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.

Sketch demand and supply in a price-quantity graph. Since we estimate the demand curve model, then an increase in the error u_i translates into a rightward shift of the demand curve. This leads to both a rise in quantity and price of butter. Then $log(P_i^{butter})$ is positively correlated with the regression error, then the OLS estimator is inconsistent and is larger wrt the true value of β_1 (the direction of the difference is given by the sign of the covariance between the error term and the dependent variable $log(P_i^{butter})$!).

Note that in the following demands, there is no right answer. You need to provide evidence that you are familiar with the type of reasoning you need to apply when assessing as an "educated guess" the validity of instruments.

2. *In the study of cigarette demand you use the following model:*

$$ln(Q_{state}) = \beta_0 + \beta_1 log(P_{state}) + u_{state} \tag{2}$$

Now suppose that we used as an instrument the number of trees per capita in the state. Is it a valid instrument?

The number of trees per capita in the state is valid if:

- exogenous: probably yes, because it is plausibly uncorrelated with the error in the demand function.
- relevant: probably no, I personally do not see why this should hold. I am open to your intuitions!

Then is not a valid IV.

3. *In the following regression model*:

$$crime\_rates_{state} = \beta_0 + \beta_1 incarceration\_rates + u_{state} \tag{3}$$

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.

The number of lawyers per capita in the state is valid if:

- exogenous: probably no, but the direction is arguable. ON one side, it is reasonable that states with higher than expected crime rates (with positive regression errors) are likely to have more lawyers (criminals must be defended and prosecuted), so the number of lawyers will be positively correlated with the regression error. On the other side, we might expect that the incentive to commit a crime is negatively correlated with the quality of the system of justice, and thus the number of lawyers.
- relevant: probably yes, the number of lawyers is arguably correlated with the incarceration rate. Then is not a valid IV.
- Even if it was valid, but we would include the number of inhabitants and layers in the state, this would make our IV not useful since it would be perfect multicollinear with the other regressors in the model: indeed recall the instruments Z has to explain a part a X not explained by the exogenous regressors W included in the second stage.

4. *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?*

   Distance to an hospital is:

   - relevant: if it is correlated with the the patient receiving cardiac catheterization. To check for non weak instruments we run an F-test on the first stage regression which in this case is:

$$cardiac\_catherization_i = \gamma_0 + \gamma_1 distance_i + u_i \tag{4}$$

     and testing the hypothesis: $H_0 = 0$ vs $H_1 \neq 0$ with a "rule of thumb type" rejection rule: reject H_0 if the F statistics is >10.
   - exogenous: not correlated with the error term, in this case this means not correlated with health status. Checking instrument exogeneity is more difficult. Only when there are more instruments than endogenous regressors the joint exogeneity of the instruments can be tested using test of overidentifying restrictions. However, when the number of instruments is equal to the number of endogenous regressors, then it is impossible to test for exogeneity statistically. In the authors' study there is one endogenous regressor (treatment) and one instrument (distance from hospital), so the J-test cannot be used. The exogeneity assumption in this case should be proposed as an educated guess.

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:

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

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{6}$$

Figure 1: Data generating process

```r
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
```

Let's imagine that we are interested in the impact of x_1 on y. Suppose that x_2 is unobservable (like the skill of a person, its innate talent, or just not measurable) but in this case we have not information over multiple years (panel database), that we can use to control for unobserved time-constant heterogeneity. This forces us to look at the relationship between y and x_1 without controlling for x_2.

$$y = b_0 + b_1 x_1 + w \tag{7}$$

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?*

   Remember that if x_2 were uncorrelated with x_1, we wouldn't have a problem. Although including x_2 in a regression would help us to predict y better, excluding x_2 would not bias our estimate of the impact of x_1 on y. But this is not the case. You can tell by looking at the covariance matrix that x_1 and x_2 are correlated. We also know that x_2 is included in y and thus x_2 will be in the error term in the equation in (7). The fact that x_2 correlates both with x_1 and y tell us that there is OVB. Which sign? When x_1 increases, two things happen. First, y increases (β_1 is positive). Second, because x_2 is positively correlated with x_1, when x_1 increases x_2 is likely to be increasing too. And when x_2 increases, y decreases, then if we do not include it we just see x_1 increasing and

y either decreasing, or at least increasing less than it should due to the hidden effect of x_2. In short: our estimate of x_1 will be biased downwards, i.e. the bias is negative and the estimated coefficient is too low.

