Normative modeling in Schizophrenia - Analysis of the 34 regions parcellation

Noemi González Lois

2021-07-07

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 = "parc35",
                                    harmonization = "lC",
                                    match = F)
```

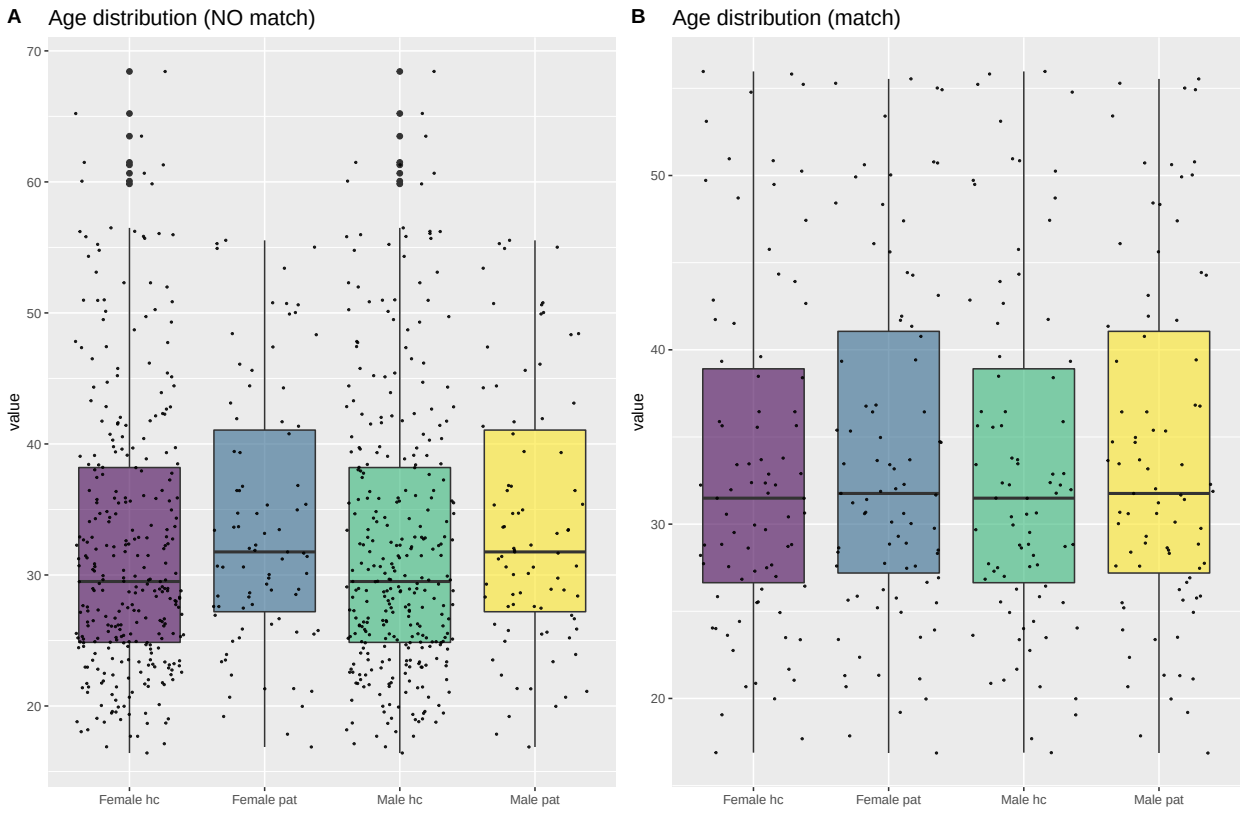
```
## File parc35 already exists, reading file...
## File with longCombat harmonization already exists, reading file...
## Done!
```

```
df_lC_matched <- DataPreparation(parc = "parc35",
                                 harmonization = "lC",
                                 match = T)
```

```
## File parc35 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

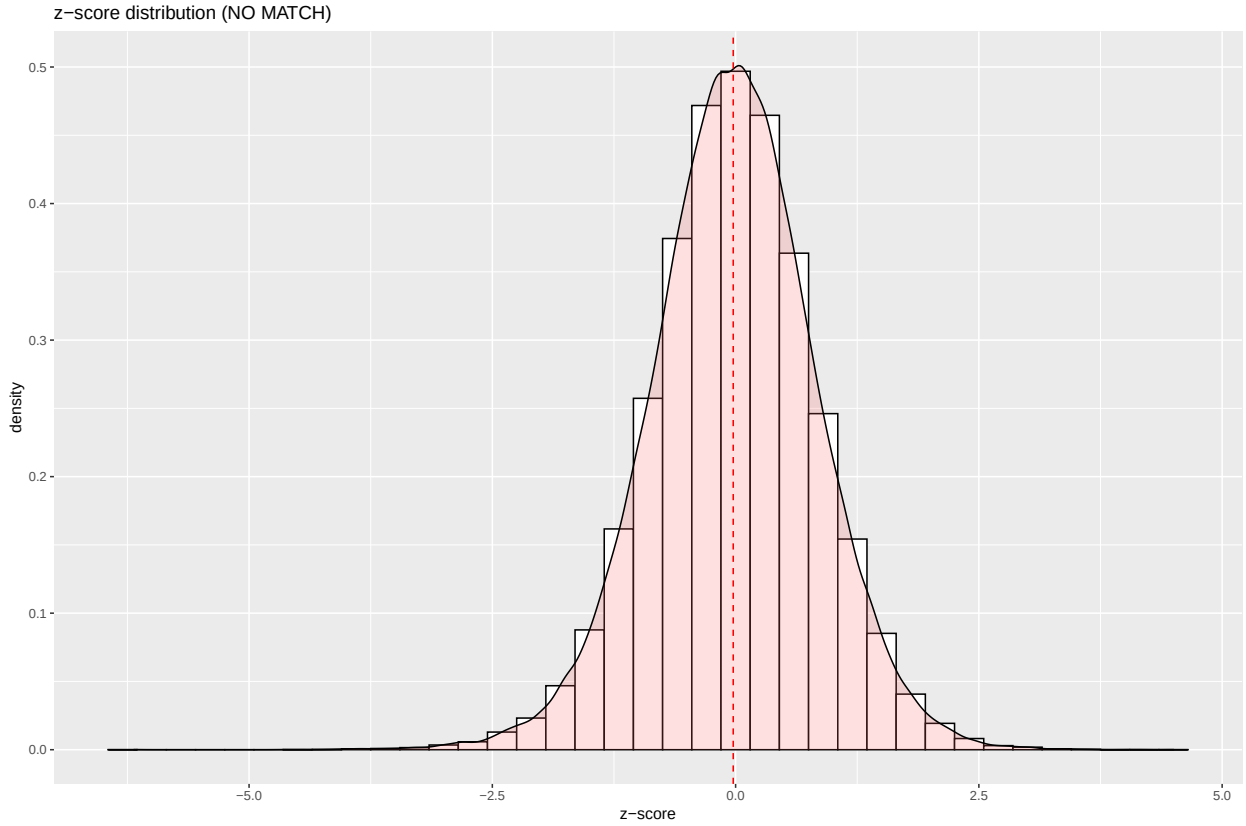
Timepoint 3	sex 0	sex 1
# controls	8	41
# patients	7	42

