Exercise Class - Econometrics Class 4

Instructor: Irene Iodice
Email: irene.iodice@malix.univ-paris1.fr

Ex.1: Review of the concepts [mostly taken from your book, end chapter 12]

1. *Take the following demand curve regression model for butter*:

$$ln(Q_i^{butter}) = \beta_0 + \beta_1 log(P_i^{butter}) + u_i \tag{1}$$

is $log(P_i^{butter})$ positively or negatively correlated with the error, u_i If β_1 is estimated by OLS, would you expect the estimated value to be larger or smaller than the true value of β_1? Explain.

2. *In the following regression model*:

$$crime\_rate_{state} = \beta_0 + \beta_1 incarceration\_rate_{state} + u_{state} \tag{2}$$

discuss the validity of the number of lawyers per capita as an instrument. Then imagine that in the original regression you add as control the number of lawyers and the number of inhabitants. Comment what changes in your considerations.

3. *In their study of the effectiveness of a treatment for cardiac catherization, McClellan, McNeil and Newhouse (1994) used as an instrument to the fact of receiving the treatment the distance of a patient to regular hospitals. How could you determine whether this instrument is relevant? And whether it is exogenous?*

Ex.2: IV knowing the data generating process [A similar ex. can be found on Prof. Nathaniel Higgins website]

Read the code below that generates an artificial dataset for x_1, x_2, x_3, x_4, u and y with certain characteristics.

Note that you have defined the true model for y (the true data generating process of y) as:

$$y = 5 + 2x_1 - 15x_2 + u \tag{3}$$

then note that a proper estimation of this model should be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{4}$$

Let's imagine that are interested in the impact of x_1 on y. Suppose that x_2 is unobservable and we look at the relationship between y and x_1 without controlling for x_2.

$$y = b_0 + b_1 x + w \tag{5}$$

Figure 1: Data generating process

```
library(MASS)
library(AER)
# 1) Set the seed for replicability
set.seed(2)

# 2) generate a covariance matrix for the RV x1_temp, x2, x3 starting from corr and sdev
  Corr_matrix = matrix(
     c(1 , 0.7 , 0.5, 0.7 , 1 , 0.5,
         0.5 , 0.5 , 1), nrow=3, ncol=3, byrow = TRUE)
  stdevs = c(0.5 ,2 ,1)
  stdevs_matrix = stdevs %*% t(stdevs)
  Cov_matrix= Corr_matrix/stdevs_matrix

# 3) Define the first moments of x1_temp, x2, x3
  mu=c(3,2,2)

# 5) Draw three random variables from a multivariate distribution
  sample = data.frame(mvrnorm(n = 100, mu, Cov_matrix, empirical = FALSE,
                              EISPACK = FALSE))
  colnames(sample) <- c("x1_temp","x2","x3")
  attach(sample)

# 6) Draw a fourth RV independently of the others
  x4 = rnorm(100, mean=1, sd=3)

# 7) Introduce some x4 and x3 into x1
  x1 = x1_temp + x3 + x4

# 8) Randomly add some "unobservable" variation in y
  u = rnorm(100, mean = 0, sd = 1)

# 9) Create the dependent variable y
  y = 5 +2*x1 - 15*x2 + u
```

1. *Why do you expect $\hat{b}_1$ to be biased? In what direction is the bias? Do you expect the estimated coefficient to be too high or too low relative to the true value?*

2. *Regress y on x_1 and record what happens, in particular look from Figure 2 at whether the true value of the coefficient is contained in the 95 % confidence interval of b_1 (assuming large sample size)?*

3. *Which type of instrumental variables do you have at hand? Explain, starting from the data generating process which is the best candidate to be an IV.*

4. *Given your answer in point 3, discuss the TSLS you would conduct to get an consistent estimate of β_1. Then look at the results of the second stage in Figure 4 and discuss.*

5. *Now let's assume you do not know the DGP, how would you test the instrument relevance of x_4? Perform a test using the info from Figure 3.*

6. *Discuss what are the consequences of having a weak instrument.*

7. *Still let's say that since you did not know the DGP you have picked both IVs x_3 and x_4. Which type of evidence may suggest you that one of the two is not exogenous? (do not perform any test, just argue why the results proposed below in figure 4 and 5 tell you something about this.)*

Figure 2: Regression estimates of (5)

```
>   summary(lm(y~x1))

Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-14.1167  -5.5232  -0.3455   5.2390  19.5283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.9166     1.4396 -11.751  < 2e-16 ***
x1            0.9327     0.1803   5.174 1.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.37 on 98 degrees of freedom
Multiple R-squared:  0.2145,    Adjusted R-squared:  0.2065
F-statistic: 26.77 on 1 and 98 DF,  p-value: 1.219e-06
```

Figure 3: 1st stage of TSLS when using x_4 as IV

```
>   summary(stage1)

Call:
lm(formula = x1 ~ x4)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5125 -2.4389 -0.0312  2.2516  5.6788

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.9180     0.3806  12.921  < 2e-16 ***
x4            0.9996     0.1142   8.755 6.06e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.093 on 98 degrees of freedom
Multiple R-squared:  0.4389,    Adjusted R-squared:  0.4332
F-statistic: 76.65 on 1 and 98 DF,  p-value: 6.059e-14
```

Ex.3: IV not knowing the data generating process but having a sample of info [A similar ex. on www.r-exercises.com]

Consider the simple Ordinary Least Squares (OLS) regression setting in which we model wages as a function of years of schooling (education):

$$log(wage_i) = \beta_0 + \beta_1 education_i + u_i \tag{6}$$

1. *From the thousands of similar ex. done in class, explain why you think that estimating this model would not give a reasonable estimate of the effect of education. What would you do if you had infinite resources available (in terms of info availability)?*

   Now we load the PSID1976 dataset provided within the AER package. This has data regarding labor force participation of married women sourced from: Mroz, T. A. (1987) The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. Econometrica 55, 765–799.

