Utrecht University

# *Tutorials*
# *Week 6*

| Pdf file on Blackboard | Dataset on Blackboard | Papers | Description |
|---|---|---|---|
| 15.1 | | | Characteristics of IVs, the meaning of a natural experiment, consequences of omitted bias. |
| 15.7 | | Rouse, C. E. (1998), "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," Quarterly Journal of Economics 113, 553–602. | Omitted bias; IV criteria |
| C.15.1 | wage2.dta | Blackburn McK. and Neumark, D. (1992): Unobserved ability, efficiency wages, and interindustry wage differentials. The Quarterly Journal of Economics, | IV criteria; identification assumption; structural and reduced form equation; comparison OLS vs IV estimator. |

Utrecht University

1 Consider a simple model to estimate the effect of personal computer (PC) ownership on college grade point average for graduating seniors at a large public university:

$$GPA = \beta_0 + \beta_1 PC + u,$$

where $PC$ is a binary variable indicating PC ownership.

(i) Why might PC ownership be correlated with $u$?

(ii) Explain why $PC$ is likely to be related to parents' annual income. Does this mean parental income is a good IV for $PC$? Why or why not?

(iii) Suppose that, four years ago, the university gave grants to buy computers to roughly one-half of the incoming students, and the students who received grants were randomly chosen. Carefully explain how you would use this information to construct an instrumental variable for $PC$.

Utrecht University

i. The theoretical reasons for such correlation can the following:

1. Measurement error in PC: If we have a random error when observing PC ownership then this error would add to the error-term and cause a correlation between the measured PC ownership dummy and the residual.
2. Simultaneous relationship: It is logical to assume that PC ownership may affect grades but would grades also affect PC ownership? If we have reason to believe this than PC and u should be correlated.
3. Omitted variable bias: If there is another variable that affects performance in school and correlates with the probability of PC ownership as well, then we should expect that u is correlated with PC. For example if students from wealthier families are more likely to have better grades and wealthier parents are more likely to buy computer for their children then we should have an upward bias in $\beta_1$.

ii. A good instrument should fulfill the following conditions:

1. **Relevance**: it is correlated with the endogeneous RHS variable. This condition is likely to be fulfilled since we already assumed that students from wealthier families are more likely to have a PC.

2. **Exogeneity**: the instrument should be uncorrelated with the error term. This means that the instrument should only affect the dependent variable through the endogenous RHS variables, not directly. This is where parental income fails as a potential instrument: it directly affects grades. Hence, it should be included directly in the model. It could not be an instrument anymore. Having parental income as an explanatory variable would solve the problem, and including an IV is no longer necessary.

iii. This would be a natural experiment. And, the grant would be a a binary variable: grant=1 for those who received a grant, otherwise, grant =0.

Grant (the potential IV) should be correlated with PC, since students with a grant are more likely to own a PC (relevance).

The IV would be uncorrelated with the error term (and any omitted variables within the error term) because the grant receivers are randomly assigned. (exogeneity).

* To know if it is a good instrument, run an OLS regression of *PC* on *grant* and test whether the IV (*grant*) coefficient is significantly different from zero.** This refers to testing one of the two conditions an IV needs to fulfill: *relevance*. However, to be a "good" instrument, *grant* must also fulfill the second condition: *exogeneity*, which can not be tested in a standard hypothesis testing sense like some other statistical parameters.

**7** The following is a simple model to measure the effect of a school choice program on standardized test performance [see Rouse (1998) for motivation and Computer Exercise C11 for an analysis of a subset of Rouse's data]:

$$score = \beta_0 + \beta_1 choice + \beta_2 faminc + u_1,$$

where *score* is the score on a statewide test, *choice* is a binary variable indicating whether a student attended a choice school in the last year, and *faminc* is family income. The IV for *choice* is *grant*, the dollar amount granted to students to use for tuition at choice schools. The grant amount differed by family income level, which is why we control for *faminc* in the equation.

(i) Even with *faminc* in the equation, why might *choice* be correlated with $u_1$?

(ii) If within each income class, the grant amounts were assigned randomly, is *grant* uncorrelated with $u_1$?

(iii) Write the reduced form equation for *choice*. What is needed for *grant* to be partially correlated with *choice*?