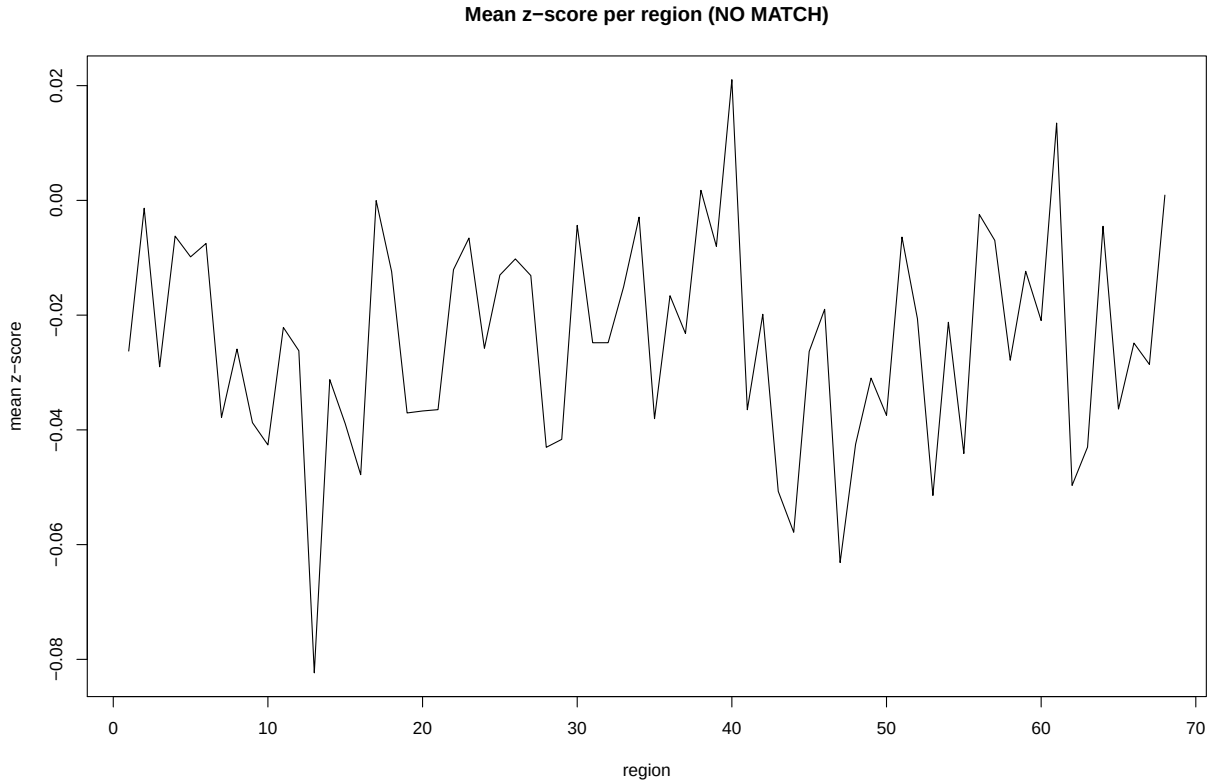
Linear Mixed Effects Model Regresion

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):

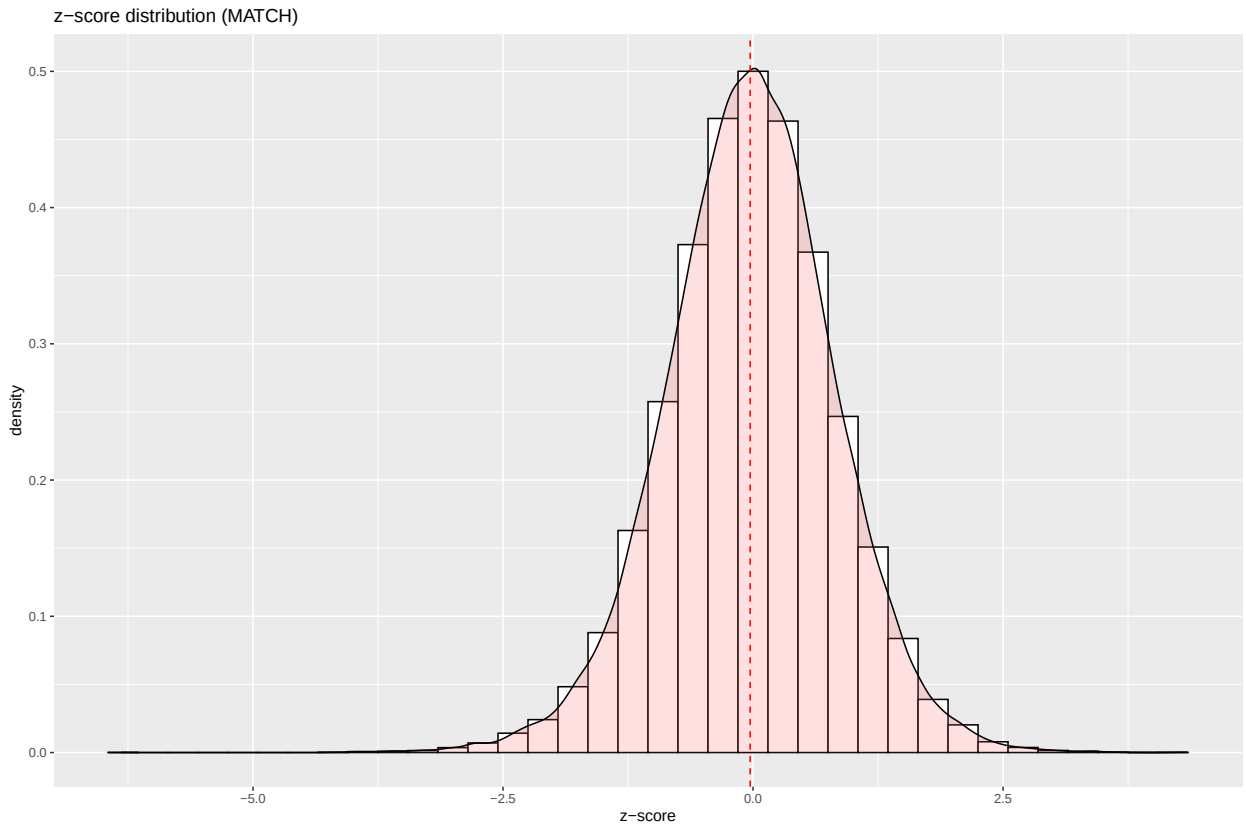
```
Zs_NOmatch <- run_NormativeModel(df_lC_NO_matched,
                                 measure = "CT_freesurfer",
                                 parc = "parc35",
                                 match = "NOmatch",
                                 harmonization = "lC")
```

