Instrumental Variables

Alex Alekseev

```
library(AER)
library(tidyverse)
library(lmtest)
library(sandwich)
library(modelsummary)
```

Intro

In this section, we will be replicating some of the results from the Card (1995) study on the effect of education on earnings. The study instruments education with the presence of a four-year college in the county where a respondent resides. The study uses the data from the National Longitudinal Survey, 1966. The assumption behind the instrument is that having a college in the county where you live reduces the costs of attending a college, e.g., due to being able to live with parents. Thus, having a college in your county should increase the likelihood of attending a college and getting more education. The validity assumption here means that we are assuming that having a college in your county affects your earnings only via the education channel.

Data

Let's load the `card_1995.csv` dataset.

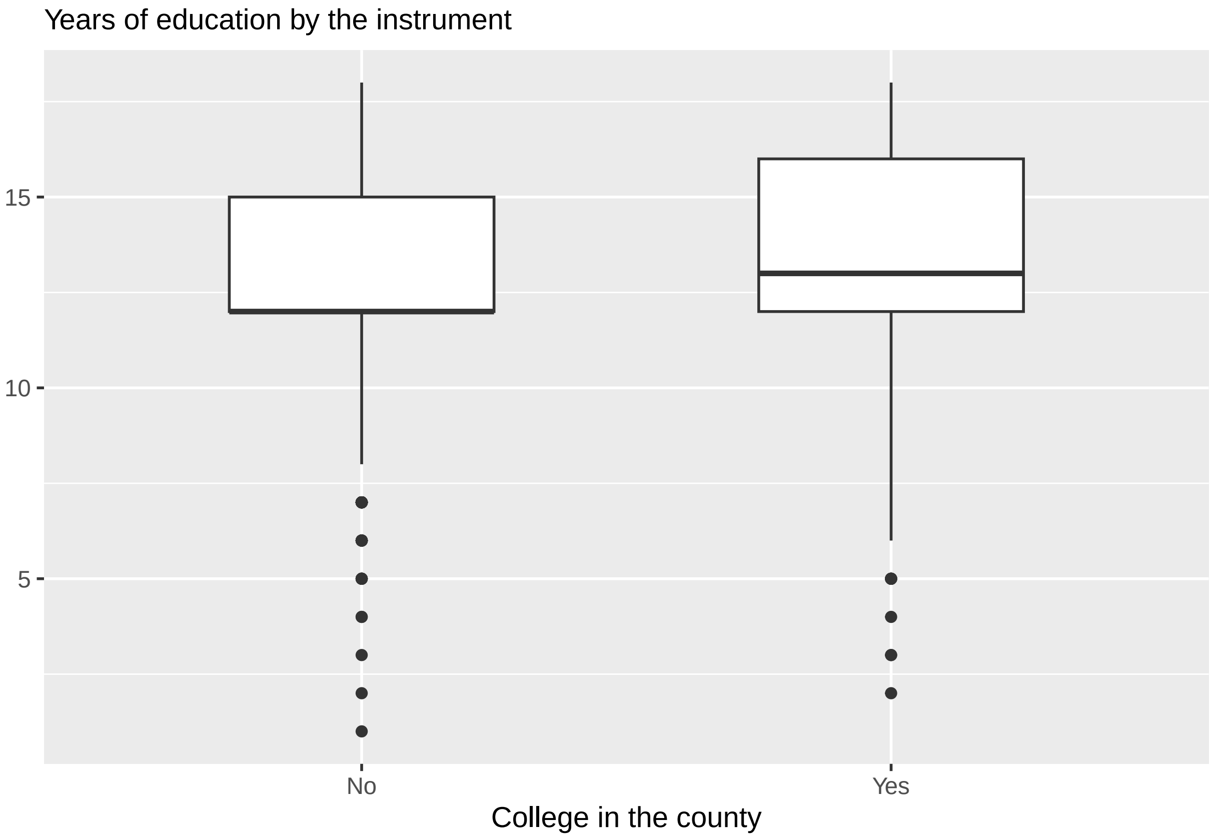
```
df <- read_csv("path_to_file/card_1995.csv")
```

The dependent variable we are interested in is log wages `lwage`. The treatment variable is the years of education `educ`. The instrument is the indicator variable for whether there is a four-year college in a respondent's county `nearc4`. The dataset also includes many other variables that we will later use as controls.