(iv) Write the reduced form equation for *score*. Explain why this is useful. (*Hint:* How do you interpret the coefficient on *grant*?)

Utrecht University

i. Even at a given income level, some students are more motivated and more able than others, and their families are more supportive (say, in terms of providing transportation) and enthusiastic about education. Therefore, there is likely to be a self-selection problem: students that would do better anyway are also more likely to attend a choice school.

ii. Assuming we have the functional form for *faminc* correct, the answer is yes. Since $u_1$ does not contain income, random assignment of grants within income class means that grant designation is not correlated with unobservables such as student ability, motivation, and family support.

iii. The reduced form is $choice_i = \gamma_0 + \gamma_1 faminc + \gamma_2 grant + v_i$ and we need $\gamma_2 \neq 0$. In other words, after accounting for income, the grant amount must have some effect on *choice*. This seems reasonable, provided the grant amounts differ within each income class.

Utrecht University

iv. Now we express the reduced form equation for score by the reduced form of choice:

$choice_i = \gamma_0 + \gamma_1 faminc + \gamma_2 grant + v_i$, into the structural equation: $score = \beta_0 + \beta_1 choice + \beta_2 faminc + u_1$,

$$score_i = \beta_0 + \beta_1(\gamma_0 + \gamma_1 faminc_i + \gamma_2 grant_i + v_i) + \beta_2 faminc_i + \varepsilon_i$$

Rewrite slightly:

$$score_i = (\beta_0 + \beta_1\gamma_0) + \beta_1\gamma_2 grant_i + (\beta_2 + \beta_1\gamma_1) faminc_i + (\beta_1 v_i + \varepsilon_i)$$

- This expression is useful since it expresses the dependent variable as a function of the exogenous variables only. The coefficients can be interpreted as the effect of a change in an exogenous variable with the multiplication effect resulting from the indirect effect of grant on a score through an effect of grant on choice ($\gamma_2$) and the effect of choice on score ($\beta_1$).
- Hence, grant is assumed to affect score only through its effect on choice.

Utrecht University

**C1** Use the data in WAGE2.RAW for this exercise.

(i) In Example 15.2, if *sibs* is used as an instrument for *educ*, the IV estimate of the return to education is .122. To convince yourself that using *sibs* as an IV for *educ* is *not* the same as just plugging *sibs* in for *educ* and running an OLS regression, run the regression of log(*wage*) on *sibs* and explain your findings.

(ii) The variable *brthord* is birth order (*brthord* is one for a first-born child, two for a second-born child, and so on). Explain why *educ* and *brthord* might be negatively correlated. Regress *educ* on *brthord* to determine whether there is a statistically significant negative correlation.

(iii) Use *brthord* as an IV for *educ* in equation (15.1). Report and interpret the results.

(iv) Now, suppose that we include number of siblings as an explanatory variable in the wage equation; this controls for family background, to some extent:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 sibs + u.$$

Suppose that we want to use *brthord* as an IV for *educ*, assuming that *sibs* is exogenous. The reduced form for *educ* is

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v.$$

State and test the identification assumption.

(v) Estimate the equation from part (iv) using *brthord* as an IV for *educ* (and *sibs* as its own IV). Comment on the standard errors for $\hat{\beta}_{educ}$ and $\hat{\beta}_{sibs}$.

(vi) Using the fitted values from part (iv), $\widehat{educ}$, compute the correlation between $\widehat{educ}$ and *sibs*. Use this result to explain your findings from part (v).

## EXAMPLE 15.2 ESTIMATING THE RETURN TO EDUCATION FOR MEN

We now use WAGE2.RAW to estimate the return to education for men. We use the variable *sibs* (number of siblings) as an instrument for *educ*. These are negatively correlated, as we can verify from a simple regression:

$$\widehat{educ} = 14.14 - .228\, sibs$$
$$(.11)\quad (.030)$$
$$n = 935,\ R^2 = .057.$$