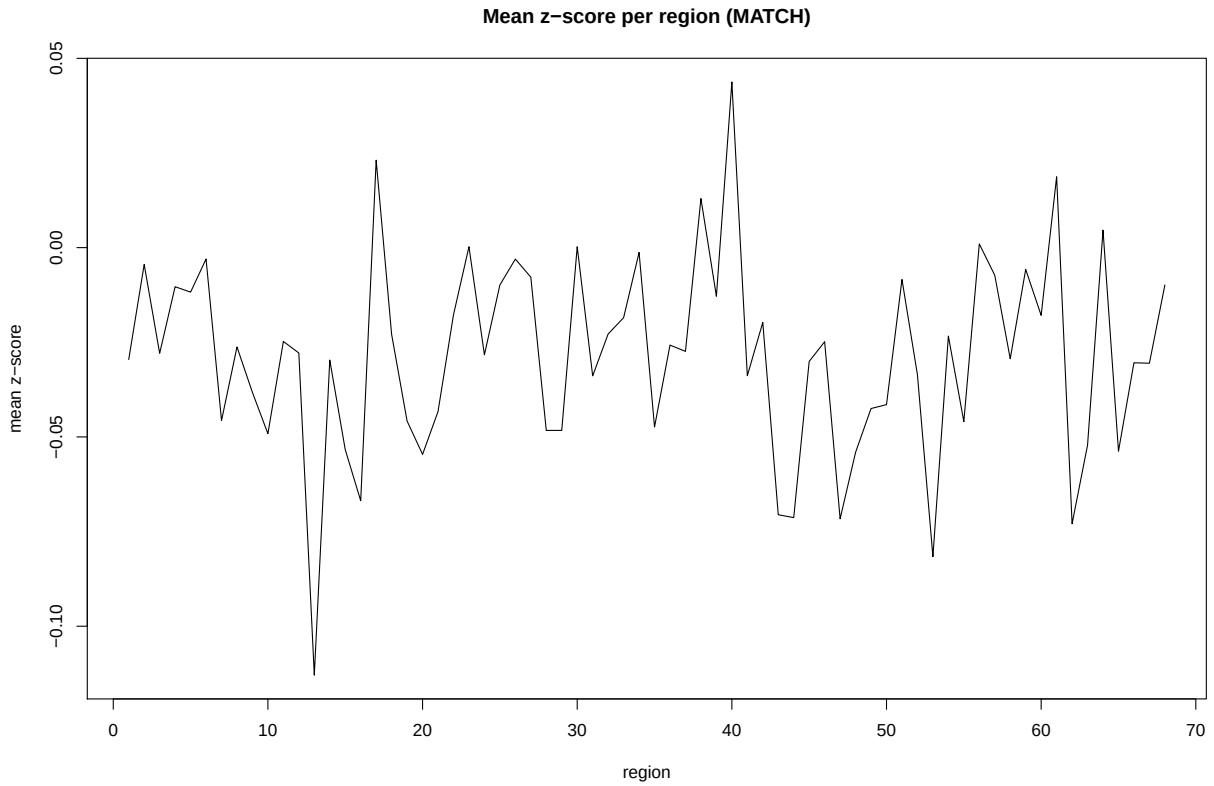

z−score distribution (NO MATCH)
density
0.0
0.1
0.2
0.3
0.4
0.5
−5.0
−2.5
0.0
2.5
5.0
z−score


Mean z−score per region (NO MATCH)
mean z−score
0.02
0.00
−0.02
−0.04
−0.06
−0.08
0
10
20
30
40
50
60
70
region

Match-it dataframe (longCombat):

```
Zs_match <- run_NormativeModel(df_lC_matched,
                               measure = "CT_freesurfer",
                               parc = "parc35",
                               match = "match",
                               harmonization = "lC")
```





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

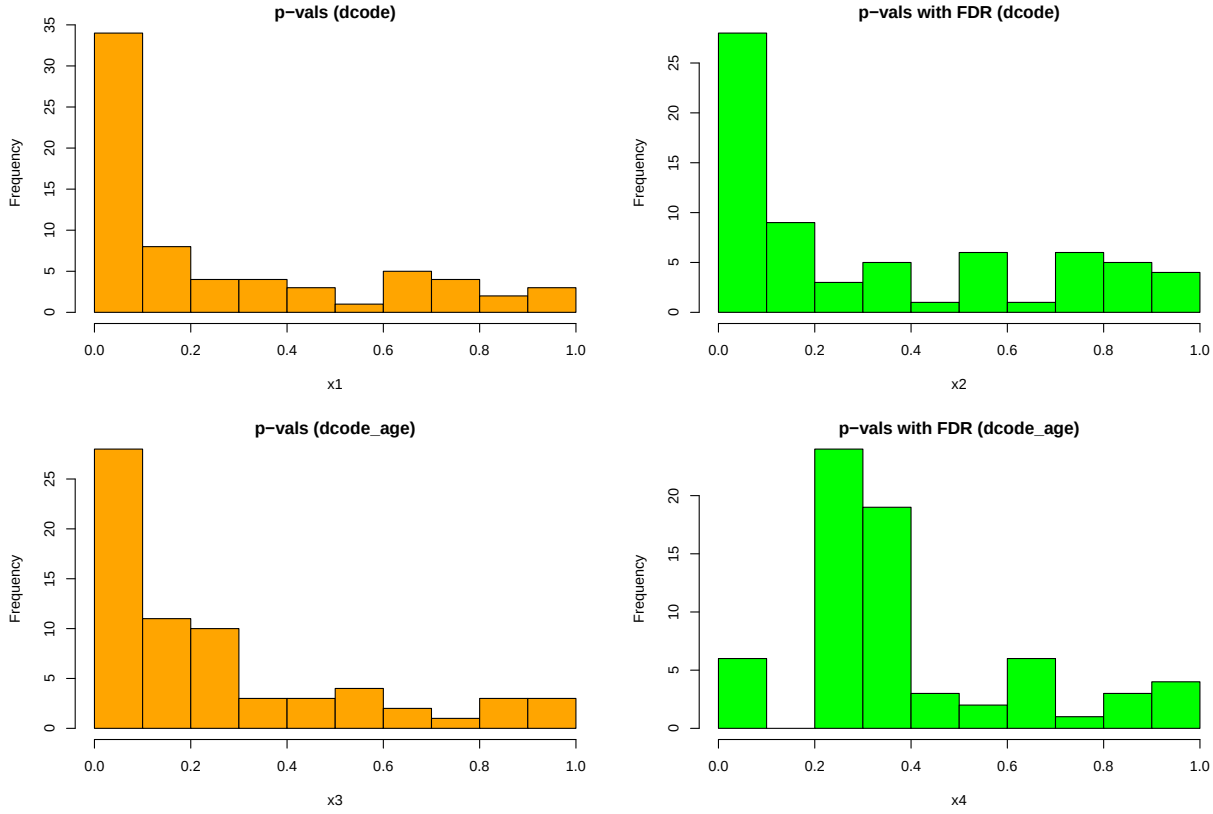
Calling the *run_AgeDiagnosisModel* function. In this case, we build the regression with the data from healthy controls as well as patients. Next we exclude the deviants from the model and run again the function

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

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

p_val_FDR <- Apply_FDR_Correction(p_val)
```

```
## Without FDR correction:
## Variable dcode has statistical significance for  28 / 68 regions
## Variable dcode_age has statistical significance for  14 / 68 regions

##      reg       scode          age        euler      dcode   dcode_age
## [1,]   0 0.902627774 1.376441e-10 1.585103e-03 0.05520999 0.311477261
## [2,]   0 0.002974282 4.919699e-02 6.363141e-02 0.69065034 0.000958305
## [3,]   0 0.011063046 0.000000e+00 1.653418e-02 0.14065095 0.242002275
## [4,]   0 0.656891364 1.284373e-06 3.036019e-05 0.47491087 0.813085773

## With FDR correction:
## Variable dcode has statistical significance for  20 / 68 regions
## Variable dcode_age has statistical significance for  3 / 68 regions
```

