MVMR LASSO analysis

Eleanor Sanderson

colnames(df) <- paste("Column" ,original_cols,sep="-")

```r
setwd(projectfolder)

linker <- read_csv("linker.csv")
```

```
## Rows: 488377 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ieu
## dbl (1): app
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
dat <- as_tibble(phenodat) %>%
  rename(app=eid) %>%
  inner_join(linker, by="app")

dat <- dat %>%
  rename(FID=ieu) %>%
       inner_join(snpdat, by="FID") %>%
      inner_join(PCs, by="FID")

setwd(datafolder)

#add in the scores created using the GWAS effect sizes
educationscore <- read_table2("education.sscore")
```

```
## Warning: `read_table2()` was deprecated in readr 2.0.0.
## Please use `read_table()` instead.
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `#FID` = col_character(),
##   IID = col_character(),
##   NMISS_ALLELE_CT = col_double(),
##   NAMED_ALLELE_DOSAGE_SUM = col_double(),
##   SCORE1_AVG = col_double()
## )
```

```r
cogscore <- read_table2("cognitive_ability.sscore")
```

```
## Warning: `read_table2()` was deprecated in readr 2.0.0.
## Please use `read_table()` instead.
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `#FID` = col_character(),
##   IID = col_character(),
##   NMISS_ALLELE_CT = col_double(),
##   NAMED_ALLELE_DOSAGE_SUM = col_double(),
##   SCORE1_AVG = col_double()
## )
```

```r
educationscore_update <- read_table2("education_new.sscore")
```

```
## Warning: `read_table2()` was deprecated in readr 2.0.0.
## Please use `read_table()` instead.
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `#FID` = col_character(),
##   IID = col_character(),
##   NMISS_ALLELE_CT = col_double(),
##   NAMED_ALLELE_DOSAGE_SUM = col_double(),
##   SCORE1_AVG = col_double()
## )
```

```r
cogscore_update <- read_table2("cognitive_ability_new.sscore")
```

```
## Warning: `read_table2()` was deprecated in readr 2.0.0.
## Please use `read_table()` instead.
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `#FID` = col_character(),
##   IID = col_character(),
##   NMISS_ALLELE_CT = col_double(),
##   NAMED_ALLELE_DOSAGE_SUM = col_double(),
##   SCORE1_AVG = col_double()
## )
```

```r
educationscore <- educationscore %>%
                          rename(FID = `#FID`) %>%
                          rename(edu_grs = SCORE1_AVG)

cogscore <- cogscore %>%
```

```
                             rename(FID = `#FID`) %>%
                             rename(cog_grs = SCORE1_AVG)

educationscore_update <- educationscore_update %>%
                             rename(FID = `#FID`) %>%
                             rename(edu_grs_update = SCORE1_AVG)

cogscore_update <- cogscore_update %>%
                             rename(FID = `#FID`) %>%
                             rename(cog_grs_update = SCORE1_AVG)

scores <- educationscore %>%
  inner_join(cogscore, by="FID") %>%
  inner_join(educationscore_update, by="FID") %>%
  inner_join(cogscore_update, by="FID") %>%
  select(FID, edu_grs, cog_grs, edu_grs_update, cog_grs_update)

dat <- dat %>%
  inner_join(scores, by="FID")


setwd(projectfolder)
snp_list_edu <- read.table("edu_snplist.txt")
snp_list_cog <- read.table("cog_snplist.txt")

source("fsw.R")
```

Data Analysis dataset: UKBiobank

Education SNPs taken from Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016

Cognitive ability SNPs taken from Sniekers S, Stringer S, Watanabe K, Jansen PR, Coleman JR, Krapohl E, et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. Nat Genet. 2017.

Get the betas for the snps, identify any overlapping SNPs and generate the scores

4 SNPS are in LD across education and IQ lists The pairs are;

IQ: rs10191758, education: rs17824247 IQ: rs13010010, education: rs12987662 IQ: rs41352752, education: rs10061788 IQ: rs78164635, education: rs1008078

Cleaning the phenotype data - rename variables and create the age variable

- create a list of SNPs for the instruments for the MR analysis
- remove the effect alleles from the column names in the SNP data
- replace edu age with 21 if highest qual is degree
- complete case data only
- remove age leaving education < 10
- standardise cognitive ability
- log bmi

Plot the distributions for each of the main variables used in the analysis

2. MVMR estimation

2SLS regression including each snp as a separate instrument

These regressions give similar results to those in Sanderson et al 2019. Differences have arrisen because: - here interim release data has not been excluded from the analysis - fewer covariates have been included in the estimation

Covariates included in each regression are; age, sex and 10 PC's.

