# Instrumental Variables Estimation and Two Stage Least Squares

2023/7/4

Learning points:

- conducting IV estimation by `ivreg()` function of `AER` package

We first load the required packages and datasets.

```
# load the required library
library(AER); library(stargazer); library(wooldridge)
# load the required dataset
data(list=c("card", "mroz", "wage2"), wooldridge)
```

## Instrumental variables in simple regression models

We are concerned with estimating the following wage equation

$$log(wage_i) = \alpha_0 + \alpha_1 educ_i + \epsilon_i$$

In words, we want to look at the effects of education on log wages. To do this we will use the `CARD` dataset, available in the `wooldridge` package in R. We know that `educ` is an endogeneous variable. As such, running a simple OLS regression on the above regression will yield a biased estimate. We will therefore need (at least) one instrument.

### Using one instrument

We use the instrument variable, namely `nearc4`, a dummy variable indicating proximity to a four-year college. In order to be a valid instrument, we need this distance measure (a) to be correlated with education (**instrument relevance**: the instrument and the endogeneous regressor are correlated, and (b) not to affect log wages other than through education (**instrument exogeneity**: the instrument and the error are uncorrelated). We cannot test our exogeneity assumption (you can try to think for yourself whether you think that it is satisfied in this case), but we can test the relevance assumption.

To test the relevance assumption, we simple run an OLS regression of `educ` on `nearc4`. In addition to the standard errors on the assumption of homoskedasticity, we alos report heteroskedastic robust standard errors.

```
# relevance assumption
card_FS1 <- lm(educ ~ nearc4, data = card)
# heteroskedastic SEs
robust.se <- coeftest(card_FS1, vcov = vcovHC(card_FS1, type = "HC1"))
# report the result
stargazer(card_FS1, card_FS1, se=list(NULL, robust.se[,2]), column.labels=c("default","robust"),
          type = "text")
```

```
##
## ===============================================================
##                                    Dependent variable:
##                                ---------------------------
##                                            educ
##                                   default        robust
##                                     (1)            (2)
## ---------------------------------------------------------------
## nearc4                             0.829***      0.829***
##                                    (0.104)        (0.107)
##
## Constant                          12.698***      12.698***
##                                    (0.086)        (0.090)
##
## ---------------------------------------------------------------
## Observations                         3,010          3,010
## R2                                   0.021          0.021
## Adjusted R2                          0.020          0.020
## Residual Std. Error (df = 3008)      2.649          2.649
## F Statistic (df = 1; 3008)         63.912***      63.912***
## ===============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

The relevance assumption requires that `educ` and `nearc4` are strongly correlated. We can test this by looking at whether the coefficient on `nearc4` (0.829 in our regression) is statistically significantly different from zero. Given the $t$-statistic of 7.77 (or the $p$-value that's essentially zero), we conclude that this coefficient is statistically significantly different from zero. Thus, the relevance assumption should be valid.

## Further examples

```
# restrict to non-missing wage observations
oursample <- subset(mroz, !is.na(wage))

# OLS slope parameter manually
with(oursample, cov(log(wage),educ) / var(educ))
```

```
## [1] 0.1086487
```

```
# IV slope parameter manually
with(oursample, cov(log(wage),fatheduc) / cov(educ,fatheduc))
```

```
## [1] 0.05917348
```

```
# OLS automatically
reg.ols <- lm(log(wage) ~ educ, data=oursample)

