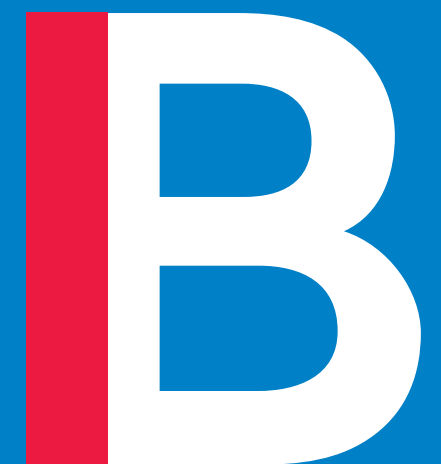




# Instrumental Variables

Beating endogeneity...with music?

by Ralf Martin ([r.martin@imperial.ac.uk](mailto:r.martin@imperial.ac.uk))



# Fixing endogeneity



Endogeneity  
Problem

**Instrumental Variable idea:**

$$Y = \beta X + \epsilon \quad \text{but} \quad E\{\epsilon|X\} \neq 0$$

$$X = X(A, B, Z)$$

- Many different factors are driving  $X$
- Suppose there is at least one factor that is independent from  $\epsilon$ :

$$E\{\epsilon|Z\} = 0$$

- We can then potentially use this variable to identify an unbiased estimate of  $\beta$

## 2 Stage Least Squares Estimator (2 SLS)

Setup:  $Y = \beta_0 + \beta X + \epsilon$

Instrument:  $Z$

Stage 1: Regress  $X$  on  $Z$ ; i.e.

$$X = \pi_0 + \pi Z + \eta$$

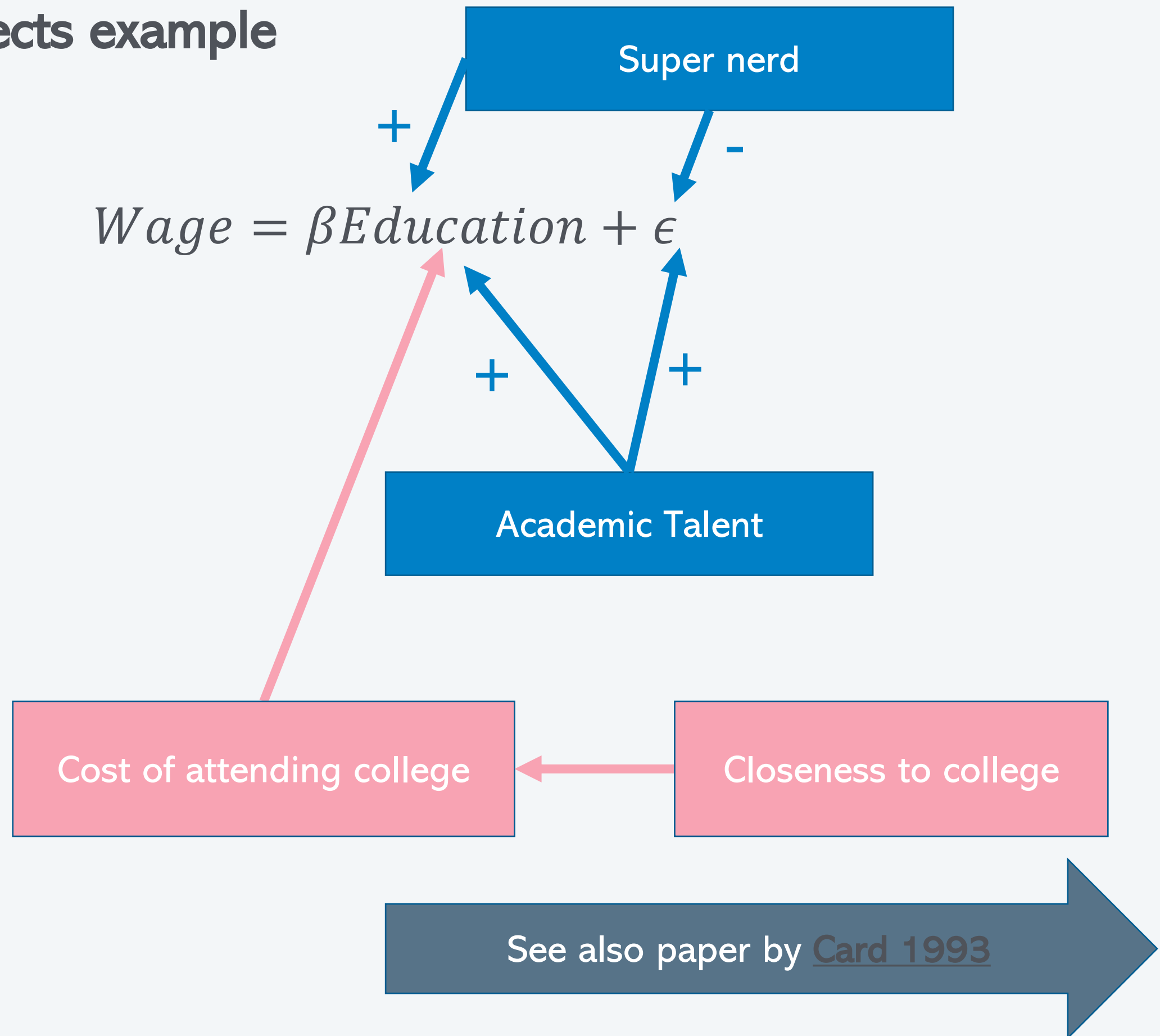
Predict  $X$ :  $\hat{X} = \pi_0 + \hat{\pi}Z$

