Nonbinaryanalyses_Ch4

V_warrier

28/03/2019

```
library(data.table)
library(dplyr)
```

```
## Warning: replacing previous import by 'rlang::dots_n' when loading 'dplyr'
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(grid)
library(epitools)
data = read.csv("data- originalchannel4 bigdata.csv", header = T, fill = T) #758916 obs
cat("the total number of IDs in this file is", nrow(data))
```

```
## the total number of IDs in this file is 758916
```

Section 1: Data cleaning and generating variables

data cleanup and recoding variables

```
#remove_repeats
data2 = subset(data, repeat. == 0) #695166

cat("the number of non-repeated participants is", nrow(data2))
```

```
## the number of non-repeated participants is 695166
```

```
#keep males of females
data2 = subset(data2, sex < 4)#672279

cat("\n")
```

```
cat("Of this, the number of participants who are either male or female is",  nrow(data2))
```

```
## Of this, the number of participants who are either male or female is 676132
```

```
##keep age bound

data2 = subset(data2, age > 15 & age < 90)

cat("\n")
```

```
cat("Of this, the number participants after all QC is ",  nrow(data2))
```

```
## Of this, the number participants after all QC is  675360
```

Recoding and defining variables

```
#recode AQ
data2[,c(51,57,58,60)] <- lapply(data2[,c(51,57,58,60)], function(x)
  recode(x,"1" = 1, "2" = 1, "3" = 0, "4" =0 ))

data2[,c(52,53,54,55,56,59)] <- lapply(data2[,c(52,53,54,55,56,59)], function(x)
   recode(x,"1" = 0, "2" = 0, "3" = 1, "4" =1))

data2$AQ_full = data2$AQ_1 + data2$AQ_2 + data2$AQ_3 + data2$AQ_4 + data2$AQ_5 + data2$AQ_6 + data2$AQ_

data2$AQ_Z = scale(data2$AQ_full, center = TRUE, scale = TRUE)


#Recode EQ
data2[,c(31, 32, 34, 36, 39)] <- lapply(data2[,c(31, 32, 34, 36, 39)], function(x)
  recode(x,"1" = 2, "2" = 1, "3" = 0, "4" =0 ))

data2[,c(33, 35, 37, 38, 40 )] <- lapply(data2[,c(33, 35, 37, 38, 40)], function(x)
  recode(x,"1" = 0, "2" = 0, "3" = 1, "4" =2))

data2$EQ_full = data2$EQ_1 + data2$EQ_2 + data2$EQ_3 + data2$EQ_4 + data2$EQ_5 + data2$EQ_6 + data2$EQ_

data2$EQ_Z = scale(data2$EQ_full, center = TRUE, scale = TRUE)


#Recode SQ
data2[,c(41,43, 44, 46, 47, 49, 50)] <- lapply(data2[,c(41,43, 44, 46, 47, 49, 50)], function(x)
  recode(x,"1" = 2, "2" = 1, "3" = 0, "4" =0 ))

data2[,c(42, 45, 48 )] <- lapply(data2[,c(42, 45, 48)], function(x)
  recode(x,"1" = 0, "2" = 0, "3" = 1, "4" =2))

data2$SQ_full = data2$SQR_1 + data2$SQR_2 + data2$SQR_3 + data2$SQR_4 + data2$SQR_5 + data2$SQR_6 + data

data2$SQ_Z = scale(data2$SQ_full, center = TRUE, scale = TRUE)


#Recode SPQ

data2[,c(21:30)] <- lapply(data2[,c(21:30)], function(x)
  recode(x,"1" = 3, "2" = 2, "3" = 1, "4" =0 ))

data2$SPQ_full = data2$SPQ_1 + data2$SPQ_2 + data2$SPQ_3 + data2$SPQ_4 + data2$SPQ_5 + data2$SPQ_6 + da


data2$SPQ_Z = scale(data2$SPQ_full, center = TRUE, scale = TRUE)

data2 = data2[!is.na(data2$AQ_full),]


#Define cases
```

```r
#define cases based on all different options
data2$autism = ifelse(data2$diagnosis_0 == "2" | data2$diagnosis_1 == "2"| data2$diagnosis_3 == "2" | da

data2$autism[is.na(data2$autism)] <- 0


data2$ADHD = ifelse(data2$diagnosis_0 == "1" | data2$diagnosis_1 == "1"| data2$diagnosis_3 == "1" | data

data2$ADHD[is.na(data2$ADHD)] <- 0

data2$bipolar = ifelse(data2$diagnosis_0 == "3" | data2$diagnosis_1 == "3"| data2$diagnosis_3 == "3" | d

data2$bipolar[is.na(data2$bipolar)] <- 0


data2$depression = ifelse(data2$diagnosis_0 == "4" | data2$diagnosis_1 == "4"| data2$diagnosis_3 == "4"

data2$depression[is.na(data2$depression)] <- 0


data2$ld = ifelse(data2$diagnosis_0 == "5" | data2$diagnosis_1 == "5"| data2$diagnosis_3 == "5" | data2

data2$ld[is.na(data2$ld)] <- 0


data2$ocd = ifelse(data2$diagnosis_0 == "6" | data2$diagnosis_1 == "6"| data2$diagnosis_3 == "6" | data1

data2$ocd[is.na(data2$ocd)] <- 0


data2$scz = ifelse(data2$diagnosis_0 == "7" | data2$diagnosis_1 == "7"| data2$diagnosis_3 == "7" | data1

data2$scz[is.na(data2$scz)] <- 0


#count the number of cases


a =  nrow(subset(data2, autism == 1))

cat("the number of autistic individuls are", a)
```

```
## the number of autistic individuls are 37545
```

```r
a =  nrow(subset(data2, ADHD == 1))

cat("the number of ADHD individuls are", a)
```

```
## the number of ADHD individuls are 25864
```

```r
a =  nrow(subset(data2, bipolar == 1))

cat("the number of bipolar individuls are", a)
```

```
## the number of bipolar individuls are 11954
```

```r
a =  nrow(subset(data2, depression == 1))

cat("the number of depression individuls are", a)
```

```
## the number of depression individuls are 159297
```

```r
a =  nrow(subset(data2, ld == 1))

cat("the number of ld individuls are", a)
```

```
## the number of ld individuls are 24077
```

```r
a =  nrow(subset(data2, scz == 1))

cat("the number of scz individuls are", a)
```

```
## the number of scz individuls are 1812
```

```r
cat("\n")
```

```r
## Education
data2$education = ifelse(data2$education == "5", 0, data2$education)
data2$binary = ifelse(data2$sex < 3, "binary", "nonbinary")
data2$sex = as.factor(data2$sex)
data3 = subset(data2, autism == "0")
```

Generating D-scores and brain types

```
#This is Wheelwright's method: https://www.ncbi.nlm.nih.gov/pubmed/16473340

controls = subset(data2, autism == "0")
cases = subset(data2, autism == "1")

meanSQ = mean(controls$SQ_full)
meanEQ = mean(controls$EQ_full)

data2$SQ_full_standardized_w = (data2$SQ_full - meanSQ)/20
data2$EQ_full_standardized_w = (data2$EQ_full - meanEQ)/20
data2$wheelwrightD = data2$SQ_full_standardized_w - data2$EQ_full_standardized_w

data2$dpercentile = ntile(data2$wheelwrightD, 100)

data2$braintype = ifelse(data2$dpercentile < 2.5, "1",
                 ifelse(between(data2$dpercentile, 2.499, 35), "2",
                    ifelse(between(data2$dpercentile, 34.99, 65), "3",
                           ifelse(between(data2$dpercentile, 64.99 , 97.5), "4", "5"))))

controls = subset(data2, autism == "0") #redefining cases and controls again so downstream analysis can
cases = subset(data2, autism == "1")
```

