Intermediate Econometrics

19th 2025 - Vincenzo Gioia

Instrumental Variable Regression

IV regression

- Can eliminate the bias when $\mathbb{E}[\varepsilon|X] \neq 0$, using an instrumental variable (that we will denote as Z)
- Cases when $\mathbb{E}[\varepsilon|X] \neq 0$

1. **Omitted variable bias**, due to a variable correlated with X
2. **Simultaneity causality bias**, that happens when X causes Y and Y causes X
3. **Errors-in-Variable bias**, when the variables are measured with errors

IV regression or OLS regression

OLS regression

- Estimate the effect of X on Y controlling for other predictors: **It works only if X is uncorrelated with errors**
- It requires exogeneity

$$\text{cov}(X, \varepsilon) = 0$$

- It is suitable when **there is no endogeneity (or it is negligible)** or you have good control variables to include in the regression

IV regression

- Estimate the effect of X on Y using an instrument Z
- It is important to consider when the X variable is endogenous, that is correlated with the errors (for the reasons explained in the previous slide)
- It should be used when **there is a strong (and reported) source of endogeneity and exist a valid instrument** (not always easy to detect)

Instrumental Variable Estimator

General Idea of IV

- The instruments allow to get an exogenous source of variation of the covariate, that is a source of variation that has nothing to do with the process of interest

1. do not have a direct effect on the response
2. have an indirect effect on the response because of their correlation with the endogenous covariates

Conditions on the instrument (Z)

1. **Relevance**: The instrument must effect (X), that is

$$\mathrm{cov}(Z, X) \neq 0$$

2. **Exclusion**: The instrument must be incorrelated with the errors (ε), that is

$$\mathrm{cov}(Z, \varepsilon) = 0$$

- If at least one of the condition is missing, the IV estimate is not valid

Simple Instrumental Variable Estimator

A key difference on variables

- **Endogenous**: variable correlated with the error ε
- **Exogeneous**: variable uncorrelated with the error ε
- IV focuses on cases where the covariate X is endogeneous and the instrument Z is exogeneous

Framework: Simple instrumental variable estimator

- Let us consider the regression

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

- Let z_i an instrumental variable, that is correlated with x_i but it is uncorrelated with ε_i
- We consider the **exactly/just identified case** (the number of instruments, one in this case, is equal to the number of regressors, one in this case)

Simple Instrumental Variable Estimator

Two Stage Least Squares (TSLS) Estimator

- Stage 1: Regress X on Z via OLS

$$x_i = \pi_1 + \pi_2 z_i + \varepsilon_i^*$$

and calculate the predicted values

$$\hat{x}_i = \hat{\pi}_1 + \hat{\pi}_2 z_i$$

- Stage 2: In the regression of interest, substitute the observed x_i with the predicted $\hat{x}_i$,

$$y_i = \beta_1 + \beta_2 \hat{x}_i + \varepsilon_i$$

Instrumental variable estimator

Two Least Square Estimator

$$Y = X\beta + \varepsilon$$

- Let's denote with Z the model matrix of the instruments and with X the model matrix of the regressors
- The estimator $\hat{\beta}^{TSLS}$ is

  $$\hat{\beta}^{TSLS} = (X^\top P_Z X)^{-1} X^\top P_Z Y$$

  where $P_Z = Z(Z^\top Z)^{-1} Z^\top$
- It is clear the reason of substituting $\hat{X}$ onto X. Indeed, one of the results of the OLS is that $\hat{X} = P_Z X$ and so

  $$\hat{\beta} = (\hat{X}\hat{X}^\top)^{-1} \hat{X}^\top Y$$

Instrumental variable estimator

Two Least Square Estimator

1. Consistent ($\hat{\beta} \overset{p}{\to} \beta$)
2. The estimate of the variance of the estimator is, assuming spherical disturbances, equal to

$$\hat{\mathbb{V}}(\hat{\beta}^{TSLS}) = \hat{\sigma}^2 (X^\top P_Z X)^{-1}$$

   where

$$\hat{\sigma}^2 = \frac{1}{n - \dim(\beta)} \sum_{i=1}^{n} e_i^2$$

   with $e_i = y_i - X\hat{\beta}^{TSLS}$

- From $\hat{\mathbb{V}}(\hat{\beta}^{TSLS})$ we can obtain the estimate of the standard error of the β's

Econometric example

Segregation effects on urban poverty and inequality

- This is an example aiming to investigate the effect of segregation on urban poverty and inequality (it is extracted from Ananat, 2011)
- As outcome variable, it is considered the **poverty rate for the black populations of 121 American cities**
- **The level of segregation is measured by a dissimilarity index**. This index vary from 0 (no segregation) to 1 (perfect segregation)
- In that work, the author suggest that the way cities were subdivided by railroads into a large number of neighborhoods can be used as an instrument.
- Moreover, the tracks were mostly built during the nineteenth century, prior to the great migration (between 1915 to 1950) of African Americans from the south.
- So, it is used an index that is 0 if the city is completely undivided and tends to 1 if the number of neighborhoods tends to infinity.