Overall the results show that education has a bmi lowering effect and cognitive ability has limited evidence of any effect. When the SNPs are included individually the Sargan statistic is large - indicating substantial heterogenetiy in the results. However the instruments are relatively weak. When the genetic risk scores are used as instruments the instruments are strong and the effect estimates are further from the null for each exposure.

```
covars <- paste(" age + male +", paste0("PC",1:10,collapse = "+"), "|",
                "age + male +", paste0("PC",1:10,collapse = "+"))
#[Note - covariates need to be included on both sides of the covars paste command]

ivformula <- as.formula(paste("lnbmi ~ edu_age + cog", covars,
                              paste(instruments, collapse="+"), sep = "+"))

indreg <- ivreg(ivformula, data=dat, model = TRUE)
summary(indreg, diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = ivformula, data = dat, model = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72077 -0.11456 -0.01342  0.09806  0.94307
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.814e+00  9.259e-02  41.190  < 2e-16 ***
## edu_age     -3.011e-02  4.795e-03  -6.280 3.41e-10 ***
## cog          3.184e-02  1.265e-02   2.516   0.0119 *
## age          2.294e-04  1.471e-04   1.559   0.1189
## male         4.129e-02  1.328e-03  31.092  < 2e-16 ***
## PC1          1.623e-04  3.785e-04   0.429   0.6681
## PC2         -6.167e-04  3.889e-04  -1.586   0.1128
## PC3         -6.375e-04  3.796e-04  -1.679   0.0931 .
## PC4          6.650e-05  2.925e-04   0.227   0.8201
## PC5          1.036e-03  1.340e-04   7.726 1.12e-14 ***
## PC6          7.809e-05  3.588e-04   0.218   0.8277
## PC7          1.357e-04  3.214e-04   0.422   0.6729
## PC8         -2.957e-04  3.216e-04  -0.919   0.3579
## PC9         -7.552e-04  1.540e-04  -4.903 9.45e-07 ***
## PC10         4.316e-04  2.827e-04   1.527   0.1268
##
```

```
## Diagnostic tests:
##                               df1   df2 statistic  p-value
## Weak instruments (edu_age)    89 86048      7.76  < 2e-16 ***
## Weak instruments (cog)        89 86048      7.21  < 2e-16 ***
## Wu-Hausman                     2 86133     16.18 9.43e-08 ***
## Sargan                        87    NA    249.57  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.168 on 86135 degrees of freedom
## Multiple R-Squared: -0.08027,    Adjusted R-squared: -0.08044
## Wald test: 110.3 on 14 and 86135 DF,  p-value: < 2.2e-16
```

```
fsw(indreg)
```

```
##
## Model sample size:  86150
##
## Sanderson-Windmeijer conditional F-statistics for first stage model:
##         F value  d.f. Residual d.f. Pr(>F)
## edu_age 1.82516    88         86048 0.1612
## cog     1.79340    88         86048 0.1664
```

2SLS regression using the weighted scores

```
grsformula <- as.formula(paste("lnbmi ~ edu_age + cog", covars,
                              "cog_grs", "edu_grs", sep = "+"))

scorereg <- ivreg(grsformula, data=dat, model=TRUE)
summary(scorereg, diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = grsformula, data = dat, model = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83436 -0.14848 -0.01108  0.13413  1.36285
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.441e+00  2.725e-01  16.299  < 2e-16 ***
## edu_age     -6.306e-02  1.447e-02  -4.359 1.31e-05 ***
## cog          1.116e-01  3.981e-02   2.804  0.00504 **
## age          2.146e-04  2.192e-04   0.979  0.32776
## male         4.067e-02  1.886e-03  21.565  < 2e-16 ***
## PC1          5.391e-04  5.186e-04   1.040  0.29854
## PC2         -7.387e-04  5.032e-04  -1.468  0.14211
## PC3         -9.667e-04  5.117e-04  -1.889  0.05889 .
## PC4         -3.465e-04  3.989e-04  -0.869  0.38497
## PC5          1.362e-03  2.208e-04   6.168 6.96e-10 ***
## PC6          2.651e-05  4.635e-04   0.057  0.95440
```

```
## PC7         -1.062e-04  4.225e-04  -0.251  0.80161
## PC8         -3.330e-05  4.284e-04  -0.078  0.93804
## PC9         -1.316e-03  2.953e-04  -4.457 8.31e-06 ***
## PC10         1.426e-04  3.792e-04   0.376  0.70678
##
## Diagnostic tests:
##                              df1   df2 statistic  p-value
## Weak instruments (edu_age)     2 86135    275.62  < 2e-16 ***
## Weak instruments (cog)         2 86135    235.51  < 2e-16 ***
## Wu-Hausman                     2 86133     25.09 1.27e-11 ***
## Sargan                         0    NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2164 on 86135 degrees of freedom
## Multiple R-Squared: -0.7919, Adjusted R-squared: -0.7922
## Wald test: 67.19 on 14 and 86135 DF,  p-value: < 2.2e-16
```

```
fsw(scorereg)
```

```
##
## Model sample size:  86150
##
## Sanderson-Windmeijer conditional F-statistics for first stage model:
##          F value  d.f. Residual d.f.     Pr(>F)
## edu_age 24.83207     1         86135 1.6545e-11 ***
## cog     24.64298     1         86135 1.9987e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

\section(Adaptive LASSO results)

Using the unclumped list of SNPs - and all overlapping SNPs in the list for both exposures.

Some SNPs are excluded from the analysis due to not being present in the UK Biobank data rs10191758 - IQ/overlapping rs13010010 - IQ/overlapping rs4728302 - Education

Adaptive lasso based on 10 fold cross validation

