# IV example

Adapted from Cunningham, S. (2021) by José Manuel Arencibia Alemán*

Spring 2022

## 0. College in County

Card (1995)[1] was interested in the causal effect of **schooling years**, `educ`, on **log wages**, `lwage`. However, we can imagine that several variables will affect both the independent (wages) and dependent (schooling) variables and, thus, that OLS will yield biased estimators. Instead, Card (1995) decided to instrumentalize schooling with a dummy variable that takes value 1 when an individual lived in a **county in which there was a 4-year college**, `nearc4`, while controlling for several *available* common covariates: **years of experience**, `exper`, and **whether or not a person is black**, `black`, **lives in the southern US**, `south`, **is married**, `married`, and lives in an urban area (**Standard Metropolitan Statistical Area**), `smsa`.

## 1. Loading packages and data

```r
# Packages
#install.packages("AER")    # function iv_reg
#install.packages("haven")

library(AER)
library(haven)

# Data
Card1995 <- read_dta("https://raw.github.com/scunning1975/mixtape/master/card.dta")

# Subset variables of interest
Card1995 <-
  Card1995[c("lwage",                                       #outcome
             "educ",                                        #treatment(endogenous)
             "exper", "black", "south", "married", "smsa",  #covariates
             "nearc4")]                                     #instrument
```

## 2. Regressions

### 2.1. Bivariate regression

A first bivariate *log-lin* regression suggests that increases by 1 year of education predict increments of 5,2% in wages.

*Cunningham, Scott (2021). Causal Inference: The Mixtape

[1] Card, David (1995). "Aspects of Labour Economics: Essays in Honour of John Vanderkamp." In. University of Toronto Press.

```r
ols1 <- lm(lwage ~ educ, data = Card1995)
summary(ols1)
```

```
##
## Call:
## lm(formula = lwage ~ educ, data = Card1995)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73799 -0.27764  0.02373  0.28839  1.46080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.57088    0.03883  143.47   <2e-16 ***
## educ         0.05209    0.00287   18.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4214 on 3008 degrees of freedom
## Multiple R-squared:  0.09874,    Adjusted R-squared:  0.09844
## F-statistic: 329.5 on 1 and 3008 DF,  p-value: < 2.2e-16
```

### 2.2. OLS with controls

As you know, we can add available control variables to account for differences between groups that might confound the effect of schooling. This step corrects our prior prediction by 2,1 percent points.

```r
ols2 <- lm(lwage ~ educ +
             exper + black + south + married + smsa,
           data = Card1995)
summary(ols2)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper + black + south + married +
##     smsa, data = Card1995)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59924 -0.23035  0.01812  0.23046  1.36797
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.063317   0.063740  79.437   <2e-16 ***
## educ         0.071173   0.003482  20.438   <2e-16 ***
## exper        0.034152   0.002214  15.422   <2e-16 ***
## black       -0.166027   0.017614  -9.426   <2e-16 ***
## south       -0.131552   0.014969  -8.788   <2e-16 ***
## married     -0.035871   0.003401 -10.547   <2e-16 ***
## smsa         0.175787   0.015458  11.372   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3702 on 2996 degrees of freedom
##   (7 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.305,  Adjusted R-squared:  0.3036
## F-statistic: 219.2 on 6 and 2996 DF,  p-value: < 2.2e-16
```

