Lab 11: LASSO for Instrumental Variable Models

Required packages are “foreign”: importing data from STATA; “hdm”: LASSO-based IV estimation; “AER”: robust standard errors and IV estimation.

```
library(foreign)
library(hdm)
library(AER)
```

We apply LASSO-based methods to the data from Angrist and Krueger (1991, Quarterly Journal of Economics) “Does Compulsory School Attendance Affect Schooling and Earnings?”. The main equation of interest is given by

$$\log(\text{Wage}_i) = \alpha \cdot (\text{Education}_i) + X_i'\beta + U_i,$$

where X_i are controls.Angrist and Krueger (1991) propose to use IV estimation with the quarter of birth dummies used as the IVs for education.

Import the data

The data from Angrist and Krueger (1991) is available on the author’s website http://economics.mit.edu/faculty/angrist/data1/data/angkru1991. We import the STATA version of the data set:

```
Angrist<-read.dta("NEW7080.dta")
```

We need to change the names of the variables. The list of the names corresponding to the `v` variables can be found at http://economics.mit.edu/files/5354

```
colnames(Angrist) <-
   c(
  "AGE",
  "AGEQ",
  "v3",
  "EDUC", #education
  "ENOCENT", #region dummy
  "ESOCENT", #region dummy
  "v7",
  "v8",
  "LWKLYWGE", # Log weekly wage
  "MARRIED", #1 if married
  "MIDATL", #region dummy
  "MT", #region dummy
  "NEWENG", #region dummy
  "v14","v15",
  "CENSUS", #70 or #80
  "v17",
  "QOB", #quarter of birth
  "RACE", #1 if black, 0 otherwise
  "SMSA", #region dummy
  "SOATL", #region dummy
  "v22","v23",
  "WNOCENT", #region dummy
  "WSOCENT", #region dummy
  "v26",
  "YOB" #year of birth
  )

Angrist$AGESQ=Angrist$AGEQ^2 #squared age
```

Following the paper, we focus on middle-aged men in the 1980 census. The number of observations in the selected sample is 486926.

```
Angrist804049<-subset(Angrist, CENSUS==80 & YOB>=40 & YOB<=49)
nrow(Angrist804049)
```

```
## [1] 486926
```

OLS and IV estimation:

We first produce OLS estimates. To compute heteroskedasticity robust standard errors, we use the `coeftest()` function from the package “AER”.

```
OLS=lm(LWKLYWGE~EDUC+RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT+SOATL
       +ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ, data=Angrist804049)
coeftest(OLS,vcov=vcovHC(OLS, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                     Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)      -1.7897e+01  4.6878e+00  -3.8178 0.0001347 ***
## EDUC              5.2066e-02  3.4729e-04 149.9211 < 2.2e-16 ***
## RACE             -2.1076e-01  3.6511e-03 -57.7256 < 2.2e-16 ***
## MARRIED           2.4444e-01  2.4765e-03  98.7052 < 2.2e-16 ***
## SMSA             -1.4186e-01  2.3930e-03 -59.2809 < 2.2e-16 ***
## NEWENG           -9.2609e-02  4.2345e-03 -21.8699 < 2.2e-16 ***
## MIDATL           -1.4299e-02  3.2519e-03  -4.3969 1.098e-05 ***
## ENOCENT           4.2969e-02  3.1275e-03  13.7391 < 2.2e-16 ***
## WNOCENT          -7.0006e-02  4.0684e-03 -17.2075 < 2.2e-16 ***
## SOATL            -1.0500e-01  3.2872e-03 -31.9428 < 2.2e-16 ***
## ESOCENT          -1.2024e-01  4.3369e-03 -27.7245 < 2.2e-16 ***
## WSOCENT          -5.8022e-02  3.7910e-03 -15.3053 < 2.2e-16 ***
## MT               -6.7554e-02  4.6095e-03 -14.6553 < 2.2e-16 ***
## as.factor(YOB)41  9.9263e-03  5.4427e-03   1.8238 0.0681848 .
## as.factor(YOB)42  1.5461e-02  7.6888e-03   2.0109 0.0443395 *
## as.factor(YOB)43  2.2842e-02  1.0385e-02   2.1995 0.0278460 *
## as.factor(YOB)44  1.0009e-02  1.3397e-02   0.7471 0.4550288
## as.factor(YOB)45  2.5042e-03  1.6447e-02   0.1523 0.8789802
## as.factor(YOB)46 -1.0406e-02  1.9597e-02  -0.5310 0.5954091
## as.factor(YOB)47 -2.5268e-02  2.2537e-02  -1.1212 0.2622101
## as.factor(YOB)48 -4.5235e-02  2.5703e-02  -1.7599 0.0784227 .
## as.factor(YOB)49 -6.2342e-02  2.8865e-02  -2.1597 0.0307941 *
## AGE              -5.9207e-03  3.2050e-03  -1.8473 0.0647020 .
## AGESQ             6.1686e-06  1.2717e-06   4.8505 1.232e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, we compute the IV estimator using the `ivreg()` command from the package “AER”. We use the quarter of birth dummies as the IVs for education. We generate more IVs by taking interactions between the quarter of birth dummies and the controls.