```
##      reg      scode          age        euler     dcode  dcode_age
## [1,]   0 0.96138159 2.752882e-10 0.0063404134 0.1294579 0.42360908
## [2,]   0 0.02596661 5.395799e-02 0.1395785747 0.7960038 0.03738268
## [3,]   0 0.05645537 0.000000e+00 0.0505620137 0.2452375 0.34434681
## [4,]   0 0.82496212 1.940831e-06 0.0002195698 0.6093196 0.87761639
```



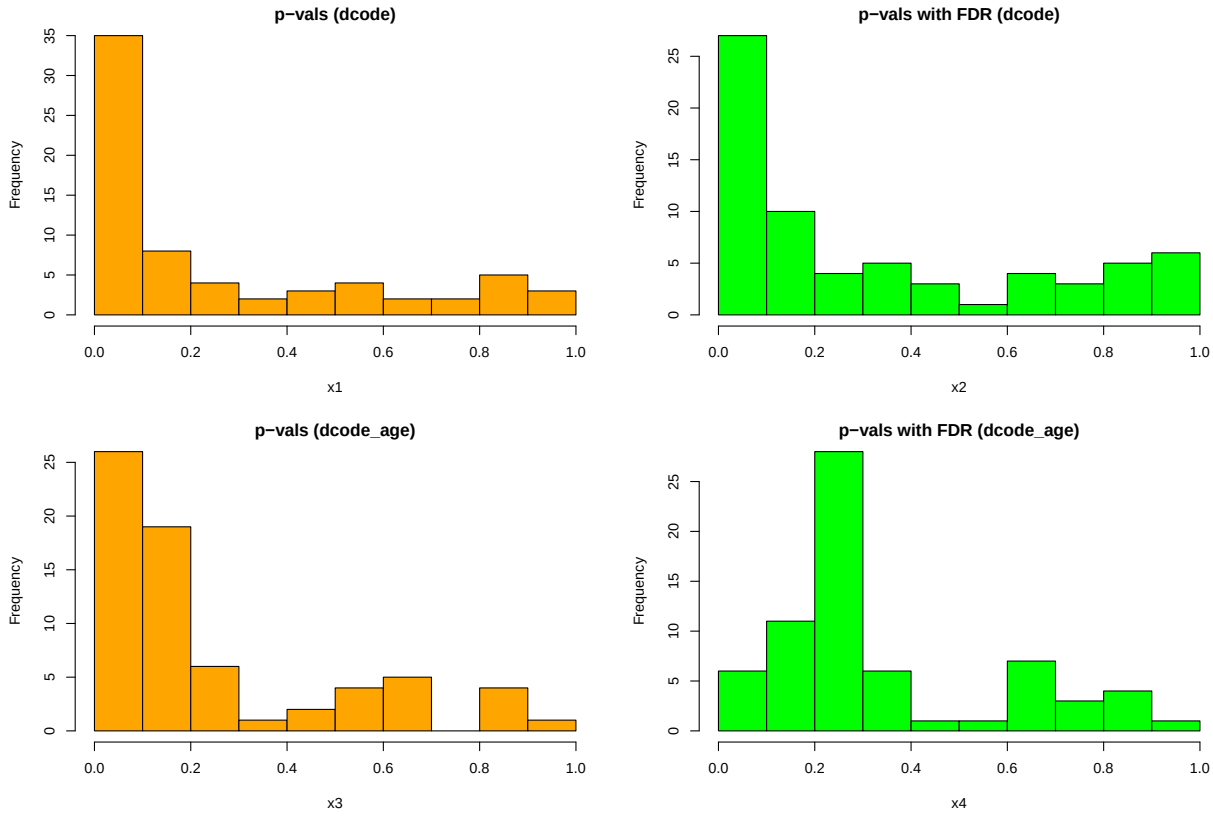
Match-it dataframe (Age*Diagnosis) AFTER EXCLUDING DEVIANTS

```
p_val <- run_AgeDiagnosisModel(df_1C_matched,
                               measure = "CT_freesurfer",
                               Z = Zs_match,
                               exclude_deviants = T)

p_val_FDR <- Apply_FDR_Correction(p_val)
```

```
## Without FDR correction:
## Variable dcode has statistical significance for  28 / 68 regions
## Variable dcode_age has statistical significance for  17 / 68 regions
```

```
##      reg       scode          age        euler      dcode    dcode_age
## [1,]   0 0.897895774 9.864110e-12 8.442243e-05 0.06502851 0.1316012362
## [2,]   0 0.003033047 6.971218e-02 1.151092e-01 0.87177220 0.0008659586
## [3,]   0 0.011344863 0.000000e+00 1.293067e-02 0.12199549 0.1847867203
## [4,]   0 0.796635433 5.294333e-06 4.887556e-05 0.52064019 0.9225434551
```

```
## With FDR correction:
## Variable dcode has statistical significance for  22 / 68 regions
## Variable dcode_age has statistical significance for  3 / 68 regions

