MVMR LASSO analysis

Eleanor Sanderson

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.

1. 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.

```r
covars <- paste(" age + male +", paste0("PC",1:10,collapse = "+"), "|",
                "age + male +", paste0("PC",1:10,collapse = "+"))
#[Note - covriates 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
## -0.70261 -0.12014 -0.01326  0.10574  1.13454
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.953e+00  8.609e-02  45.920  < 2e-16 ***
## edu_age     -3.549e-02  4.330e-03  -8.196 2.52e-16 ***
## cog          3.136e-02  1.089e-02   2.878  0.00400 **
## age         -4.305e-04  1.522e-04  -2.828  0.00469 **
## male         3.948e-02  1.258e-03  31.386  < 2e-16 ***
## PC1          4.831e-04  3.533e-04   1.367  0.17155
## PC2         -3.964e-04  3.660e-04  -1.083  0.27888
## PC3         -2.130e-04  3.545e-04  -0.601  0.54795
## PC4          1.728e-04  2.725e-04   0.634  0.52617
## PC5          7.197e-04  1.243e-04   5.789 7.08e-09 ***
## PC6         -8.772e-04  3.364e-04  -2.608  0.00911 **
## PC7          2.889e-05  3.024e-04   0.096  0.92389
## PC8         -4.074e-04  3.076e-04  -1.325  0.18531
## PC9         -8.112e-04  1.467e-04  -5.530 3.21e-08 ***
## PC10        -1.027e-04  2.694e-04  -0.381  0.70317
##
## Diagnostic tests:
##                               df1    df2 statistic  p-value
## Weak instruments (edu_age)     89 107269     9.574  < 2e-16 ***
## Weak instruments (cog)         89 107269    10.657  < 2e-16 ***
## Wu-Hausman                      2 107354    35.434 4.13e-16 ***
## Sargan                         87     NA   220.021 1.69e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1764 on 107356 degrees of freedom
## Multiple R-Squared: -0.1331, Adjusted R-squared: -0.1332
## Wald test:   115 on 14 and 107356 DF,  p-value: < 2.2e-16
```

```r
fsw(indreg)
```

```
##
```

```
## Model sample size:  107371
##
## Sanderson-Windmeijer conditional F-statistics for first stage model:
##         F value   d.f. Residual d.f.   Pr(>F)
## edu_age 2.57475     88        107269 0.076177 .
## cog     2.64630     88        107269 0.070918 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2SLS regression using the weighted scores

```r
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
## -0.72894 -0.13220 -0.01335  0.11788  1.35415
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.235e+00  1.661e-01  25.490  < 2e-16 ***
## edu_age     -4.972e-02  8.422e-03  -5.903 3.57e-09 ***
## cog          6.290e-02  2.082e-02   3.021 0.002517 **
## age         -8.040e-04  2.377e-04  -3.383 0.000718 ***
## male         3.895e-02  1.473e-03  26.435  < 2e-16 ***
## PC1          5.804e-04  3.919e-04   1.481 0.138616
## PC2         -3.214e-04  4.037e-04  -0.796 0.426012
## PC3         -2.618e-04  3.911e-04  -0.669 0.503229
## PC4         -1.531e-05  3.109e-04  -0.049 0.960721
## PC5          8.249e-04  1.469e-04   5.616 1.96e-08 ***
## PC6         -8.967e-04  3.697e-04  -2.425 0.015297 *
## PC7          5.560e-05  3.332e-04   0.167 0.867458
## PC8         -2.007e-04  3.538e-04  -0.567 0.570435
## PC9         -1.044e-03  1.967e-04  -5.309 1.10e-07 ***
## PC10        -1.231e-04  2.964e-04  -0.415 0.677998
##
## Diagnostic tests:
##                                 df1    df2 statistic  p-value
## Weak instruments (edu_age)        2 107356    337.81  < 2e-16 ***
## Weak instruments (cog)            2 107356    389.54  < 2e-16 ***
## Wu-Hausman                        2 107354     32.94 4.99e-15 ***
## Sargan                            0     NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1937 on 107356 degrees of freedom
```

```
## Multiple R-Squared: -0.3671, Adjusted R-squared: -0.3673
## Wald test: 94.41 on 14 and 107356 DF,  p-value: < 2.2e-16
```

```r
fsw(scorereg)
```

```
##
## Model sample size:  107371
##
## Sanderson-Windmeijer conditional F-statistics for first stage model:
##          F value   d.f. Residual d.f.     Pr(>F)
## edu_age 67.73296      1        107356 < 2.22e-16 ***
## cog     68.64690      1        107356 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```