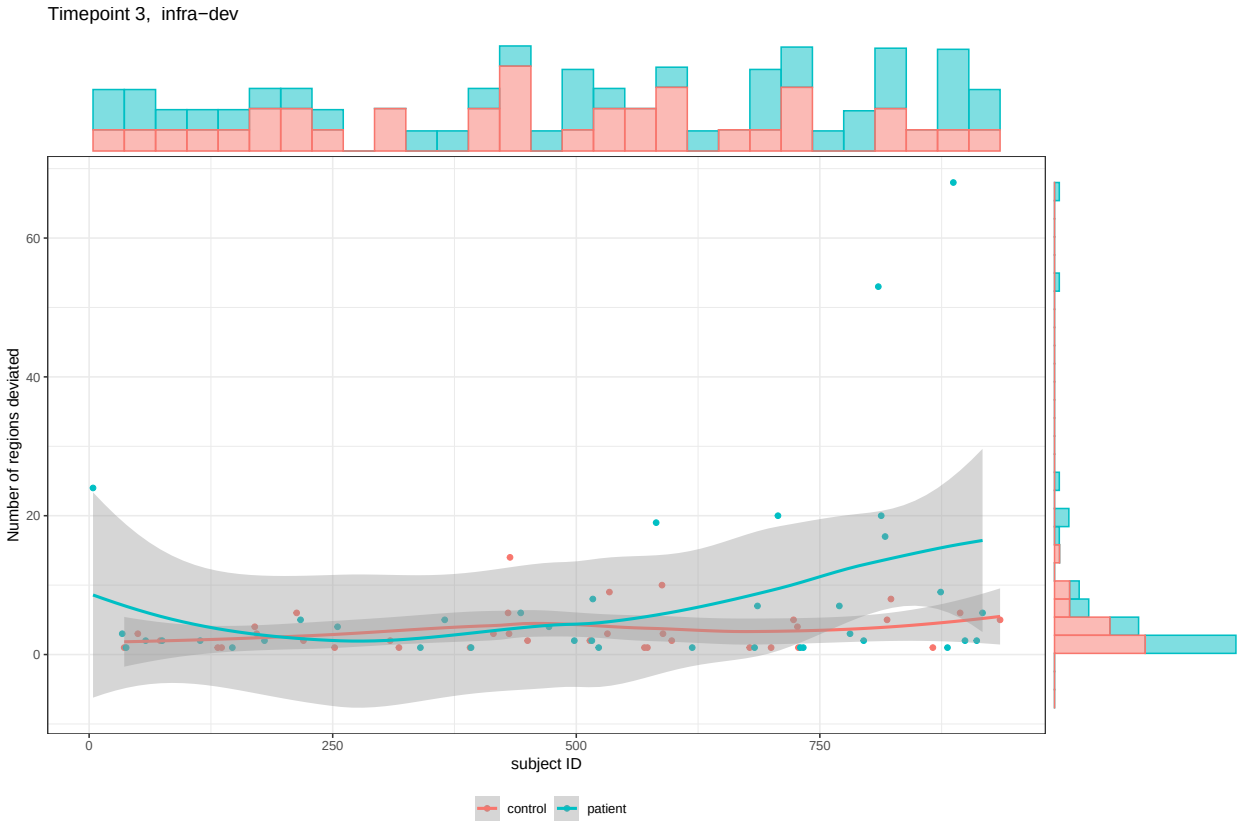
We compute the number of deviant samples ($Z < -1.96$ and $Z > 1.96$):