Stage 2: Regress  $Y$  on  $\hat{X}$  ; i.e.

$$Y = \beta_0 + \beta \hat{X} + \epsilon$$

Unlike  $X$ ,  $\hat{X}$  is independent of  $\epsilon$  because it is only driven by  $Z$   
Hence Stage 2 provides consistent (although not unbiased)  
estimate of  $\beta$ .

# Schooling effects example

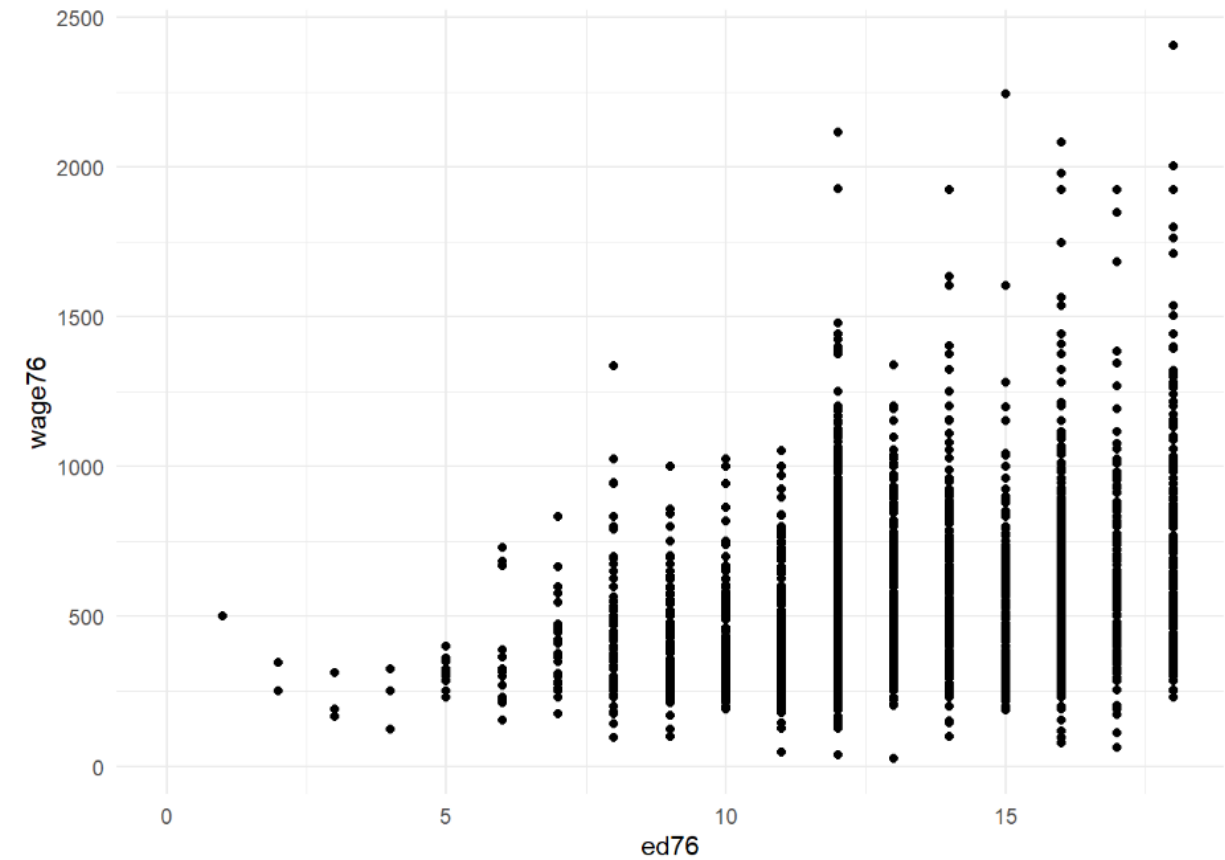


# Schooling example

```
library(AER)
library(dplyr)
df=read.csv("https://www.dropbox.com/s/diecbkq03gfid0p/card1993.csv?dl=1")
#from https://davidcard.berkeley.edu/data_sets.html

lm(wage76 ~ ed76, data = df) %>% summary()
```

```
##
## Call:
## lm(formula = wage76 ~ ed76, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -623.6  -173.7   -33.3   128.3  1687.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   183.934     23.160   7.942 2.78e-15 ***
## ed76          29.566      1.712  17.274 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 251.7 on 3015 degrees of freedom
## (596 observations deleted due to missingness)
## Multiple R-squared:  0.09006, Adjusted R-squared:  0.08875
## F-statistic: 298.4 on 1 and 3015 DF, p-value: 2.78e-15
```



For every additional year of education \$30 more

# Schooling example: The distance instrument

```
first=lm(ed76~nearc4a, data = df)
first %>% summary()
```

```
