Modelos Computacionais em Economia

Marcleiton Morais

Universidade Federal do Tocantins (UFT)

July 23, 2021

Instrumental Variables [Adams, 2020]

OLS requires that the unobserved characteristic of the individual enters into the model **independently and additively**.

- The **IV method** allows the estimation of causal effects when the **independence assumption does not hold**.

Making public colleges free

Does public colleges will encourage more people to attend college? Does people who attend college earn more money? But do the newly encouraged college attendees earn more money?
The policy question is whether encouraging more people to attend college will lead these people to earn more.

A Confounded Model

The problem with comparing earnings from college attendees with those that have not attended college is **confounding**. People who attend college may earn more than people who do not attend college for reasons that have nothing at all to do with attending college.

Confounded Model DAG*

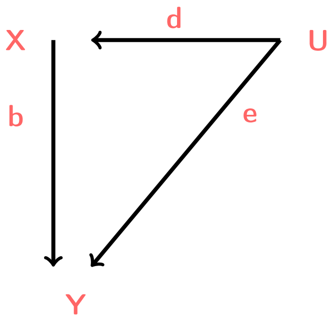


Figure: Confounded graph.

The backdoor problem: estimate b by regressing Y on X will not give an estimate of b. Rather it will give an estimate of $b + \frac{e}{c}$.
*Directed acyclic graphs.

Instrumental Variables

Confounded Linear Model

$$y_i = a + bx_i + ev_{1i}$$

$$x_i = f + dz_i + v_{2i} + cv_{1i}$$

- y_i represents individual i's income.
- x_i is their education level.
- v_{1i} and v_{2i} are unobserved characteristics that determine income and education level respectively.

To see the problem with OLS:

$$y_i = a + bx_i + e\left(\frac{x_i - f - dz_i - v_{2i}}{c}\right)$$

If we run OLS, we estimate the coefficient on x as $b + \frac{e}{c}$, not b.

Instrumental Variables

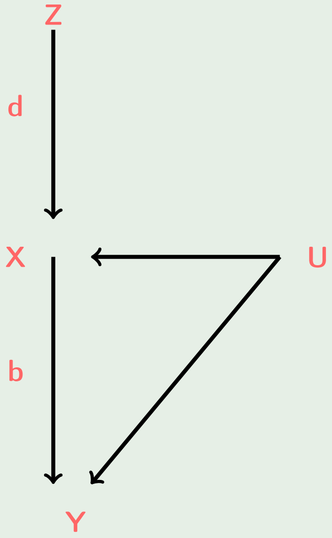
Simulation of Confounded Data

```
set.seed(123456789)
N <- 1000
a <- 2
b <- 3
c <- 2
e <- 3
f <- -1
d <- 4
z <- runif(N)
u_1 <- rnorm(N, mean=0, sd=3)
u_2 <- rnorm(N, mean=0, sd=1)
#We can change the mean and standard
#deviation of the normal distribution
x <- f + d*z + u_2 + c*u_1
y <- a + b*x + e*u_1
lm1 <- lm(y ~ x)
```

OLS estimates b as 4.41 diferent of 3. 4.41 is closer to the **backdoor relationship** of $b + \frac{e}{c} = 4.5$.

Instrumental Variables

Graph Algebra of IV Estimator



The **IV** (Z) has a direct causal effect on X but is not determined by U.

b is **estimated by the relationship between** Z and Y. That effect is given by $b \cdot b$.

Running a regression of X on Z and **dividing** the result of the **first regression** by the result of the **second regression**.

Instrumental Variables

Properties of IV Estimator

- The variable directly affects the policy variable of interest ($Z \rightarrow X$).
- The variable is independent of the unobserved characteristics that affect the policy variable and the outcome of interest ($U \nrightarrow Z$).
- The variable affects the policy variable independently of the unobserved effect ($X = dZ + U$).

