Prep_for_hyrda_with_matching

R Markdown

The purpose of this script is to read in CNB data, subset it, merge it, and write files to use in Hydra. Before doing hydra, however, we must address the vastly different ages, maternal education and race between depressed and control groups. Unfortunately, the matching removes a high number of our patients.

```
library(visreg)
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-22. For overview type 'help("mgcv-package")'.
```

```
library(tableone)
library(dplyr)
```

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

```
## The following object is masked from 'package:nlme':
##
##     collapse
```

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

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

```
library(plm)
```

```
## Loading required package: Formula
```

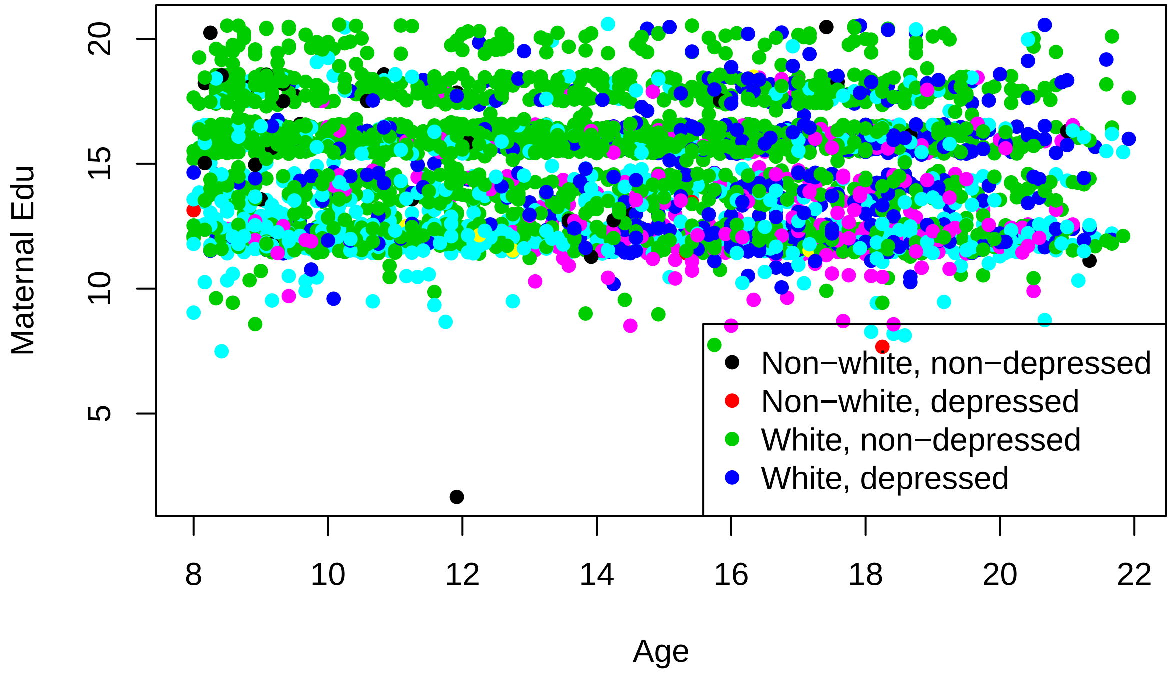
```
##
## Attaching package: 'plm'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

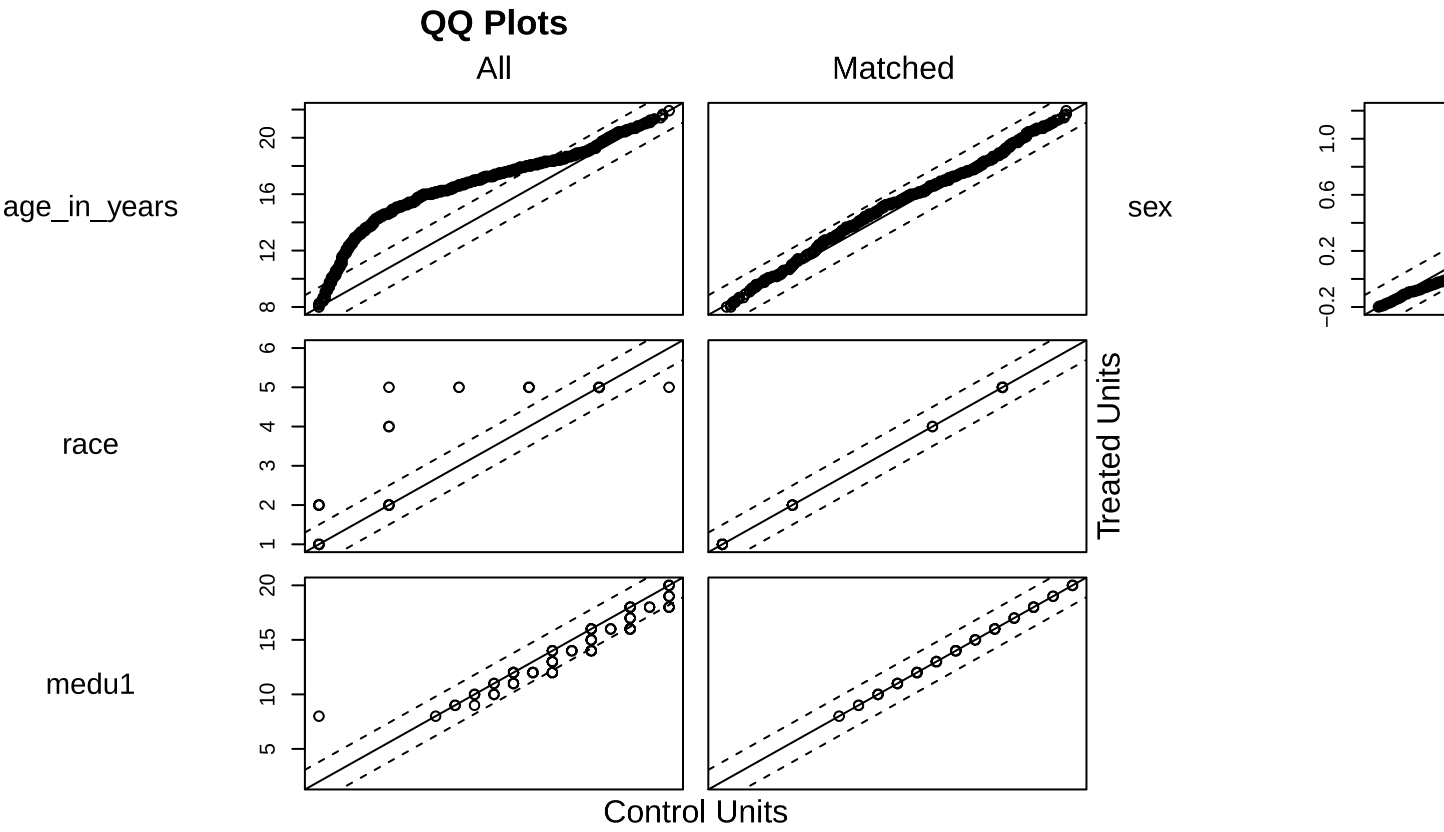
```
library(MatchIt)