##
## Call:
## lm(formula = ed76 ~ nearc4a, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6575  -1.6575  -0.6575   2.3425   5.1935
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.80654    0.06343  201.911  <2e-16 ***
## nearc4a       0.85094    0.09041   9.412   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.717 on 3611 degrees of freedom
## Multiple R-squared:  0.02394,    Adjusted R-squared:  0.02367
## F-statistic 88.58 on 1 and 3611 DF,  p-value: < 2.2e-16
```

- If you grow up near a public 4 year college you have nearly an additional year of education on average
- Note that this is super significant (including high  $F > 10$ )

# Implementing IV

New regression command to implement 2SLS  
[part of library(AER)]

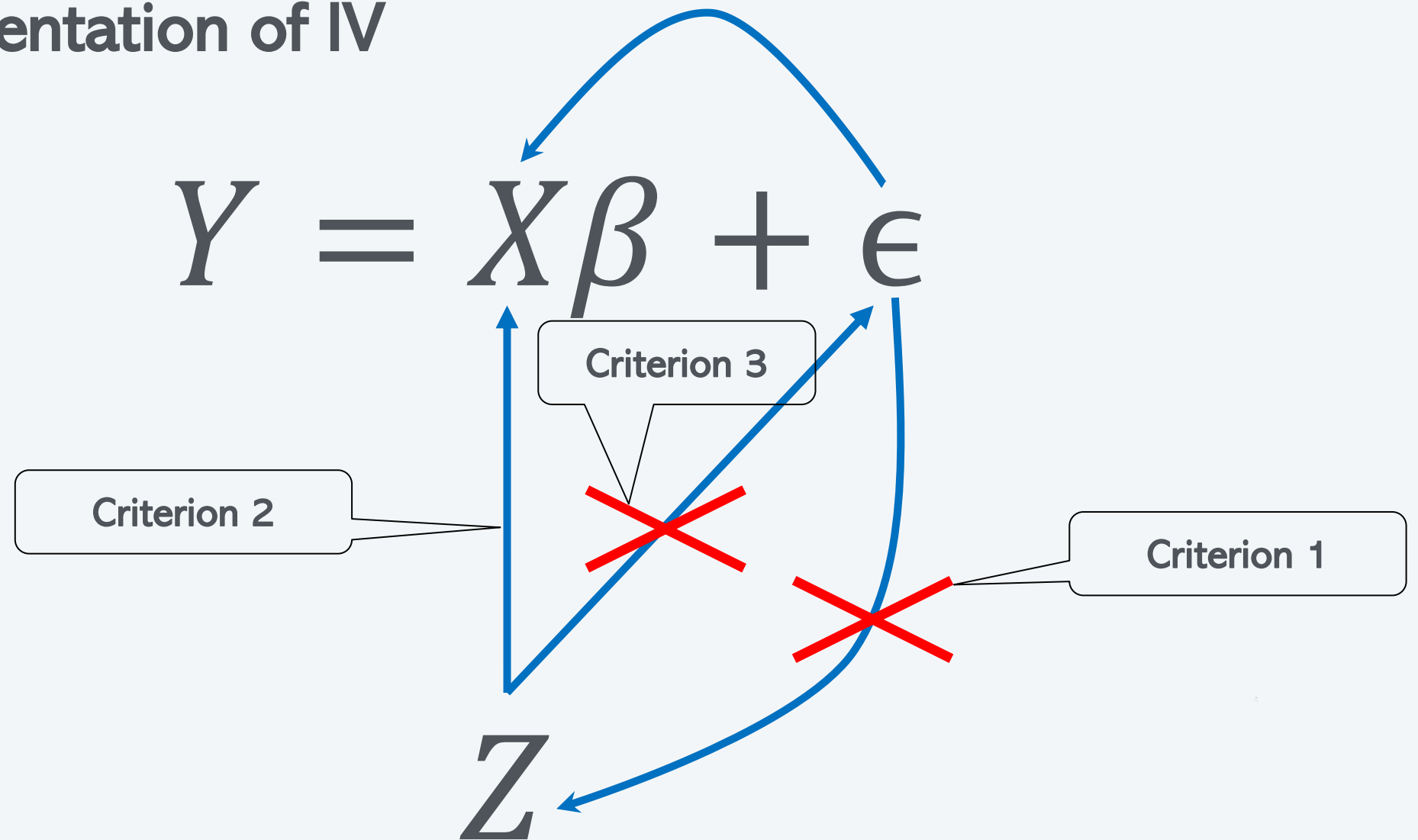
```
iv=ivreg(wage76 ~ ed76 | nearc4a, data=df)
iv %>% summary()
```

```
##
## Call:
## ivreg(formula = wage76 ~ ed76 | nearc4a, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -941.253 -211.663  -6.304  204.517 1683.696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -943.77     198.71    -4.749 2.13e-06 ***
## ed76          114.59      14.97     7.652 2.64e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 339.4 on 3015 degrees of freedom
## Multiple R-Squared:  -0.6547, Adjusted R-squared:  -0.6552
## Wald test: 58.56 on 1 and 3015 DF, p-value: 2.639e-14
```