Econometric example

Segregation effects on urban poverty and inequality

- The dataset is named **tracks_side**, and it is in .csv format
- From the summary we can see:

1. the poverty rate (*povb*) ranges from 0.09 to 0.50 with a median value of 0.26;
2. the segregation index (*segregation*) ranges from 0.33 to 0.87, with a median value of 0.57;
3. the rail tracks index (*raildiv*) ranges from 0.24 to 0.99 in the sample, with a median value of 0.74;

```r
tracks_side <- read.csv("tracks_side.csv")
summary(tracks_side[,-1])
```

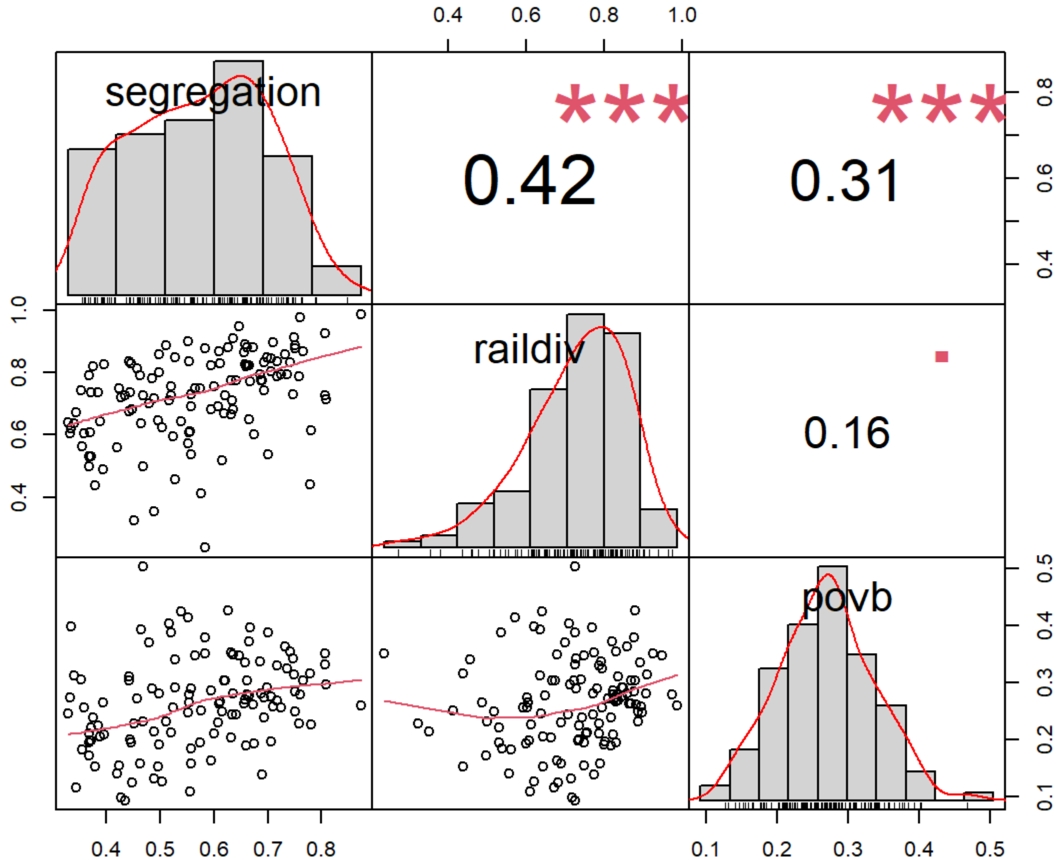
```
      povb            segregation         raildiv      
 Min.   :0.09258   Min.   :0.3289   Min.   :0.2376  
 1st Qu.:0.20886   1st Qu.:0.4569   1st Qu.:0.6383  
 Median :0.26401   Median :0.5745   Median :0.7420  
 Mean   :0.26410   Mean   :0.5687   Mean   :0.7233  
 3rd Qu.:0.31252   3rd Qu.:0.6731   3rd Qu.:0.8299  
 Max.   :0.50422   Max.   :0.8728   Max.   :0.9868  
```