##      reg      scode          age        euler     dcode  dcode_age
## [1,]   0 0.96915734 2.235865e-11 0.0004415942 0.1473980 0.26320247
## [2,]   0 0.03033302 7.524489e-02 0.2206451814 0.9335941 0.03076526
## [3,]   0 0.05558303 0.000000e+00 0.0399675197 0.2167181 0.29551921
## [4,]   0 0.90303250 7.826406e-06 0.0002838175 0.6679912 0.92254346
```



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	31.688	31.348	10.744
# samples dev (%)	795 (2.51%)	696 (2.22%)	275 (2.56%)
# controls samples	20.264	19.924	7.412
# controls dev (%)	450 (2.22%)	335 (1.68%)	143 (1.93%)
# patients samples	11.424	11.424	3.332
# patients dev (%)	345 (3.02%)	361 (3.16%)	132 (3.96%)

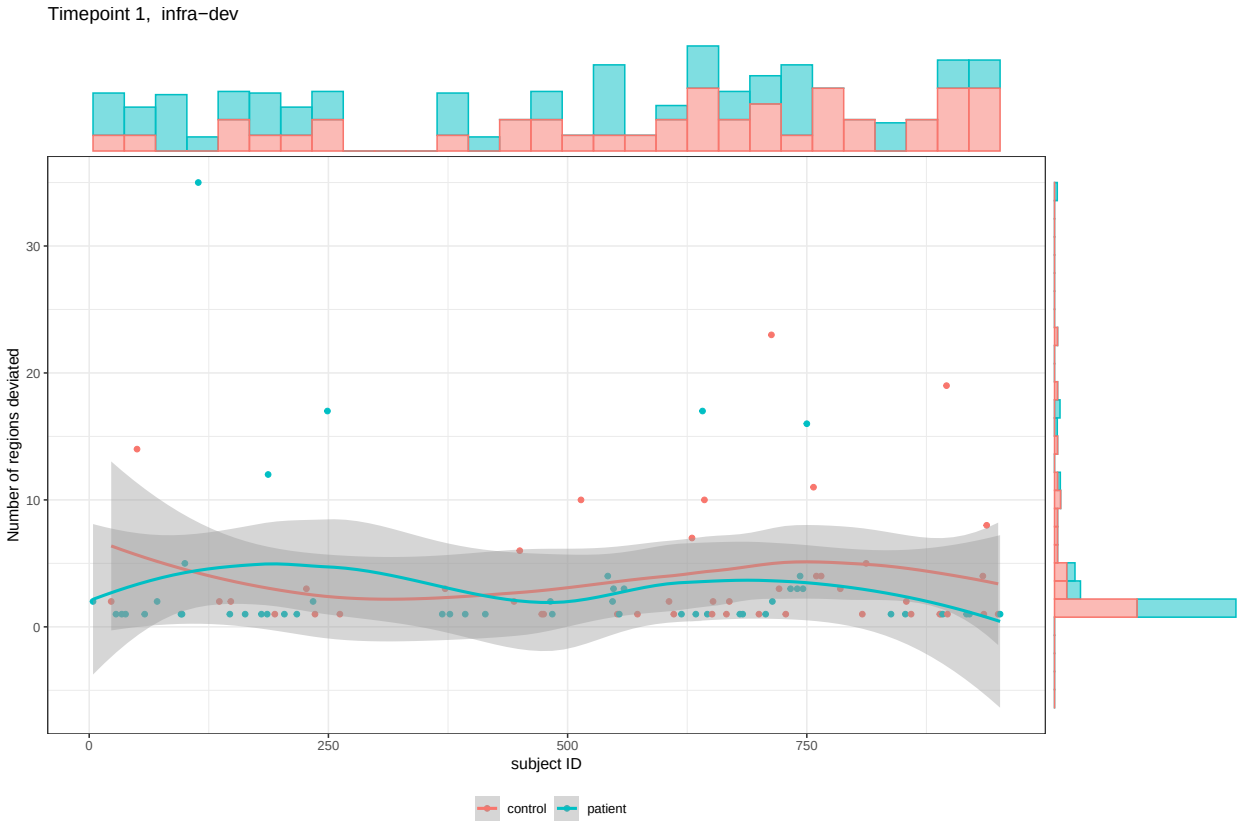
MATCHED	timepoint 1	timepoint 2	timepoint 3
# total samples	22.916	22.304	6.664
# samples dev (%)	617 (2.69%)	519 (2.33%)	187 (2.81%)
# controls samples	11.492	11.152	3.332
# controls dev (%)	269 (2.34%)	176 (1.58%)	71 (2.13%)
# patients samples	11.424	11.152	3.332
# patients dev (%)	348 (3.05%)	343 (3.08%)	116 (3.48%)

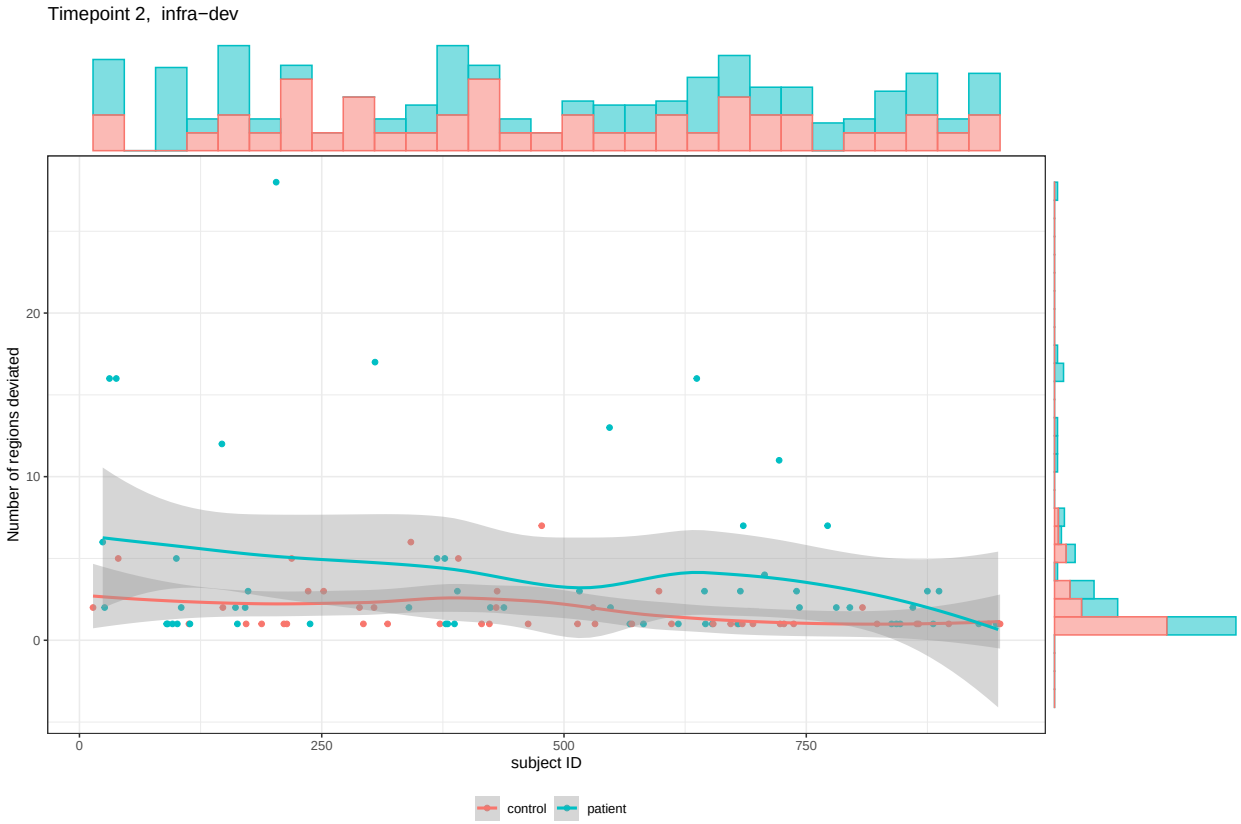
In the following, we chose to work with the matched dataset and the Zs derived from its lme model, providing a better statistical support for the analysis.

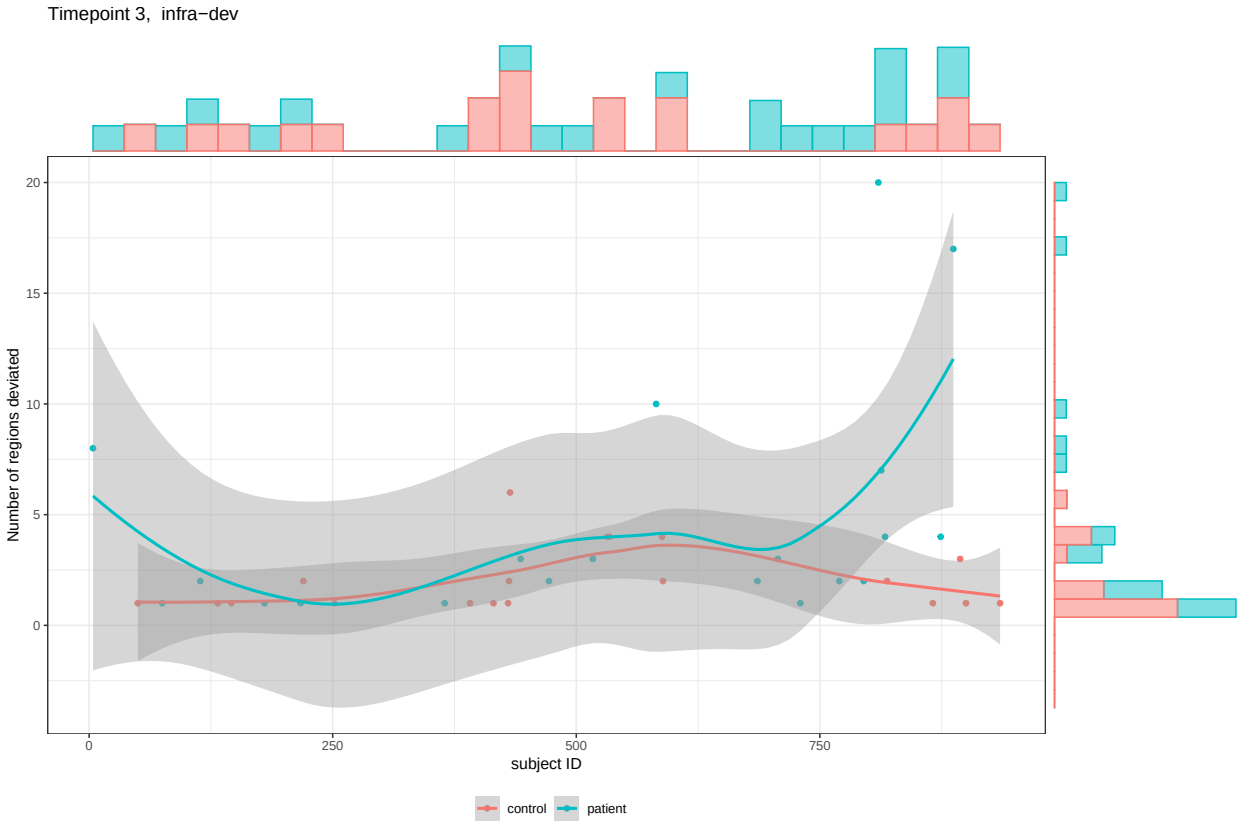
We computed the number of deviant samples ($Z < -1.96$ and $Z > 1.96$):

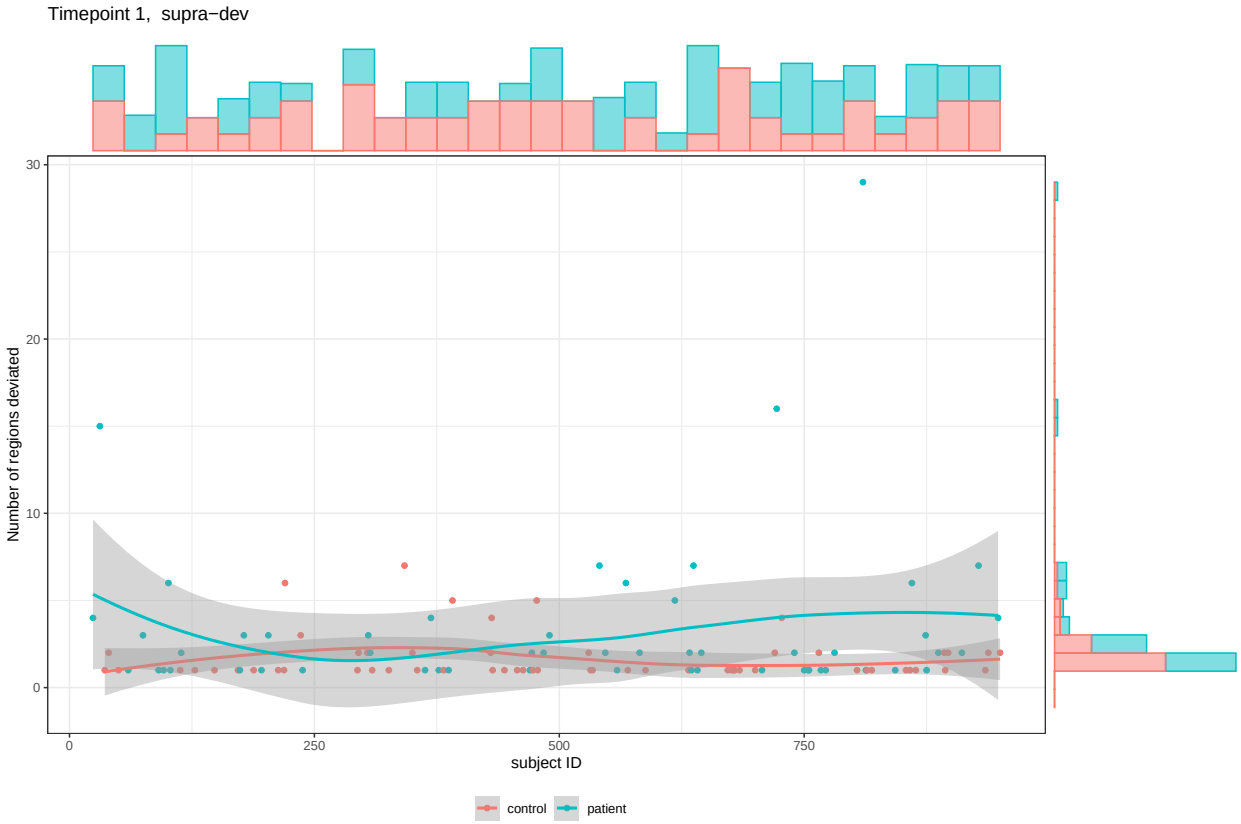
In terms of regions deviated:

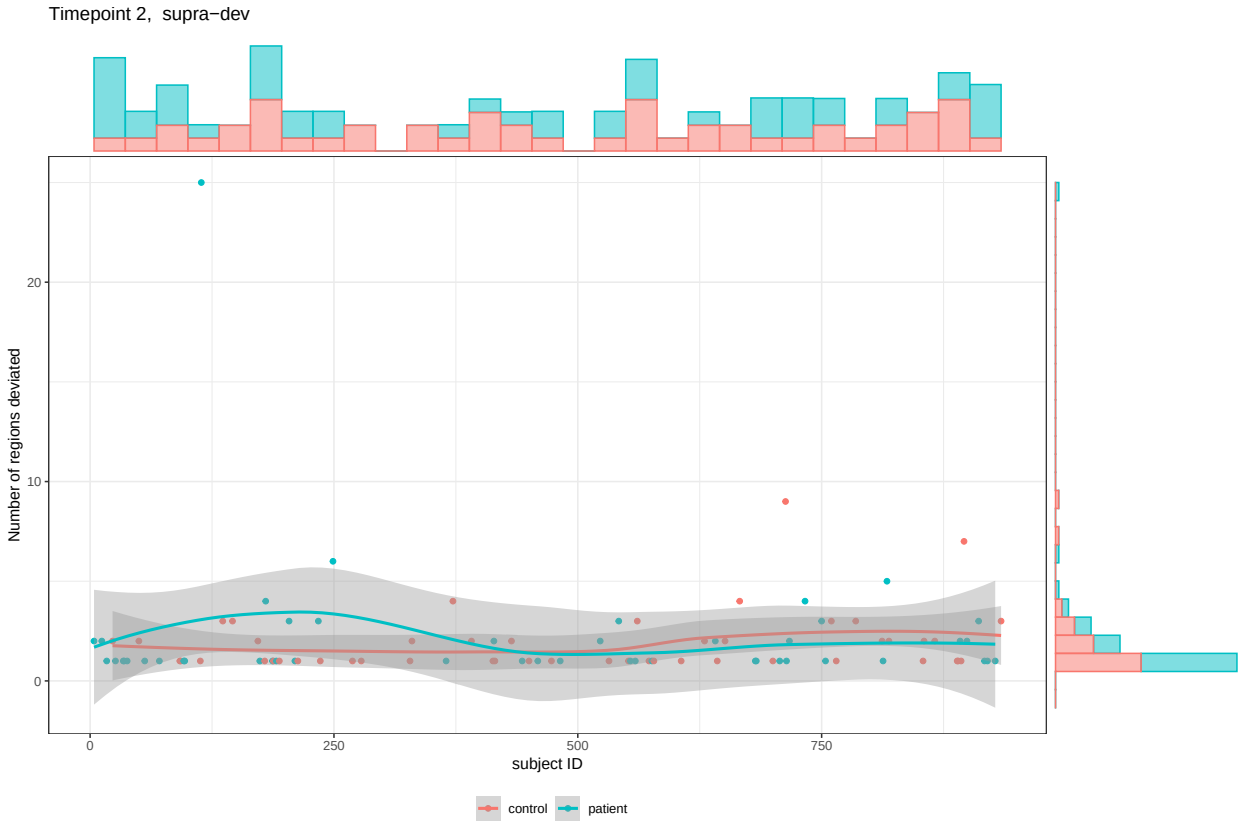
	Infra-normal deviants	Non deviants	Supra-normal deviants	Deviants