Seems the effect of education is stronger  
than thought: an extra \$114 for every year

Could mean  
(a) In OLS case we had  
downward bias  
(b) Could be indicative of  
problem with IV

# A graphical representation of IV



## 3 Criteria for instrumental variables

1. Must be independent of shocks
2. Must be driver of variable of interest
3. Must not affect outcome variable other than through variable of interest

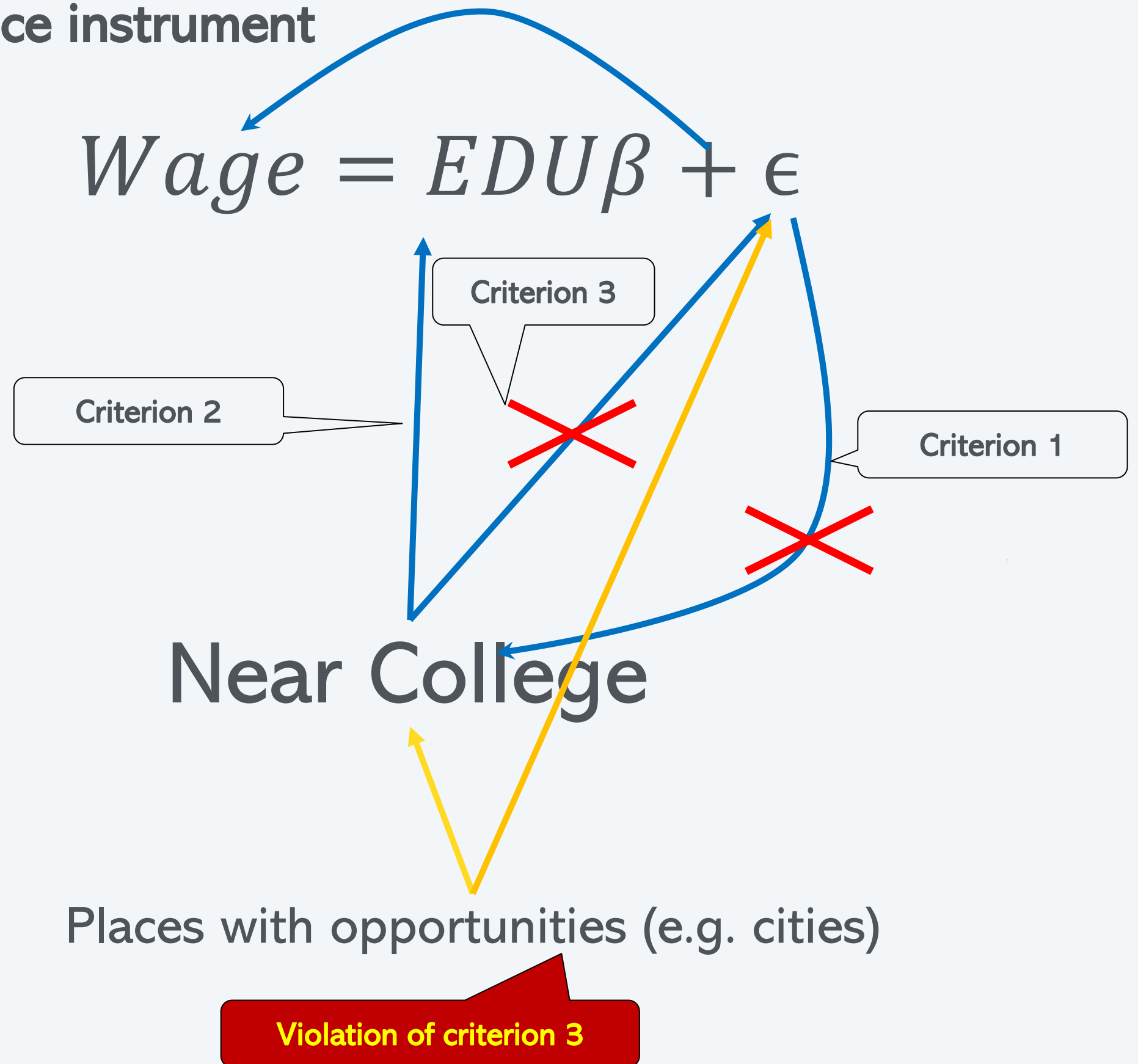
Must be argued

Can be checked

Must be argued



# The College distance instrument



# Conditional IV

Conditioning/Control Variables e.g. controls for nice places

$$Wage = \beta EDU + \beta_2 X + \eta$$

Criterion 2

Criterion 3

Criterion 1

Near College

Places with opportunities (e.g. cities)

# Conditional IV in action

```
iv=ivreg(wage76 ~ ed76+factor(region) +nearc4b+nearc2 | nearc4a+nearc4b
+nearc2+factor(region)
        , data=df )
iv %>% summary()
```

```
##
## Call:
## ivreg(formula = wage76 ~ ed76 + factor(region) + nearc4b + nearc2,
##       data = df, fixed = FALSE, method = "2SLS")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -853.100 -198.498   -6.864  187.265 1630.953
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -729.039    291.987  -2.497  0.01258 *
## ed76           98.811     20.916   4.724 2.42e-06 ***
## factor(region)2 -126.976     40.111  -3.166  0.00156 **
## factor(region)3  -36.559     34.319  -1.065  0.28683
## factor(region)4  -21.313     42.946  -0.496  0.61974
## factor(region)5  -14.037     38.400  -0.366  0.71474
## factor(region)6  -62.001     30.296  -2.047  0.04079 *
## factor(region)7   17.031     24.201   0.704  0.48166
## factor(region)8  -11.016     23.908  -0.461  0.64499
## factor(region)9  -19.767     35.121  -0.563  