#read in csvs
demographics <- read.csv("/Users/eballer/BBL/from_chead/ballerDepHeterogen/data/n9498_demographics_go1_
cnb_scores <- read.csv("/Users/eballer/BBL/from_chead/ballerDepHeterogen/data/n9498_cnb_zscores_fr_2017
health <- read.csv("/Users/eballer/BBL/from_chead/ballerDepHeterogen/data/n9498_health_20170405.csv", h
psych_summary <- read.csv("/Users/eballer/BBL/from_chead/ballerDepHeterogen/data/n9498_goassess_psych_s

#remove people with NA for race, age, or sex.  START WITH N = 9498
demographics_noNA_race <- demographics[!is.na(demographics$race),] #everyone has a race, N = 9498
demographics_noNA_race_age <- demographics_noNA_race[!is.na(demographics_noNA_race$ageAtClinicalAssess1
demographics_noNA_race_age_sex <- demographics_noNA_race_age[!is.na(demographics_noNA_race_age$sex),] #
demographics_noNA_race_age_andCNBage_sex <- demographics_noNA_race_age_sex[!is.na(demographics_noNA_rac
#remove people with NA for depression or total psych score, START WITH N = 9498
psych_summary_no_NA_dep <- psych_summary[!is.na(psych_summary$smry_dep),] #take out those with NA for d
psych_summary_no_NA_dep_and_smry_psych_overall <- psych_summary_no_NA_dep[!is.na(psych_summary_no_NA_dep

#merge the csvs
#merge demographics and cnb #this is if we want to include people without full demographic data
dem_cnb <- merge(demographics_noNA_race_age_andCNBage_sex, cnb_scores, by = "bblid") #merge demographic
psych_health <- merge(psych_summary_no_NA_dep_and_smry_psych_overall, health, by = "bblid") #merge psyc
dem_cnb_psych_health_merged <- merge (dem_cnb, psych_health, by = "bblid") #merge all 4 csvs, lost 1 pe