### 2.3. 2SLS

However, the inclusion of controls does not solve the problem of endogeinity in `educ`. *What if there are other unobservable variables biasing our estimate?* If `nearc4` satisfies *relevance* (i.e., our first stage coefficient is strong), *independence* (i.e., that individuals' reasons for being settled in a county are related to whether or not `nearc4` is 1 or 0 and, thus, it is as good as *random*), and *exclusion* (i.e., that `nearc4` is associated with `lwage`, only through `educ`—or at least is uncorrelated with any unobservable variable biasing estimates), we can use it as *imperfect* instrument of `educ`.

```
iv_reg = ivreg(
  #Second stage
  lwage ~ educ +
    exper + black + south + married + smsa |
    #First stage (treatment omitted)
    nearc4 +
    exper + black + south + married + smsa,
  #Data
  data = Card1995)
summary(iv_reg)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + black + south + married +
##     smsa | nearc4 + exper + black + south + married + smsa, data = Card1995)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81301 -0.23805  0.01766  0.24727  1.32278
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.162476   0.849590   4.899 1.01e-06 ***
## educ         0.124164   0.049956   2.485  0.01299 *
## exper        0.055588   0.020286   2.740  0.00618 **
## black       -0.115686   0.050741  -2.280  0.02268 *
## south       -0.113165   0.023244  -4.869 1.18e-06 ***
## married     -0.031975   0.005087  -6.286 3.73e-10 ***
## smsa         0.147707   0.030895   4.781 1.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3843 on 2996 degrees of freedom
## Multiple R-Squared: 0.2513,  Adjusted R-squared: 0.2498
## Wald test: 139.8 on 6 and 2996 DF,  p-value: < 2.2e-16
```

```
# Compare coefficients and SE from OLS and IV
out_ols2 <- summary(ols2)
out_iv_reg <- summary(iv_reg)
out_ols2$coefficients[,1:2]
```

```
##                Estimate  Std. Error
## (Intercept)  5.06331654 0.063740191
## educ         0.07117285 0.003482405
```

```
## exper        0.03415182 0.002214445
## black       -0.16602745 0.017613671
## south       -0.13155177 0.014969061
## married     -0.03587071 0.003401161
## smsa         0.17578712 0.015457781
```

```r
out_iv_reg$coefficients[,1:2]
```

```
##                Estimate  Std. Error
## (Intercept)  4.16247588 0.849590453
## educ         0.12416424 0.049955802
## exper        0.05558822 0.020286089
## black       -0.11568555 0.050741489
## south       -0.11316470 0.023243878
## married     -0.03197537 0.005086886
## smsa         0.14770651 0.030895151
```

LATE is 12,4% (that is 5.3 percent points or 75% larger than our OLS estimate).

The external validity of our LATE (i.e., whether or not LATE is (approximately) equal to TOT) depends on our assumptions. We can interpret this result the following way: among compliers in our sample (i.e., those who extend their education because of living in a county with a college, for example, because it reduces the cost of schooling), an extra year of schooling increases wages 12,4% on average. If we can argue that this effect is indeed independent from being a complier, we might suggest that this is an unbiased estimator of years of education on wages for the whole population (as opposed to the OLS).

# 3. Standard Errors

**Manual estimation** of IV only takes into account the regression error from the second stage, i.e., it ignores the regression error from the first stage and, thus, **provides incorrect standard errors**. Modern software (e.g., R and Stata) provides robust standard errors automatically, so… even if the coefficients are the same, **let sofware do the hard calculations for you!!**