IV by hand

Let's first compute the IV effect by hand. We begin with the first stage: the effect of college-in-the-county on years of schooling. Here is a boxplot of the first stage.

```
ggplot(df, aes(factor(nearc4, labels = c("No", "Yes")), educ)) +
  geom_boxplot(width = 0.5) +
  labs(
    x = "College in the county"
    , y = NULL
    , subtitle = "Years of education by the instrument")
```



It does look like that having college in the county positively affects years of schooling. In particular, respondents without a college in the county on average have 12.7 years of schooling, while respondents with a college in the county have 13.5 years of schooling.

We can estimate the effect of the college-in-the-county on education using a simple regression.

```
first_stage <- lm(educ ~ nearc4, data = df)
summary(first_stage)
```

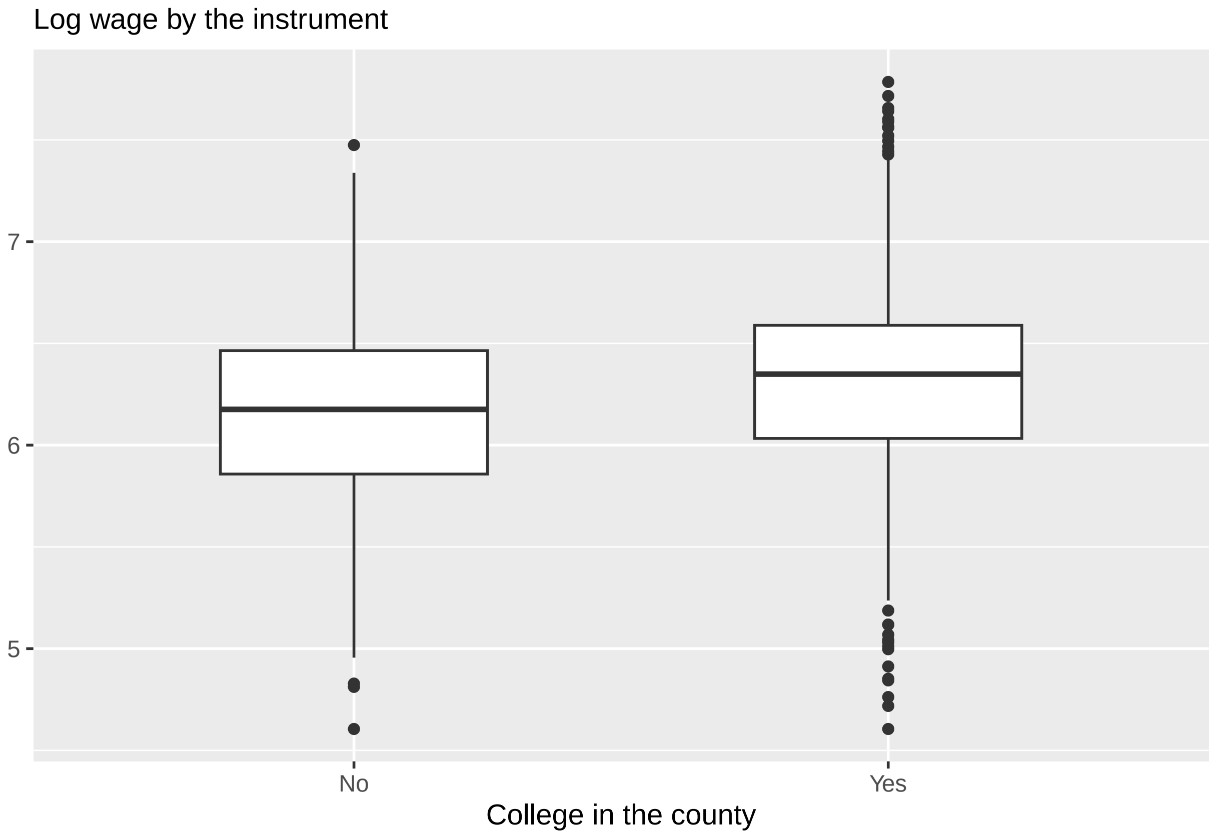
```
##
## Call:
## lm(formula = educ ~ nearc4, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.698  -1.527  -0.527   2.473   5.302
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.69801    0.08564 148.269  < 2e-16 ***
## nearc4       0.82902    0.10370   7.994 1.84e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.649 on 3008 degrees of freedom
## Multiple R-squared:  0.02081,    Adjusted R-squared:  0.02048
## F-statistic: 63.91 on 1 and 3008 DF,  p-value: 1.838e-15
```