Chisquare case-control

```
control_males = subset(controls, sex == "1")
control_females = subset(controls, sex == "2")
control_nonbinary = subset(controls, sex == "3")

autism_males = subset(cases, sex == "1")
autism_females = subset(cases, sex == "2")
autism_nonbinary = subset(cases, sex == "3")

sex_3_way = matrix(c(nrow(control_males), nrow(autism_males), nrow(control_females), nrow(autism_females

colnames(sex_3_way) = c("Controls", "Cases")

rownames(sex_3_way) = c("Males", "Females", "Nonbinary")

chisq.test(sex_3_way)
```

```
##
##  Pearson's Chi-squared test
##
## data:  sex_3_way
## X-squared = 4355.2, df = 2, p-value < 2.2e-16
```

```
oddsratio(sex_3_way)
```

```
## $data
##           Controls Cases  Total
## Males       241355 18188 259543
## Females     393600 18460 412060
## Nonbinary     2857   897   3754
## Total       637812 37545 675357
##
## $measure
##                           NA
## odds ratio with 95% C.I.  estimate     lower     upper
##                 Males     1.0000000        NA        NA
##                 Females   0.6223753 0.6093699 0.6356417
##                 Nonbinary 4.1667977 3.8579952 4.4960510
##
## $p.value
##            NA
## two-sided   midp.exact  fisher.exact chi.square
##   Males             NA            NA         NA
##   Females            0  0.000000e+00          0
##   Nonbinary          0 1.383791e-226          0
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "median-unbiased estimate & mid-p exact CI"
```

```
#############


male_nonbinary = matrix(c(nrow(control_males), nrow(autism_males), nrow(control_nonbinary), nrow(autism_

colnames(male_nonbinary) = c("Controls", "Cases")

rownames(male_nonbinary) = c("Males", "Nonbinary")

male_nonbinary
```

```
##           Controls Cases
## Males       241355 18188
## Nonbinary     2857   897
```

```
chisq.test(male_nonbinary)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  male_nonbinary
## X-squared = 1567.1, df = 1, p-value < 2.2e-16
```

```
oddsratio(male_nonbinary)
```

```
## $data
##           Controls Cases  Total
## Males       241355 18188 259543
## Nonbinary     2857   897   3754
## Total       244212 19085 263297
##
## $measure
##                          NA
## odds ratio with 95% C.I. estimate    lower    upper
##                Males     1.000000       NA       NA
##                Nonbinary 4.166798 3.857995 4.496051
##
## $p.value
##            NA
## two-sided   midp.exact  fisher.exact chi.square
##   Males             NA            NA         NA
##   Nonbinary          0 1.383791e-226          0
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "median-unbiased estimate & mid-p exact CI"
```

```
##########

female_nonbinary = matrix(c(nrow(control_females), nrow(autism_females), nrow(control_nonbinary), nrow(a

colnames(female_nonbinary) = c("Controls", "Cases")

rownames(female_nonbinary) = c("Females", "Nonbinary")
```

```
female_nonbinary
```

```
##           Controls Cases
## Females     393600 18460
## Nonbinary     2857   897
```

```
chisq.test(female_nonbinary)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  female_nonbinary
## X-squared = 3154.8, df = 1, p-value < 2.2e-16
```

```
oddsratio(female_nonbinary)
```

```
## $data
##           Controls Cases  Total
## Females     393600 18460 412060
## Nonbinary     2857   897   3754
## Total       396457 19357 415814
##
## $measure
##                          NA
## odds ratio with 95% C.I. estimate    lower   upper
##                Females   1.000000       NA      NA
##                Nonbinary 6.695063 6.199259 7.22356
##
## $p.value
##            NA
## two-sided   midp.exact fisher.exact chi.square
##   Females           NA           NA         NA
##   Nonbinary          0            0          0
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "median-unbiased estimate & mid-p exact CI"
```

```
##########

binary_nonbinary = matrix(c(nrow(control_females) + nrow(control_males), nrow(autism_females) + nrow(au

colnames(binary_nonbinary) = c("Controls", "Cases")

rownames(binary_nonbinary) = c("Binary", "Nonbinary")

binary_nonbinary
```

```
##           Controls Cases
## Binary      634955 36648
## Nonbinary     2857   897
```

```
chisq.test(binary_nonbinary)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  binary_nonbinary
## X-squared = 2413.7, df = 1, p-value < 2.2e-16
```

```
oddsratio(binary_nonbinary)
```

```
## $data
##           Controls Cases  Total
## Binary      634955 36648 671603
## Nonbinary     2857   897   3754
## Total       637812 37545 675357
##
## $measure
##                          NA
## odds ratio with 95% C.I. estimate    lower    upper
##                Binary    1.000000       NA       NA
##                Nonbinary 5.440371 5.040969 5.865217
##
## $p.value
##            NA
## two-sided   midp.exact  fisher.exact chi.square
##   Binary            NA            NA         NA
##   Nonbinary          0 1.303697e-305          0
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "median-unbiased estimate & mid-p exact CI"
```

```
summary(glm(autism ~ relevel(sex, ref = "3") + scale(age) + scale(education), data = data2, family = bi
```

```
##
## Call:
## glm(formula = autism ~ relevel(sex, ref = "3") + scale(age) +
##     scale(education), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1327  -0.3838  -0.3115  -0.2294   3.3837
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.511861   0.039555  -38.22   <2e-16 ***
## relevel(sex, ref = "3")1 -1.328363   0.040066  -33.16   <2e-16 ***
## relevel(sex, ref = "3")2 -1.636131   0.040071  -40.83   <2e-16 ***
## scale(age)                -0.509211   0.007302  -69.74   <2e-16 ***
## scale(education)          -0.307837   0.005408  -56.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 289950  on 675353  degrees of freedom
## Residual deviance: 276093  on 675349  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 276103
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(autism ~ as.factor(binary) + scale(age) + scale(education), data = data2, family = binomial)
```

```
##
## Call:
## glm(formula = autism ~ as.factor(binary) + scale(age) + scale(education),
##     family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1415  -0.3769  -0.3197  -0.2344   3.3828
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.020027   0.006286 -480.47   <2e-16 ***
## as.factor(binary)nonbinary  1.491292   0.039747   37.52   <2e-16 ***
## scale(age)                 -0.538953   0.007253  -74.30   <2e-16 ***
## scale(education)           -0.309932   0.005396  -57.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 289950  on 675353  degrees of freedom
## Residual deviance: 276876  on 675350  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 276884
##
## Number of Fisher Scoring iterations: 6
```

Section 2: Autistic traits

Sex differences in controls

```
#Basic statistics

# Sex differences - AQ
summary(aov(AQ_full ~sex, data = controls))
```

```
##                 Df  Sum Sq Mean Sq F value Pr(>F)
## sex              2   37514   18757    3709 <2e-16 ***
## Residuals   637809 3225664       5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Post-hoc t tests

t.test(control_males$AQ_full, control_nonbinary$AQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_males$AQ_full and control_nonbinary$AQ_full
## t = -38.251, df = 2905.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.023514 -1.826177
## sample estimates:
## mean of x mean of y
##  3.570429  5.495275
```

```
t.test(control_females$AQ_full, control_nonbinary$AQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_females$AQ_full and control_nonbinary$AQ_full
## t = -46.305, df = 2884.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.424504 -2.227514
## sample estimates:
## mean of x mean of y
##  3.169266  5.495275
```

```
t.test(control_females$AQ_full, control_males$AQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_females$AQ_full and control_males$AQ_full
## t = -68.682, df = 501320, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4126117 -0.3897157
## sample estimates:
## mean of x mean of y
##  3.169266  3.570429
```

```r
# Sex differences - EQ
summary(aov(EQ_full ~sex, data = controls))
```

```
##                 Df   Sum Sq Mean Sq F value Pr(>F)
## sex              2   573043  286522   12358 <2e-16 ***
## Residuals   637809 14787512      23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Post-hoc t tests

t.test(control_males$EQ_full, control_nonbinary$EQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_males$EQ_full and control_nonbinary$EQ_full
## t = 15.931, df = 2916.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.325391 1.697437
## sample estimates:
## mean of x mean of y
##  8.875432  7.364018
```

```r
t.test(control_females$EQ_full, control_nonbinary$EQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_females$EQ_full and control_nonbinary$EQ_full
## t = 36.251, df = 2894.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.247030 3.618374
## sample estimates:
## mean of x mean of y
## 10.796720  7.364018
```

```r
t.test(control_females$EQ_full, control_males$EQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_females$EQ_full and control_males$EQ_full
## t = 155.09, df = 518030, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.897007 1.945568
## sample estimates:
## mean of x mean of y
## 10.796720  8.875432
```

```r
# Sex differences - SQ
summary(aov(SQ_full ~sex, data = controls))
```

```
##                 Df   Sum Sq Mean Sq F value Pr(>F)
```

```
## sex              2   274647  137323    8588 <2e-16 ***
## Residuals   637809 10198360      16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Post-hoc t tests

t.test(control_males$SQ_full, control_nonbinary$SQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_males$SQ_full and control_nonbinary$SQ_full
## t = -26.433, df = 2909.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.511778 -2.164873
## sample estimates:
## mean of x mean of y
##  6.734478  9.072804
```

```
t.test(control_females$SQ_full, control_nonbinary$SQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_females$SQ_full and control_nonbinary$SQ_full
## t = -41.039, df = 2884.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.795517 -3.449369
## sample estimates:
## mean of x mean of y
##  5.450361  9.072804
```

```
t.test(control_females$SQ_full, control_males$SQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_females$SQ_full and control_males$SQ_full
## t = -122.11, df = 481110, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.304729 -1.263506
## sample estimates:
## mean of x mean of y
##  5.450361  6.734478
```

```
# Sex differences - SPQ
summary(aov(SPQ_full ~sex, data = controls))
```

```
##                 Df   Sum Sq Mean Sq F value Pr(>F)
## sex              2   125766   62883    1961 <2e-16 ***
## Residuals   637809 20451742      32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Post-hoc t tests

t.test(control_males$SPQ_full, control_nonbinary$SPQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_males$SPQ_full and control_nonbinary$SPQ_full
## t = -29.435, df = 2911.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.602510 -3.152524
## sample estimates:
## mean of x mean of y
##  13.99455  17.37207
```

```r
t.test(control_females$SPQ_full, control_nonbinary$SPQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_females$SPQ_full and control_nonbinary$SPQ_full
## t = -22.259, df = 2892.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.774738 -2.325471
## sample estimates:
## mean of x mean of y
##  14.82196  17.37207
```

```r
t.test(control_females$SPQ_full, control_males$SPQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  control_females$SPQ_full and control_males$SPQ_full
## t = 57.099, df = 526580, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7990105 0.8558141
## sample estimates:
## mean of x mean of y
##  14.82196  13.99455
```

```r
## Generate means and SDs

controls2 = controls[,c ("sex", "AQ_full", "EQ_full", "SQ_full", "SPQ_full" )]
controls2 = na.omit(controls2)
controls2_melt =  melt(controls2, id.vars=c("sex"))
ddply(controls2_melt, c("sex", "variable"), summarise,
      mean = mean(value), sd = sd(value),
      sem = sd(value)/sqrt(length(value)))
```

```
##    sex variable      mean       sd         sem
## 1    1  AQ_full  3.570429 2.278934 0.004638778
## 2    1  EQ_full  8.875432 4.756196 0.009681254
```

```
## 3    1  SQ_full  6.734478 4.180116 0.008508640
## 4    1 SPQ_full 13.994552 5.515901 0.011227636
## 5    2  AQ_full  3.169266 2.226789 0.003549372
## 6    2  EQ_full 10.796720 4.849119 0.007729213
## 7    2  SQ_full  5.450361 3.877523 0.006180545
## 8    2 SPQ_full 14.821964 5.747510 0.009161196
## 9    3  AQ_full  5.495275 2.678252 0.050106762
## 10   3  EQ_full  7.364018 5.044529 0.094376862
## 11   3  SQ_full  9.072804 4.706396 0.088050801
## 12   3 SPQ_full 17.372069 6.103894 0.114196260
```