IV Estimator with Standard Algebra

$$y_i = a + b(f + dz_i + v_{2i} + cv_{1i}) + ev_{1i}$$

or

$$y_i = a + bf + bdz_i + bv_{2i} + bcv_{1i} + bev_{1i}$$

Instrumental Variables

Simulation of an IV Estimator

```
bd_hat <- lm(y ~ z)$coef[2]
d_hat <- lm(x ~ z)$coef[2]

# picking the slope coefficient from each regression
bd_hat/d_hat
```

If we take the coefficient estimate from the first regression and divide that number by the coefficient estimate from the second regression, we get an estimate that is close to the true relationship.

Instrumental Variables

IV Estimator with Matrix Algebra

$$y = X\beta + v$$

- y is a $100x1$.
- X is a $100x2$ matrix of the observed explanatory variables $\{1, x_i\}$.
- β is a $2x1$ vector of the model parameters $\{a, b\}$.
- v is a $100x1$ vector of the error term v_i.

In addition,

$$X = Z\Delta + E$$

- Z is a $100x2$ matrix of the instrumental variables $\{1, z_i\}$.
- Δ is a $2x2$ matrix.
- E is a $100x2$ matrix of unobserved characteristic.

Instrumental Variables

IV Estimator with Matrix Algebra

So,

$$\Delta = (Z'Z)^{-1}Z'X - (Z'Z)^{-1}Z'E$$

- Z is full-column rank.

Our intent to treat regression:

$$y = Z\Delta\beta + E\beta + v$$

Rearranging this equation we can get an estimator for the coefficients $\Delta\beta$:

$$\Delta\beta = (Z'Z)^{-1}Z'y - (Z'Z)^{-1}Z'E\beta - (Z'Z)^{-1}Z'v$$

Instrumental Variables

IV Estimator with Matrix Algebra

Substituting $\Delta = (Z'Z)^{-1}Z'Z - (Z'Z)^{-1}Z'E$ into this equation and simplifying,

$$\beta = (Z'X)^{-1}Z'y - (Z'X)^{-1}Z'v$$

Our instrumental variable estimator is

$$\hat{\beta_{IV}} = (Z'X)^{-1}Z'y$$

Two-Stage Least Squares

A **common algorithm** for IV is two-stage least squares.

- First-stage estimator:

$$\hat{\Delta} = (Z'Z)^{-1}Z'X$$

- Second-stage regression: we replace the $Z\Delta$ with $Z\hat{\Delta}$ in the equation $y = Z\Delta\beta + E\beta + v$.

Instrumental Variables

IV Estimator in R

We can use $\hat{\beta_{IV}}$ as **pseudo-code** for the instrumental variavle estimator.

```
X <- cbind(1,x) # remember the column of 1's for the i
Z <- cbind(1,z) # remember Z same size as X
beta_hat_ols <- solve(t(X)%*%X)%*%t(X)%*%y
beta_hat_iv <- solve(t(Z)%*%X)%*%t(Z)%*%y
beta_hat_ols
beta_hat_iv
```

Instrumental Variables

Bootstrap IV Estimator for R

The following bootstrap IV estimator defaults to an OLS estimator.

```
lm_iv <- function(y, X_in, Z_in = X_in, Reps = 100,
min_in = 0.05, max_in = 0.95) {
        # takes in the y, x variables and the z if available.
        # Set up
        set.seed(123456789)
        X <- cbind(1,X_in) # adds a column of 1's
        Z <- cbind(1,Z_in)

        # Bootstrap
        bs_mat <- matrix(NA,Reps,dim(X)[2])
        # dim gives the number of rows and columns
        # the second element is the number of columns.
        N <- length(y) # number of observations
        for (r in 1:Reps) {
                index_bs <- round(runif(N, min = 1, max = N))
                y_bs <- y[index_bs] # note Y is a vector
                X_bs <- X[index_bs,]
                Z_bs <- Z[index_bs,]
                bs_mat[r,] <- solve(t(Z_bs)%*%X_bs)%*%t(Z_bs)%*%y_bs
        }
        ...
}
```