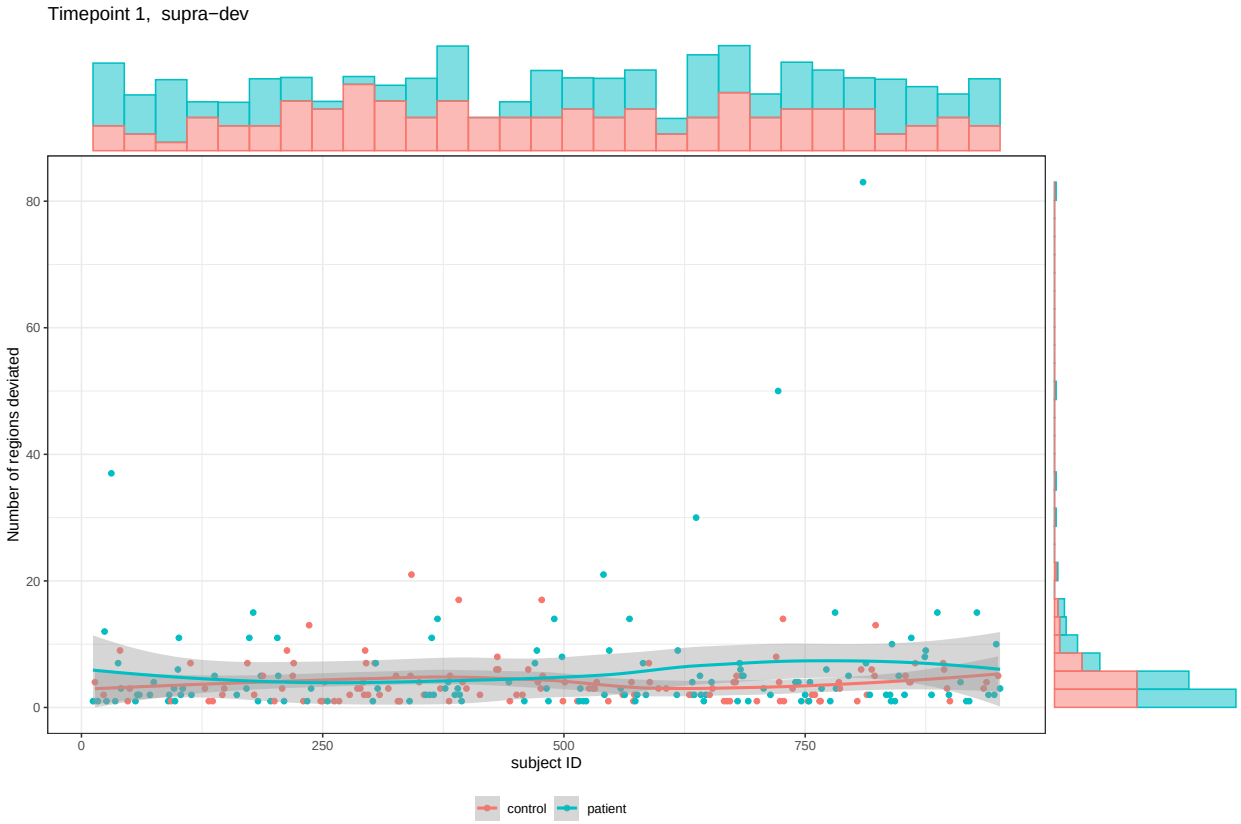
In terms of regions deviated:

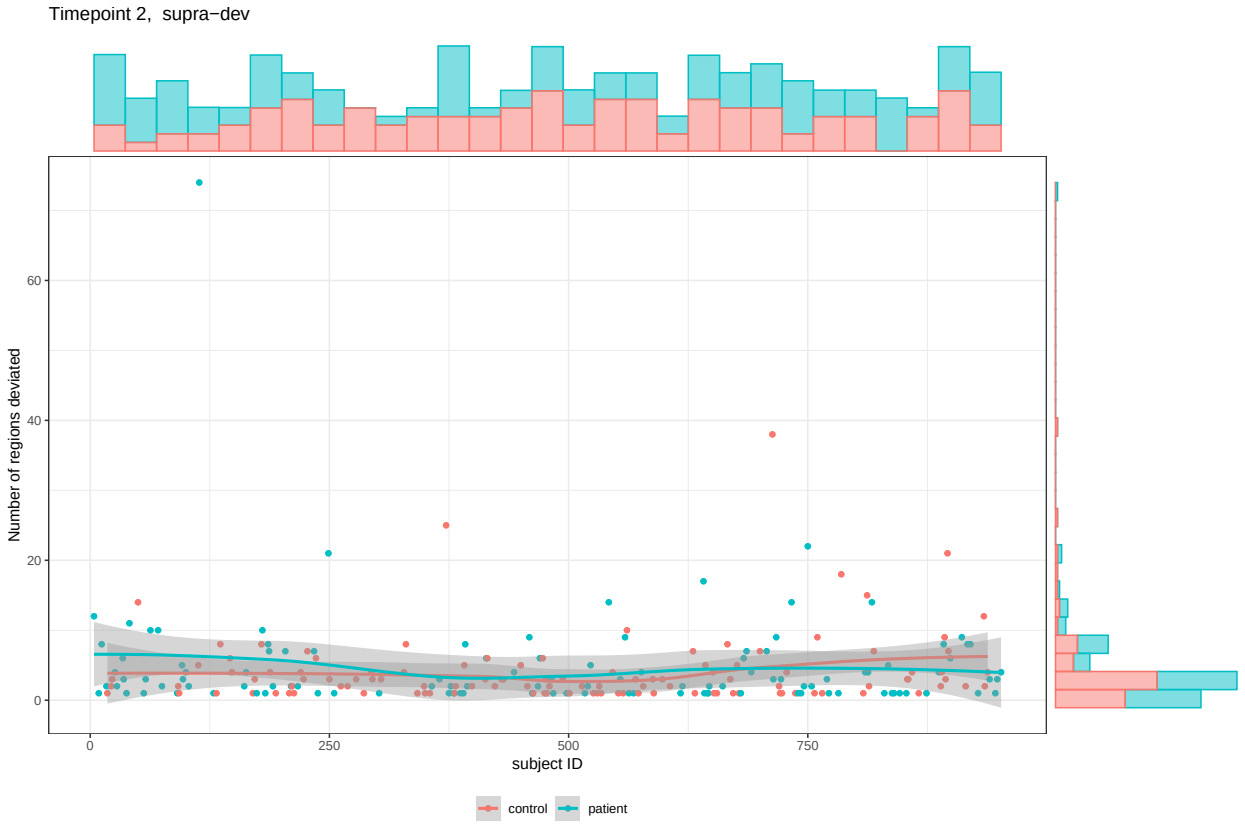
	Infra-normal deviants	Non deviants	Supra-normal deviants	Deviants
tp 1	1.261 (1.21%)	101.321 (97.62%)	1.214 (1.17%)	2.475 (2.38%)
controls	631 (50.04%)	50.941 (50.28%)	480 (39.54%)	1.111 (44.89%)
patients	630 (49.96%)	50.380 (49.72%)	734 (60.46%)	1.364 (55.11%)
tp 2	1.294 (1.28%)	98.725 (97.72%)	1.005 (0.99%)	2.299 (2.28%)
controls	444 (34.31%)	49.602 (50.24%)	466 (46.37%)	910 (39.58%)
patients	850 (65.69%)	49.123 (49.76%)	539 (53.63%)	1.389 (60.42%)
tp 3	440 (1.46%)	29.456 (97.59%)	288 (0.95%)	728 (2.41%)
controls	124 (28.18%)	14.804 (50.26%)	164 (56.94%)	288 (39.56%)
patients	316 (71.82%)	14.652 (49.74%)	124 (43.06%)	440 (60.44%)
total	2.995 (1.27%)	22.9502 (97.66%)	2.507 (1.07%)	5.502 (2.34%)