```
TSLS=ivreg(LWKLYWGE~EDUC+RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT
           +SOATL+ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ
           | as.factor(QOB)+RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT
           +SOATL+ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ
          +as.factor(QOB)*(RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT+SOATL
                              +ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ)
,data=Angrist804049)
coeftest(TSLS,vcov=vcovHC(TSLS, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                      Estimate  Std. Error  t value  Pr(>|t|)    
## (Intercept)      -2.3748e+01  5.0514e+00  -4.7013 2.586e-06 ***
## EDUC              9.8769e-02  1.4100e-02   7.0050 2.473e-12 ***
## RACE             -1.5267e-01  1.7886e-02  -8.5358 < 2.2e-16 ***
## MARRIED           2.4553e-01  2.5348e-03  96.8608 < 2.2e-16 ***
## SMSA             -9.9066e-02  1.3166e-02  -7.5243 5.310e-14 ***
## NEWENG           -7.7379e-02  6.2959e-03 -12.2905 < 2.2e-16 ***
## MIDATL            8.0082e-03  7.4993e-03   1.0679 0.2855850    
## ENOCENT           8.0285e-02  1.1697e-02   6.8638 6.714e-12 ***
## WNOCENT          -5.0155e-02  7.3009e-03  -6.8696 6.444e-12 ***
## SOATL            -6.6016e-02  1.2232e-02  -5.3969 6.782e-08 ***
## ESOCENT          -6.2286e-02  1.8058e-02  -3.4492 0.0005623 ***
## WSOCENT          -2.7775e-02  9.9150e-03  -2.8013 0.0050904 ** 
## MT               -6.6090e-02  4.7802e-03 -13.8260 < 2.2e-16 ***
## as.factor(YOB)41  1.1531e-02  5.5488e-03   2.0782 0.0376937 *  
## as.factor(YOB)42  1.5977e-02  7.8543e-03   2.0342 0.0419278 *  
## as.factor(YOB)43  2.6969e-02  1.0643e-02   2.5339 0.0112806 *  
## as.factor(YOB)44  1.8465e-02  1.3835e-02   1.3347 0.1819896    
## as.factor(YOB)45  1.0304e-02  1.6891e-02   0.6100 0.5418470    
## as.factor(YOB)46 -2.2098e-03  2.0109e-02  -0.1099 0.9124943    
## as.factor(YOB)47 -1.4135e-02  2.3160e-02  -0.6103 0.5416478    
## as.factor(YOB)48 -2.6778e-02  2.6643e-02  -1.0051 0.3148609    
## as.factor(YOB)49 -3.7488e-02  3.0143e-02  -1.2437 0.2136237    
## AGE              -5.2158e-03  3.2943e-03  -1.5833 0.1133573    
## AGESQ             7.5427e-06  1.3519e-06   5.5792 2.418e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can also investigate the first-stage equation. There is a large number of parameters with none of the quarter of birth variables or their interactions significant. This may be due to the fact that there are a large number of regressors.

```
FirstStage=lm(
  EDUC~as.factor(QOB)
  +RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT+SOATL
  +ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ
  + as.factor(QOB)*(RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT
                    +WNOCENT+SOATL+ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ),
  data = Angrist804049)
coeftest(FirstStage,vcov=vcovHC(FirstStage, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                                     Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)                      -4.1416e+09  4.4312e+15  0.0000 0.9999993