Simple instrumental variable estimator

Two Least Square Estimator

- We can notice a strong correlation between segregation and raildiv and povb, while a low correlation between raildiv and povb

```
library(PerformanceAnalytics)
chart.Correlation(tracks_side[,c(3,4,2)])
```



segregation
raildiv
povb
0.42
0.31
0.16
***
***
0.4
0.5
0.6
0.7
0.8
1.0
0.1
0.2
0.3

Simple instrumental variable estimator

Two Least Square Estimator

- We'll focus on the effect of segregation on the poverty rate of black people: let's fit the model

$$povb_i = \beta_1 + \beta_2 segregation_i + \varepsilon_i$$

- The coefficient of segregation is positive and significant. It indicates that a 1 point increase of the segregation index raises the poverty rate of black people by about 0.18 point

```
lm_yx <- lm(povb ~ segregation, tracks_side)
summary(lm_yx)$coefficients
```

```
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 0.1607271 0.03003485 5.351352 4.300269e-07
segregation 0.1817783 0.05139239 3.537067 5.777075e-04
```

Simple instrumental variable estimator

Two Least Square Estimator

- The correlation between segregation and bad economic outcome for black people is well established but, according to the author, **the OLS estimator cannot easily be considered as a measure of the causal relationship of segregation on income, as there are some other variables that both influence segregation and outcome for black people**
- As an example, the situation of Detroit is described, which is a highly segregated city with poor economic outcomes, but other characteristics of the city (political corruption, legacy of a manufacturing economy) can be the cause of these two phenomena
- Therefore, **the OLS estimator is suspected to be biased and inconsistent because of the omitted variable bias**.
- The instrumental variable estimator can be used in this context, but it requires the use of a good instrumental variable, i.e., a variable which is correlated with the endogenous covariate (segregation), but not directly with the response (the poverty rate).

Simple instrumental variable estimator

Two Least Square Estimator: First stage regression

- Regression of segregation on raildiv results in a coefficient of raildiv positive and highly significant

```r
lm_xz <- lm(segregation ~ raildiv, tracks_side)
summary(lm_xz)
```

```
Call:
lm(formula = segregation ~ raildiv, data = tracks_side)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.23135 -0.10322  0.00791  0.08834  0.32227 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.27970    0.05866   4.768 5.32e-06 ***
raildiv      0.39954    0.07961   5.019 1.84e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1234 on 119 degrees of freedom
Multiple R-squared:  0.1747,	Adjusted R-squared:  0.1678 
```

Simple instrumental variable estimator

Two Least Square Estimator: First stage regression

- It is important to check that the correlation between the covariate and the instrument is strong enough to get a precise IV estimator. This can be performed by computing

1. the coefficient of correlation
2. the R^2 or the F statistic of the first stage regression

```
with(tracks_side, cor(segregation, raildiv))
```

```
[1] 0.417972
```

```
summary(lm_xz)$r.squared
```

```
[1] 0.1747006
```

```
summary(lm_xz)$fstatistic
```

```
    value     numdf     dendf 
 25.19009   1.00000 119.00000 
```

```
pf(25.19009, 1, 119, lower.tail = F)
```

```
[1] 1.840434e-06
```

Simple instrumental variable estimator

Two Least Square Estimator: Second stage regression

- The IV estimator can be obtained by regressing the response on the fitted values of the first stage regression
- The IV estimator is larger than the OLS estimator (0.23 vs. 0.18).
- An increase of 1 point of segregation would increase povb of 0.23 points

```
lm_yhx <- lm(povb ~ fitted(lm_xz), tracks_side)
summary(lm_yhx)$coefficients
```

```
               Estimate Std. Error  t value   Pr(>|t|)
(Intercept)   0.1326782 0.07286888 1.820780 0.07115287
fitted(lm_xz) 0.2310998 0.12751058 1.812397 0.07244604
```

Simple instrumental variable estimator

Two Least Square Estimator: Note

- The IV estimate can be obtained by dividing the OLS coefficients of the regressions of y on z and of x on z.
- A 1 point increase of raildiv is associated with a 0.4 point increase of the discrimination index and with a 0.09 point of the poverty rate. Therefore, the 0.4 point increase of segregation increases povb by 0.09, which means that an increase of 1 point of segregation would increase povb by 0.09/0.4 = 0.23

```r
lm_yz <- lm(povb ~ raildiv, tracks_side)
coef(lm_yz)[2]
```

```
   raildiv 
0.09233439 
```

```r
coef(lm_yz)[2] / coef(lm_xz)[2]
```

```
  raildiv 
0.2310998 
```