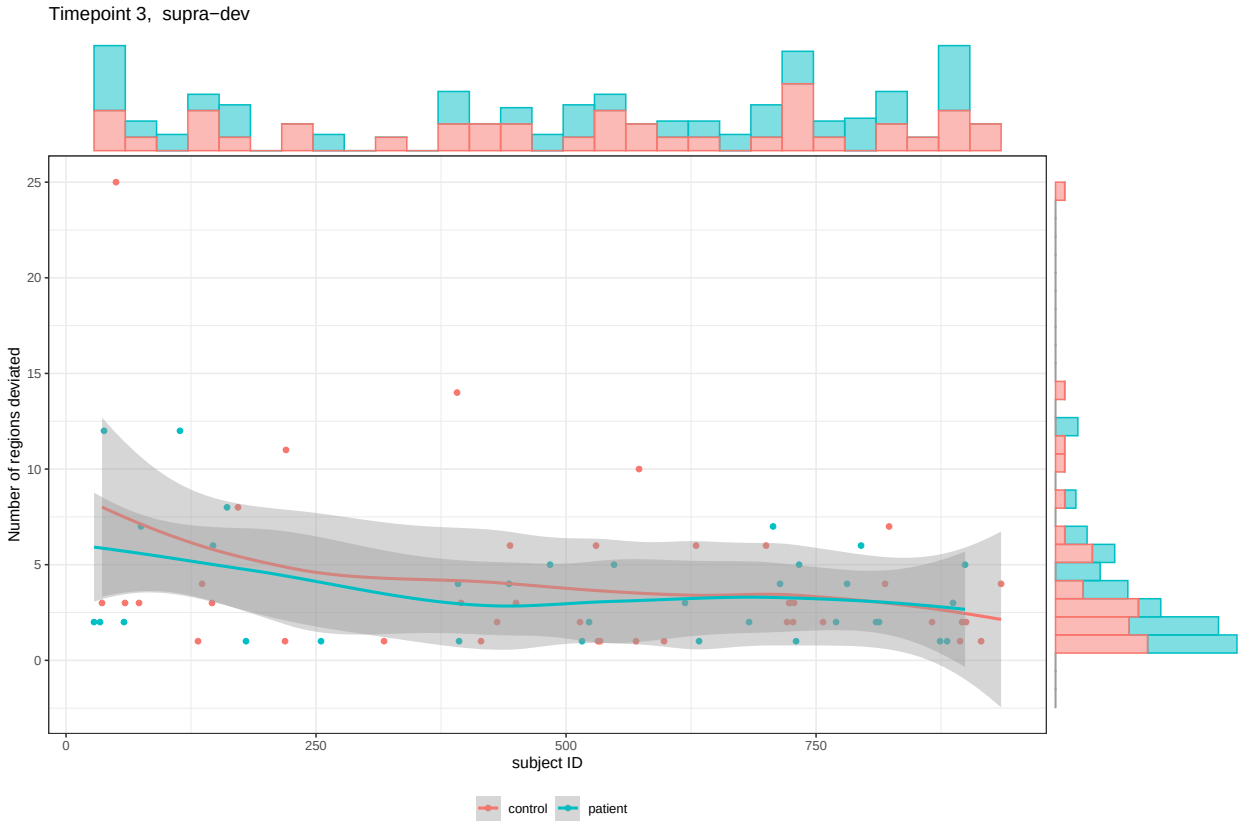

Timepoint 1, infra−dev
Number of regions deviated
subject ID
control
patient


Timepoint 2, infra−dev
Number of regions deviated
subject ID
control
patient


Timepoint 3, infra−dev
Number of regions deviated
0
20
40
60
0
250
500
750
subject ID
control
patient


Timepoint 1, supra−dev
Number of regions deviated
0
20
40
60
80
0
250
500
750
subject ID
control
patient





We computed the number of subjects that are **deviant at timepoint 1** and stay **deviant in subsequent**

timepoints, for each region. A maximum of 6 subjects (1.78% of total subjects) fulfill that condition for any region.

	Condition
lh_cuneus_part2_thickness	6 (1.78%)
lh_precentral_part5_thickness	6 (1.78%)
lh_lateraloccipital_part5_thickness	5 (1.48%)

We computed the number of subjects that are **not deviant at timepoint 1** but become **deviant in subsequent timepoints**, for each region. A maximum of 23 subjects (6.82% of total subjects) fulfill that condition for any region.

	Condition
lh_superiorfrontal_part9_thickness	23 (6.82%)
rh_medialorbitofrontal_part1_thickness	22 (6.53%)
rh_temporalpole_part1_thickness	22 (6.53%)

---

1. Percentage of patients deviated from the normative range for any single cortical region

- Timepoint 1: No more than 9.52% of patients deviated from the normative range for any single cortical region.
- Timepoint 2: No more than 9.76% of patients deviated from the normative range for any single cortical region.
- Timepoint 3: No more than 10.2% of patients deviated from the normative range for any single cortical region.