0.57360
## nearc4b         -8.513     15.013  -0.567  0.57075
## nearc2          27.357     12.357   2.214  0.02691 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 311.4 on 3005 degrees of freedom
## Multiple R-Squared:  -0.3884, Adjusted R-squared:  -0.3935
## Wald test: 17.66 on 11 and 3005 DF, p-value: < 2.2e-16
```

Conditioning/Control Variables e.g. controls for nice places

- Education coefficient remains high and significant

# Checking first stage in conditional IV case

```
first=lm(ed76~nearc4a+factor(region) +nearc4b+nearc2, data = df)
first %>% summary()
```

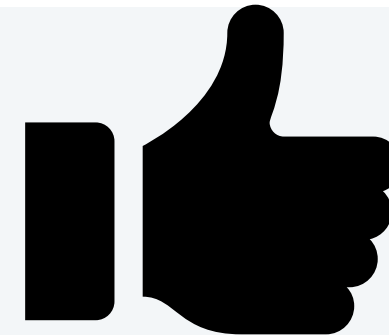
```
##
## Call:
## lm(formula = ed76 ~ nearc4a + factor(region) + nearc4b + nearc2,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3438  -1.6767  -0.2662   2.0324   5.9761
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    13.41923     0.18020   74.469  < 2e-16 ***
## nearc4a         0.66714     0.10785    6.186 6.88e-10 ***
## factor(region)2  0.25742     0.29739    0.866  0.3868
## factor(region)3 -0.88954     0.20580   -4.322 1.58e-05 ***
## factor(region)4 -1.39531     0.21202   -6.581 5.35e-11 ***
## factor(region)5 -1.31319     0.18230   -7.204 7.11e-13 ***
## factor(region)6 -0.17377     0.22849   -0.761  0.4470
## factor(region)7 -0.15300     0.18226   -0.839  0.4013
## factor(region)8 -0.15192     0.18456   -0.823  0.4105
## factor(region)9 -0.54352     0.25847   -2.103  0.0355 *
## nearc4b         0.25475     0.13183    1.932  0.0534 .
## nearc2          0.03420     0.09664    0.354  0.7234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.662 on 3601 degrees of freedom
## Multiple R-squared:  0.06556,    Adjusted R-squared:  0.0627
## F-statistic: 22.97 on 11 and 3601 DF,  p-value: < 2.2e-16
```