Simple instrumental variable estimator

Two Least Square Estimator: The ivreg() function

- The IV regression can be obtained easily using the ivreg() function (in the homonimous package)

```r
library(ivreg)
ivFit1 <- ivreg(povb ~ segregation | raildiv, data = tracks_side)
summary(ivFit1)
```

```
Call:
ivreg(formula = povb ~ segregation | raildiv, data = tracks_side)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.15319 -0.04358 -0.01011  0.04000  0.26335 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  0.13268    0.07054   1.881   0.0624 .
segregation  0.23110    0.12343   1.872   0.0636 .

Diagnostic tests:
                 df1 df2 statistic  p-value    
Weak instruments   1 119    25.190 1.84e-06 ***
Wu-Hausman         1 118     0.194    0.661    
```

Simple instrumental variable estimator

Two Least Square Estimator: By hand

1. $\hat{\beta}^{TSLS} = (X^\top P_Z X)^{-1} X^\top P_Z Y$, where $P_Z = Z(Z^\top Z)^{-1} Z^\top$
2. $\hat{\mathbb{V}}(\hat{\beta}) = \hat{\sigma}^2 (X^\top P_Z X)^{-1}$, where $\hat{\sigma}^2 = \frac{1}{n - \dim(\beta)} \sum_{i=1}^n e_i^2$ with $e_i = y_i - X\hat{\beta}^{TSLS}$

```
X <- model.matrix(~segregation, data = tracks_side)
Z <- model.matrix(~raildiv, data = tracks_side)
PZ <- Z%*%solve(t(Z)%*%Z)%*%t(Z)
B <- solve(t(X)%*%PZ%*%X)%*%t(X)%*%PZ%*%tracks_side$povb
B
```

```
                 [,1]
(Intercept) 0.1326782
segregation 0.2310998
```

```
M <-  solve(t(X)%*%PZ%*%X)
V1 <- sqrt(diag(sum(residuals(ivFit1)^2)/(nrow(tracks_side)-2)  * M))
V1
```

```
(Intercept) segregation
 0.07053777  0.12343145
```

Instrumental variable estimator

A special case: the Wald estimator

- The Wald estimator is the special case of the instrumental variable estimator where the instrument is a binary variable and therefore defines two groups ($z = 0$ and $z = 1$).
- The Wald estimate is then

$$\hat{\beta}_2 = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$$

Instrumental variable estimator

Example: long-term effects of slave trade

- **Africa experienced poor economic performances during the second half of the twentieth century, which can be explained by its experience of slave trade and colonialism**
- In particular, **slave trade may induce long-term negative effects on the economic development of African countries** because of induced corruption, ethnic fragmentation and weakening of established states
- Africa experienced, between 1400 and 1900 four slave trades: the trans-Atlantic slave trade (the most important), but also the trans-Saharan, the Red Sea and the Indian Ocean slave trades.
- Not including those who died during the slave trade process, **about 18 millions slaves were exported from Africa**
- Nunn (2008) conducted a quantitative analysis of the effects of slave trade on economic performances, by regressing the 2000 GDP per capita of 52 African countries on a measure of the level of slave extraction.