---

2. Most common regions with infra-normal deviations. Percentage of patients.

- Timepoint 1: Infra-normal deviations in CT of subjects were most commonly located in **lh_superiortemporal_part2_thickness** cortices, although only 4.17% of patients showed significant deviations in these regions.
- Timepoint 2: Infra-normal deviations in CT of subjects were most commonly located in **lh_superiorfrontal_part9_thickness** cortices, although only 6.1% of patients showed significant deviations in these regions.
- Timepoint 3: Infra-normal deviations in CT of subjects were most commonly located in **lh_insula_part1_thickness** cortices, although only 8.16% of patients showed significant deviations in these regions.

---

3. Most common regions with supra-normal deviations. Percentage of individuals.

- Timepoint 1: Supra-normal deviations in CT were most common in the **lh__superiorfrontal__part7__thickness** regions, 3.26% of individuals.
- Timepoint 2: Supra-normal deviations in CT were most common in the **rh__medialorbitofrontal__part1__thickness** regions, 4.88% of individuals.
- Timepoint 3: Supra-normal deviations in CT were most common in the **lh__inferiorparietal__part6__thickness** regions, 6.12% of individuals.

---

4. Percentage of subjects with at least one region infra-normal deviated. Patients vs Healthy controls.

- Timepoint 1: Infra-normal deviations for at least one region were evident in 60.71% of patients, whereas this was the case for 60.36% of healthy individuals.
- Timepoint 2: Infra-normal deviations for at least one region were evident in 63.41% of patients, whereas this was the case for 62.2% of healthy individuals.
- Timepoint 3: Infra-normal deviations for at least one region were evident in 77.55% of patients, whereas this was the case for only 73.47% of healthy individuals.

---

5. Percentage of subjects with at least one region supra-normal deviated. Patients vs Healthy controls.

- Timepoint 1: Supra-normal deviations for at least one region were evident in 76.19% of patients and 71.6% of healthy individuals.
- Timepoint 2: Supra-normal deviations for at least one region were evident in 67.68% of patients and 69.51% of healthy individuals.
- Timepoint 3: Supra-normal deviations for at least one region were evident in 67.35% of patients and 81.63% of healthy individuals.

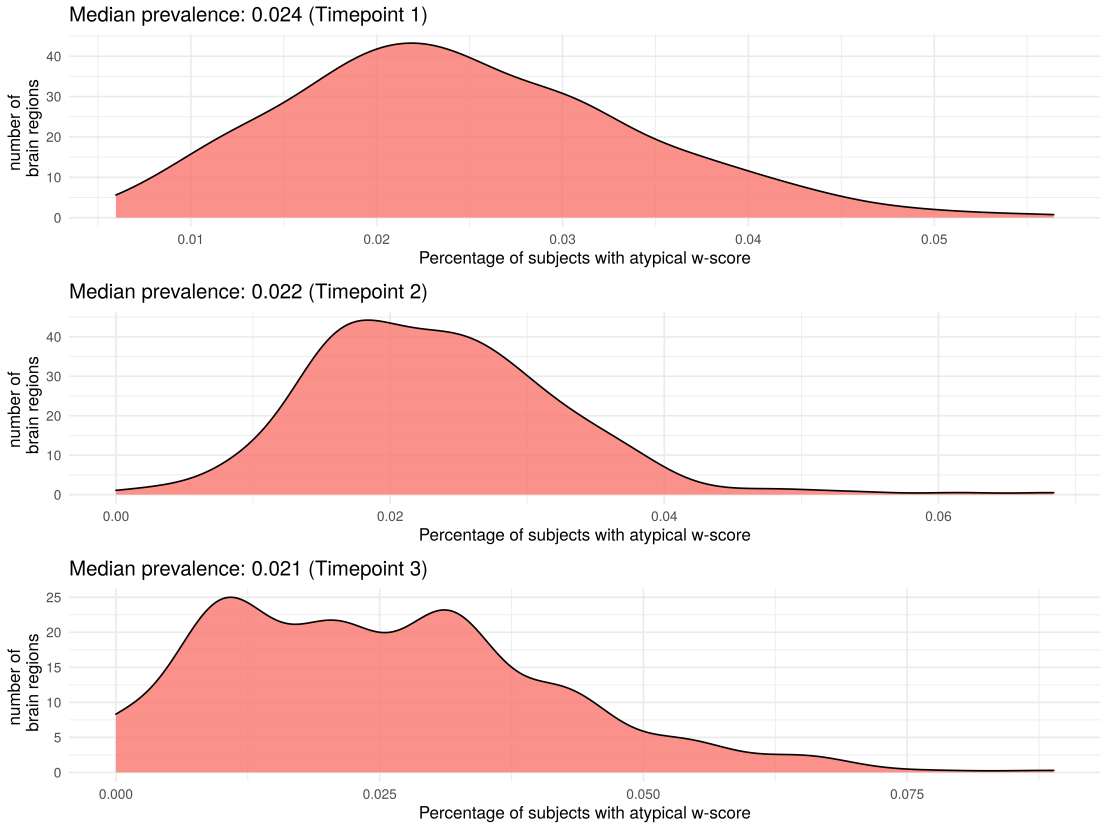
Percentage of deviant subjects for number of regions