## as.factor(QOB)2                   1.0673e+06  1.1455e+12  0.0000 0.9999993
## as.factor(QOB)3                   2.1346e+06  2.2877e+12  0.0000 0.9999993
## as.factor(QOB)4                  -3.5498e+08  4.5528e+14  0.0000 0.9999994
## RACE                             -1.3072e+00  8.7140e+00 -0.1500 0.8807541
## MARRIED                          -1.1299e-02  7.4696e+00 -0.0015 0.9987931
## SMSA                             -9.5187e-01  3.0714e-01 -3.0991 0.0019412 **
## NEWENG                           -3.4996e-01  1.1681e+00 -0.2996 0.7644848
## MIDATL                           -4.9607e-01  8.4864e-01 -0.5846 0.5588476
## ENOCENT                          -8.5748e-01  1.7106e-01 -5.0128 5.366e-07 ***
## WNOCENT                          -4.3332e-01  2.5480e-01 -1.7006 0.0890116 .
## SOATL                            -9.3539e-01  1.1595e-01 -8.0675 7.194e-16 ***
## ESOCENT                          -1.3046e+00  3.8768e-01 -3.3652 0.0007651 ***
## WSOCENT                          -7.7195e-01  6.0025e-01 -1.2860 0.1984297
## MT                               -8.8526e-02  1.3941e-01 -0.6350 0.5254332
## as.factor(YOB)41                  4.2686e+06  4.5711e+12  0.0000 0.9999993
## as.factor(YOB)42                  8.5349e+06  9.1386e+12  0.0000 0.9999993
## as.factor(YOB)43                  1.2799e+07  1.3750e+13  0.0000 0.9999993
## as.factor(YOB)44                  1.7061e+07  1.8236e+13  0.0000 0.9999993
## as.factor(YOB)45                  2.1321e+07  2.2782e+13  0.0000 0.9999993
## as.factor(YOB)46                  2.5578e+07  2.7394e+13  0.0000 0.9999993
## as.factor(YOB)47                  2.9834e+07  3.2041e+13  0.0000 0.9999993
## as.factor(YOB)48                  3.4087e+07  3.6467e+13  0.0000 0.9999993
## as.factor(YOB)49                  3.8338e+07  4.0996e+13  0.0000 0.9999993
## AGESQ                             1.1004e+03  1.1772e+09  0.0000 0.9999993
## as.factor(QOB)2:RACE              7.0317e-03  3.6477e+01  0.0002 0.9998462
## as.factor(QOB)3:RACE              1.3415e-01  1.5873e+02  0.0008 0.9993257
## as.factor(QOB)4:RACE              1.0944e-01  4.2687e+01  0.0026 0.9979543
## as.factor(QOB)2:MARRIED           1.6362e-02  1.0274e+01  0.0016 0.9987293
## as.factor(QOB)3:MARRIED          -3.0879e-02  3.4380e+01 -0.0009 0.9992834
## as.factor(QOB)4:MARRIED          -3.1047e-02  9.4983e+00 -0.0033 0.9973919
## as.factor(QOB)2:SMSA              2.3464e-02  3.2871e-01  0.0714 0.9430943
## as.factor(QOB)3:SMSA              5.0543e-02  5.8058e+00  0.0087 0.9930540
## as.factor(QOB)4:SMSA              6.6162e-02  1.3279e+00  0.0498 0.9602611
## as.factor(QOB)2:NEWENG            3.6949e-02  1.9920e+01  0.0019 0.9985200
## as.factor(QOB)3:NEWENG            2.4745e-02  9.0658e+01  0.0003 0.9997822
## as.factor(QOB)4:NEWENG            2.8934e-02  2.4225e+01  0.0012 0.9990470
## as.factor(QOB)2:MIDATL            3.8055e-02  2.1626e+01  0.0018 0.9985959
## as.factor(QOB)3:MIDATL            3.9182e-02  9.7364e+01  0.0004 0.9996789
## as.factor(QOB)4:MIDATL           -4.9977e-03  2.6110e+01 -0.0002 0.9998473
## as.factor(QOB)2:ENOCENT           5.1180e-02  2.1738e+01  0.0024 0.9981214
## as.factor(QOB)3:ENOCENT           8.9164e-02  9.8671e+01  0.0009 0.9992790
## as.factor(QOB)4:ENOCENT           8.9481e-02  2.6533e+01  0.0034 0.9973091
## as.factor(QOB)2:WNOCENT          -2.1540e-02  1.9661e+01 -0.0011 0.9991259
## as.factor(QOB)3:WNOCENT          -2.9218e-02  9.2733e+01 -0.0003 0.9997486
## as.factor(QOB)4:WNOCENT           8.5247e-02  2.4828e+01  0.0034 0.9972605
## as.factor(QOB)2:SOATL             5.9663e-02  2.4244e+01  0.0025 0.9980365
## as.factor(QOB)3:SOATL             1.7246e-01  1.0982e+02  0.0016 0.9987470