IV still significant

- However we need a super strong IV
- Rule of thumb: F-stat of  $H_0: Z=0 > 10$

```
linearHypothesis(first,c("nearc4a=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## nearc4a = 0
##
## Model 1: restricted model
## Model 2: ed76 ~ nearc4a + factor(region) + nearc4b + nearc2
##
##      Res.Df    RSS Df Sum of Sq      F      Pr(>F)
## 1      3602 25791
## 2      3601 25520   1    271.15 38.261 6.882e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# Reduced form: regressing outcome on instrument

```
reduced=lm(wage76~nearc4a+factor(region)+nearc4b+nearc2, data = df)
reduced %>% summary()
```

```
##
## Call:
## lm(formula = wage76 ~ nearc4a + factor(region) + nearc4b + nearc2,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -631.80 -164.44  -37.61  120.43 1735.20
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    600.144    18.934   31.697 < 2e-16 ***
## nearc4a         65.301    11.192    5.835 5.97e-09 ***
## factor(region)2  -79.575    31.983   -2.488  0.01290 *
## factor(region)3 -122.686    21.488   -5.709 1.24e-08 ***
## factor(region)4 -152.045    22.014   -6.907 6.03e-12 ***
## factor(region)5 -142.303    19.136   -7.436 1.34e-13 ***
## factor(region)6  -72.354    24.341   -2.973  0.00298 **
## factor(region)7   -1.002    19.112   -0.052  0.95819
## factor(region)8  -25.757    19.240   -1.339  0.18075
## factor(region)9  -75.947    26.576   -2.858  0.00430 **
## nearc4b          18.110    13.632    1.329  0.18409
## nearc2           29.110     9.963    2.922  0.00351 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 252.2 on 3005 degrees of freedom
## (596 observations deleted due to missingness)
## Multiple R-squared:  0.08978,    Adjusted R-squared:  0.08644
## F-statistic: 26.94 on 11 and 3005 DF,  p-value: < 2.2e-16
```

If reduced form is not significant IV won't be either

# IV health warning

Don't mess with weak instruments

- Estimator is consistent if the probability limit is equal to true parameter.
- Note 2SLS IV is consistent but not unbiased

## - Properties of IV with a poor instrumental variable

- IV may be much more inconsistent than OLS if the instrumental variable is not completely exogenous and only weakly related to explanatory var x

$$\hat{\beta}_{OLS} \approx \beta + \text{Corr}(X, \epsilon) \frac{\sigma_{\epsilon}}{\sigma_X}$$

$$\hat{\beta}_{IV} \approx \beta + \frac{\text{Corr}(\epsilon, Z) \sigma_{\epsilon}}{\text{Corr}(X, Z) \sigma_X}$$

There is no problem if the instrumental variable is really exogenous. If not, the bias will be the larger the weaker the correlation with x.

If  $\text{Corr}(X, Z)$  is small = weak instrument IV could be more biased than OLS even if  $\text{Corr}(\epsilon, Z) < \text{Corr}(\epsilon, X)$

That's why we want the first stage F-statistic to be large!



# Extra Slides



# More instrumental variable examples





# Additional controls many IV's example: Family size

$$Y = \beta FamSize + \epsilon$$

Are large families good or bad?



Outcome	Regression (1)
Highest Grade Completed	-0.145 (0.005)
Years of Schooling $\geq 12$	-0.029 (0.001)
Some College (age $\geq 24$ )	-0.023 (0.001)
College Graduate (age $\geq 24$ )	-0.015 (0.001)

What do you think?

- Many potential confounders. Family size is by and large a choice
- Poorer and less educated parents tend to have more kids
- The same factors could drive outcomes

# Family size and children outcomes

(Study by Angrist, Lavy & Schlosser)