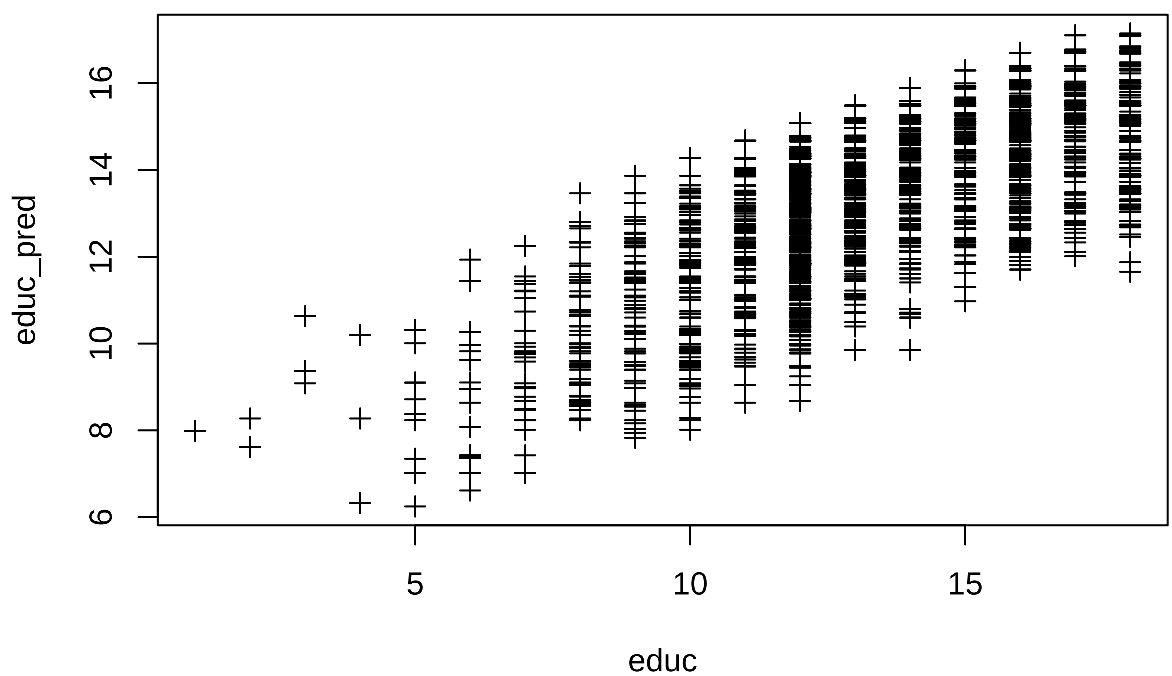
See the example below:

```r
# Manual First stage
fs <- lm(educ ~
           nearc4 +
           exper + black + south + married + smsa,
         data = Card1995)
summary(fs)
```

```
##
## Call:
## lm(formula = educ ~ nearc4 + exper + black + south + married +
##     smsa, data = Card1995)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6308 -1.4454 -0.0526  1.2986  6.3449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.83070    0.13075 128.727  < 2e-16 ***
## nearc4       0.32728    0.08242   3.971 7.33e-05 ***
## exper       -0.40443    0.00894 -45.238  < 2e-16 ***
```

```
## black       -0.94753    0.09053 -10.467  < 2e-16 ***
## south       -0.29735    0.07906  -3.761 0.000173 ***
## married     -0.07269    0.01775  -4.096 4.31e-05 ***
## smsa         0.42090    0.08487   4.959 7.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.937 on 2996 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.4774, Adjusted R-squared:  0.4764
## F-statistic: 456.1 on 6 and 2996 DF,  p-value: < 2.2e-16
```

```r
# Add a column with predicted treatment from first-stage
Card1995$educ_pred <- with(Card1995, 16.83069650 + 0.32728259*nearc4 -
                           0.40443401*exper - 0.94752809*black -
                           0.29735279*south - 0.07269361*married +
                           0.42089454*smsa)
# Plotting First-Stage
with(Card1995, plot(educ,educ_pred,xlab="educ",ylab="educ_pred",pch=3))
```




```r
# Manual second stage
ss <- lm(lwage ~ educ_pred +
           exper + black + south + married + smsa,
         data = Card1995)
invisible(summary(ss))

# Compare coefficients and SE
```

```r
out_ss <- summary(ss)
out_ss$coefficients[,1:2]
```

```
##                Estimate  Std. Error
## (Intercept)  4.16247586 0.872895305
## educ_pred    0.12416424 0.051326124
## exper        0.05558822 0.020842550
## black       -0.11568555 0.052133364
## south       -0.11316470 0.023881473
## married     -0.03197537 0.005226423
## smsa         0.14770650 0.031742626
```

```r
out_iv_reg$coefficients[,1:2]
```

```
##                Estimate  Std. Error
## (Intercept)  4.16247588 0.849590453
## educ         0.12416424 0.049955802
## exper        0.05558822 0.020286089
## black       -0.11568555 0.050741489
## south       -0.11316470 0.023243878
## married     -0.03197537 0.005086886
## smsa         0.14770651 0.030895151
```