Instrumental variable estimator

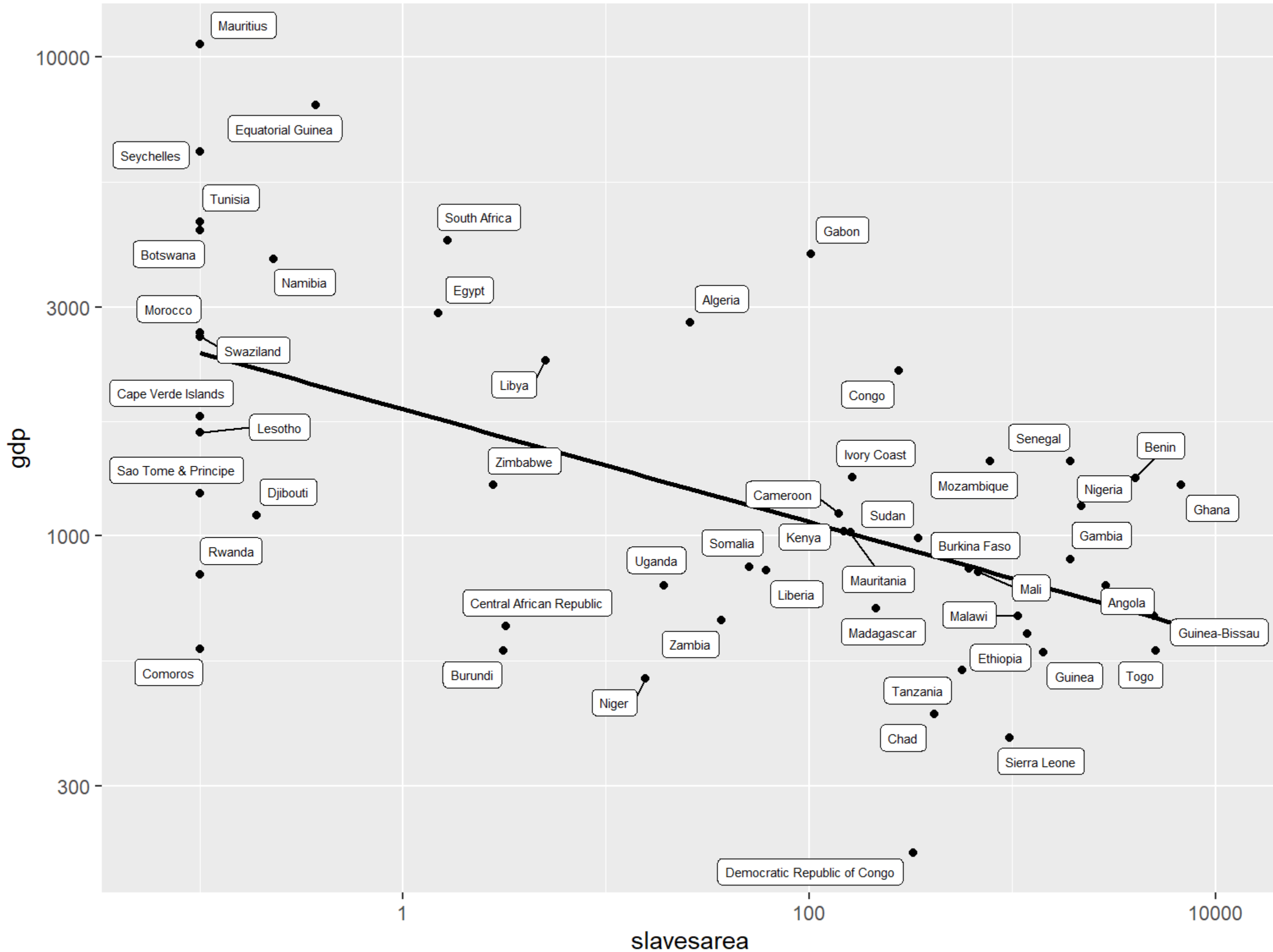
Example: long-term effects of slave trade

- The slave_trade data set is available in the **necountries** R package
- **The response is gdp and the main covariate is a measure of the level of slave extraction**, which is the number of slaves normalized by the area of the country
- Let's explore a scatterplot, using log scales for both variables, which clearly indicates a negative relationship between slaves extraction and per capita GDP in 2000

```
library(necountries)
sltd <- as.data.frame(slave_trade)
```

Example: long-term effects of slave trade

```r
library(ggplot2)
sltd %>% ggplot(aes(slavesarea, gdp)) + geom_point() +
  scale_x_continuous(trans = "log10", expand = expansion(mult = c(.1))) +
  scale_y_log10() + geom_smooth(method = "lm", se = FALSE, color = "black") +
  ggrepel::geom_label_repel(aes(label = country), size = 2, max.overlaps = Inf)
```



Mauritius
Equatorial Guinea
Seychelles
Tunisia
South Africa
Gabon
Botswana
Namibia
Egypt
Morocco
Algeria
Swaziland
Libya
Cape Verde Islands
Congo
Lesotho
Senegal
Benin
Ivory Coast
Zimbabwe
Sao Tome & Principe
Mozambique
Nigeria
Djibouti
Cameroon
Ghana
Sudan
Gambia
Kenya
Somalia
Burkina Faso
Rwanda
Uganda
Mauritania
Mali
Liberia
Central African Republic
Angola
Malawi
Madagascar
Guinea-Bissau
Zambia
Ethiopia
Comoros
Burundi
Guinea
Togo
Tanzania
Niger
Chad
Sierra Leone
Democratic Republic of Congo
10000
3000
1000
300
gdp
1
100
10000
slavesarea