Sex differences in cases

```
# Sex differences - AQ
summary(aov(AQ_full ~sex, data = cases))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## sex             2   6374    3187   438.1 <2e-16 ***
## Residuals   37542 273146       7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Post-hoc t tests

t.test(autism_males$AQ_full, autism_nonbinary$AQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_males$AQ_full and autism_nonbinary$AQ_full
## t = -30.086, df = 1005.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.666689 -2.340129
## sample estimates:
## mean of x mean of y
##  4.875632  7.379041
```

```
t.test(autism_females$AQ_full, autism_nonbinary$AQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_females$AQ_full and autism_nonbinary$AQ_full
## t = -32.587, df = 1011.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.878446 -2.551468
## sample estimates:
## mean of x mean of y
##  4.664085  7.379041
```

```
t.test(autism_females$AQ_full, autism_males$AQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_females$AQ_full and autism_males$AQ_full
## t = -7.4906, df = 36638, p-value = 7.008e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2669022 -0.1561934
## sample estimates:
## mean of x mean of y
##  4.664085  4.875632
```

```r
# Sex differences - EQ
summary(aov(EQ_full ~sex, data = cases))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## sex             2  25147   12573   532.3 <2e-16 ***
## Residuals   37542 886795      24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Post-hoc t tests

t.test(autism_males$EQ_full, autism_nonbinary$EQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_males$EQ_full and autism_nonbinary$EQ_full
## t = 17.043, df = 1011.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.169916 2.734623
## sample estimates:
## mean of x mean of y
##  6.920497  4.468227
```

```r
t.test(autism_females$EQ_full, autism_nonbinary$EQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_females$EQ_full and autism_nonbinary$EQ_full
## t = 26.293, df = 1026.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.513774 4.080541
## sample estimates:
## mean of x mean of y
##  8.265385  4.468227
```

```r
t.test(autism_females$EQ_full, autism_males$EQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_females$EQ_full and autism_males$EQ_full
## t = 26.415, df = 36550, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.245096 1.444680
## sample estimates:
## mean of x mean of y
##  8.265385  6.920497
```

```r
# Sex differences - SQ
summary(aov(SQ_full ~sex, data = cases))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
```

```
## sex             2  25063   12531   610.5 <2e-16 ***
## Residuals   37542 770560      21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Post-hoc t tests

t.test(autism_males$SQ_full, autism_nonbinary$SQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_males$SQ_full and autism_nonbinary$SQ_full
## t = -20.817, df = 962.95, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.166244 -3.448395
## sample estimates:
## mean of x mean of y
##  8.077854 11.885173
```

```
t.test(autism_females$SQ_full, autism_nonbinary$SQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_females$SQ_full and autism_nonbinary$SQ_full
## t = -26.251, df = 954.36, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.148551 -4.432304
## sample estimates:
## mean of x mean of y
##  7.094745 11.885173
```

```
t.test(autism_females$SQ_full, autism_males$SQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_females$SQ_full and autism_males$SQ_full
## t = -20.866, df = 36441, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0754565 -0.8907598
## sample estimates:
## mean of x mean of y
##  7.094745  8.077854
```

```
# Sex differences - SPQ
summary(aov(SPQ_full ~sex, data = cases))
```

```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## sex             2   25065   12532   322.2 <2e-16 ***
## Residuals   37542 1460251      39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Post-hoc t tests

t.test(autism_males$SPQ_full, autism_nonbinary$SPQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_males$SPQ_full and autism_nonbinary$SPQ_full
## t = -21.409, df = 967.92, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.583058 -4.645479
## sample estimates:
## mean of x mean of y
##  16.33055  21.44482
```

```r
t.test(autism_females$SPQ_full, autism_nonbinary$SPQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_females$SPQ_full and autism_nonbinary$SPQ_full
## t = -18.172, df = 964.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.805258 -3.868556
## sample estimates:
## mean of x mean of y
##  17.10791  21.44482
```

```r
t.test(autism_females$SPQ_full, autism_males$SPQ_full)
```

```
##
##  Welch Two Sample t-test
##
## data:  autism_females$SPQ_full and autism_males$SPQ_full
## t = 11.968, df = 36607, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6500477 0.9046750
## sample estimates:
## mean of x mean of y
##  17.10791  16.33055
```

```r
cases2 = cases[,c ("sex", "AQ_full", "EQ_full", "SQ_full", "SPQ_full" )]
cases2 = na.omit(cases2)
cases2_melt =  melt(cases2, id.vars=c("sex"))
ddply(cases2_melt, c("sex", "variable"), summarise,
      mean = mean(value), sd = sd(value),
      sem = sd(value)/sqrt(length(value)))
```

```
##    sex variable      mean       sd        sem
## 1    1  AQ_full  4.875632 2.662900 0.01974524
## 2    1  EQ_full  6.920497 4.710717 0.03492967
## 3    1  SQ_full  8.077854 4.642268 0.03442213
## 4    1 SPQ_full 16.330548 6.272512 0.04651029
```

```
## 5    2  AQ_full  4.664085 2.743430 0.02019194
## 6    2  EQ_full  8.265385 5.032808 0.03704201
## 7    2  SQ_full  7.094745 4.371088 0.03217168
## 8    2 SPQ_full 17.107909 6.160577 0.04534251
## 9    3  AQ_full  7.379041 2.420874 0.08083064
## 10   3  EQ_full  4.468227 4.180544 0.13958430
## 11   3  SQ_full 11.885173 5.379895 0.17962947
## 12   3 SPQ_full 21.444816 7.017637 0.23431208
```

Linear regression

```
summary(lm(scale(AQ_full) ~ relevel(sex, ref = "3") + as.character(autism) + education + age, data = da
```

```
##
## Call:
## lm(formula = scale(AQ_full) ~ relevel(sex, ref = "3") + as.character(autism) +
##     education + age, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0710 -0.7295 -0.1388  0.6524  3.1113
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.1789576  0.0164837  71.523   <2e-16 ***
## relevel(sex, ref = "3")1 -0.8659974  0.0161247 -53.706   <2e-16 ***
## relevel(sex, ref = "3")2 -1.0239831  0.0160970 -63.613   <2e-16 ***
## as.character(autism)1     0.5755212  0.0052547 109.526   <2e-16 ***
## education                -0.0897679  0.0013197 -68.019   <2e-16 ***
## age                      -0.0008745  0.0001001  -8.735   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9793 on 675348 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.04096,    Adjusted R-squared:  0.04095
## F-statistic:  5768 on 5 and 675348 DF,  p-value: < 2.2e-16
```

```
summary(lm(scale(EQ_full) ~ relevel(sex, ref = "3") + as.character(autism) + education + age, data = da
```

```
##
## Call:
## lm(formula = scale(EQ_full) ~ relevel(sex, ref = "3") + as.character(autism) +
##     education + age, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36040 -0.72165  0.01259  0.71770  2.92382
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -7.999e-01  1.634e-02 -48.968   <2e-16 ***
## relevel(sex, ref = "3")1  3.221e-01  1.598e-02  20.159   <2e-16 ***
## relevel(sex, ref = "3")2  6.927e-01  1.595e-02  43.423   <2e-16 ***
## as.character(autism)1    -4.206e-01  5.207e-03 -80.766   <2e-16 ***
## education                 9.817e-02  1.308e-03  75.058   <2e-16 ***
## age                       9.491e-04  9.922e-05   9.566   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9705 on 675348 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.05814,    Adjusted R-squared:  0.05813
```

```
## F-statistic:  8337 on 5 and 675348 DF,  p-value: < 2.2e-16
```

```
summary(lm(scale(SQ_full) ~ relevel(sex, ref = "3") + as.character(autism) + education + age, data = da
```

```
##
## Call:
## lm(formula = scale(SQ_full) ~ relevel(sex, ref = "3") + as.character(autism) +
##     education + age, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8330 -0.7602 -0.1495  0.5925  3.6622
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.6335807  0.0164898   38.42   <2e-16 ***
## relevel(sex, ref = "3")1 -0.6531061  0.0161307  -40.49   <2e-16 ***
## relevel(sex, ref = "3")2 -0.9846514  0.0161030  -61.15   <2e-16 ***
## as.character(autism)1     0.4028954  0.0052566   76.64   <2e-16 ***
## education                 0.0137508  0.0013202   10.41   <2e-16 ***
## age                       0.0055163  0.0001002   55.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9797 on 675348 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.04024,    Adjusted R-squared:  0.04023
## F-statistic:  5663 on 5 and 675348 DF,  p-value: < 2.2e-16
```

```
summary(lm(scale(SPQ_full) ~ relevel(sex, ref = "3") + as.character(autism) + education + age, data = da
```

```
##
## Call:
## lm(formula = scale(SPQ_full) ~ relevel(sex, ref = "3") + as.character(autism) +
##     education + age, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0826 -0.7059 -0.0262  0.6876  3.0274
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.5488396  0.0165653   33.13   <2e-16 ***
## relevel(sex, ref = "3")1 -0.6463591  0.0162045  -39.89   <2e-16 ***
## relevel(sex, ref = "3")2 -0.5288013  0.0161767  -32.69   <2e-16 ***
## as.character(autism)1     0.4183973  0.0052807   79.23   <2e-16 ***
## education                -0.0978069  0.0013263  -73.75   <2e-16 ***
## age                       0.0084765  0.0001006   84.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9842 on 675348 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.03144,    Adjusted R-squared:  0.03143
## F-statistic:  4384 on 5 and 675348 DF,  p-value: < 2.2e-16
```