## as.factor(QOB)4:SOATL             1.6199e-01  2.9501e+01  0.0055 0.9956188
## as.factor(QOB)2:ESOCENT          -5.5170e-02  2.2336e+01 -0.0025 0.9980292
## as.factor(QOB)3:ESOCENT           1.2630e-01  1.0536e+02  0.0012 0.9990435
## as.factor(QOB)4:ESOCENT           1.6981e-01  2.8390e+01  0.0060 0.9952276
## as.factor(QOB)2:WSOCENT           1.2281e-01  2.2157e+01  0.0055 0.9955775
## as.factor(QOB)3:WSOCENT           1.9691e-01  1.0278e+02  0.0019 0.9984714
## as.factor(QOB)4:WSOCENT           1.6845e-01  2.7629e+01  0.0061 0.9951355
## as.factor(QOB)2:MT                8.2836e-02  1.9700e+01  0.0042 0.9966449
## as.factor(QOB)3:MT                7.1403e-02  8.9605e+01  0.0008 0.9993642
## as.factor(QOB)4:MT                7.3564e-02  2.4173e+01  0.0030 0.9975718
## as.factor(QOB)2:as.factor(YOB)41 -5.5013e+02  5.8907e+08  0.0000 0.9999993
## as.factor(QOB)3:as.factor(YOB)41 -1.1004e+03  1.1757e+09  0.0000 0.9999993
## as.factor(QOB)4:as.factor(YOB)41  3.6766e+05  4.7091e+11  0.0000 0.9999994
## as.factor(QOB)2:as.factor(YOB)42 -1.1003e+03  1.1744e+09  0.0000 0.9999993
## as.factor(QOB)3:as.factor(YOB)42 -2.2009e+03  2.3567e+09  0.0000 0.9999993
## as.factor(QOB)4:as.factor(YOB)42  7.3513e+05  9.4460e+11  0.0000 0.9999994
## as.factor(QOB)2:as.factor(YOB)43 -1.6507e+03  1.7701e+09  0.0000 0.9999993
## as.factor(QOB)3:as.factor(YOB)43 -3.3014e+03  3.5386e+09  0.0000 0.9999993
## as.factor(QOB)4:as.factor(YOB)43  1.1024e+06  1.4124e+12  0.0000 0.9999994
## as.factor(QOB)2:as.factor(YOB)44 -2.2009e+03  2.3517e+09  0.0000 0.9999993
## as.factor(QOB)3:as.factor(YOB)44 -4.4018e+03  4.7116e+09  0.0000 0.9999993
## as.factor(QOB)4:as.factor(YOB)44  1.4695e+06  1.8821e+12  0.0000 0.9999994
## as.factor(QOB)2:as.factor(YOB)45 -2.7510e+03  2.9502e+09  0.0000 0.9999993
## as.factor(QOB)3:as.factor(YOB)45 -5.5022e+03  5.9128e+09  0.0000 0.9999993
## as.factor(QOB)4:as.factor(YOB)45  1.8364e+06  2.3513e+12  0.0000 0.9999994
## as.factor(QOB)2:as.factor(YOB)46 -3.3013e+03  3.5253e+09  0.0000 0.9999993
## as.factor(QOB)3:as.factor(YOB)46 -6.6025e+03  7.0796e+09  0.0000 0.9999993
## as.factor(QOB)4:as.factor(YOB)46  2.2031e+06  2.8152e+12  0.0000 0.9999994
## as.factor(QOB)2:as.factor(YOB)47 -3.8514e+03  4.1244e+09  0.0000 0.9999993
## as.factor(QOB)3:as.factor(YOB)47 -7.7031e+03  8.2586e+09  0.0000 0.9999993
## as.factor(QOB)4:as.factor(YOB)47  2.5696e+06  3.2949e+12  0.0000 0.9999994
## as.factor(QOB)2:as.factor(YOB)48 -4.4017e+03  4.7155e+09  0.0000 0.9999993
## as.factor(QOB)3:as.factor(YOB)48 -8.8035e+03  9.4240e+09  0.0000 0.9999993
## as.factor(QOB)4:as.factor(YOB)48  2.9360e+06  3.7660e+12  0.0000 0.9999994
## as.factor(QOB)2:as.factor(YOB)49 -4.9520e+03  5.3014e+09  0.0000 0.9999993
## as.factor(QOB)3:as.factor(YOB)49 -9.9040e+03  1.0613e+10  0.0000 0.9999993
## as.factor(QOB)4:as.factor(YOB)49  3.3021e+06  4.2138e+12  0.0000 0.9999994
## as.factor(QOB)4:AGESQ             9.5245e+01  1.2194e+08  0.0000 0.9999994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LASSO-based estimation

We use LASSO to select the relevant controls and IVs, and then post-LASSO to partial out the effect of controls. We use the `rlassoIV()` command from the package "hdm". The command performs LASSO selection, post-LASSO partialling out, and post-LASSO IV estimation. We use the options `select.X=TRUE` for performing selection over controls and `select.Z=TRUE` for performing selection over IVs.