However, there is randomness in family size as well

Two factors that are out of control of controlling parents:

- Occurrence of twins
- Sex of baby

Ok those might meet criteria 1 from earlier but how about criteria 2?

- Twins: families might only have planned for 2 kids, but when they had twins they un-intentionally had 3
- Many families have preference for a sex mix (a boy and a girl)
- Hence, if they have two kids of the same sex they are more likely to carry on having more kids

# Second stage

Outcome	Regression Estimates (1)	2SLS Estimates		
		Twins Instruments (2)	Same-sex Instruments (3)	Twins & Samesex (4)
Years of Schooling	-0.145 (0.005)	0.174 (0.166)	0.318 (0.210)	0.237 (0.128)
High School Graduate	-0.029 (0.001)	0.030 (0.028)	0.001 (0.033)	0.017 (0.021)
Some College (for age $\geq 24$ )	-0.023 (0.001)	0.017 (0.052)	0.078 (0.054)	0.048 (0.037)
College graduate (for age $\geq 24$ )	-0.015 (0.001)	-0.021 (0.045)	0.125 (0.053)	0.052 (0.032)

Notes: This table reports OLS and 2SLS estimates of the effect of family size on schooling. OLS estimates appear in Column (1). Columns (2), (3) and (4) show 2SLS estimates constructed using the instruments indicated in column headings. Standard errors are reported in parentheses.

# 2SLS with multiple instruments & multiple endogenous explanatory variables

Setup:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_c X_c + \epsilon$

Instruments:  $Z_1, Z_2, \dots$

At least as many Z as endogenous X

Stage 1: Regress all X on all Z (and  $X_c$ ); e.g.

$$X_1 = \pi_{01} + \pi_{11}Z_1 + \pi_{21}Z_2 + \pi_{c1}X_c + \eta_1$$

$$X_2 = \pi_{02} + \pi_{12}Z_1 + \pi_{22}Z_2 + \pi_{c2}X_c + \eta_2$$

Predict all endogenous X:  $\hat{X}_1, \hat{X}_2$

Stage 2: Regress Y on predicted endogenous X; i.e.

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 \hat{X}_2 + \beta_c \hat{X}_c + \epsilon$$

# Another example: Subsidies and Unemployment

$$\Delta \ln Unemp = \beta \Delta NGE + \epsilon$$

Change in (log)  
unemployment between  
2002 and 1997

Change in support  
rate

## Some Causal Effects of an Industrial Policy

Chiara Criscuolo  
Ralf Martin  
Henry G. Overman  
John Van Reenen

AMERICAN ECONOMIC REVIEW  
VOL. 109, NO. 1, JANUARY 2019  
(pp. 48-85)

Download Full Text PDF  
(Complimentary)

The 10 worst places to live in England revealed - and the results may surprise you!

23 July 2019, 13:04 | Updated: 23 July 2019, 15:49



The worst towns in the UK have been revealed in this 2019 survey. Picture: Getty



# Instrument construction

Know your history



How much support an area gets is based on how

- a) Various indicators of past deprivation (GDP, unemployment, etc.)
- b) Weightings for different indicators

In reality not linear but let's use as approximation

$$NGE_t = f(D_t, W_{r(t)}) \approx \sum_k D_{kt} W_{kr(t)}$$

Indicators of Deprivation

Weights of different indicators  
(Determined by EU rules)

Rules that apply in period t.  
Change every 7 years; e.g. in  
2000

# Instrument construction



Conditional IV

$$\Delta \ln Unemp = \beta \Delta NGE + \beta D_{pre2000} + v$$

$$\Delta NGE = \sum_k D_{k,2002} W_{k,post2000} - D_{k,1997} W_{k,pre2000}$$

$$\Delta Z = \sum_k D_{k,pre2000} W_{k,post2000} - D_{k,pre2000} W_{k,pre2000}$$

After controlling for  $D_{k,pre2000}$ ,  $\Delta Z$  does no longer depend on  $v$  and thus becomes valid instrument

# Implementation