This equation implies that every sibling is associated with, on average, about .23 less of a year of education. If we assume that *sibs* is uncorrelated with the error term in (15.14), then the IV estimator is consistent. Estimating equation (15.14) using *sibs* as an IV for *educ* gives

$$\log(wage) = \beta_0 + \beta_1 educ + u. \qquad [15.14]$$

$$\widehat{\log(wage)} = 5.13 + .122\, educ$$
$$(.36)\quad (.026)$$
$$n = 935.$$

(The *R*-squared is computed to be negative, so we do not report it. A discussion of *R*-squared in the context of IV estimation follows.) For comparison, the OLS estimate of $\beta_1$ is .059 with a standard error of .006. Unlike in the previous example, the IV estimate is now much higher than the OLS estimate. While we do not know whether the difference is statistically significant, this does not mesh with the omitted ability bias from OLS. It could be that *sibs* is also correlated with ability: more siblings means, on average, less parental attention, which could result in lower ability. Another interpretation is that the OLS estimator is biased toward zero because of measurement error in *educ*. This is not entirely convincing because, as we discussed in Section 9.3, *educ* is unlikely to satisfy the classical errors-in-variables model.

i. We are required to replace *educ* by *sibs* and observe that the results are not the same as with the IV estimator.

```
. ivreg lwage (educ=sibs), first

First-stage regressions
```

| Source | SS | df | MS | Number of obs | = | 935 |
|---|---|---|---|---|---|---|
| | | | | F(1, 933) | = | 56.67 |
| Model | 258.055048 | 1 | 258.055048 | Prob > F | = | 0.0000 |
| Residual | 4248.7642 | 933 | 4.55387374 | R-squared | = | 0.0573 |
| | | | | Adj R-squared | = | 0.0562 |
| Total | 4506.81925 | 934 | 4.82528828 | Root MSE | = | 2.134 |

| educ | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| sibs | -.2279164 | .0302768 | -7.53 | 0.000 | -.287335 | -.1684979 |
| _cons | 14.13879 | .1131382 | 124.97 | 0.000 | 13.91676 | 14.36083 |

```
Instrumental variables 2SLS regression
```

| Source | SS | df | MS | Number of obs | = | 935 |
|---|---|---|---|---|---|---|
| | | | | F(1, 933) | = | 21.59 |
| Model | -1.51973315 | 1 | -1.51973315 | Prob > F | = | 0.0000 |
| Residual | 167.176016 | 933 | .179181154 | R-squared | = | . |
| | | | | Adj R-squared | = | . |
| Total | 165.656283 | 934 | .177362188 | Root MSE | = | .4233 |

| lwage | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| educ | .1224326 | .0263506 | 4.65 | 0.000 | .0707194 | .1741459 |
| _cons | 5.130026 | .3551712 | 14.44 | 0.000 | 4.432999 | 5.827053 |

```
Endogenous: educ
Exogenous:  sibs
```

```
. reg lwage sibs
```

| Source | SS | df | MS | Number of obs | = | 935 |
|---|---|---|---|---|---|---|
| | | | | F(1, 933) | = | 22.31 |
| Model | 3.86818074 | 1 | 3.86818074 | Prob > F | = | 0.0000 |
| Residual | 161.788103 | 933 | .173406326 | R-squared | = | 0.0234 |
| | | | | Adj R-squared | = | 0.0223 |
| Total | 165.656283 | 934 | .177362188 | Root MSE | = | .41642 |

| lwage | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| sibs | -.0279044 | .0059082 | -4.72 | 0.000 | -.0394992 | -.0163096 |
| _cons | 6.861076 | .0220776 | 310.77 | 0.000 | 6.817748 | 6.904403 |

- This is a simple regression equation. It shows that, controlling for no other factors, one more sibling in the family is associated with a monthly salary that is about 2.8% lower. The *t*-statistic on *sibs* is –4.72.
- *sibs* can be correlated with many things that should have a bearing on wage including, as we saw, years of education.

ii. It could be that older children are given priority for higher education, and families may hit budget constraints and may not be able to afford as much education for children born later. The simple regression of *educ* on *brthord* gives:

```
. reg educ brthord
```

| Source | SS | df | MS | Number of obs | = | 852 |
|---|---|---|---|---|---|---|
| | | | | F(1, 850) | = | 37.29 |
| Model | 173.087012 | 1 | 173.087012 | Prob > F | = | 0.0000 |
| Residual | 3945.88364 | 850 | 4.64221605 | R-squared | = | 0.0420 |
| | | | | Adj R-squared | = | 0.0409 |
| Total | 4118.97066 | 851 | 4.84015353 | Root MSE | = | 2.1546 |

| educ | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| brthord | -.2826441 | .0462882 | -6.11 | 0.000 | -.3734967 | -.1917915 |
| _cons | 14.14945 | .1286754 | 109.96 | 0.000 | 13.89689 | 14.40201 |

$$\underset{}{educ} = \underset{(0.13)}{14.15} - \underset{(.046)}{.283}\, brthord$$

$$n = 852, \quad R^2 = .042.$$

The equation predicts that every one-unit increase in *brthord* reduces predicted education by about .28 years.