Example: long-term effects of slave trade

Example: long-term effects of slave trade

- Nunn (2008) presents a series of linear regressions, with different sets of controls
- We must work with the covariate *colony*: grouping the levels having low observations

```
table(sltd$colony)
```

```
    none       uk   france portugal  belgium    spain  germany    italy 
       2       18       21        5        3        1        1        1 
```

```
levels(sltd$colony) <- c(levels(sltd$colony), "other")
sltd[sltd$colony == "spain" |
     sltd$colony == "germany" |
     sltd$colony == "italy" |
      sltd$colony ==  "none", ]$colony <- "other"
sltd$colony <- factor(sltd$colony,
                      levels = c("other", "uk", "france", "portugal","belgium"))
table(sltd$colony)
```

```
   other       uk   france portugal  belgium 
       5       18       21        5        3 
```

Example: long-term effects of slave trade

Example: long-term effects of slave trade

- We just consider Nunn's first specification, which includes only dummies for the colonizer as supplementary covariates
- The coefficient is negative and highly significant: it implies that a 10% increase of (log) slave extraction induces a reduction of 1% of (log) GDP per capita

```r
slaves_ols <- lm(log(gdp) ~ log(slavesarea) + colony, data = sltd)
summary(slaves_ols)$coefficients
```

```
                  Estimate Std. Error    t value     Pr(>|t|)
(Intercept)      7.8853270 0.29154346 27.0468322 6.883311e-30
log(slavesarea) -0.1231455 0.02343012 -5.2558652 3.707242e-06
colonyuk        -0.2716053 0.32610941 -0.8328657 4.092240e-01
colonyfrance    -0.3150688 0.32213495 -0.9780648 3.331583e-01
colonyportugal  -0.4176166 0.40835556 -1.0226787 3.118082e-01
colonybelgium   -1.5420188 0.46960874 -3.2836246 1.962937e-03
```

Example: long-term effects of slave trade

Example: long-term effects of slave trade

- As noticed by Nunn (2008), the estimation of the effect of slave trade on GDP can be inconsistent for two reasons:

1. **the level of slave extraction**, which is based on information of the ethnicity of individual slaves and then aggregated at the current countries' level **can be prone to error of measurement**; moreover, for countries inside the continent (compared to coastal countries), a lot of slaves died during the journey to the coastal port of export, so the **level of extraction may be underestimated for these countries**
2. **the average economic conditions may be different for countries that suffered a large extraction, compared to the others**; in particular, if countries where the trade was particularly important were poor, their current poor economic conditions can be explained by their poor economic conditions 600 years ago and not by slave trades

Example: long-term effects of slave trade

Example: long-term effects of slave trade

1. **Measurement error induces an attenuation bias, which means that without measurement error, the negative effect of slave trades on GDP per capita would be stronger**
2. **The second effect would induce an upward-bias (in absolute value) of the coefficient on slave trades**

- But, actually, Nunn (2008) showed that areas of Africa that suffered the most slave trade were in general not the poorest areas, but the most developed ones
- In this case, **the OLS estimator would underestimate the effect of slave trades on GDP per capita**
- Nunn (2008) then performs instrumental variable regressions, **using as instruments the distance between the centroid of the countries and the closest major market for the four slave trades** (for example Mauritius and Oman for the Indian Ocean slave trade and Massawa, Suakin and Djibouti for the Red Sea slave trade)

Example: long-term effects of slave trade

Example: long-term effects of slave trade

- The IV regression can be performed by first regressing the endogenous covariate on the external instruments (atlantic, indian, redsea and sahara) and on the exogenous covariates (here colony, the factor indicating the previous colonizer)

```r
slaves_first <- lm(log(slavesarea) ~ colony + atlantic + indian +
                       redsea + sahara, sltd)
summary(slaves_first)$coefficients
```

```
                  Estimate Std. Error    t value     Pr(>|t|)
(Intercept)     32.2166845  7.9986187  4.0277810 0.0002249973
colonyuk         1.3600990  1.9461550  0.6988647 0.4883995569
colonyfrance    -0.5131590  1.8397132 -0.2789342 0.7816324508
colonyportugal   1.0661060  2.3156511  0.4603915 0.6475550983
colonybelgium   -1.8997964  2.6699544 -0.7115464 0.4805882807
atlantic        -1.5157978  0.3820335 -3.9677086 0.0002706148
indian          -1.1080040  0.4100331 -2.7022306 0.0098190753
redsea          -0.3836068  0.7646345 -0.5016865 0.6184471653
sahara          -2.5691599  0.8809979 -2.9161931 0.0056097945
```

