EC402_week_7

YULIA

2/22/2018

Questions for IV and Panel IV in R

Preliminaries

Load and install packages, note the new packages!! (toggles are there for future reference if running code on remote servers etc):

```
#installation_needed  <- TRUE
#loading_needed <- TRUE
#package_list <- c('foreign', 'xtable', 'plm','gmm', 'AER','stargazer','readstata13', 'boot', 'arm', 'l
#if(installation_needed){install.packages(package_list, repos='http://cran.us.r-project.org')}
#if(loading_needed){lapply(package_list, require, character.only = TRUE)}
```

Clear the global workspace.

```
rm(list=ls())
library(AER)
```

```
## Loading required package: car

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

```
library(ivpack)
library(plm)
```

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

Let's first understand how IV works before we move on to panel IV. This example is from https://www.r-bloggers.com/a-simple-instrumental-variables-problem/ with edits and comments by RM. It is about the effect of college on wages. As we go through you may notice that there is a mistake in the first part of this example! We'll explore that as we go on here.

```
#Load the data
data("CollegeDistance")
```

Suppose you never took any econometrics and you regress wages on education plus controls:

```
reg_wages_on_education <- lm(wage ~ urban + gender + ethnicity + unemp + education , data=CollegeDistanc
summary(reg_wages_on_education)
```

```
##
## Call:
## lm(formula = wage ~ urban + gender + ethnicity + unemp + education,
##     data = CollegeDistance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3484 -0.8408  0.1808  0.8119  3.9875
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.641490   0.157008  55.039   <2e-16 ***
## urbanyes           0.070117   0.044727   1.568   0.1170
## genderfemale      -0.085242   0.037069  -2.300   0.0215 *
## ethnicityafam     -0.556056   0.052167 -10.659   <2e-16 ***
## ethnicityhispanic -0.544007   0.048670 -11.177   <2e-16 ***
## unemp              0.133101   0.006711  19.834   <2e-16 ***
## education          0.005369   0.010362   0.518   0.6044
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.268 on 4732 degrees of freedom
## Multiple R-squared:  0.1098, Adjusted R-squared:  0.1087
## F-statistic: 97.27 on 6 and 4732 DF,  p-value: < 2.2e-16
```

We know that education is endogenous here.Because of the selection problem our estimate will be biased (with the direction dependent on the effect of the selection. We do not have panel data. So next idea: we need an instrument.
How about the distance between your childhood home and a college campus?
The logic goes something like this:
Distance from a college will strongly predict a decision to pursue a college degree but may not predict wages apart from increased education.

Note: This is a terrible instrument.
We can imagine some problems with an instrument like college distance:
Families who value education may move into neighborhoods close to colleges or neighborhoods near colleges may have stronger job markets. Both of those features may invalidate the instrument by introducing unobserved variables which influence lifetime earnings but cannot be captured in our measure of schooling.

Q1: Let's see if education is "predicted" by childhood distance from college. Use linear regression to check the correlation.

```
first_stage_education_on_distance <- lm(education ~ distance, data=CollegeDistance)
summary(first_stage_education_on_distance)
```

```
##
## Call:
## lm(formula = education ~ distance, data = CollegeDistance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9386 -1.7935 -0.6483  2.0686  4.4968
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.93861    0.03290 423.683  < 2e-16 ***
## distance    -0.07258    0.01127  -6.441  1.3e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.782 on 4737 degrees of freedom
## Multiple R-squared:  0.008683,   Adjusted R-squared:  0.008474
## F-statistic: 41.49 on 1 and 4737 DF,  p-value: 1.301e-10
```

If you like the F test you will notice that the F test is looking good. Let's be mindful of the associated problems with the F test, of course.

Q2. Construct the predicted education level from our instrument

```
CollegeDistance$predicted_education <- predict(first_stage_education_on_distance)
```

Q3. Now perform a manual second stage using the OLS command "lm" again

```
second_stage_wage_education <- lm(wage ~ urban + gender + ethnicity + unemp + predicted_education , dat
summary(second_stage_wage_education)
```

```
##
## Call:
## lm(formula = wage ~ urban + gender + ethnicity + unemp + predicted_education,
##     data = CollegeDistance)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1692 -0.8294  0.1502  0.8482  3.9537
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.053604   1.675314  -1.226   0.2203
## urbanyes            -0.013588   0.046403  -0.293   0.7697
## genderfemale        -0.086700   0.036909  -2.349   0.0189 *
## ethnicityafam       -0.566524   0.051686 -10.961  < 2e-16 ***
## ethnicityhispanic   -0.529088   0.048429 -10.925  < 2e-16 ***
## unemp                0.145806   0.006969  20.922  < 2e-16 ***
## predicted_education  0.774340   0.120372   6.433 1.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.263 on 4732 degrees of freedom
## Multiple R-squared:  0.1175, Adjusted R-squared:  0.1163
## F-statistic:   105 on 6 and 4732 DF,  p-value: < 2.2e-16
```

Now there were actually some issues with what we have just done. To see them, we will use a canned routine to do the right thing.
Introducing: the ivreg command!
The following two formulations are equivalent:

formulation 1