Figura B del paper de Bethlehem. En el eje de las x se representa el porcentaje de sujetos y en el eje y el número de regiones con el mismo ratio $\frac{|Z|>1.96}{|Z|<1.96}$. Se calcula para cada timepoint ($timepoint = 1, 2, 3$).

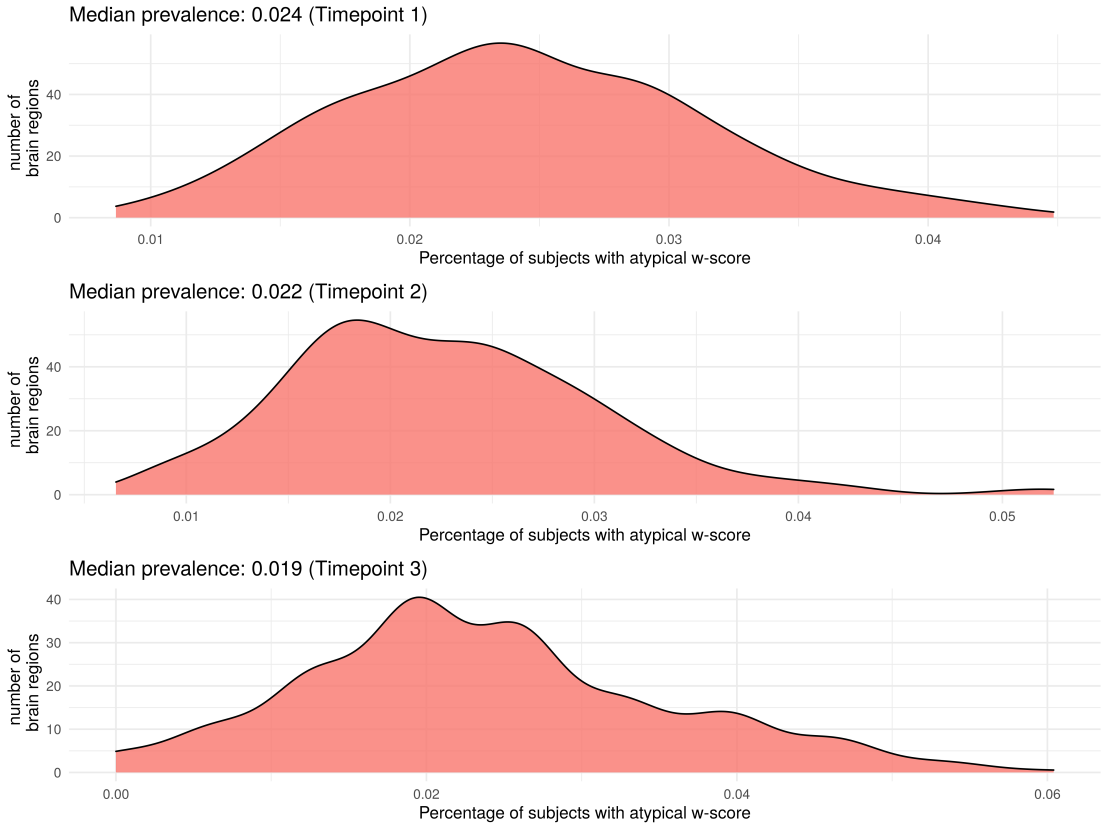
```
getStatistics(Zs= Zs_match,  lab= "1C_match_parc308", parc= "parc308")
getStatistics(Zs= Zs_NOmatch, lab= "1C_NOmatch_parc308", parc= "parc308")
```