Example: long-term effects of slave trade

Example: long-term effects of slave trade

- The 4 instruments and the exogenous covariate explain more than 30% of the variance of slave extraction; the F-test just suggest a moderate improvement of considering covariates for explaining the slave extraction

```r
summary(slaves_first)$r.squared
```

```
[1] 0.3314668
```

```r
summary(slaves_first)$fstatistic
```

```
   value    numdf    dendf 
 2.66499  8.00000 43.00000 
```

```r
pf(2.66499, 8,  43, lower.tail = FALSE)
```

```
[1] 0.01802744
```

Instrumental variable regression

Example: long-term effects of slave trade

- The second stage is obtained by regressing the response on the fitted values of the first-step estimation
- The coefficient has almost doubled, compared to the OLS estimator, which confirms that this latter estimator is biased, with an attenuation bias due to measurement error and a downward-bias caused by the fact that the most developed African regions were more affected by slave trade

```r
slaves_second <- lm(log(gdp) ~ predict(slaves_first) + colony, data = sltd)
summary(slaves_second)$coefficients
```

```
                        Estimate Std. Error    t value     Pr(>|t|)
(Intercept)            8.0362404 0.31844538 25.2358520 1.369834e-28
predict(slaves_first) -0.1959978 0.04442276 -4.4121040 6.126686e-05
colonyuk              -0.1864887 0.34849637 -0.5351238 5.951419e-01
colonyfrance          -0.1965713 0.34682675 -0.5667708 5.736252e-01
colonyportugal        -0.2983837 0.43722399 -0.6824503 4.983767e-01
colonybelgium         -1.5806014 0.49838949 -3.1714179 2.700532e-03
```

Instrumental variable regression

Example: long-term effects of slave trade

- Same results must be extracted via the **ivreg()** function

```r
library(ivreg)
fit_iv<-ivreg(log(gdp) ~ log(slavesarea) + colony| colony +
                redsea + atlantic + sahara + indian, data = sltd)
summary(fit_iv)
```

```
Call:
ivreg(formula = log(gdp) ~ log(slavesarea) + colony | colony +
    redsea + atlantic + sahara + indian, data = sltd)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9262 -0.4487  0.0698  0.4580  1.3327 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      8.03624    0.33034  24.327  < 2e-16 ***
log(slavesarea) -0.19600    0.04608  -4.253 0.000102 ***
colonyuk        -0.18649    0.36151  -0.516 0.608424    
colonyfrance    -0.19657    0.35978  -0.546 0.587455    
```

colonyportugal	-0.29838	0.45355	-0.658	0.513894

Instrumental Variable Regression

References

- Ananat, Elizabeth Oltmans (2011). "The Wrong Side(s) of the Tracks: The Causal Effects of Racial Segregation on Urban Poverty and Inequality." American Economic Journal: Applied Economics 3 (2): 34–66.
- Nunn, Nathan. 2008. "The Long-Term Effects of Africa's Slave Trades." The Quarterly Journal of Economics 123 (1): 139–76.

Instrumental Variable Regression

Weak instruments

- Instruments should be not only uncorrelated with the error of the model, but they should also be correlated with the covariates
- Therefore, if the correlation between the endogenous variable and the instruments is weak, i.e., if the IV is performed using weak instruments, the estimator will not only be highly imprecise, but it will also be seriously biased, in the direction of the OLS estimator
- While performing IV estimation it is therefore important to check that the instruments are sufficiently correlated with the endogenous covariate
- This can be performed using an F test for the first stage regression, comparing the fit for the regression of the endogenous covariates on the set of the exogenous covariates, and on the external instruments.
- A rule of thumb often used is that the F statistic should be at least equal to 10 (a less strict rule is at least equal to 5)

Instrumental Variable Regression

Example segregation effect

- The F statistic of the first stage regression is the same reported in the weak instruments raw
- In such a case ($F > 10$ and p-value extremely low), we can conclude that the instrument is not weak

```
summary(lm_xz)$fstatistic
```

```
    value     numdf     dendf 
 25.19009   1.00000 119.00000 
```

```
pf(25.190, 1, 119, lower.tail = FALSE)
```

```
[1] 1.840505e-06
```

```
summary(ivFit1, diagnostics = T)$diagnostics
```

```
                 df1 df2  statistic      p-value
Weak instruments   1 119 25.1900948 1.840430e-06
Wu-Hausman         1 118  0.1936444 6.607055e-01
Sargan             0  NA         NA           NA
```