#make subsets
subset_just_dep_and_no_medicalratingExclude <- subset.data.frame(dem_cnb_psych_health_merged, (medicalra
subset_no_psych_no_medicalratingExclude <- subset.data.frame(dem_cnb_psych_health_merged, (medicalrating
subset_dep_or_no_psych_and_no_medicalratingExclude <- subset.data.frame(dem_cnb_psych_health_merged, (me

#would binarize depression smry score to -1 (less than 4, not depressed) and 1 (score 4 , depressed)
dep_binarized <- ifelse(subset_dep_or_no_psych_and_no_medicalratingExclude$smry_dep == 4, 1, -1)
subset_dep_or_no_psych_and_no_medicalratingExclude_DEPBINARIZED <- cbind(subset_dep_or_no_psych_and_no_

#make depression and gender into factor scores
subset_dep_or_no_psych_and_no_medicalratingExclude_DEPBINARIZED$dep_binarized <- as.factor(subset_dep_o
subset_dep_or_no_psych_and_no_medicalratingExclude_DEPBINARIZED$sex <- as.factor(subset_dep_or_no_psych

#divide ageAtCNB by 12 for age
subset_dep_or_no_psych_and_no_medicalratingExclude_DEPBINARIZED$age_in_years <- subset_dep_or_no_psych_

#age demeaned and squared, from Toni
subset_dep_or_no_psych_and_no_medicalratingExclude_DEPBINARIZED$ageSq <- as.numeric(I(scale(subset_dep_

#Subset only variables needed for hydra analysis
#(BBLID, cognitive variables, depression), also do by males(1555)/females(1729) separately
subset_bblidAndCog_features <- data.frame(cbind(subset_dep_or_no_psych_and_no_medicalratingExclude_DEPB
subset_bblidAndCog_features_males <- subset.data.frame(data.frame(cbind(subset_dep_or_no_psych_and_no_me
subset_bblidAndCog_features_females <- subset.data.frame(data.frame(cbind(subset_dep_or_no_psych_and_no_

#subset of covariates (BBLID, sex, age in years), also do by males/females
subset_bblidAndCovariates <- data.frame(cbind(subset_dep_or_no_psych_and_no_medicalratingExclude_DEPBINA
subset_bblidAndCovariates_males <- subset.data.frame(data.frame(cbind(subset_dep_or_no_psych_and_no_medi
subset_bblidAndCovariates_females <- subset.data.frame(data.frame(cbind(subset_dep_or_no_psych_and_no_me

#save files for hyd
write.csv(subset_bblidAndCog_features, file="/Users/eballer/BBL/from_chead/ballerDepHeterogen/results/hy
write.csv(subset_bblidAndCovariates, file="/Users/eballer/BBL/from_chead/ballerDepHeterogen/results/hyd
write.csv(subset_bblidAndCog_features_males, file="/Users/eballer/BBL/from_chead/ballerDepHeterogen/resu
write.csv(subset_bblidAndCovariates_males, file="/Users/eballer/BBL/from_chead/ballerDepHeterogen/result
write.csv(subset_bblidAndCog_features_females, file="/Users/eballer/BBL/from_chead/ballerDepHeterogen/re
write.csv(subset_bblidAndCovariates_females, file="/Users/eballer/BBL/from_chead/ballerDepHeterogen/resu
```

Including Plots

```
#match with matchit (starting n = 3284, males = 1555, females 1729)
data.unmatched = subset_dep_or_no_psych_and_no_medicalratingExclude_DEPBINARIZED
data.unmatched$unmatchedRows =rownames(data.unmatched)
dataset = data.unmatched
# Some preprocessing
dataset = dplyr::select(dataset, sex, age_in_years, ageSq, medu1, race, dep_binarized, unmatchedRows)
#dataset = dplyr::filter(dataset, !is.na(group))
# Dep: 1, Health = 0
dataset$dep_binarized = 1*(dataset$dep_binarized==1)
#"male": 1, "female": 0
dataset$sex = 1*(dataset$sex==1)
# Remove subjects with NA for maternal edu, new N = 3256, males = 1539, females 1717
dataset <- dataset[!is.na(dataset$medu1),]

# Plot prematch
plot(dataset$age_in_years,jitter(dataset$medu1, factor=3), col=2*dataset$race+dataset$dep_binarized+1,p
legend("bottomright",c("Non-white, non-depressed", "Non-white, depressed", "White, non-depressed", "Whi
```



```
#GAM for propensity score
ps.model =gam(dep_binarized ~s(age_in_years) +s(medu1) + race + sex, data=dataset, family=binomial)
ps =exp(predict(ps.model))/(1 +exp(predict(ps.model)))
m.out <-matchit(dep_binarized ~ age_in_years, data=dataset, method="nearest", exact=c("race", "medu1",
plot(m.out)
```