```
IV_wages_education <- ivreg(wage ~ urban + gender + ethnicity + unemp + education | urban + gender + et
                      data = CollegeDistance, x= TRUE)
```

formulation 2

```
IV_wages_education <- ivreg(wage ~ urban + gender + ethnicity + unemp + education |  . - education + di
                            data = CollegeDistance, x= TRUE)

summary(IV_wages_education )
```

```
##
## Call:
## ivreg(formula = wage ~ urban + gender + ethnicity + unemp + education |
##     . - education + distance, data = CollegeDistance, x = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.20896 -1.14578 -0.02361  1.33303  4.77571
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.35903    1.90830  -0.188   0.8508
## urbanyes           0.04614    0.06039   0.764   0.4449
## genderfemale      -0.07075    0.04997  -1.416   0.1569
## ethnicityafam     -0.22724    0.09863  -2.304   0.0213 *
## ethnicityhispanic -0.35129    0.07706  -4.559 5.28e-06 ***
## unemp              0.13916    0.00912  15.259  < 2e-16 ***
## education          0.64710    0.13594   4.760 1.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 4732 degrees of freedom
## Multiple R-Squared: -0.6118, Adjusted R-squared: -0.6138
## Wald test: 57.48 on 6 and 4732 DF,  p-value: < 2.2e-16
```

```
summary(IV_wages_education , vcov = sandwich, df = Inf, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = wage ~ urban + gender + ethnicity + unemp + education |
##     . - education + distance, data = CollegeDistance, x = TRUE)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.20896 -1.14578 -0.02361  1.33303  4.77571
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.35903    1.91755  -0.187   0.8515
## urbanyes           0.04614    0.05926   0.779   0.4362
## genderfemale      -0.07075    0.04974  -1.422   0.1549
## ethnicityafam     -0.22724    0.09539  -2.382   0.0172 *
## ethnicityhispanic -0.35129    0.07577  -4.636 3.55e-06 ***
## unemp              0.13916    0.00934  14.899  < 2e-16 ***
## education          0.64710    0.13691   4.727 2.28e-06 ***
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    1 4732     50.19 1.60e-12 ***
## Wu-Hausman          1 4731     40.30 2.38e-10 ***
## Sargan              0   NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on Inf degrees of freedom
## Multiple R-Squared: -0.6118, Adjusted R-squared: -0.6138
## Wald test: 342.5 on 6 DF,  p-value: < 2.2e-16
```

Ah but wait… this doesnt match our manual work?
It’s because the example missed out something crucial!!

Q4: What was wrong with what we just did manually? There are two issues.
First issue: the OLS doesn’t realise the predicted variable is constructed, so it understates the uncertainty.
Second (more important) issue: for a first stage you must regress the endogenous variable on ALL THE EXOGENOUS ONES

Q5(A): Implement the manual procedure which fixes the most serious issue.

Q5(B): Use IV reg to see if the results change when you don’t condition on gender.
What does it mean to not condition on gender here? Do I remove it from both sides or do I instrumnet it?

```
IV_wages_education_nogender <- ivreg(wage ~ urban + ethnicity + unemp + education |  urban + ethnicity +
                                     data = CollegeDistance, x= TRUE)
summary(IV_wages_education_nogender)
```

```
##
## Call:
## ivreg(formula = wage ~ urban + ethnicity + unemp + education |
##     urban + ethnicity + unemp + distance, data = CollegeDistance,
##     x = TRUE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1662 -1.1327 -0.0349  1.3385  4.7409
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.379445   1.906040  -0.199   0.8422    
## urbanyes           0.047427   0.060354   0.786   0.4320    
## ethnicityafam     -0.231751   0.098667  -2.349   0.0189 *  
## ethnicityhispanic -0.350596   0.076989  -4.554 5.40e-06 ***
## unemp              0.138756   0.009112  15.228  < 2e-16 ***
## education          0.646012   0.135901   4.754 2.06e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.705 on 4733 degrees of freedom
## Multiple R-Squared: -0.61,   Adjusted R-squared: -0.6117 
## Wald test: 68.46 on 5 and 4733 DF,  p-value: < 2.2e-16
```

Q6: Implement the Anderson Rubin Confidence Interval for the beta of interest (education) (Hint: this is easy in the ivpack package)