The coefficient on the instrument is statistically significant and implies that having a college in the county increases the years of schooling by 0.83, on average. Thus the instrument does seem to be relevant.

Now let's look at the reduced form, the effect of the instrument on the earnings. We will start with the graph and then use a regression.

```
ggplot(df, aes(factor(nearc4, labels = c("No", "Yes")), lwage)) +
  geom_boxplot(width = 0.5) +
  labs(
    x = "College in the county"
    , y = NULL
    , subtitle = "Log wage by the instrument")
```



We again see a positive effect of the instrument on earnings.

```
reduced_form <- lm(lwage ~ nearc4, data = df)
summary(reduced_form)
```

```
##
## Call:
## lm(formula = lwage ~ nearc4, data = df)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -1.70623 -0.28795  0.02866  0.28860  1.47349 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  6.15549    0.01415 434.865   <2e-16 ***
## nearc4       0.15591    0.01714   9.096   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4379 on 3008 degrees of freedom
## Multiple R-squared:  0.02677,    Adjusted R-squared:  0.02645 
## F-statistic: 82.74 on 1 and 3008 DF,  p-value: < 2.2e-16
```

The regression shows that having a college in the county leads to 0.16 increase in log wages.

To get the effect of the treatment on the outcome, we can use several methods. First, we can use the formula

$$\delta = \frac{Cov(Y,Z)/V(Z)}{Cov(D,Z)/V(Z)}.$$

In other words, we need to divide the coefficient from the reduced form by the coefficient from the first stage. This corresponds to the following code

```r
coef(reduced_form)["nearc4"]/coef(first_stage)["nearc4"]
```

```
##    nearc4 
## 0.1880626
```

Alternatively, we can use the formula

$$\delta = \frac{Cov(Y,\hat{D})}{V(\hat{D})}.$$

that says that we can get the IV effect by regression the outcome on the fitted values from the first stage (two-stage least squares, 2SLS).

```r
second_stage <- lm(lwage ~ fitted(first_stage), data = df)
summary(second_stage)
```

```
## 
## Call:
## lm(formula = lwage ~ fitted(first_stage), data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.70623 -0.28795  0.02866  0.28860  1.47349 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          3.76747    0.27433  13.733   <2e-16 ***
## fitted(first_stage)  0.18806    0.02067   9.096   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4379 on 3008 degrees of freedom
## Multiple R-squared:  0.02677,    Adjusted R-squared:  0.02645
## F-statistic: 82.74 on 1 and 3008 DF,  p-value: < 2.2e-16
```

Notice that the coefficient (0.188) is exactly the same as the one that we derived before.

Before we proceed further, it is worth reflecting on what exactly are we estimating here. As discussed during the lecture, the IV produces a LATE: the average treatment effect for compliers. In our case, the compliers would be the respondents whose behavior is affected by the instrument. These are the people who would have less education without a college in the county but more education with a college in the county. Our estimated effect applies only to this specific slice of the sample.

IV regression and OLS

In practice, we typically use special functions for estimating IV, because, for one thing, they adjust the standard errors in the 2SLS procedure. One option to run IV is to use the `ivreg` function from the AER package. But before we run the IV, let's have a look at the results we would obtain using OLS.

```
ols <- lm(lwage ~ educ, data = df)
```

Now let's run the IV

```
iv <- ivreg(lwage ~ educ | nearc4, data = df)
```

Let's compare the results in a single table, also including the results from the first stage. Notice that we are using the robust standard errors (`HC3` option).

```
options(modelsummary_factory_default = 'kableExtra')
msummary(
  models = list(
    "OLS" = ols
    , "IV" = iv
    , "IV First Stage" = first_stage
  )
  , coef_map = c(educ = "Education", nearc4 = "College in the county")
  , vcov = "HC3"
  , gof_map = "nobs"
  , fmt = 3
)
```

The OLS estimate is smaller than the IV estimate. Notice, however, that the OLS estimate is more precise. This is the consequence of IV using only a fraction of the available variation, the variation due to the variation in the instrument.