```
df=read.csv( "https://www.dropbox.com/s/8pdffaq268v7m8o/unempprep.csv?dl=1")
# Simple OLS
regOLS=lm( DDDln1Punemp ~ DDDNGE ,df)
summary(regOLS)

##
## Call:
## lm(formula = DDDln1Punemp ~ DDDNGE, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5451  -0.1907   0.0165   0.2109   2.8601
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  -0.462220    0.003648 -126.703  < 2e-16 ***
## DDDNGE       -0.221062    0.036012  -6.138  8.62e-10 ***
```

A change in support from 0 to 10% (0-0.1) will reduce unemployment by 2.2%



# IV

List of historic deprivation controls

```
controls=c("gdp91","manufshare_1991","popdens_1981","current_unemprate1991",  
           "actrate_1991","resid_emp_rate92")  
fff=paste(controls, collapse = "+")
```

Now run iv:

```
fffviv =paste0( "DDDln1Punemp ~ DDDNGE+",fff, "| DDDxnivav +",fff)  
regIV=ivreg( fffviv ,data=df)  
summary(regIV)
```

```
##  
## Call:  
## ivreg(formula = fffviv, data = df)  
##  
## Residuals:  
##      Min      1Q   Median      3Q      Max   
## -3.52984 -0.18444  0.01339  0.20575  2.74181   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)   -0.7851275   0.0820543  -9.568 < 2e-16 ***  
## DDDNGE        -0.4546888   0.1131435  -4.019 5.89e-05 ***  
## gdp91          0.0030923   0.0003235   9.559 < 2e-16 ***  
## manufshare_1991 0.5714105   0.0481121  11.877 < 2e-16 ***  
## popdens_1981    0.0004555   0.0004222   1.079  0.281      
## current_unemprate1991 -2.1341897  0.4972815  -4.292 1.79e-05 ***  
## actrate_1991    0.4900480   0.0628717   7.794 7.07e-15 ***  
## resid_emp_rate92 -0.3915193   0.0876794  -4.465 8.08e-06 ***  
## ---
```

$\Delta Z$

A change in support from 0 to 10% (0-0.1) will reduce unemployment by 4.5%

# First Stage

Check first stage

```
fffffs =paste0( "DDDNGE~DDDxnivav+",fff)
```

```
regFS=lm( fffffs ,df)
```

```
summary(regFS)
```

```
##
```

```
## Call:
```

```
## lm(formula = fffffs, data = df)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.38214 -0.01537  0.00344  0.02657  0.43048
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.210e-01  2.303e-02  -9.599  < 2e-16 ***
## DDDxnivav      1.015e+00  2.932e-02  34.627  < 2e-16 ***
## gdp91         -4.575e-04  8.145e-05  -5.617  2.00e-08 ***
## manufshare_1991 -1.038e-01  1.146e-02  -9.059  < 2e-16 ***
## popdens_1981     1.969e-04  1.070e-04   1.839  0.06592 .
## current_unemprate1991 1.061e+00  1.372e-01   7.732  1.16e-14 ***
## actrate_1991    -4.786e-02  1.592e-02  -3.007  0.00265 **
## resid_emp_rate92  3.674e-01  2.482e-02  14.803  < 2e-16 ***
## ---
```

The IV variable is called DDDxnivav

# F-test of $Z=0$



```
linearHypothesis(regFS, "DDDxnivav=0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## DDDxnivav = 0
##
## Model 1: restricted model
## Model 2: DDDNGE ~ DDDxnivav + gdp91 + manufshare_1991 +
popdens_1981 +
##      current_unemprate1991 + actrate_1991 + resid_emp_rate92
##
##      Res.Df      RSS Df Sum of Sq    F    Pr(>F)
## 1   10757  104.652
## 2   10756   94.156  1    10.496 1199 < 2.2e-16 ***
## ---
```

F statistic is larger  
10...hurray!!!

# Reduced Form

Reduced Form:

```
fffrf =paste0( "DDDln1Punemp~DDDxnivav+",fff)  
summary(lm( fffrf ,df))
```

```
##  
## Call:  
## lm(formula = fffrf, data = df)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -3.5300 -0.1834  0.0138  0.2023  2.7342   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)   -0.6846189   0.0898471   -7.620 2.75e-14 ***  
## DDDxnivav     -0.4616909   0.1144068   -4.036 5.49e-05 ***  
## gdp91          0.0033003   0.0003178   10.386 < 2e-16 ***  
## manufshare_1991 0.6185964   0.0446947   13.840 < 2e-16 ***  
## popdens_1981    0.0003659   0.0004176    0.876  0.381        
## current_unemprate1991 -2.6164480   0.5352012   -4.889 1.03e-06 ***  
## actrate_1991    0.5118111   0.0621070    8.241 < 2e-16 ***  
## resid_emp_rate92 -0.5585603   0.0968272   -5.769 8.21e-09 ***  
## ---
```

IF reduced form  
is not significant  
IV will not be  
either

# Panic on the streets of London

(Study by Draca, Machin and Witt)



Does more police on the street lead to less crime?

Simple regression?

$$Crime_{Area} = \beta Policetime_{Area} + \epsilon_{Area}$$



Police go where the crime is

# Panic on the streets of London

(Study by Draca, Machin and Witt)



Solution:

Use



July 7 bombings of 2005 led to increases in police in certain areas (central London, near tube stations) which had nothing to do with crimes such as pick pocketing etc.

Results: 10% more police activity reduces crime by 3 to 4 percent.

# Panic on the streets of London

$$Crime_{Post7/7} - Crime_{Pre\ 7/7}$$

$$\Delta Crime_a = \beta \Delta Police_a + \epsilon_a$$

$$Z_a = \text{area Close to terror target}$$



# Summary

- Endogeneity is often a problem:  $X(\epsilon)$
- However,  $X$  is also driven by other factors
- If we can find data on at least one other factor  $Z$  which is independent of  $\epsilon$  we can do 2SLS IV
- Can combine with using various other controls to make it more plausible that remaining error  $\epsilon$  is indeed independent of  $Z$
- Need to ensure strong first stage
- Finding IVs is a bit of an art