Show the results from the global ratios obtained:

Match-it dataframe (longCombat):



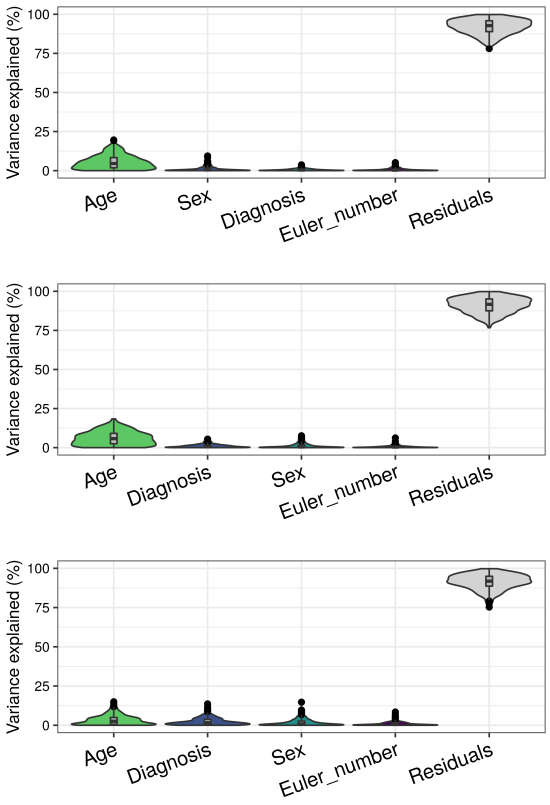
NO Match-it dataframe (longCombat):



Variance contribution across measures

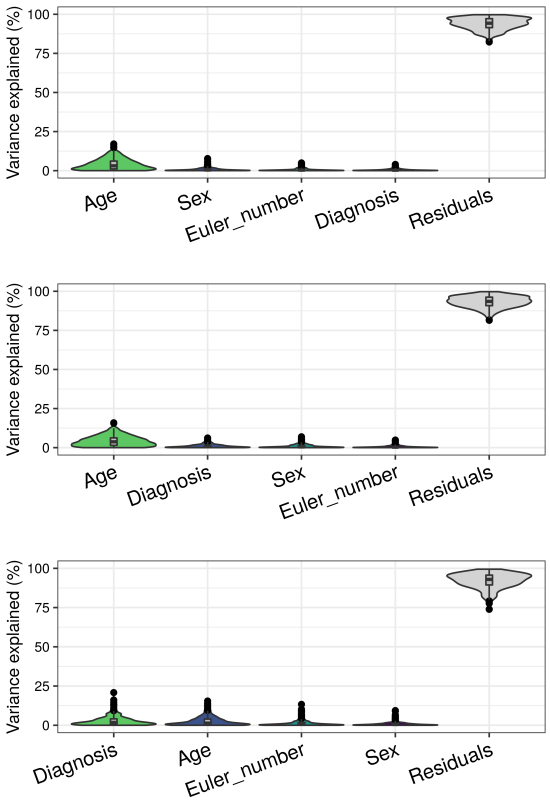
NO Match-it dataframe (longCombat):

```
variancePartition(df = df_lC_NO_matched,
                  measure = "thickness",
                  lab = "lC_NO_matched_parc308",
                  par = "parc308")
```



Match-it dataframe (longCombat):

```
variancePartition(df = df_lC_matched,
                  measure = "thickness",
                  lab = "lC_matched_parc308",
                  par = "parc308")
```


Variance explained (%)
100
75
50
25
0
Age
Sex
Euler_number
Diagnosis
Residuals
Variance explained (%)
100
75
50
25
0
Age
Diagnosis
Sex
Euler_number
Residuals
Variance explained (%)
100
75
50
25
0
Diagnosis
Age
Euler_number
Sex
Residuals