tp 1	339 (1.48%)	22.299 (97.31%)	278 (1.21%)	617 (2.69%)
controls	172 (50.74%)	11.223 (50.33%)	97 (34.89%)	269 (43.60%)
patients	167 (49.26%)	11.076 (49.67%)	181 (65.11%)	348 (56.40%)
tp 2	322 (1.44%)	21.785 (97.67%)	197 (0.88%)	519 (2.33%)
controls	83 (25.78%)	10.976 (50.38%)	93 (47.21%)	176 (33.91%)
patients	239 (74.22%)	10.809 (49.62%)	104 (52.79%)	343 (66.09%)
tp 3	133 (2.00%)	6.477 (97.19%)	54 (0.81%)	187 (2.81%)
controls	39 (29.32%)	3.261 (50.35%)	32 (59.26%)	71 (37.97%)
patients	94 (70.68%)	3.216 (49.65%)	22 (40.74%)	116 (62.03%)
total	794 (1.53%)	50.561 (97.45%)	529 (1.02%)	1.323 (2.55%)

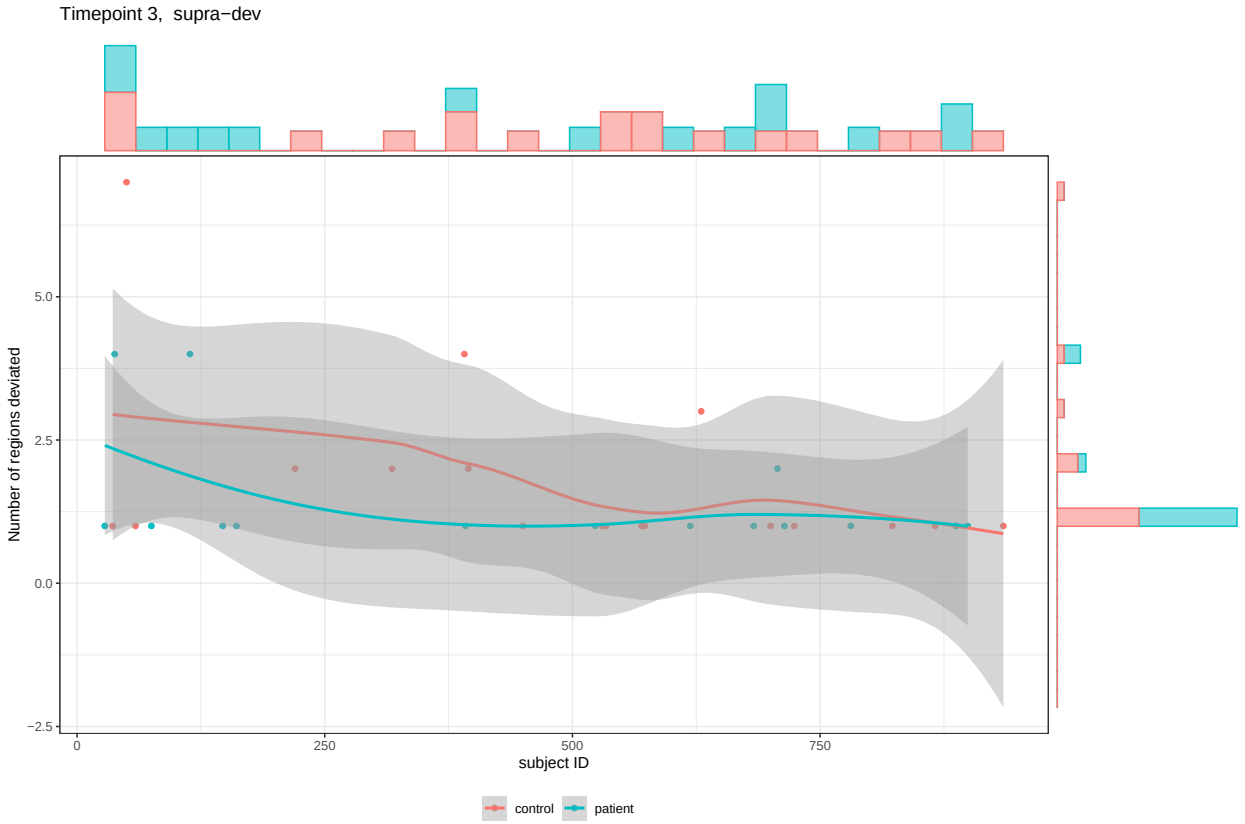

Timepoint 1, infra−dev
Number of regions deviated
subject ID
0
250
500
750
0
10
20
30
control
patient


Timepoint 2, infra−dev
Number of regions deviated
subject ID
0
250
500
750
0
10
20
control
patient


Timepoint 3, infra−dev
Number of regions deviated
subject ID
control
patient


Timepoint 1, supra−dev
Number of regions deviated
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 5 subjects (1.48% of total subjects) fulfill that condition for any region.

	Condition
rh_cuneus_CT_freesurfer	5 (1.48%)
lh_rostralmiddlefrontal_CT_freesurfer	4 (1.19%)
lh_temporalpole_CT_freesurfer	4 (1.19%)

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 20 subjects (5.93% of total subjects) fulfill that condition for any region.

	Condition
lh_lateraloccipital_CT_freesurfer	20 (5.93%)
rh_temporalpole_CT_freesurfer	20 (5.93%)
lh_posteriorcingulate_CT_freesurfer	18 (5.34%)

---

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