```
# MVadap.cv(Y,D,ivs,X,alpha = 0.05)
```

Adaptive lasso based on Sargan testing downward selection

```
# MVadap.dt(Y,D,ivs,X,alpha = 0.05, tuning = 0.1/log(length(Y)))
```

Adaptive lasso based on Sargan testing downward selection with a block structure applied to the SNPs

```
# MVadap.dtblock(Y,D,ivs,index1 = c(1:(lengthedu+lengthboth)), index2 = c((lengthedu+1):ncol(ivs)),X,al
```

2SLS regression with the score excluding the identified SNPs 9 SNPs were removed from the education score and 3 from the cognitive ability score

```r
grsformula <- as.formula(paste("lnbmi ~ edu_age + cog", covars,
                               "cog_grs_update", "edu_grs_update", sep = "+"))

scorereg <- ivreg(grsformula, data=dat, model=TRUE)
summary(scorereg, diagnostics=TRUE)
```

```
##
## Call:
## ivreg(formula = grsformula, data = dat, model = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68613 -0.12599 -0.01267  0.11089  1.12316
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.091e+00  2.243e-01  18.240  < 2e-16 ***
## edu_age     -4.441e-02  1.204e-02  -3.687 0.000227 ***
## cog          6.086e-02  3.626e-02   1.678 0.093281 .
## age          1.186e-04  2.286e-04   0.519 0.603932
## male         4.157e-02  1.782e-03  23.323  < 2e-16 ***
## PC1          2.931e-04  4.461e-04   0.657 0.511224
## PC2         -6.692e-04  4.271e-04  -1.567 0.117141
## PC3         -7.560e-04  4.380e-04  -1.726 0.084387 .
## PC4         -1.618e-04  3.329e-04  -0.486 0.626881
## PC5          1.167e-03  1.894e-04   6.161 7.26e-10 ***
## PC6          7.248e-05  3.944e-04   0.184 0.854215
## PC7          1.473e-05  3.570e-04   0.041 0.967093
## PC8         -1.883e-04  3.641e-04  -0.517 0.605099
## PC9         -1.004e-03  2.454e-04  -4.091 4.30e-05 ***
## PC10         2.943e-04  3.197e-04   0.921 0.357307
##
## Diagnostic tests:
##                               df1   df2 statistic  p-value
## Weak instruments (edu_age)      2 86135    235.81  < 2e-16 ***
## Weak instruments (cog)          2 86135    168.22  < 2e-16 ***
## Wu-Hausman                      2 86133     19.43 3.66e-09 ***
## Sargan                          0    NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1837 on 86135 degrees of freedom
## Multiple R-Squared: -0.2913, Adjusted R-squared: -0.2915
## Wald test: 92.42 on 14 and 86135 DF,  p-value: < 2.2e-16
```

```r
fsw(scorereg)
```

```
##
## Model sample size:  86150
##
## Sanderson-Windmeijer conditional F-statistics for first stage model:
##          F value  d.f. Residual d.f.     Pr(>F)
```

```
## edu_age 23.52768     1         86135 6.0933e-11 ***
## cog     23.06532     1         86135 9.6725e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```