Instrumental Variable Regression

Example slave trade

- Here, the problem is a bit more complex, because we are considering in the main regression the regressor *colony*, which is also considered as an instrument
- In such a case, let

1. RSS_r= residual sum of square of the regression of (log) slaveareas against colony (having in total 5 parameters)
2. RSS_u= residual sum of square of the regression of (log) slaveareas against colony and the remaining instruments (having in total 9 parameters)

- The F statistic is then

$$F = \frac{(RSS_r - RSSu)/q}{RSSu/(n-k)}$$

where $q = 4$ (the number of instruments excluded in the reduced model) and $n - k = 52 - 9 = 43$

- Although the instruments are statistically significant (p = 0.0024), the first-stage $F = 4.89$ is low, indicating that the instruments are weak (they are not unuseful but weak, meaning that they are not generating an IV reliable)

Instrumental Variable Regression

```r
slaves_first_com <- lm(log(slavesarea) ~ colony + atlantic + indian +
                         redsea + sahara, sltd)

slaves_first_red <- lm(log(slavesarea) ~ colony, sltd)

RSS_r <- sum(resid(slaves_first_red)^2)
RSS_u <- sum(resid(slaves_first_com)^2)

q <- 4
k <- 9
n <- nrow(sltd)

F <- ((RSS_r - RSS_u)/q) / (RSS_u/(n - k))
F
```

```
[1] 4.894355
```

```r
pf(F, q, n - k, lower.tail = FALSE)
```

```
[1] 0.002424174
```

```r
summary(fit_iv, diagnostics = T)$diagnostics
```

```
                 df1 df2 statistic     p-value
Weak instruments   4  43  4.894355 0.002424174
Wu-Hausman         1  45  4.761698 0.034360994
Sargan             3  NA  3.630492 0.304227942
```

Instrumental Variable Regression

Hausman test

- The Hausman test evaluates whether regressors treated as endogenous actually are endogenous. It compares:

1. IV/2SLS estimates (consistent even if regressors are endogenous)
2. OLS estimates (more efficient but inconsistent if regressors are endogenous)

- If the two sets of estimates differ significantly there is evidence of endogeneity.

Instrumental Variable Regression

Hausman test

- Suppose the structural model is

$$y = X\beta + \varepsilon$$

- Some regressors in X may be endogenous, then let's denote the two estimators with
  1. $\hat{\beta}_{\mathrm{OLS}}$: OLS estimator,
  2. $\hat{\beta}_{\mathrm{IV}}$ IV/2SLS estimator.

- The Hausman statistic is

$$H = (\hat{\beta}_{\mathrm{IV}} - \hat{\beta}_{\mathrm{OLS}})' \left[\mathrm{Var}(\hat{\beta}_{\mathrm{IV}}) - \mathrm{Var}(\hat{\beta}_{\mathrm{OLS}}) \right]^{-1} (\hat{\beta}_{\mathrm{IV}} - \hat{\beta}_{\mathrm{OLS}})$$

- Under the null hypothesis H_0, the regressors are exogenous, so OLS and IV are both consistent and the difference between the estimators should be close to zero
- The **ivreg()** reports a different test called Wu-Hausman, which is obtained in a slightly different way

Instrumental Variable Regression

Hausman test

1. For the first example (on segregation): the instruments are not weak but we do not need to use them, because the Wu-Hausman does not reject exogeneity $\implies$ OLS preferred to IV, because it is more efficient and unbiased
2. For the second example (on slave trade): the instruments are not weak and we have moderately evidence to use them, because the Wu-Hausman test ("partially") reject the exogeneity; it is plausible to consider the IV regression

```
summary(ivFit1, diagnostics = T)$diagnostics #segregation
```

```
                 df1 df2  statistic      p-value
Weak instruments   1 119 25.1900948 1.840430e-06
Wu-Hausman         1 118  0.1936444 6.607055e-01
Sargan             0  NA         NA           NA
```

```
summary(fit_iv, diagnostics = T)$diagnostics #slave trade
```

```
                 df1 df2 statistic     p-value
Weak instruments   4  43  4.894355 0.002424174
Wu-Hausman         1  45  4.761698 0.034360994
Sargan             3  NA  3.630492 0.304227942
```