Case-control AQ, EQ, SQ, and SPQ

```
males = subset(data2, sex == "1")
females = subset(data2, sex == "2")
others = subset(data2, sex == "3")

t.test(others$AQ_full ~ others$autism)
```

```
##
##  Welch Two Sample t-test
##
## data:  others$AQ_full by others$autism
## t = -19.808, df = 1640.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.070299 -1.697234
## sample estimates:
## mean in group 0 mean in group 1
##        5.495275        7.379041
```

```
t.test(others$EQ_full ~ others$autism)
```

```
##
##  Welch Two Sample t-test
##
## data:  others$EQ_full by others$autism
## t = 17.186, df = 1785.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.565321 3.226260
## sample estimates:
## mean in group 0 mean in group 1
##        7.364018        4.468227
```

```
t.test(others$SQ_full ~ others$autism)
```

```
##
##  Welch Two Sample t-test
##
## data:  others$SQ_full by others$autism
## t = -14.058, df = 1353.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.204809 -2.419929
## sample estimates:
## mean in group 0 mean in group 1
##        9.072804       11.885173
```

```
t.test(others$SPQ_full ~ others$autism)
```

```
##
##  Welch Two Sample t-test
##
## data:  others$SPQ_full by others$autism
## t = -15.625, df = 1348.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -4.584088 -3.561407
## sample estimates:
## mean in group 0 mean in group 1
##        17.37207        21.44482
```

```
data2melt = data2[,c ("sex", "autism", "AQ_full", "EQ_full", "SQ_full", "SPQ_full" )]
data2melt =  data2melt[!is.na(data2melt$AQ_full),]
data2melt =  melt(data2melt, id.vars=c("sex", "autism"))


head(data2melt)
```

```
##   sex autism variable value
## 1   2      1  AQ_full     6
## 2   1      1  AQ_full     6
## 3   1      0  AQ_full     6
## 4   1      0  AQ_full     3
## 5   3      1  AQ_full     6
## 6   2      1  AQ_full     1
```

```
ddply(data2melt, c("sex", "autism", "variable"), summarise,
      mean = mean(value), sd = sd(value),
      sem = sd(value)/sqrt(length(value)))
```

```
##    sex autism variable      mean       sd         sem
## 1    1      0  AQ_full  3.570429 2.278934 0.004638778
## 2    1      0  EQ_full  8.875432 4.756196 0.009681254
## 3    1      0  SQ_full  6.734478 4.180116 0.008508640
## 4    1      0 SPQ_full 13.994552 5.515901 0.011227636
## 5    1      1  AQ_full  4.875632 2.662900 0.019745239
## 6    1      1  EQ_full  6.920497 4.710717 0.034929673
## 7    1      1  SQ_full  8.077854 4.642268 0.034422132
## 8    1      1 SPQ_full 16.330548 6.272512 0.046510289
## 9    2      0  AQ_full  3.169266 2.226789 0.003549372
## 10   2      0  EQ_full 10.796720 4.849119 0.007729213
## 11   2      0  SQ_full  5.450361 3.877523 0.006180545
## 12   2      0 SPQ_full 14.821964 5.747510 0.009161196
## 13   2      1  AQ_full  4.664085 2.743430 0.020191941
## 14   2      1  EQ_full  8.265385 5.032808 0.037042007
## 15   2      1  SQ_full  7.094745 4.371088 0.032171680
## 16   2      1 SPQ_full 17.107909 6.160577 0.045342508
## 17   3      0  AQ_full  5.495275 2.678252 0.050106762
## 18   3      0  EQ_full  7.364018 5.044529 0.094376862
## 19   3      0  SQ_full  9.072804 4.706396 0.088050801
## 20   3      0 SPQ_full 17.372069 6.103894 0.114196260
## 21   3      1  AQ_full  7.379041 2.420874 0.080830635
## 22   3      1  EQ_full  4.468227 4.180544 0.139584297
## 23   3      1  SQ_full 11.885173 5.379895 0.179629475
## 24   3      1 SPQ_full 21.444816 7.017637 0.234312083
```

Brain type analysis

Others control percentage

```
#Sex distribution

controls = subset(data2, autism == "0")
other_controls = subset(controls, sex == "3")

cat("total number of nonbinary individuals is", nrow(other_controls))
```

```
## total number of nonbinary individuals is 2857
```

```
extremeE = subset(other_controls, braintype == "1")
E = subset(other_controls, braintype == "2")
B = subset(other_controls, braintype == "3")
S = subset(other_controls, braintype == "4")
extremeS = subset(other_controls, braintype == "5")

cat("percentage of nonbinary individuals with extreme E is", (nrow(extremeE)/nrow(other_controls))*100)
```

```
## percentage of nonbinary individuals with extreme E is 0.3850193
```

```
cat("percentage of nonbinary individuals with E is", (nrow(E)/nrow(other_controls))*100 )
```

```
## percentage of nonbinary individuals with E is 13.40567
```

```
cat("percentage of nonbinary individuals with B is", (nrow(B)/nrow(other_controls))*100)
```

```
## percentage of nonbinary individuals with B is 20.86104
```

```
cat("percentage of nonbinary individuals with S is", (nrow(S)/nrow(other_controls))*100)
```

```
## percentage of nonbinary individuals with S is 52.78264
```

```
cat("percentage of nonbinary individuals with Extreme S is", (nrow(extremeS)/nrow(other_controls))*100)
```

```
## percentage of nonbinary individuals with Extreme S is 12.56563
```

Brain type analysis

Other cases percentage

```
#Sex distribution

cases = subset(data2, autism == "1")
other_cases = subset(cases, sex == "3")

cat("total number of nonbinary individuals is", nrow(other_cases))
```

```
## total number of nonbinary individuals is 897
```

```
extremeE = subset(other_cases, braintype == "1")
E = subset(other_cases, braintype == "2")
B = subset(other_cases, braintype == "3")
S = subset(other_cases, braintype == "4")
extremeS = subset(other_cases, braintype == "5")

cat("percentage of nonbinary individuals with extreme E is", (nrow(extremeE)/nrow(other_cases))*100)
```

```
## percentage of nonbinary individuals with extreme E is 0.1114827
```

```
cat("percentage of nonbinary individuals with E is", (nrow(E)/nrow(other_cases))*100 )
```

```
## percentage of nonbinary individuals with E is 3.567447
```

```
cat("percentage of nonbinary individuals with B is", (nrow(B)/nrow(other_cases))*100)
```

```
## percentage of nonbinary individuals with B is 10.47938
```

```
cat("percentage of nonbinary individuals with S is", (nrow(S)/nrow(other_cases))*100)
```

```
## percentage of nonbinary individuals with S is 49.94426
```

```
cat("percentage of nonbinary individuals with Extreme S is", (nrow(extremeS)/nrow(other_cases))*100)
```

```
## percentage of nonbinary individuals with Extreme S is 35.89744
```