```
PLIV=rlassoIV(LWKLYWGE~EDUC+RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT
              +SOATL+ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ
           | as.factor(QOB)+RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT
           +SOATL+ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ
            + as.factor(QOB)*(RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT
                            +SOATL+ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ)
           ,data=Angrist804049,
           select.X=TRUE,select.Z=TRUE)
```

The post-LASSO IV estimates are:

```
summary(PLIV)
```

```
## Estimates and Significance Testing of the effect of target variables in the IV regression model
##       coeff.     se. t-value p-value
## EDUC 0.06282 0.02141   2.934 0.00334 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The corresponding 95% confidence interval is:

```
PLIVCI=confint(PLIV)
```

```
##           2.5 %    97.5 %
## EDUC 0.02086042 0.1047888
```

In the first stage, LASSO selected only some of the interaction terms for IVs. The following commands find the identities of the LASSO-selected IVs.

```
First=rlasso(EDUC~as.factor(QOB)+RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT
             +SOATL+ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ
          + as.factor(QOB)*(RACE+MARRIED+SMSA+NEWENG+MIDATL+ENOCENT+WNOCENT
        +SOATL+ESOCENT+WSOCENT+MT+as.factor(YOB)+AGE+AGESQ),data=Angrist804049)
which(First$index==TRUE)
```

```
##                             RACE                             SMSA 
##                                4                                6 
##                           MIDATL                          ENOCENT 
##                                8                                9 
##                          WNOCENT                            SOATL 
##                               10                               11 
##                          ESOCENT                          WSOCENT 
##                               12                               13 
##                               MT                 as.factor(YOB)46 
##                               14                               20 
##                 as.factor(YOB)47                 as.factor(YOB)49 
##                               21                               23 
##                              AGE                            AGESQ 
##                               24                               25 
##             as.factor(QOB)2:RACE          as.factor(QOB)3:WNOCENT 
##                               26                               45 
##          as.factor(QOB)2:ESOCENT as.factor(QOB)4:as.factor(YOB)42 
##                               50                               64 
## as.factor(QOB)4:as.factor(YOB)46 as.factor(QOB)2:as.factor(YOB)47 
##                               76                               77
```