We can run some diagnostics test for IV using the following code

	OLS	IV	IV First Stage
Education	0.052	0.188	
	(0.003)	(0.026)	
College in the county			0.829
			(0.107)
Num.Obs.	3010	3010	3010

```
summary(iv, diagnostics = T)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ | nearc4, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.24604 -0.36219  0.01269  0.37938  2.25907
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.76747    0.34886  10.799  < 2e-16 ***
## educ         0.18806    0.02629   7.153 1.06e-12 ***
##
## Diagnostic tests:
##                   df1  df2 statistic  p-value
## Weak instruments    1 3008     63.91 1.84e-15 ***
## Wu-Hausman          1 3007     48.45 4.14e-12 ***
## Sargan              0   NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5569 on 3008 degrees of freedom
## Multiple R-Squared: -0.5739, Adjusted R-squared: -0.5744
## Wald test: 51.17 on 1 and 3008 DF,  p-value: 1.061e-12
```

The bottom table shows the results of two tests. First, it shows the first stage F-statistic under `Weak instruments`. The value of the statistic is high enough to rule out the weak instruments issue. Second, it shows the results of the Durbin-Wu-Hausman test. The test is used to compare OLS (inconsistent) to IV (less precise). If the results are different (which corresponds to a low p-value), that means that we should probably be using IV.

Controls

Researchers often run the IV regression with additional controls. The purpose of controls is typically to increase the precision of the estimates or to close any backdoor paths between the instrument and the outcome.

	OLS	IV	IV First Stage
Education	0.071	0.124	
	(0.004)	(0.049)	
College in the county			0.327
			(0.081)
Num.Obs.	3003	3003	3003

One way to implement the control variables is to explicitly list them in the formula. This, however, can get too repetitive and can lead to errors. Here we will be using the convenient `reformulate` function. We will define our controls once as a group of variables and then simply refer to that group in the regressions. First, let's define the control variables.

```
controls <- c("exper", "black", "south", "married", "smsa")
```

Now we can create a formula with these controls as follows.

```
reformulate(c("educ", controls), "lwage")
```

```
## lwage ~ educ + exper + black + south + married + smsa
```

Let's use this trick to first estimate the OLS with additional controls and then the IV plus the first stage.

```
ols_controls <-
  lm(reformulate(c("educ", controls), "lwage"), data = df)

iv_controls <-
  ivreg(
    reformulate(c("educ", controls), "lwage")
    , reformulate(c("nearc4", controls))
    , data = df
  )

first_stage_controls <-
  lm(reformulate(c("nearc4", controls), "educ"), data = df)
```

Here is the summary of the results with controls.

```
msummary(
  models = list(
    "OLS" = ols_controls
    , "IV" = iv_controls
    , "IV First Stage" = first_stage_controls
  )
  , coef_map = c(educ = "Education", nearc4 = "College in the county")
  , vcov = "HC3"
  , gof_map = c("nobs", "F")
  , fmt = 3
)
```

The OLS estimate increased relative to the case without controls, while the IV estimate decreased. At the same time, the IV estimate became less precise. The results of the first stage changed, too. The estimated effect of the instrument on the treatment is much lower.

We can run the diagnostic tests using the same command as before.

```r
summary(iv_controls, diagnostics = T)
```

```
##
## Call:
## ivreg(formula = lwage ~ educ + exper + black + south + married +
##     smsa | nearc4 + exper + black + south + married + smsa, data = df)
##
## 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 ***
##
## Diagnostic tests:
##                    df1  df2 statistic  p-value
## Weak instruments     1 2996    15.767 7.33e-05 ***
## Wu-Hausman           1 2995     1.219     0.27
## Sargan               0   NA        NA       NA
## ---
## 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
```

The F-statistic is now lower, although not too low to worry about the weak instrument problem. Interestingly, the Durbin-Wu-Hausman test now suggests that the OLS and IV results are not statistically different.