Other conditions

```
#ADHD

summary(glm(autism ~ binary + scale(age) + scale(education), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = autism ~ binary + scale(age) + scale(education),
##     family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1415  -0.3769  -0.3197  -0.2344   3.3828
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)      -3.020027   0.006286 -480.47   <2e-16 ***
## binarynonbinary   1.491292   0.039747   37.52   <2e-16 ***
## scale(age)       -0.538953   0.007253  -74.30   <2e-16 ***
## scale(education) -0.309932   0.005396  -57.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 289950  on 675353  degrees of freedom
## Residual deviance: 276876  on 675350  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 276884
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ADHD ~ as.factor(binary), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = ADHD ~ as.factor(binary), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6761  -0.2760  -0.2760  -0.2760   2.5640
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.248834   0.006434 -504.97   <2e-16 ***
## as.factor(binary)nonbinary  1.889296   0.040987   46.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 219481  on 675356  degrees of freedom
## Residual deviance: 218029  on 675355  degrees of freedom
## AIC: 218033
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ADHD ~ scale(age) + as.factor(education) + as.factor(binary), data = data2, family = binomia
```

```
##
## Call:
## glm(formula = ADHD ~ scale(age) + as.factor(education) + as.factor(binary),
##     family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9303  -0.3012  -0.2700  -0.2409   2.9850
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.02564    0.03245 -93.248  < 2e-16 ***
```

```
## scale(age)                 -0.24142    0.00738 -32.711  < 2e-16 ***
## as.factor(education)1       0.41657    0.03806  10.945  < 2e-16 ***
## as.factor(education)2      -0.20630    0.03375  -6.113 9.77e-10 ***
## as.factor(education)3      -0.40994    0.03420 -11.987  < 2e-16 ***
## as.factor(education)4      -0.41143    0.03814 -10.789  < 2e-16 ***
## as.factor(binary)nonbinary  1.73490    0.04157  41.735  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 219481  on 675353  degrees of freedom
## Residual deviance: 215525  on 675347  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 215539
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ADHD ~ age + as.factor(education) + as.factor(binary), data = data3, family = binomial))
```

```
##
## Call:
## glm(formula = ADHD ~ age + as.factor(education) + as.factor(binary),
##     family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7506  -0.2787  -0.2571  -0.2303   2.9831
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.794747   0.043297 -64.548  < 2e-16 ***
## age                        -0.017644   0.000654 -26.979  < 2e-16 ***
## as.factor(education)1       0.471333   0.046374  10.164  < 2e-16 ***
## as.factor(education)2      -0.045647   0.041316  -1.105    0.269
## as.factor(education)3      -0.210193   0.041649  -5.047 4.49e-07 ***
## as.factor(education)4      -0.231379   0.045311  -5.106 3.28e-07 ***
## as.factor(binary)nonbinary  1.482851   0.054369  27.274  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 186019  on 637808  degrees of freedom
## Residual deviance: 183946  on 637802  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 183960
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ADHD ~ relevel(sex, ref = "3"), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = ADHD ~ relevel(sex, ref = "3"), family = binomial,
```

```
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6761  -0.3240  -0.2413  -0.2413   2.6650
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.35954    0.04048  -33.59   <2e-16 ***
## relevel(sex, ref = "3")1 -1.56108    0.04145  -37.66   <2e-16 ***
## relevel(sex, ref = "3")2 -2.16247    0.04154  -52.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 219481  on 675356  degrees of freedom
## Residual deviance: 215862  on 675354  degrees of freedom
## AIC: 215868
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ADHD ~ scale(age) + as.factor(education) + relevel(sex, ref = "3"), data = data2, family = b
```

```
##
## Call:
## glm(formula = ADHD ~ scale(age) + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9140  -0.3081  -0.2589  -0.2270   2.9939
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.265658   0.051611 -24.523  < 2e-16 ***
## scale(age)                -0.195703   0.007415 -26.392  < 2e-16 ***
## as.factor(education)1      0.397590   0.038146  10.423  < 2e-16 ***
## as.factor(education)2     -0.212184   0.033805  -6.277 3.46e-10 ***
## as.factor(education)3     -0.413193   0.034261 -12.060  < 2e-16 ***
## as.factor(education)4     -0.394866   0.038215 -10.333  < 2e-16 ***
## relevel(sex, ref = "3")1 -1.462044   0.041926 -34.872  < 2e-16 ***
## relevel(sex, ref = "3")2 -1.994073   0.042109 -47.355  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 219481  on 675353  degrees of freedom
## Residual deviance: 213878  on 675346  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 213894
##
## Number of Fisher Scoring iterations: 6
```

```r
summary(glm(ADHD ~ age + as.factor(education) + relevel(sex, ref = "3"), data = data3, family = binomial
```

```
##
## Call:
## glm(formula = ADHD ~ age + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7372  -0.2894  -0.2434  -0.2180   2.9912
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.3904452  0.0678981 -20.478  < 2e-16 ***
## age                        -0.0142315  0.0006571 -21.657  < 2e-16 ***
## as.factor(education)1       0.4541438  0.0464504   9.777  < 2e-16 ***
## as.factor(education)2      -0.0510966  0.0413642  -1.235    0.217
## as.factor(education)3      -0.2131103  0.0416991  -5.111 3.21e-07 ***
## as.factor(education)4      -0.2157886  0.0453815  -4.755 1.98e-06 ***
## relevel(sex, ref = "3")1 -1.2233147  0.0547310 -22.351  < 2e-16 ***
## relevel(sex, ref = "3")2 -1.7174025  0.0548405 -31.316  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 186019  on 637808  degrees of freedom
## Residual deviance: 182764  on 637801  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 182780
##
## Number of Fisher Scoring iterations: 6
```

```r
#ADHD


summary(glm(bipolar ~ as.factor(binary), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = bipolar ~ as.factor(binary), family = binomial,
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4110  -0.1871  -0.1871  -0.1871   2.8476
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -4.036862   0.009346 -431.92   <2e-16 ***
## as.factor(binary)nonbinary  1.607761   0.060553   26.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120144  on 675356  degrees of freedom
## Residual deviance: 119674  on 675355  degrees of freedom
## AIC: 119678
##
## Number of Fisher Scoring iterations: 7
```

```
summary(glm(bipolar ~ scale(age) + as.factor(education) + as.factor(binary), data = data2, family = bin
```

```
##
## Call:
## glm(formula = bipolar ~ scale(age) + as.factor(education) + as.factor(binary),
##     family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5610  -0.1994  -0.1670  -0.1642   2.9568
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.829570   0.048306 -79.277   <2e-16 ***
## scale(age)                  0.014931   0.009255   1.613    0.107
## as.factor(education)1       0.481561   0.055420   8.689   <2e-16 ***
## as.factor(education)2      -0.073744   0.050200  -1.469    0.142
## as.factor(education)3      -0.459437   0.051221  -8.970   <2e-16 ***
## as.factor(education)4      -0.513047   0.056308  -9.111   <2e-16 ***
## as.factor(binary)nonbinary  1.505416   0.061053  24.658   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120144  on 675353  degrees of freedom
## Residual deviance: 118667  on 675347  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 118681
##
## Number of Fisher Scoring iterations: 7
```

```
summary(glm(bipolar ~ age + as.factor(education) + as.factor(binary), data = data3, family = binomial))
```

```
##
## Call:
## glm(formula = bipolar ~ age + as.factor(education) + as.factor(binary),
##     family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5577  -0.1975  -0.1650  -0.1629   2.9616
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.9232022  0.0580771 -67.552  < 2e-16 ***
## age                         0.0009964  0.0007903   1.261    0.207
```

```
## as.factor(education)1       0.5109669  0.0608800   8.393  < 2e-16 ***
## as.factor(education)2      -0.0290739  0.0552937  -0.526    0.599
## as.factor(education)3      -0.4120638  0.0562323  -7.328 2.34e-13 ***
## as.factor(education)4      -0.4656962  0.0610591  -7.627 2.40e-14 ***
## as.factor(binary)nonbinary  1.5442986  0.0703940  21.938  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110760  on 637808  degrees of freedom
## Residual deviance: 109536  on 637802  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 109550
##
## Number of Fisher Scoring iterations: 7
```

```
summary(glm(bipolar ~ relevel(sex, ref = "3"), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = bipolar ~ relevel(sex, ref = "3"), family = binomial,
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4110  -0.2041  -0.2041  -0.1565   2.9695
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.42910    0.05983  -40.60   <2e-16 ***
## relevel(sex, ref = "3")1 -1.96764    0.06245  -31.51   <2e-16 ***
## relevel(sex, ref = "3")2 -1.43212    0.06082  -23.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120144  on 675356  degrees of freedom
## Residual deviance: 118976  on 675354  degrees of freedom
## AIC: 118982
##
## Number of Fisher Scoring iterations: 7
```

```
summary(glm(bipolar ~ scale(age) + as.factor(education) + relevel(sex, ref = "3"), data = data2, family
```

```
##
## Call:
## glm(formula = bipolar ~ scale(age) + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5492  -0.2172  -0.1782  -0.1665   3.1063
##
```

```
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.346599   0.076219 -30.788  < 2e-16 ***
## scale(age)               -0.024381   0.009435  -2.584  0.00977 **
## as.factor(education)1     0.505021   0.055445   9.108  < 2e-16 ***
## as.factor(education)2    -0.066992   0.050233  -1.334  0.18233
## as.factor(education)3    -0.455909   0.051251  -8.896  < 2e-16 ***
## as.factor(education)4    -0.527840   0.056315  -9.373  < 2e-16 ***
## relevel(sex, ref = "3")1 -1.864394   0.062875 -29.653  < 2e-16 ***
## relevel(sex, ref = "3")2 -1.302408   0.061442 -21.197  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 120144  on 675353  degrees of freedom
## Residual deviance: 117917  on 675346  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 117933
##
## Number of Fisher Scoring iterations: 7
```

```
summary(glm(bipolar ~ age + as.factor(education) + relevel(sex, ref = "3"), data = data3, family = bino
```

```
##
## Call:
## glm(formula = bipolar ~ age + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5510  -0.2140  -0.1764  -0.1648   3.1129
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.3076849  0.0889056 -25.957  < 2e-16 ***
## age                        -0.0021365  0.0008053  -2.653  0.00798 **
## as.factor(education)1       0.5335010  0.0609020   8.760  < 2e-16 ***
## as.factor(education)2      -0.0218373  0.0553290  -0.395  0.69308
## as.factor(education)3      -0.4078247  0.0562657  -7.248 4.22e-13 ***
## as.factor(education)4      -0.4795721  0.0610671  -7.853 4.06e-15 ***
## relevel(sex, ref = "3")1 -1.9046764  0.0721647 -26.393  < 2e-16 ***
## relevel(sex, ref = "3")2 -1.3477789  0.0707468 -19.051  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110760  on 637808  degrees of freedom
## Residual deviance: 108863  on 637801  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 108879
##
## Number of Fisher Scoring iterations: 7
```

```r
#ADHD

summary(glm(depression ~ as.factor(binary), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = depression ~ as.factor(binary), family = binomial,
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2461  -0.7305  -0.7305  -0.7305   1.7039
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.184907   0.002881 -411.22   <2e-16 ***
## as.factor(binary)nonbinary  1.345078   0.032874   40.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 737859  on 675356  degrees of freedom
## Residual deviance: 736243  on 675355  degrees of freedom
## AIC: 736247
##
## Number of Fisher Scoring iterations: 4
```

```r
summary(glm(depression ~ scale(age) + as.factor(education) + as.factor(binary), data = data2, family = b
```

```
##
## Call:
## glm(formula = depression ~ scale(age) + as.factor(education) +
##     as.factor(binary), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9105  -0.7304  -0.6840  -0.6230   1.9179
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.087603   0.016348  -66.53   <2e-16 ***
## scale(age)                  0.217706   0.002841   76.63   <2e-16 ***
## as.factor(education)1       0.330490   0.019501   16.95   <2e-16 ***
## as.factor(education)2      -0.010848   0.016960   -0.64    0.522
## as.factor(education)3      -0.219670   0.017062  -12.88   <2e-16 ***
## as.factor(education)4      -0.343507   0.018228  -18.84   <2e-16 ***
## as.factor(binary)nonbinary  1.374098   0.033122   41.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 737857  on 675353  degrees of freedom
```

```
## Residual deviance: 726953  on 675347  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 726967
##
## Number of Fisher Scoring iterations: 4
```

```
summary(glm(depression ~ age + as.factor(education) + as.factor(binary), data = data3, family = binomial
```

```
##
## Call:
## glm(formula = depression ~ age + as.factor(education) + as.factor(binary),
##     family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9080  -0.7316  -0.6868  -0.6244   1.9144
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.5877184  0.0188474 -84.241   <2e-16 ***
## age                         0.0174329  0.0002386  73.078   <2e-16 ***
## as.factor(education)1       0.3218271  0.0206632  15.575   <2e-16 ***
## as.factor(education)2      -0.0120543  0.0180098  -0.669    0.503
## as.factor(education)3      -0.2246294  0.0180980 -12.412   <2e-16 ***
## as.factor(education)4      -0.3492554  0.0192108 -18.180   <2e-16 ***
## as.factor(binary)nonbinary  1.3928264  0.0379012  36.749   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 698123  on 637808  degrees of freedom
## Residual deviance: 688247  on 637802  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 688261
##
## Number of Fisher Scoring iterations: 4
```

```
summary(glm(depression ~ relevel(sex, ref = "3"), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = depression ~ relevel(sex, ref = "3"), family = binomial,
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2461  -0.8156  -0.5839  -0.5839   1.9252
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.16017    0.03275   4.891    1e-06 ***
## relevel(sex, ref = "3")1 -1.84279    0.03319 -55.524   <2e-16 ***
## relevel(sex, ref = "3")2 -1.09011    0.03293 -33.105   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 737859  on 675356  degrees of freedom
## Residual deviance: 721520  on 675354  degrees of freedom
## AIC: 721526
##
## Number of Fisher Scoring iterations: 4
```

```
summary(glm(depression ~ scale(age) + as.factor(education) + relevel(sex, ref = "3"), data = data2, fami
```

```
##
## Call:
## glm(formula = depression ~ scale(age) + as.factor(education) +
##     relevel(sex, ref = "3"), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8207  -0.7849  -0.6145  -0.5133   2.1112
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.2612492  0.0366189   7.134 9.73e-13 ***
