Exercise Class - Econometrics Class 5

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

1 Part 1: Conclude exercises from last class

Ex.1: Review of the concepts [mostly taken from your book, end chapter 12]
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:
$$cardiac\_catherization_i = \gamma_0 + \gamma_1 distance_i + u_i \tag{1}$$
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]

1. *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
$$x_1 = \pi_0 + \pi_1 x_4 + v \tag{2}$$
is homoskedastic. In our case $R^2 = 0.44$ and $n = 100$ and thus the F-statistic:
$$F = \frac{(R^2 - 0)/q}{(1 - R^2)/(n - (k + 1))} = \frac{0.4389}{0.6611/(100 - 2)} = 76.65 \tag{3}$$
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.

Figure 1: 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
```

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

$$\hat{\beta}_1^{TSLS} = \frac{S_{x_4,y}}{S_{x4,x_1}} \rightarrow \frac{Cov(x_4, y)}{Cov(x_4, x_1)} = \beta_1 \tag{4}$$

   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).

3. *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.

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):

$$log(wage_i) = \beta_0 + \beta_1 education_i + u_i \tag{5}$$

Figure 2: 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
```

2 Part 2: Exercise from last year exam on IV

Consider the following regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{6}$$

where $E[u_i|X_i] \neq 0$, that is with X is endogenous. Consider that in the database you have another variable Z.

1. Under which conditions is Z a valid instrument for X to be used in a TSLS regression?

2. Why having a valid instrument is important?
   Because it is able to generate an exogenous variation in X, that is a variation that is not related to error term.

3. Imagine that the R^2 associated with $X_i = \pi_0 + \pi_1 Z_i + v_i$, with $E[v_i|Z_i] = 0$ is $R^2 = 0.1$ with $n = 50$. Assume that v is conditionally homoskedastic and test if Z is a weak instrument.
   In our case $R^2 = 0.1$ and $n = 50$ and thus the F-statistic:

$$F = \frac{(R^2 - 0)/q}{(1 - R^2)/(n - (k + 1))} = \frac{0.1}{0.9/(50 - 2)} = 5.3 \tag{7}$$

   Since $5.3 < 10$ we can conclude that the instrument is weak.

4. If n=1000 would your answer to the previous question change? Comment.
   With n=1000, F=110.8 and the the same instrument is not weak.

5. Describe how would you test in this case the exogeneity of Z.
   With only one instrument you cannot!

3 Part 3: We look to OVB exercise we started in the review class (those of you acquainted with the topic may leave!)

Pampilio Piratta is deciding if it is worthwhile to work an extra year at his enterprise or to go back to study for a master. With this aim he is exploring the relation between wage, education and tenure. Sadly Pampilio Piratta's research efforts are limited by the fact that he knows how to estimate linear regression models only if they contain one single regressor. Then he estimates the following three models:

i. $log(WAGE_i) = b_0 + b_1 EDUC_i + u_i$

```
> summary(lm(lwage ~ educ, data=wage1))

Call:
lm(formula = lwage ~ educ, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21158 -0.36393 -0.07263  0.29712  1.52339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.583773   0.097336   5.998 3.74e-09 ***
educ        0.082744   0.007567  10.935  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4801 on 524 degrees of freedom
Multiple R-squared:  0.1858,    Adjusted R-squared:  0.1843
F-statistic: 119.6 on 1 and 524 DF,  p-value: < 2.2e-16
```

ii. $log(WAGE_i) = a_0 + a_1 TENURE_i + w_i$

```
> summary(lm(lwage ~ tenure, data=wage1))

Call:
lm(formula = lwage ~ tenure, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15984 -0.38530 -0.04478  0.32696  1.46072

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.501007   0.026866  55.870  < 2e-16 ***
tenure      0.023951   0.003039   7.881 1.89e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5031 on 524 degrees of freedom
Multiple R-squared:  0.106,     Adjusted R-squared:  0.1043
F-statistic: 62.11 on 1 and 524 DF,  p-value: 1.89e-14
```

iii. $TENURE_I = c_0 + c_1 EDUC_i + v_i$

```
> summary(lm(tenure ~ educ, data=wage1))

Call:
lm(formula = tenure ~ educ, data = wage1)

Residuals:
   Min     1Q Median     3Q    Max
-6.946 -4.894 -2.601  1.520 38.813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9457     1.4638   4.745 2.69e-06 ***
educ         -0.1466     0.1138  -1.288    0.198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.22 on 524 degrees of freedom
Multiple R-squared:  0.003155,  Adjusted R-squared:  0.001253
F-statistic: 1.659 on 1 and 524 DF,  p-value: 0.1984
```