- Timepoint 1: No more than 6.548% of patients deviated from the normative range for any single cortical region.
- Timepoint 2: No more than 6.707% of patients deviated from the normative range for any single cortical region.
- Timepoint 3: No more than 10.204% 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 **rh_fusiform_CT_freesurfer** cortices, although only 1.786% of patients showed significant deviations in these regions.
- Timepoint 2: Infra-normal deviations in CT of subjects were most commonly located in **rh_temporalpole_CT_freesurfer** cortices, although only 4.268% of patients showed significant deviations in these regions.
- Timepoint 3: Infra-normal deviations in CT of subjects were most commonly located in **lh_fusiform_CT_freesurfer** cortices, although only 8.163% 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__lateraloccipital__CT__freesurfer** regions, 2.967% of individuals.
- Timepoint 2: Supra-normal deviations in CT were most common in the **lh__lateraloccipital__CT__freesurfer** regions, 2.134% of individuals.
- Timepoint 3: Supra-normal deviations in CT were most common in the **lh__parahippocampal__CT__freesurfer** regions, 3.061% 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 29.167% of patients, whereas this was the case for 26.036% of healthy individuals.
- Timepoint 2: Infra-normal deviations for at least one region were evident in 35.366% of patients, whereas this was the case for 27.439% of healthy individuals.