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

Figure 2: Regression results in (7)

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

Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2057  -5.0886  -0.3442   4.9111  18.2277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.4804     1.3463 -11.498  < 2e-16 ***
x1            1.0061     0.1686   5.967 3.85e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.893 on 98 degrees of freedom
Multiple R-squared:  0.2665,    Adjusted R-squared:  0.259
F-statistic: 35.61 on 1 and 98 DF,  p-value: 3.851e-08
```

We see that $\hat{b_1}$ is smaller than the true β_1. The CI is given by: $\hat{b_1} \pm SE(\hat{b_1})t_{2.5\%} = 0.9327 \pm 1.97 \times 0.18035$, i.e. $0.577 < b_1 < 1.29$, this to say that the CI of b_1 does not contain the true β_1.

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.*

There are two viable options: x_3 and x_4. Are they both valid?

- relevance: both x_3 and x_4 enters in the definition of x_1, this means that they are correlated.
- exogeneity: Looking at the code, we see that x_3 is correlated with x_2. In this case, that's not good. The omission of x_2 is what is causing the endogeneity. That is, x_2 is in u. And if x_3 is correlated with u, x_3 is not exogenous. x_4 on the other hand, is drawn independently from x_2.

This makes x_4 the only valid instrument. Indeed, x_4 is a component of x_1, and importantly a component that has nothing to do with the part of x_1 that is correlated with x_2. So x_4 is the candidate instrument. Indeed recall from slides that a good IV is able to generate an exogenous variation in X, that is a variation that is not related to error term.

4. *Given the choice in the point above discuss the TSLS you would conduct to get a consistent estimate of β_1. Then look at the results of the second stage and discuss.*

   (a) the first stage decomposes x_1 into two components: a problematic component that can be correlated with u and another problem-free component that is not correlated with u

   $$x_1 = \pi_0 + \pi_1 x_4 + v \tag{8}$$

   then $\hat{x}_1 = \hat{\pi_0} + \hat{\pi_1} x_4$ is the part of x_1 that can be predicted by x_4, and since x_4 is not correlated with x_2, then $\hat{x}_1$ is not correlated with u (it is exogenous part).

(b) the second stage uses the problem-free component to estimate β_1.

$$y = \beta_0 + \beta_1 \hat{x}_1 + u \tag{9}$$

We can see that the estimate of β_1 through TSLS, also called $\hat{\beta}^{TSLS}$ is closer to the true β_1, and indeed the CI is $\hat{\beta}_1^{TSLS} \pm SE(\hat{\beta}_1^{TSLS}) \times t_{\alpha/2} = 2.23 \pm 1.97 \times 0.2081$ and thus $1.82 < \beta_1 < 2.64$, which contains the true β_1.

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
```

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
```

5. *Now let's assume you do not know the DGP and you are given the results below. How would you test the instrument relevance of x_4? Perform a test using the info in fig. 3*

With one endogenous regressor a commonly implemented way to check for weak instruments is to do and F-test on the first stage regression testing the hypothesis: H_0 the instrument is weak vs H_1 the instrument is not weak, with a "rule of thumb type" rejection rule: reject H_0 if the F statistics is >10. In addition we have to assume that v from

$$x_1 = \pi_0 + \pi_1 x_4 + v \tag{10}$$

is homoskedastic. In our case $R^2 = 0.44$ and $n = 100$ and thus the F-statistic:

$$F = \frac{(R^2 - 0)/q}{(1 - R^2)/(n - (k + 1))} = \frac{0.4389}{0.6611/(100 - 2)} = 76.65 \tag{11}$$

Where q is 1 (is equal to the number of instruments m) and k is one (equal to number of endogenous variables. Since 76.65>10 we can conclude that the instrument is not weak.

6. *Discuss what are the consequences of having a weak instrument.*
   In class you have seen that the TSLS estimator is consistent, that is:

$$\hat{\beta}_1^{TSLS} = \frac{S_{x_4,y}}{S_{x4,x_1}} \rightarrow \frac{Cov(x_4, y)}{Cov(x_4, x_1)} = \beta_1 \tag{12}$$

   If you have an extremely weak instrument Z, $Cov(x_4, x_1) \approx 0$, then the consistency argument breaks and moreover β^{TSLS} is not asymptotically Normal anymore (look slide 35 for this).

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 tell you something about this.)*
   We have 1 endogenous regressor x_1 and 2 instruments, x_3 and x_4 . Since you have two instruments you can compute two different estimates: β^{TSLS} and β^{TSLS}. Then:

   (a) if x_3 and x_4 are both exogenous, then the two β^{TSLS} will tend to be close to each other;

   (b) if they are very different there must something wrong in either one of the two or both.

   Now look at the estimate of β^{TSLS} when using x_3 as IV. You can see that the value (0.07) is very different from the one estimated when using x_4 as an IV (2.26). As expected this tells us that one of the two variables is not reliable, which we know, as authors of of this DGP, being x_3.

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
```