## scale(age)                  0.1712682  0.0029033  58.991  < 2e-16 ***
## as.factor(education)1       0.3726527  0.0197133  18.904  < 2e-16 ***
## as.factor(education)2       0.0008459  0.0171383   0.049    0.961
## as.factor(education)3      -0.2147128  0.0172382 -12.456  < 2e-16 ***
## as.factor(education)4      -0.3649197  0.0184010 -19.831  < 2e-16 ***
## relevel(sex, ref = "3")1 -1.8261760  0.0334058 -54.667  < 2e-16 ***
## relevel(sex, ref = "3")2 -1.1103445  0.0331791 -33.465  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 737857  on 675353  degrees of freedom
## Residual deviance: 714142  on 675346  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 714158
##
## Number of Fisher Scoring iterations: 4
```

```
summary(glm(depression ~ age + as.factor(education) + relevel(sex, ref = "3"), data = data3, family = bi
```

```
##
## Call:
## glm(formula = depression ~ age + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8199  -0.7886  -0.6181  -0.5114   2.1127
##
## Coefficients:
```

```
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -0.1076641  0.0418893  -2.570   0.0102 *
## age                         0.0136563  0.0002438  56.009   <2e-16 ***
## as.factor(education)1       0.3636888  0.0208921  17.408   <2e-16 ***
## as.factor(education)2       0.0005899  0.0182047   0.032   0.9741
## as.factor(education)3      -0.2186142  0.0182912 -11.952   <2e-16 ***
## as.factor(education)4      -0.3698969  0.0193994 -19.067   <2e-16 ***
## relevel(sex, ref = "3")1 -1.8592006  0.0381700 -48.708   <2e-16 ***
## relevel(sex, ref = "3")2 -1.1318099  0.0379460 -29.827   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 698123  on 637808  degrees of freedom
## Residual deviance: 675824  on 637801  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 675840
##
## Number of Fisher Scoring iterations: 4
```

```
#ADHD

summary(glm(ld ~ as.factor(binary), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = ld ~ as.factor(binary), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4963  -0.2677  -0.2677  -0.2677   2.5871
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.310806   0.006621 -500.03   <2e-16 ***
## as.factor(binary)nonbinary  1.278733   0.051420   24.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 207830  on 675356  degrees of freedom
## Residual deviance: 207377  on 675355  degrees of freedom
## AIC: 207381
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ld ~ scale(age) + as.factor(education) + as.factor(binary), data = data2, family = binomial
```

```
##
## Call:
## glm(formula = ld ~ scale(age) + as.factor(education) + as.factor(binary),
##     family = binomial, data = data2)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6583  -0.2891  -0.2726  -0.2319   3.0596
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.018807   0.032171 -93.838  < 2e-16 ***
## scale(age)                 -0.321452   0.008104 -39.668  < 2e-16 ***
## as.factor(education)1       0.107409   0.039411   2.725  0.00642 **
## as.factor(education)2      -0.497133   0.033800 -14.708  < 2e-16 ***
## as.factor(education)3      -0.297876   0.033667  -8.848  < 2e-16 ***
## as.factor(education)4      -0.280418   0.037495  -7.479  7.5e-14 ***
## as.factor(binary)nonbinary  1.145296   0.051851  22.088  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 207830  on 675353  degrees of freedom
## Residual deviance: 205070  on 675347  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 205084
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ld ~ age + as.factor(education) + as.factor(binary), data = data3, family = binomial))
```

```
##
## Call:
## glm(formula = ld ~ age + as.factor(education) + as.factor(binary),
##     family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4744  -0.2818  -0.2623  -0.2232   3.0842
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.465295   0.041940 -58.781  < 2e-16 ***
## age                        -0.026325   0.000709 -37.129  < 2e-16 ***
## as.factor(education)1       0.123105   0.046749   2.633  0.00846 **
## as.factor(education)2      -0.360345   0.039970  -9.015  < 2e-16 ***
## as.factor(education)3      -0.117050   0.039721  -2.947  0.00321 **
## as.factor(education)4      -0.109330   0.043250  -2.528  0.01148 *
## as.factor(binary)nonbinary  0.635838   0.076222   8.342  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 184175  on 637808  degrees of freedom
## Residual deviance: 182247  on 637802  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 182261
##
```

```
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ld ~ relevel(sex, ref = "3"), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = ld ~ relevel(sex, ref = "3"), family = binomial,
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4963  -0.2764  -0.2621  -0.2621   2.6031
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.03207    0.05099  -39.85   <2e-16 ***
## relevel(sex, ref = "3")1 -1.21385    0.05203  -23.33   <2e-16 ***
## relevel(sex, ref = "3")2 -1.32170    0.05172  -25.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 207830  on 675356  degrees of freedom
## Residual deviance: 207313  on 675354  degrees of freedom
## AIC: 207319
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ld ~ scale(age) + as.factor(education) + relevel(sex, ref = "3"), data = data2, family = bi
```

```
##
## Call:
## glm(formula = ld ~ scale(age) + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6579  -0.2882  -0.2716  -0.2314   3.0548
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.872769   0.059974 -31.226  < 2e-16 ***
## scale(age)                -0.320057   0.008188 -39.090  < 2e-16 ***
## as.factor(education)1      0.106888   0.039414   2.712  0.00669 **
## as.factor(education)2     -0.497235   0.033800 -14.711  < 2e-16 ***
## as.factor(education)3     -0.297864   0.033667  -8.847  < 2e-16 ***
## as.factor(education)4     -0.279849   0.037498  -7.463 8.46e-14 ***
## relevel(sex, ref = "3")1 -1.136186   0.052419 -21.675  < 2e-16 ***
## relevel(sex, ref = "3")2 -1.152291   0.052191 -22.078  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 207830  on 675353  degrees of freedom
## Residual deviance: 205069  on 675346  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 205085
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ld ~ age + as.factor(education) + relevel(sex, ref = "3"), data = data3, family = binomial))
```

```
##
## Call:
## glm(formula = ld ~ age + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4751  -0.2817  -0.2614  -0.2234   3.0941
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.8239468  0.0857090 -21.281  < 2e-16 ***
## age                      -0.0265601  0.0007167 -37.060  < 2e-16 ***
## as.factor(education)1     0.1240804  0.0467497   2.654  0.00795 **
## as.factor(education)2    -0.3601478  0.0399697  -9.011  < 2e-16 ***
## as.factor(education)3    -0.1170963  0.0397202  -2.948  0.00320 **
## as.factor(education)4    -0.1105415  0.0432521  -2.556  0.01060 *
## relevel(sex, ref = "3")1 -0.6554197  0.0767040  -8.545  < 2e-16 ***
## relevel(sex, ref = "3")2 -0.6216852  0.0764680  -8.130 4.29e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 184175  on 637808  degrees of freedom
## Residual deviance: 182241  on 637801  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 182257
##
## Number of Fisher Scoring iterations: 6
```

```
#ADHD

summary(glm(ocd ~ as.factor(binary), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = ocd ~ as.factor(binary), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5036  -0.2245  -0.2245  -0.2245   2.7178
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -3.668035   0.007832 -468.32   <2e-16 ***
```

```
## as.factor(binary)nonbinary  1.666797   0.050999   32.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 159955  on 675356  degrees of freedom
## Residual deviance: 159235  on 675355  degrees of freedom
## AIC: 159239
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ocd ~ scale(age) + as.factor(education) + as.factor(binary), data = data2, family = binomial
```

```
##
## Call:
## glm(formula = ocd ~ scale(age) + as.factor(education) + as.factor(binary),
##     family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6526  -0.2375  -0.2141  -0.2068   2.9010
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.338989   0.038108 -87.620  < 2e-16 ***
## scale(age)                 -0.084824   0.008237 -10.298  < 2e-16 ***
## as.factor(education)1       0.255990   0.045222   5.661 1.51e-08 ***
## as.factor(education)2      -0.278678   0.039863  -6.991 2.73e-12 ***
## as.factor(education)3      -0.496284   0.040433 -12.274  < 2e-16 ***
## as.factor(education)4      -0.507352   0.044777 -11.331  < 2e-16 ***
## as.factor(binary)nonbinary  1.552956   0.051413  30.205  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 159955  on 675353  degrees of freedom
## Residual deviance: 158328  on 675347  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 158342
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ocd ~ age + as.factor(education) + as.factor(binary), data = data3, family = binomial))
```

```
##
## Call:
## glm(formula = ocd ~ age + as.factor(education) + as.factor(binary),
##     family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4588  -0.2310  -0.2094  -0.2021   2.8801
##
```

```
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.2997525  0.0477802 -69.061  < 2e-16 ***
## age                        -0.0054594  0.0007097  -7.693 1.44e-14 ***
## as.factor(education)1       0.2254755  0.0516828   4.363 1.28e-05 ***
## as.factor(education)2      -0.2013353  0.0453493  -4.440 9.01e-06 ***
## as.factor(education)3      -0.4060144  0.0458209  -8.861  < 2e-16 ***
## as.factor(education)4      -0.4464215  0.0499784  -8.932  < 2e-16 ***
## as.factor(binary)nonbinary  0.9630933  0.0774669  12.432  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142770  on 637808  degrees of freedom
## Residual deviance: 142077  on 637802  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 142091
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ocd ~ relevel(sex, ref = "3"), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = ocd ~ relevel(sex, ref = "3"), family = binomial,
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5036  -0.2450  -0.2450  -0.1878   2.8448
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.00124    0.05039  -39.71   <2e-16 ***
## relevel(sex, ref = "3")1 -2.02756    0.05257  -38.57   <2e-16 ***
## relevel(sex, ref = "3")2 -1.49022    0.05123  -29.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 159955  on 675356  degrees of freedom
## Residual deviance: 158233  on 675354  degrees of freedom
## AIC: 158239
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ocd ~ scale(age) + as.factor(education) + relevel(sex, ref = "3"), data = data2, family = bi
```

```
##
## Call:
## glm(formula = ocd ~ scale(age) + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data2)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6661  -0.2408  -0.2205  -0.1885   3.0636
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.808965   0.062669 -28.865  < 2e-16 ***
## scale(age)               -0.127988   0.008398 -15.240  < 2e-16 ***
## as.factor(education)1     0.277942   0.045263   6.141 8.22e-10 ***
## as.factor(education)2    -0.273650   0.039907  -6.857 7.02e-12 ***
## as.factor(education)3    -0.495133   0.040474 -12.233  < 2e-16 ***
## as.factor(education)4    -0.525090   0.044797 -11.722  < 2e-16 ***
## relevel(sex, ref = "3")1 -1.930024   0.052951 -36.449  < 2e-16 ***
## relevel(sex, ref = "3")2 -1.339127   0.051751 -25.876  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 159955  on 675353  degrees of freedom
## Residual deviance: 157142  on 675346  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 157158
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm(ocd ~ age + as.factor(education) + relevel(sex, ref = "3"), data = data3, family = binomial)
```

```
##
## Call:
## glm(formula = ocd ~ age + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4708  -0.2383  -0.2179  -0.1771   3.1017
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.2416306  0.0893650 -25.084  < 2e-16 ***
## age                       -0.0095538  0.0007244 -13.188  < 2e-16 ***
## as.factor(education)1      0.2507753  0.0517371   4.847 1.25e-06 ***
## as.factor(education)2     -0.1947281  0.0454138  -4.288 1.80e-05 ***
## as.factor(education)3     -0.4039453  0.0458829  -8.804  < 2e-16 ***
## as.factor(education)4     -0.4666981  0.0500135  -9.331  < 2e-16 ***
## relevel(sex, ref = "3")1 -1.4344666  0.0788174 -18.200  < 2e-16 ***
## relevel(sex, ref = "3")2 -0.7199334  0.0776966  -9.266  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 142770  on 637808  degrees of freedom
## Residual deviance: 140595  on 637801  degrees of freedom
##   (3 observations deleted due to missingness)
```

```
## AIC: 140611
##
## Number of Fisher Scoring iterations: 7
```

```
#ADHD


summary(glm(scz ~ as.factor(binary), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = scz ~ as.factor(binary), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3643  -0.0684  -0.0684  -0.0684   3.4808
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -6.05561    0.02526 -239.74   <2e-16 ***
## as.factor(binary)nonbinary  3.37618    0.07122   47.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25076  on 675356  degrees of freedom
## Residual deviance: 23962  on 675355  degrees of freedom
## AIC: 23966
##
## Number of Fisher Scoring iterations: 9
```

```
summary(glm(scz ~ scale(age) + as.factor(education) + as.factor(binary), data = data2, family = binomial
```

```
##
## Call:
## glm(formula = scz ~ scale(age) + as.factor(education) + as.factor(binary),
##     family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6683  -0.0787  -0.0571  -0.0506   3.9740
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -5.18507    0.08460 -61.292  < 2e-16 ***
## scale(age)                 -0.40024    0.02957 -13.537  < 2e-16 ***
## as.factor(education)1       0.36840    0.09773   3.770 0.000164 ***
## as.factor(education)2      -0.89248    0.09013  -9.902  < 2e-16 ***
## as.factor(education)3      -1.56771    0.09882 -15.865  < 2e-16 ***
## as.factor(education)4      -1.28534    0.12157 -10.572  < 2e-16 ***
## as.factor(binary)nonbinary  2.99923    0.07392  40.576  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25076  on 675353  degrees of freedom
## Residual deviance: 23013  on 675347  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 23027
##
## Number of Fisher Scoring iterations: 9
```

```
summary(glm(scz ~ age + as.factor(education) + as.factor(binary), data = data3, family = binomial))
```

```
##
## Call:
## glm(formula = scz ~ age + as.factor(education) + as.factor(binary),
##     family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2627  -0.0676  -0.0477  -0.0429   3.9615
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -5.008444   0.140512 -35.644  < 2e-16 ***
## age                        -0.021157   0.002865  -7.385 1.53e-13 ***
## as.factor(education)1       0.224335   0.145041   1.547    0.122
## as.factor(education)2      -0.649929   0.128537  -5.056 4.27e-07 ***
## as.factor(education)3      -1.366597   0.137298  -9.953  < 2e-16 ***
## as.factor(education)4      -1.462764   0.171826  -8.513  < 2e-16 ***
## as.factor(binary)nonbinary  1.773234   0.176610  10.040  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15545  on 637808  degrees of freedom
## Residual deviance: 15103  on 637802  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 15117
##
## Number of Fisher Scoring iterations: 9
```

```
summary(glm(scz ~ relevel(sex, ref = "3"), data = data2, family = binomial))
```

```
##
## Call:
## glm(formula = scz ~ relevel(sex, ref = "3"), family = binomial,
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3643  -0.0850  -0.0555  -0.0555   3.5987
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.67943    0.06659  -40.24   <2e-16 ***
```

```
## relevel(sex, ref = "3")1 -2.94202    0.07420  -39.65   <2e-16 ***
## relevel(sex, ref = "3")2 -3.79433    0.07753  -48.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25076  on 675356  degrees of freedom
## Residual deviance: 23681  on 675354  degrees of freedom
## AIC: 23687
##
## Number of Fisher Scoring iterations: 9
```

```
summary(glm(scz ~ scale(age) + as.factor(education) + relevel(sex, ref = "3"), data = data2, family = bi
```