- Timepoint 3: Infra-normal deviations for at least one region were evident in 40.816% of patients, whereas this was the case for only 38.776% 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 32.143% of patients and 34.32% of healthy individuals.
- Timepoint 2: Supra-normal deviations for at least one region were evident in 28.659% of patients and 29.878% of healthy individuals.
- Timepoint 3: Supra-normal deviations for at least one region were evident in 30.612% of patients and 36.735% of healthy individuals.

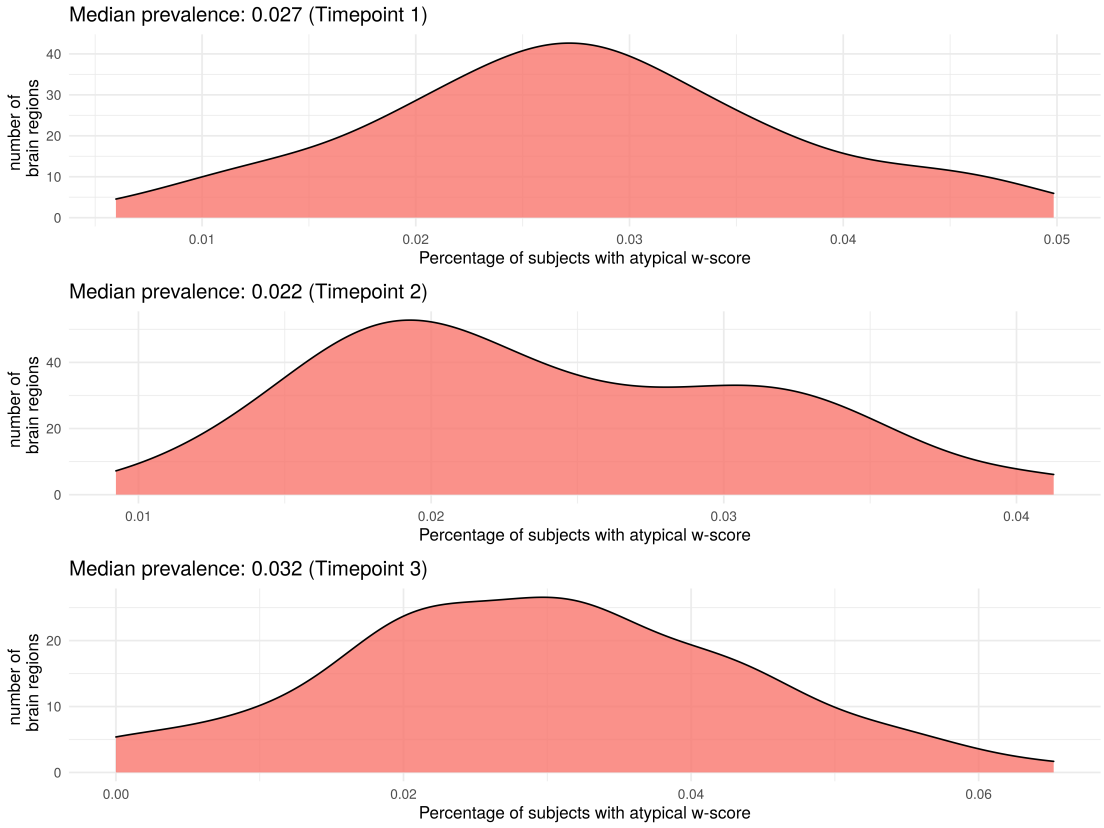
Percentage of deviant subjects for number of regions

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 ($timepoint = 1, 2, 3$).

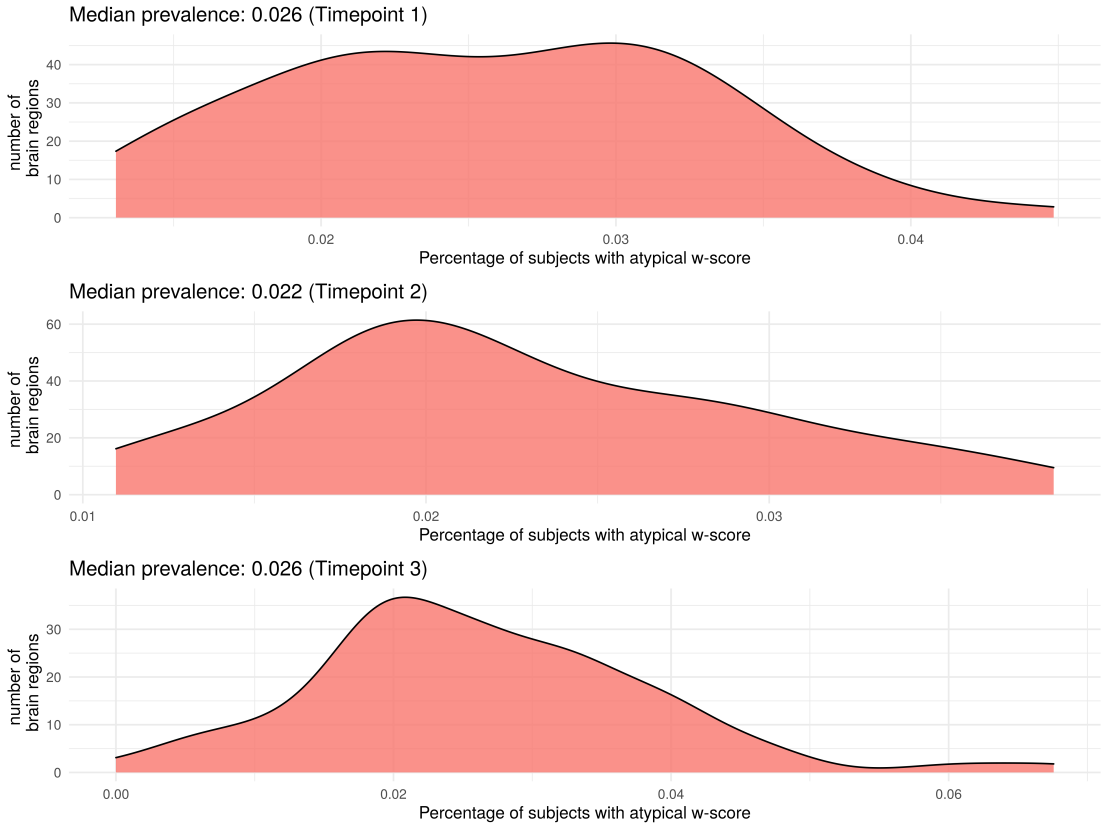
```
getStatistics(Zs= Zs_match,  lab= "lC_match_parc35", parc= "parc35")
getStatistics(Zs= Zs_NOmatch, lab= "lC_NOmatch_parc35", parc= "parc35")
```

Show the results from the global ratios obtained:

Match-it dataframe (longCombat):



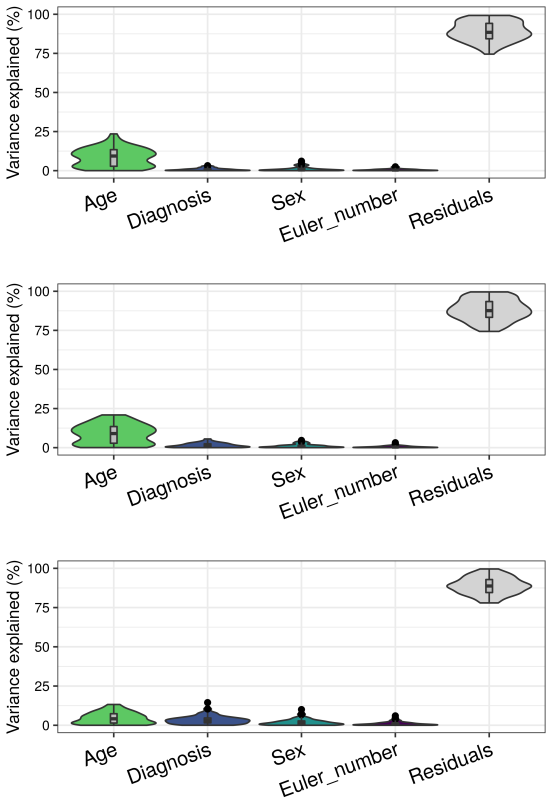
NO Match-it dataframe (longCombat):



Variance contribution across measures

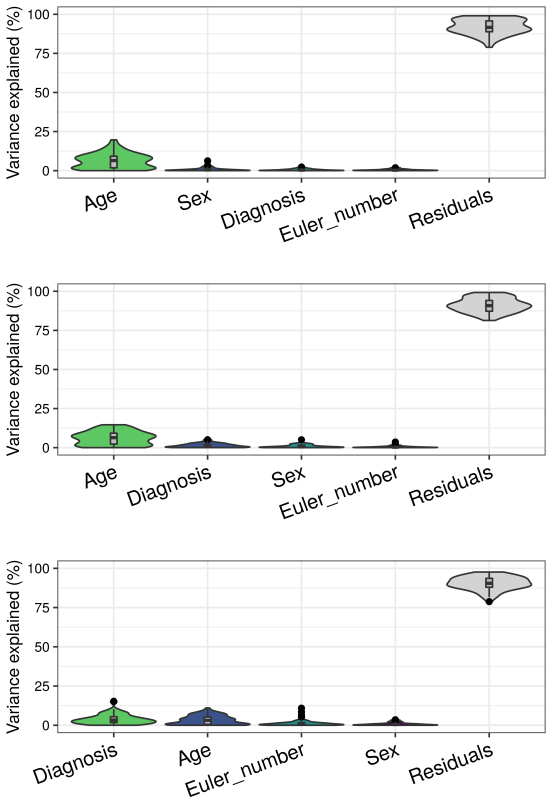
NO Match-it dataframe (longCombat):

```
variancePartition(df = df_lC_NO_matched,
                  measure = "CT_freesurfer",
                  lab = "lC_NO_matched_parc35",
                  par = "parc35")
```



Match-it dataframe (longCombat):

```
variancePartition(df = df_lC_matched,
                  measure = "CT_freesurfer",
                  lab = "lC_matched_parc35",
                  par = "parc35")
```


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