1. Explain why the OLS estimator of b_1 fails to produce an unbiased estimation of the true value of this parameter. Which of the OLS assumption is likely to be violated? Explain the meaning of this assumption.

   Sol. The ZCM assumption is likely to be violated here inducing an OVB. This is because in the error term u_i we have $TENURE_i$ (nb this is because it correlates significantly with $log(WAGE_i)$, in simple words it has an effect on the wage of a person) which correlates significantly with $EDUC_i$, this makes $E[u|EDUC] \neq E[u] = 0$.

2. Based on the results obtained above discuss, using the OVB formula asses whether b_1 is likely to be upward or downward biased. Then compute the value of the bias for the data at hand through the same formula and the results below from estimating the long model:

$$log(WAGE_i) = \beta_0 + \beta_1 EDUC_i + \beta_2 TENURE_i + \epsilon_i \tag{8}$$

   Sol. The sign of the OVB depends on the sign of the correlation between TENURE and EDUC and between WAGE and TENURE. Since they are positive and negatively, respectively, then $b_1 = 0.08$ is likely to be a downward biased estimate of β_1. In particular, recall the equation for the OVB you have in the slide reads:

$$\begin{aligned} \hat{b_1} &= \hat{\beta}_1 + \hat{\beta}_2 \times \frac{Cov(tenure_i, educ_i)}{Var(tenure_i)} \\ &= \hat{\beta}_1 + \hat{\beta}_2 \times \hat{c}_1 = 0.0865 + 0.0258 \times (-0.1466) = 0.0827 \end{aligned}$$

```
> summary(lm(lwage ~ educ+tenure, data=wage1))

Call:
lm(formula = lwage ~ educ + tenure, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.10350 -0.29287 -0.04081  0.28672  1.44967

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.404474   0.091696   4.411 1.25e-05 ***
educ        0.086528   0.006991  12.377  < 2e-16 ***
tenure      0.025814   0.002680   9.634  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4428 on 523 degrees of freedom
Multiple R-squared:  0.3085,    Adjusted R-squared:  0.3059
F-statistic: 116.7 on 2 and 523 DF,  p-value: < 2.2e-16
```

The bias for the data at hand is $\hat{\beta}_1 - \hat{b_1} = 0.0865 - 0.0827 = -0.005$, which is a negative bias as expected.

3. Would it be possible for Pampilios Piratta to obtain an estimate of b_1 not affected by this OVB but without estimating a linear model with both EDUC and TENURE? Check your answer with the results provided.

   Sol. We could use the following two stages approach, first we estimate:

$$EDUC_i = d_0 + d_1 TENURE_i + \varepsilon_i$$

   And we call $\hat{\varepsilon}_i = u\_eductenure.hat$, second we estimate

$$lwage_i = \delta_0 + \delta_1 \hat{\varepsilon}_i + n_i$$

   Indeed, as shown in the following results $\delta_1 = \beta_1 = 0.08653$

```
> summary(lm(lwage ~ u_eductenure.hat, data=wage1))

Call:
lm(formula = lwage ~ u_eductenure.hat, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.20181 -0.35600 -0.06182  0.30338  1.50708

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.62327    0.02072   78.36   <2e-16 ***
u_eductenure.hat 0.08653    0.00750   11.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4751 on 524 degrees of freedom
Multiple R-squared:  0.2025,    Adjusted R-squared:  0.201
F-statistic: 133.1 on 1 and 524 DF,  p-value: < 2.2e-16
```