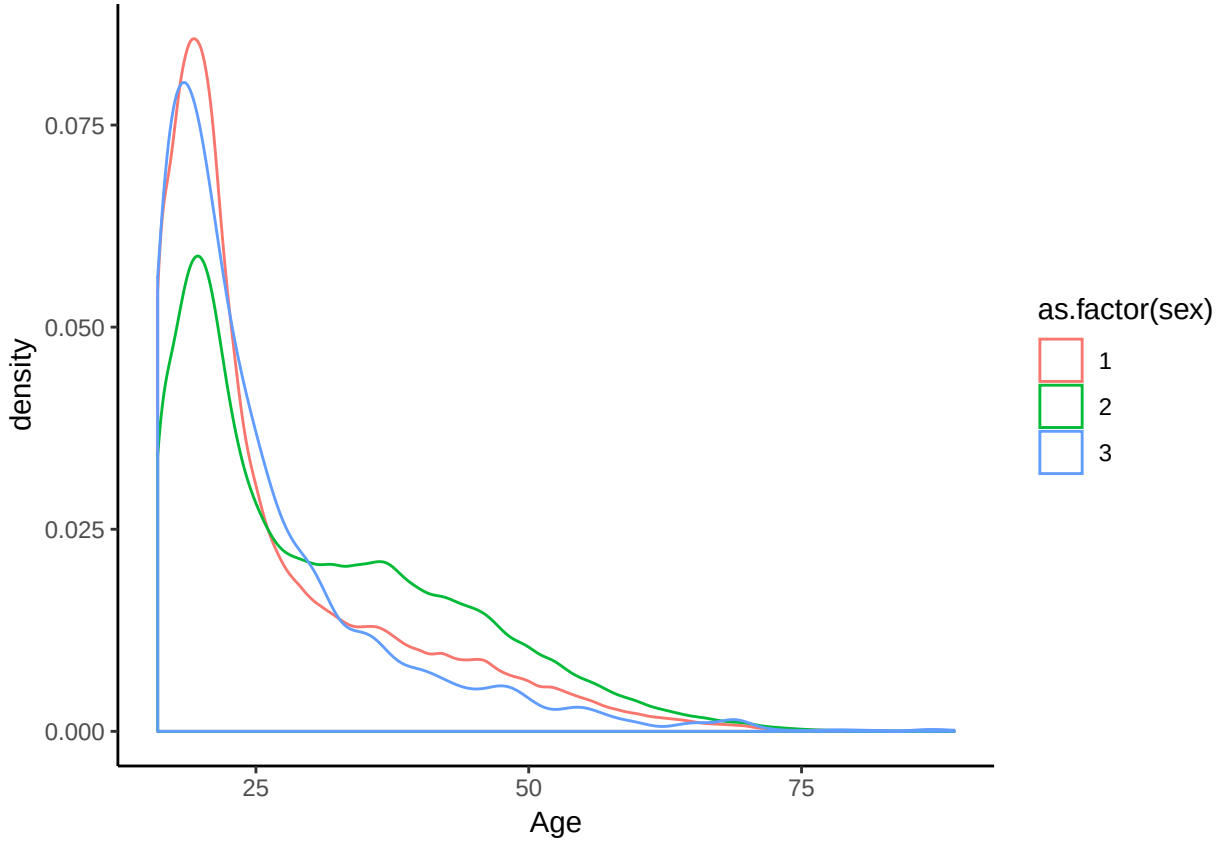
```
##
## Call:
## glm(formula = scz ~ scale(age) + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6556  -0.0691  -0.0602  -0.0450   4.0159
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.14209    0.10620 -20.170  < 2e-16 ***
## scale(age)               -0.33744    0.02958 -11.409  < 2e-16 ***
## as.factor(education)1     0.34977    0.09783   3.575  0.00035 ***
## as.factor(education)2    -0.89767    0.09016  -9.957  < 2e-16 ***
## as.factor(education)3    -1.56923    0.09887 -15.872  < 2e-16 ***
## as.factor(education)4    -1.26158    0.12170 -10.367  < 2e-16 ***
## relevel(sex, ref = "3")1 -2.66046    0.07632 -34.858  < 2e-16 ***
## relevel(sex, ref = "3")2 -3.38813    0.08047 -42.104  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25076  on 675353  degrees of freedom
## Residual deviance: 22814  on 675346  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 22830
##
## Number of Fisher Scoring iterations: 9
```

```
summary(glm(scz ~ age + as.factor(education) + relevel(sex, ref = "3"), data = data3, family = binomial)
```

```
##
## Call:
## glm(formula = scz ~ age + as.factor(education) + relevel(sex,
##     ref = "3"), family = binomial, data = data3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -0.2594  -0.0633  -0.0510  -0.0415   3.9510
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -3.291183   0.215791 -15.252  < 2e-16 ***
## age                        -0.018499   0.002886  -6.410 1.45e-10 ***
## as.factor(education)1       0.212069   0.145144   1.461    0.144
## as.factor(education)2      -0.653261   0.128546  -5.082 3.74e-07 ***
## as.factor(education)3      -1.367709   0.137318  -9.960  < 2e-16 ***
## as.factor(education)4      -1.450566   0.171916  -8.438  < 2e-16 ***
## relevel(sex, ref = "3")1 -1.579618   0.179096  -8.820  < 2e-16 ***
## relevel(sex, ref = "3")2 -1.953118   0.179691 -10.869  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15545  on 637808  degrees of freedom
## Residual deviance: 15069  on 637801  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 15085
##
## Number of Fisher Scoring iterations: 9
```

```
ggplot(data2, aes(x=age, colour=as.factor(sex))) +   geom_density() + theme_classic() + xlab("Age")
```



```r
summary(aov(age ~ as.factor(sex), data = data2))
```