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):

$$log(wage_i) = \beta_0 + \beta_1 education_i + u_i \tag{13}$$

1. *From the thousands of 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 (collaborators for data athering etc.)?*

   As we already know,in this setting the ZCM assumption is likely to be violated since the level of education (the independent variable) is reasonable correlated with the error term (in general this is called endogeneity problem). To break this correlation we could:

   - include the omitted regressors in the equation, like IQ etc. or do a multiple stage procedure to wash out the part of education that correlates with the error (and thus for example IQ).
   - identify a variable (called instrumental variable) that is not part of the model, has high correlation with education (the endogenous variable) and is uncorrelated with the error term (wage).
   - as we discussed in class you cannot perform a within transformation to get rid of the fixed effect (ex. innate ability) to estimate the effect of education on wage. This is so at least as far as education is a constant variable at the individual level over the sample of years (which is usually true when individuals earn a wage, and thus are in the labor force), then demeaning this variable with the within transformation would cancel out education from the model!

   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 and identify possible candidates as instrumental variables for education.*

```
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
```

3. *Look at the summary statistics for the data and identify possible candidates as instrumental variables for education.*

The possible candidates for instrumental variable (IV) is father's and mother's education, which are very likely relevant instruments, since we expect the education of a person to be correlated with those of their relatives. Exogeneity is more debatable: we need parents' education not to be related with innate ability, such as IQ etc.

4. *Let's say you choose the variable mother and father education, so that you have an overidentified system. You are a bit skeptical on their exogeneity, how can you check if they are valid instruments? (In case you need,* $\chi^2_{1,5\%} = 3.84$*.)*

   We want to check exogeneity of the two instruments and since we have an overidentified system, i.e. m=2 IVs for k=1 endogenous variable, *education*, we can use the following procedure:

   (a) We estimate the residual $\hat{u}^{TSLS}$ from TSLS estimation as:

   $$log(wage) = \beta_0 + \beta_i \hat{education} + u \tag{14}$$

   (b) We regress the residuals on all the exogenous variables.

   $$\hat{u}^{TSLS} = \delta_0 + \delta_1 feducation + \delta_1 meducation + e \tag{15}$$

   and let F denote the homoskedastic-only F statistic testing the hypothesis $\delta_1 = \delta_0 = 0$. Then under H_0 (+ large sample + meducation and feducation are not weak (you might test this with single tests as done in the exercise above) + homoskedasticity of e) $J = mF \to \chi^2_{m-k}$ where (m-k) is known as the degree of overidentification, i.e. number of instruments minus the number of endogenous regressors. Take the $R^2_{unrestr}$ from results above and recall that the $R^2_{restr} = 0$ since it is the model with no explanatory variables.

   $$F = \frac{(R^2_{unrestr} - R^2_{restr})/q}{(1 - R^2_{unrestr})/(n - (k+1))} = \frac{0.007/2}{0.993/(427-2)} = 1.49 \tag{16}$$

   with $q = m = 2$ and $k = 1$, and then $J = 2 \times 1.49 = 3$, which is lower than the critical value $\chi^2_{1,5\%} = 3.84$, which means that we fail to reject the null hypothesis, suggesting that all instruments are exogenous.

```
> 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
```