2. *Look at the summary statistics for the data in Figure 6 and identify possible candidates as instrumental variables for education.*

3. *Let's say you pick mother and father education as IVs, so that you have an overidentified system. You are a skeptical on their exogeneity, while you know they are relavant*

Figure 4: 2nd stage of TSLS when using x_4 as IV

```
> summary(stage2)

Call:
lm(formula = y ~ x1_hat)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8822  -3.8017  -0.1132   3.3671  14.6207

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.8492     1.5349  -16.84   <2e-16 ***
x1_hat        2.2349     0.2081   10.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.637 on 98 degrees of freedom
Multiple R-squared:  0.5406,    Adjusted R-squared:  0.5359
F-statistic: 115.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

Figure 5: 2nd stage of TSLS when using x_3 as IV

```
> summary(stage2_x3)

Call:
lm(formula = y ~ x1_hat)

Residuals:
     Min       1Q   Median       3Q      Max
-22.8370  -5.5652  -0.4302   5.6891  17.6420

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.98854    2.48729  -4.418 2.57e-05 ***
x1_hat        0.06851    0.34174   0.200    0.842
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.314 on 98 degrees of freedom
Multiple R-squared:  0.0004099, Adjusted R-squared:  -0.00979
F-statistic: 0.04019 on 1 and 98 DF,  p-value: 0.8415
```

instruments, how can you check their validity? (In case you need, $\chi^2_{1,5\%} = 3.84$.*)*

Figure 6: Summary info on PSID1976 dataset

```
participation     hours          youngkids        oldkids          age          education         wage
no :325       Min.   :   0.0   Min.   :0.0000   Min.   :0.000   Min.   :30.00   Min.   : 5.00   Min.   : 0.000
yes:428       1st Qu.:   0.0   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:36.00   1st Qu.:12.00   1st Qu.: 0.000
              Median : 288.0   Median :0.0000   Median :1.000   Median :43.00   Median :12.00   Median : 1.625
              Mean   : 740.6   Mean   :0.2377   Mean   :1.353   Mean   :42.54   Mean   :12.29   Mean   : 2.375
              3rd Qu.:1516.0   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:49.00   3rd Qu.:13.00   3rd Qu.: 3.788
              Max.   :4950.0   Max.   :3.0000   Max.   :8.000   Max.   :60.00   Max.   :17.00   Max.   :25.000
   repwage         hhours          hage         heducation        hwage            fincome          tax
Min.   :0.00   Min.   : 175   Min.   :30.00   Min.   : 3.00   Min.   : 0.4121   Min.   : 1500   Min.   :0.4415
1st Qu.:0.00   1st Qu.:1928   1st Qu.:38.00   1st Qu.:11.00   1st Qu.: 4.7883   1st Qu.:15428   1st Qu.:0.6215
Median :0.00   Median :2164   Median :46.00   Median :12.00   Median : 6.9758   Median :20880   Median :0.6915
Mean   :1.85   Mean   :2267   Mean   :45.12   Mean   :12.49   Mean   : 7.4822   Mean   :23081   Mean   :0.6789
3rd Qu.:3.58   3rd Qu.:2553   3rd Qu.:52.00   3rd Qu.:15.00   3rd Qu.: 9.1667   3rd Qu.:28200   3rd Qu.:0.7215
Max.   :9.98   Max.   :5010   Max.   :60.00   Max.   :17.00   Max.   :40.5090   Max.   :96000   Max.   :0.9415
  meducation       feducation         unemp         city       experience    college   hcollege
Min.   : 0.000   Min.   : 0.000   Min.   : 3.000   no :269   Min.   : 0.00   no :541   no :458
1st Qu.: 7.000   1st Qu.: 7.000   1st Qu.: 7.500   yes:484   1st Qu.: 4.00   yes:212   yes:295
Median :10.000   Median : 7.000   Median : 7.500             Median : 9.00
Mean   : 9.251   Mean   : 8.809   Mean   : 8.624             Mean   :10.63
3rd Qu.:12.000   3rd Qu.:12.000   3rd Qu.:11.000             3rd Qu.:15.00
Max.   :17.000   Max.   :17.000   Max.   :14.000             Max.   :45.00
```

Figure 7: Exogeneity test results

```
> test_overidentif <- lm(hat_u~meducation+feducation, data= subset(PSID1976,
+                                                    participation =="yes")) #2nd stage
> summary(test_overidentif)

Call:
lm(formula = hat_u ~ meducation + feducation, data = subset(PSID1976,
    participation == "yes"))

Residuals:
    Min      1Q  Median      3Q     Max
-3.2146 -0.3758  0.0574  0.4141  2.0623

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.010914   0.113840   0.096    0.924
meducation  -0.006599   0.012700  -0.520    0.604
feducation   0.005772   0.011924   0.484    0.629

Residual standard error: 0.7227 on 425 degrees of freedom
Multiple R-squared:  0.0007654, Adjusted R-squared:  -0.003937
F-statistic: 0.1628 on 2 and 425 DF,  p-value: 0.8498
```