```
ar_ci <- anderson.rubin.ci(IV_wages_education, conflevel = 0.95)
ar_ci
```

```
## $confidence.interval
## [1] "[ 0.418294456012369 , 0.983198866413292 ]"
```

Now how about panel IV?
We will look at this in an empirical application of linear panel data methods to the effect of democracy on economic growth based on Acemoglu, Naidu, Restrepo and Robinson (2005), “Democracy Does Cause Growth” in the JPE.
This code is based on Chernozhukov, Fernandez-Val, Demirer and Semenova.
Make sure we download the data into the working directory! And/or set the right working directory in the console.

```
library(foreign)
setwd("/Users/yuliav/Downloads/")
data_democracy <- read.dta("democracy-balanced-l4.dta")
#data_democracy <- pdata.frame(data, index = c("id","year"))
```

Q6. Perform OLS regression of log gdp on democracy plus 4 lags of log gdp and year dummies.

```
lgdp_democracy <- lm(lgdp ~ dem + year + lag(data_democracy$lgdp, k = 4), data = data_democracy)
summary(lgdp_democracy)
```

```
## Warning in summary.lm(lgdp_democracy): essentially perfect fit: summary may
## be unreliable
```

```
##
## Call:
## lm(formula = lgdp ~ dem + year + lag(data_democracy$lgdp, k = 4),
##     data = data_democracy)
##
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -3.645e-15 -1.260e-16 -4.700e-17  2.800e-17  1.583e-13
```

```
##
## Coefficients:
##                                Estimate Std. Error   t value Pr(>|t|)
## (Intercept)                   7.032e-15  1.433e-14 4.910e-01    0.624
## dem                           2.261e-15  1.070e-16 2.113e+01   <2e-16
## year                          2.628e-18  7.179e-18 3.660e-01    0.714
## lag(data_democracy$lgdp, k = 4) 1.000e+00  3.191e-17 3.133e+16   <2e-16
##
## (Intercept)
## dem                           ***
## year
## lag(data_democracy$lgdp, k = 4) ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.729e-15 on 3377 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 3.91e+32 on 3 and 3377 DF,  p-value: < 2.2e-16
```

Q7. Perform Fixed Effects estimation of log gdp on democracy plus 4 lags of log gdp, with two way FEs.

```
linear_model_lgdp_democracy <- lgdp ~ dem + lag(data_democracy$lgdp, k = 4)

fixed_effects_fit <- plm(linear_model_lgdp_democracy, data_democracy, model="within", effect = "twoways

summary(fixed_effects_fit)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = linear_model_lgdp_democracy, data = data_democracy,
##     effect = "twoways", model = "within", index = c("id", "year"))
##
## Balanced Panel: n = 147, T = 23, N = 3381
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -2.07e-14 -4.32e-18  7.81e-20  4.93e-18  3.10e-16
##
## Coefficients:
##                                   Estimate  Std. Error     t-value
## dem                            -1.9171e-16  2.6527e-17 -7.2271e+00
## lag(data_democracy$lgdp, k = 4)  1.0000e+00  3.6212e-17  2.7615e+16
##                                  Pr(>|t|)
## dem                            6.135e-13 ***
## lag(data_democracy$lgdp, k = 4) < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    103.66
## Residual Sum of Squares: 4.311e-28
## R-Squared:      1
## Adj. R-Squared: 1
## F-statistic: 3.85921e+32 on 2 and 3210 DF, p-value: < 2.22e-16
```

Q8. Perform Anderson-Hsiao estimation by instrumenting for democracy with 1 lag of democracy, and instrumenting for 1 lag of growth with the 2nd lag of growth

Hint: use pgmm not plm for this one, and use model = "twosteps"

```
AH_estimation <- pgmm(lgdp ~ dem +  lag(data_democracy$lgdp, k = 1)| lag(data_democracy$dem, k = 1) + la
           data = data_democracy, effect = "twoways", model = "twosteps", index=c("id", "year"))
```

```
## Warning in pgmm(lgdp ~ dem + lag(data_democracy$lgdp, k = 1) |
## lag(data_democracy$dem, : the second-step matrix is singular, a general
## inverse is used
```

Q9. Perform Arellano-Bond estimation using all possible lags as instruments assuming sequential exogeneity

hint: continue with the general estimation protocol of Q8

Notice that the second step matrix was singular and the computer implemented a general inverse. This can mean there was too little variation in the instruments. We will discuss further when we do GMM.