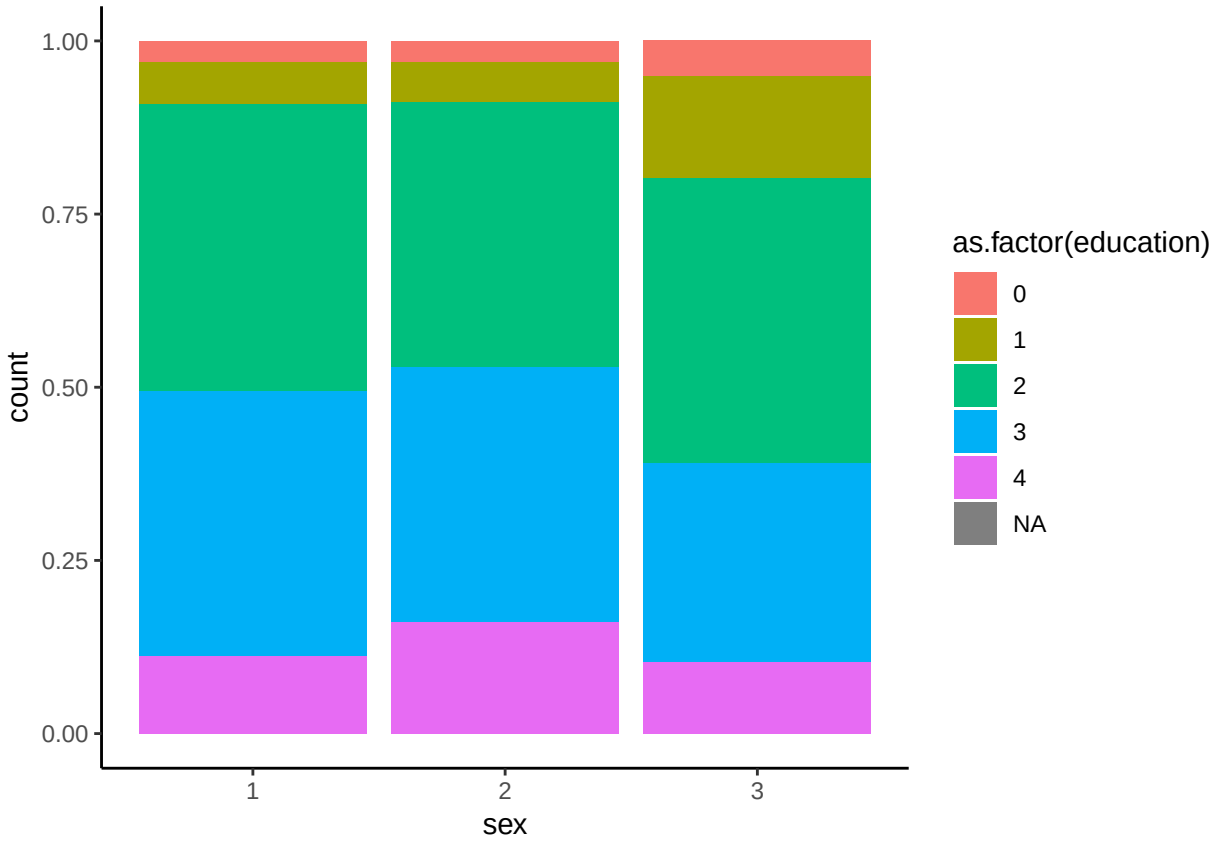
```
##                     Df   Sum Sq Mean Sq F value Pr(>F)
## as.factor(sex)       2  2744483 1372241    9484 <2e-16 ***
## Residuals       675354 97720984     145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary_edu = setDT(data2)[, list(count=.N) , list(sex, education)]

summary_edu
```

```
##     sex education  count
##  1:   2         4  66633
##  2:   1         0   7704
##  3:   1         1  15767
##  4:   1         3  98952
##  5:   3         1    553
##  6:   2         3 151606
##  7:   3         2   1543
##  8:   2         2 157298
##  9:   1         4  29394
## 10:   1         2 107725
## 11:   2         1  24290
## 12:   2         0  12231
## 13:   3         3   1077
## 14:   3         4    391
## 15:   3         0    190
## 16:   2        NA      2
## 17:   1        NA      1
```

```r
ggplot(data2, aes(sex, fill = as.factor(education))) + geom_bar(position = 'fill') + theme_classic()
```


1.00
0.75
0.50
0.25
0.00
count
1
2
3
sex
as.factor(education)
0
1
2
3
4
NA