# IV automatically
reg.iv <- ivreg(log(wage) ~ educ | fatheduc, data=oursample)

# pretty regression table
stargazer(reg.ols, reg.iv, type="text")
```

```
##
## ========================================================================
##                                        Dependent variable:
##                                    ------------------------------------
##                                               log(wage)
##                                      OLS           instrumental
##                                                      variable
##                                      (1)               (2)
## ------------------------------------------------------------------------
## educ                                0.109***           0.059*
##                                      (0.014)           (0.035)
##
## Constant                            -0.185             0.441
##                                      (0.185)           (0.446)
##
## ------------------------------------------------------------------------
## Observations                          428               428
## R2                                    0.118             0.093
## Adjusted R2                           0.116             0.091
## Residual Std. Error (df = 426)        0.680             0.689
## F Statistic                    56.929*** (df = 1; 426)
## ========================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

```
# Example 15.2 on p. 502

# IV automatically
reg.iv2 <- ivreg(log(wage) ~ educ | sibs, data=wage2)
stargazer(reg.iv2, type="text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                               log(wage)
## ------------------------------------------------
## educ                          0.122***
##                                (0.026)
##
## Constant                      5.130***
##                                (0.355)
##
## ------------------------------------------------
## Observations                    935
## R2                             -0.009
## Adjusted R2                    -0.010
## Residual Std. Error      0.423 (df = 933)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```
# Complete Example 15.2

# variable educ and sibs are correlated
```

```
reg <- lm(educ ~ sibs, data = wage2)
stargazer(reg, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                  educ
## ------------------------------------------------
## sibs                           -0.228***
##                                 (0.030)
##
## Constant                       14.139***
##                                 (0.113)
##
## ------------------------------------------------
## Observations                      935
## R2                               0.057
## Adjusted R2                      0.056
## Residual Std. Error       2.134 (df = 933)
## F Statistic           56.667*** (df = 1; 933)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
# run the ols and iv estimation
olsreg <- lm(log(wage) ~ educ, data = wage2)
ivreg1 <- ivreg(log(wage) ~ educ | sibs, data = wage2)

# compare the result
stargazer(olsreg, ivreg1, type = "text")
```

```
##
## ===============================================================
##                                       Dependent variable:
##                                   ----------------------------
##                                          log(wage)
##                                     OLS            instrumental
##                                                      variable
##                                     (1)                (2)
## ---------------------------------------------------------------
## educ                              0.060***           0.122***
##                                   (0.006)            (0.026)
##
## Constant                          5.973***           5.130***
##                                   (0.081)            (0.355)
##
## ---------------------------------------------------------------
## Observations                        935                935
## R2                                 0.097              -0.009
## Adjusted R2                        0.096              -0.010
## Residual Std. Error (df = 933)     0.400              0.423
## F Statistic                 100.700*** (df = 1; 933)
```

```
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

## More exogenous regressors (Example 15.4 of Wooldridge)

We use `card` data to estimate the return to education. Education is allowed to be endogenous and instrumented with the dummy variable `near4` which indicates whether the individual grew up close to a college.

We first check for relevance by regressing the endogenous independent variable `educ` on all exogenous variables including the instrument `near4`.

```
# reduced form equation: check for relevance
redf <- lm(educ ~ nearc4 + exper + I(exper^2) + black + smsa + south + smsa66 +
            reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + reg669,
           data = card)
stargazer(redf, keep=c("nearc4"), type="text", title = "Reduced form equation")
```

```
##
## Reduced form equation
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                  educ
## ------------------------------------------------
## nearc4                         0.320***
##                                 (0.088)
##
## ------------------------------------------------
## Observations                     3,010
## R2                               0.477
## Adjusted R2                      0.474
## Residual Std. Error       1.941 (df = 2994)
## F Statistic          182.129*** (df = 15; 2994)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The parameter for `nearc4` is highly significantly different from zero, so relevance is supported. We then estimate the log wage equation with OLS and IV.

```
# OLS
ols <- lm(log(wage) ~ educ + exper + I(exper^2) + black + smsa + south + smsa66 +
          reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + reg669,
          data = card)

# IV
iv <- ivreg(log(wage) ~ educ + exper + I(exper^2) + black + smsa + south + smsa66 +
            reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 + reg669 |
            nearc4 + exper + I(exper^2) + black + smsa + south + smsa66 +
            reg662 + reg663 + reg664 + reg665 + reg666 + reg667 + reg668 +
            reg669, data = card)

# table of the results
stargazer(ols, iv, keep=c("ed", "exp", "bl"), type="text", title = "OLS vs IV estimation")
```

```
##
## OLS vs IV estimation
## ================================================================
##                                        Dependent variable:
##                                    ----------------------------
##                                           log(wage)
##                                      OLS            instrumental
##                                                       variable
##                                      (1)                (2)
## ----------------------------------------------------------------
## educ                                0.075***           0.132**
##                                      (0.003)           (0.055)
##
## exper                               0.085***           0.108***
##                                      (0.007)           (0.024)
##
## I(exper2)                           -0.002***          -0.002***
##                                     (0.0003)           (0.0003)
##
## black                               -0.199***          -0.147***
##                                      (0.018)           (0.054)
##
## ----------------------------------------------------------------
## Observations                          3,010              3,010
## R2                                    0.300              0.238
## Adjusted R2                           0.296              0.234
## Residual Std. Error (df = 2994)       0.372              0.388
## F Statistic                  85.476*** (df = 15; 2994)
## ================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Notes on R

In `ivreg()`, we have to include the exogenous variables both to the list of regressors left of the `|` symbol and to the list of exogenous instrument to the right of the `|` symbol.