Utrecht University

$$\log(wage) = \beta_0 + \beta_1 educ + u, \qquad [15.1]$$

```
. ivreg lwage (educ=brthord), first

First-stage regressions
-----------------------

      Source |       SS       df       MS      Number of obs   =       852
-------------+------------------------------   F(1, 850)       =     37.29
       Model |  173.087012     1  173.087012   Prob > F        =    0.0000
    Residual |  3945.88364   850  4.64221605   R-squared       =    0.0420
-------------+------------------------------   Adj R-squared   =    0.0409
       Total |  4118.97066   851  4.84015353   Root MSE        =    2.1546

------------------------------------------------------------------------------
        educ | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     brthord |  -.2826441   .0462882    -6.11   0.000    -.3734967   -.1917915
       _cons |   14.14945   .1286754   109.96   0.000     13.89689    14.40201
------------------------------------------------------------------------------


Instrumental variables 2SLS regression
--------------------------------------

      Source |       SS       df       MS      Number of obs   =       852
-------------+------------------------------   F(1, 850)       =     16.63
       Model | -4.20185534     1 -4.20185534   Prob > F        =    0.0000
    Residual |  151.01792   850  .177668141   R-squared       =         .
-------------+------------------------------   Adj R-squared   =         .
       Total |  146.816065   851  .172521815   Root MSE        =    .42151

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |  .1306448   .0320385     4.08   0.000     .0677609    .1935287
       _cons |  5.030397    .432949    11.62   0.000     4.180622    5.880171
------------------------------------------------------------------------------
Endogenous: educ
Exogenous:  brthord
```

iii. When *brthord* is used as an IV for *educ* in the simple wage equation, we get

$$\log(wage) = \underset{(0.43)}{5.03} + \underset{(.032)}{.131}\, educ$$

$$n = 852.$$

- This is much higher than the OLS estimate (.060) and even above the estimate when *sibs* is used as an IV for *educ* (.122).
- Because of missing data on *brthord*, we are using fewer observations than in the previous analyses.

- We find that the rate of return to education is estimated at 13% by 2SLS, while the OLS estimate was 5.98%.

- This is not what was expected since the omitted bias (ability) was expected to be upward. However, it can still be possible.

iv. What we need now is $\pi_2 \neq 0$ from the reduced form equation:

$$educ = \pi_0 + \pi_1 sibs + \pi_2 brthord + v,$$

```
. reg educ sibs brthord
```

| Source | SS | df | MS | Number of obs | = | 852 |
|---|---|---|---|---|---|---|
| | | | | F(2, 849) | = | 26.29 |
| Model | 240.246365 | 2 | 120.123183 | Prob > F | = | 0.0000 |
| Residual | 3878.72429 | 849 | 4.56857985 | R-squared | = | 0.0583 |
| | | | | Adj R-squared | = | 0.0561 |
| Total | 4118.97066 | 851 | 4.84015353 | Root MSE | = | 2.1374 |

| educ | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| sibs | -.1528673 | .0398705 | -3.83 | 0.000 | -.2311236 | -.0746109 |
| brthord | -.1526742 | .0570764 | -2.67 | 0.008 | -.2647017 | -.0406467 |
| _cons | 14.2965 | .1332881 | 107.26 | 0.000 | 14.03489 | 14.55811 |

```
. test brthord

 ( 1)  brthord = 0

       F(  1,   849) =    7.16
            Prob > F =    0.0076
```

- We take the null to $H_0$: $\pi_2 = 0$ , and look to reject $H_0$ at a small significance level.
- We reject Ho
- $H_1$: $\pi_2 \neq 0$
- Birthorder is correlated to education.
- Therefore, the identification assumptions appears to hold.

- The regression of *educ* on *sibs* and *brthord* (using 852 observations) yields $\hat{\pi}_2$ = -0.153 and s.e. ( $\hat{\pi}_2$ ) = 0.057
- The *t* statistic is about –2.68, which rejects $H_0$.