```
#return the matched dataset. N = 1518, males = 518, females = 1000
m.data <- match.data(m.out)

# Test for significant difference in age between groups
t.test(age_in_years~dep_binarized, data=m.data)
```

```
##
##  Welch Two Sample t-test
##
## data:  age_in_years by dep_binarized
## t = -1.1805, df = 1512.8, p-value = 0.238
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4996978  0.1242037
## sample estimates:
## mean in group 0 mean in group 1
##        15.83366        16.02141
```

```
t.test(race~dep_binarized, data=m.data)
```

```
##
##  Welch Two Sample t-test
##
## data:  race by dep_binarized
## t = 0, df = 1516, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1196981  0.1196981
## sample estimates:
## mean in group 0 mean in group 1
##        1.768116        1.768116
```

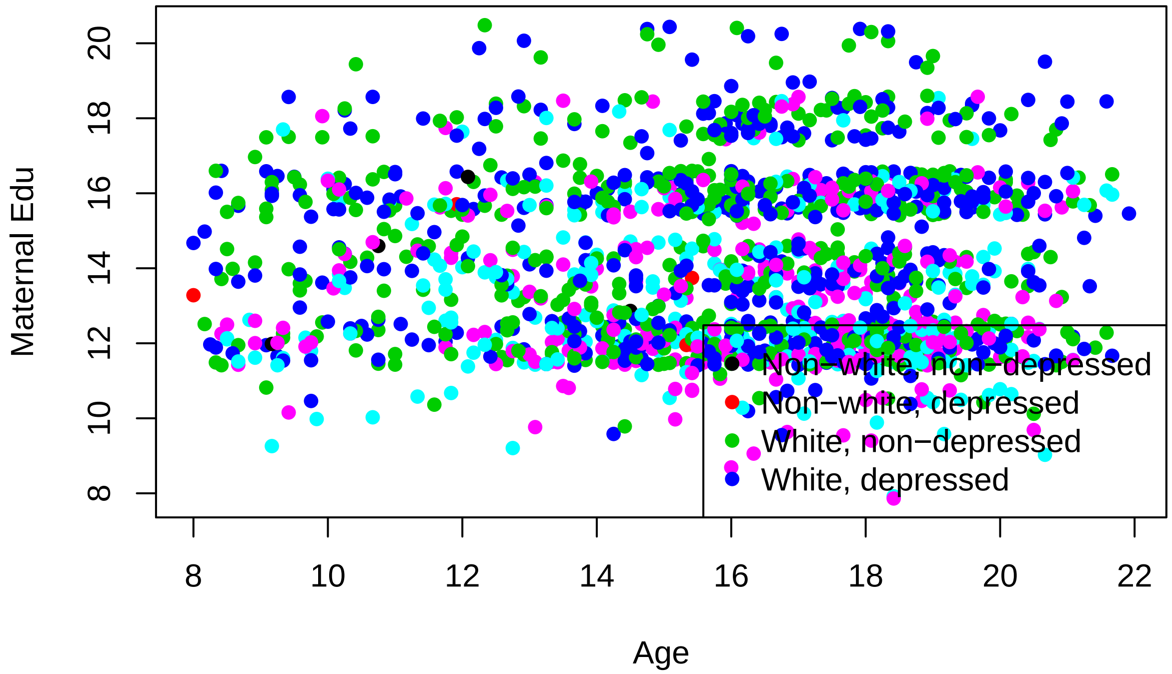
```
t.test(sex~dep_binarized, data=m.data)
```

```
##
##  Welch Two Sample t-test
##
## data:  sex by dep_binarized
## t = 0, df = 1516, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04777145  0.04777145
## sample estimates:
## mean in group 0 mean in group 1
##       0.3412385       0.3412385
```

```
t.test(medu1~dep_binarized, data=m.data)
```

```
##
##  Welch Two Sample t-test
##
## data:  medu1 by dep_binarized
## t = 0, df = 1516, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2277971  0.2277971
## sample estimates:
## mean in group 0 mean in group 1
##        14.12253        14.12253
```

```
# Re-plot
plot(m.data$age_in_years,jitter(m.data$medu1, factor=3), col=2*m.data$race+m.data$dep_binarized+1,pch=1
legend("bottomright",c("Non-white, non-depressed", "Non-white, depressed", "White, non-depressed", "Whi
```



```
# Make the final matched data set
data.matched = data.unmatched[data.unmatched$unmatchedRows%in%m.data$unmatchedRows,]
data.matched$unmatchedRows = NULL

saveRDS(data.matched, file='/Users/eballer/BBL/from_chead/ballerDepHeterogen/results/hydraMatched_age_ra
```