- A practical concern: rule of thumb: the exclusion test F-statistics is less than 10 (threshold taken in empirical studies)

Utrecht University

v. The equation estimated by IV is

$$\log(wage) = \underset{(1.06)}{4.94} + \underset{(.075)}{.137\, educ} + \underset{(.0174)}{.0021\, sibs}$$

$$n = 852.$$

- The standard error on $\widehat{\beta_{educ}}$ is much larger than in iii)
- The 95% CI for $\beta_{educ}$ is roughly -0.010 to 0.284, which is very wide and zero value alue zero.
- The s.e. of $\hat{\beta}_{sibs}$ is very large relative to the coefficient estimates, rendering sibs insignificant.

```
. ivreg lwage (educ=brthord) sibs, first

First-stage regressions
```

| Source | SS | df | MS | Number of obs | = | 852 |
|---|---|---|---|---|---|---|
| | | | | F(2, 849) | = | 26.29 |
| Model | 240.246365 | 2 | 120.123183 | Prob > F | = | 0.0000 |
| Residual | 3878.72429 | 849 | 4.56857985 | R-squared | = | 0.0583 |
| | | | | Adj R-squared | = | 0.0561 |
| Total | 4118.97066 | 851 | 4.84015353 | Root MSE | = | 2.1374 |

| educ | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| sibs | -.1528673 | .0398705 | -3.83 | 0.000 | -.2311236 | -.0746109 |
| brthord | -.1526742 | .0570764 | -2.67 | 0.008 | -.2647017 | -.0406467 |
| _cons | 14.2965 | .1332881 | 107.26 | 0.000 | 14.03489 | 14.55811 |

```
Instrumental variables 2SLS regression
```

| Source | SS | df | MS | Number of obs | = | 852 |
|---|---|---|---|---|---|---|
| | | | | F(2, 849) | = | 10.90 |
| Model | -7.96903706 | 2 | -3.98451853 | Prob > F | = | 0.0000 |
| Residual | 154.785102 | 849 | .182314607 | R-squared | = | . |
| | | | | Adj R-squared | = | . |
| Total | 146.816065 | 851 | .172521815 | Root MSE | = | .42698 |

| lwage | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| educ | .1369941 | .0746812 | 1.83 | 0.067 | -.0095873 | .2835756 |
| sibs | .0021108 | .0173717 | 0.12 | 0.903 | -.0319858 | .0362073 |
| _cons | 4.938527 | 1.05569 | 4.68 | 0.000 | 2.866458 | 7.010596 |

```
Endogenous: educ
Exogenous:  sibs brthord
```

Utrecht University

vi. Save the fitted values from the reduced form in iv)

```
. reg educ sib brthord
```

| Source | SS | df | MS | Number of obs | = | 852 |
|---|---|---|---|---|---|---|
| | | | | F(2, 849) | = | 26.29 |
| Model | 240.246365 | 2 | 120.123183 | Prob > F | = | 0.0000 |
| Residual | 3878.72429 | 849 | 4.56857985 | R-squared | = | 0.0583 |
| | | | | Adj R-squared | = | 0.0561 |
| Total | 4118.97066 | 851 | 4.84015353 | Root MSE | = | 2.1374 |

| educ | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| sibs | -.1528673 | .0398705 | -3.83 | 0.000 | -.2311236 | -.0746109 |
| brthord | -.1526742 | .0570764 | -2.67 | 0.008 | -.2647017 | -.0406467 |
| _cons | 14.2965 | .1332881 | 107.26 | 0.000 | 14.03489 | 14.55811 |

```
. predict educhat
(option xb assumed; fitted values)
(83 missing values generated)
```

Look at the correlation between $\widehat{educ}$ and sibs

```
. corr educhat sib
(obs=852)
```

| | educhat | sibs |
|---|---|---|
| educhat | 1.0000 | |
| sibs | -0.9295 | 1.0000 |

Letting $educ_i$ be the first-stage fitted values, the correlation between $educ_i$ and $sibs_i$ is about -0.930, which is a very strong negative correlation. This means that, for the purposes of using IV, multicollinearity is a problem here, and is not allowing to estimate $\beta_{educ}$ with much precision.