Instrumental Variables

Bootstrap IV Estimator for R

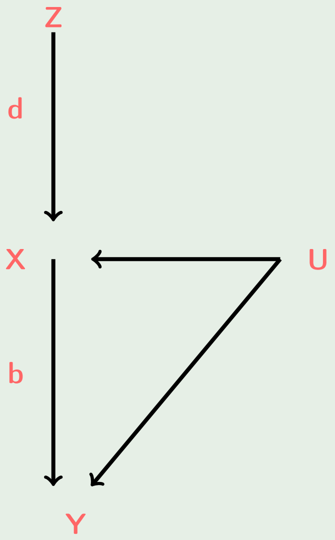
The following bootstrap IV estimator defaults to an OLS estimator.

```
lm_iv <- function(y, X_in, Z_in = X_in, Reps = 100,
        min_in = 0.05, max_in = 0.95) {
        ...
        # Present results
        tab_res <- matrix(NA,dim(X)[2],4)
        tab_res[,1] <- colMeans(bs_mat)
        for (j in 1:dim(X)[2]) {
                tab_res[j,2] <- sd(bs_mat[,j])
                tab_res[j,3] <- quantile(bs_mat[,j],min_in)
                tab_res[j,4] <- quantile(bs_mat[,j],max_in)
        }
        colnames(tab_res) <-c("coef","sd",as.character(min_in),
        as.character(max_in))
        return(tab_res)
}

print(lm_iv(y,x), digits = 3) # OLS
print(lm_iv(y,x,z), digits = 3) # IV
```

Instrumental Variables

Returns of Schooling



[Card, 1993] finds that an extra year of schooling increases income by approximately 7.5%.
Unobserved characteristics:

- Uoung men from wealthier families.
- Young men may go into well paying jobs due to family connections.

Instrumental Variables

Distance to College as an Instrument

[Card, 1993] argues that:

- Young men who grow up near a 4 year college.
- Growing up close to a 4 year college is unlikely to be determined by unobserved characteristics.

Formally,

$$log \quad wage76_i = \alpha_1 + \beta\delta nearCollege_i + \gamma_i observables_i + unobservables_{i1}$$

$$ed_i = \alpha_2 + \delta nearCollege_i + \gamma_2 observables + unobservables_{i2}$$

The estimated effect is made up of two effects, the return to schooling effect (β) and the effect of the instrumental variable on the propensity to get another year of education (δ).

Instrumental Variables

NLSM Data

```
x <- read.xlsx("nls.xlsx")
x$lwage76 <- as.numeric(x$lwage76)
x1 <- x[is.na(x$lwage76)==0,]
# working years after school
x1$exp <- x1$age76 - x1$ed76 - 6
# experienced squared divided by 100
x1$exp2 <- (x1$exp^2)/100
```

Instrumental Variables

OLS model of returns to schooling

```
# OLS Estimate
lm4 <- lm(lwage76 ~ ed76 + exp +exp2 +black +reg76r+
smsa76r + smsa66r + reg662 + reg663 + reg664 +
reg665 + reg666 + reg667 + reg668 + reg669,
data=x1)
# smsa refers to urban or rural.
# reg region of the US - North, South, West etc.
lm4$coefficients[2]
```

Instrumental Variables

IV Estimate of Returns to Schooling

```
# Intent To Treat Estimate
lm5 <- lm(lwage76 ~ nearc4 +exp +exp2 +black +reg76r+
smsa76r + smsa66r + reg662 + reg663 + reg664 +
reg665 + reg666 + reg667 + reg668 + reg669,
data=x1)
# nearc4 is a dummy for distance to a 4 year college.
lm5$coefficients[2]

# Effect of instrument on explanatory variable
lm6 <- lm(ed76 ~ nearc4+ exp+ exp2+ black+ reg76r+
smsa76r + smsa66r + reg662 + reg663 + reg664 +
reg665 + reg666 + reg667 + reg668 + reg669,
data=x1)
lm6$coefficients[2]

# IV Estimate of Returns to Schooling
lm5$coefficients[2]/lm6$coefficients[2]
```

Instrumental Variables

Matrix Algebra IV Estimates of Returns to Schooling

[Card, 1993] uses multiple instruments.

```
y <- x1$lwage76
X <- cbind(x1$ed76, x1$exp, x1$exp2, x1$black, x1$reg76r,
x1$smsa76r, x1$smsa66r, x1$reg662, x1$reg663,
x1$reg664, x1$reg665, x1$reg666, x1$reg667,
x1$reg668, x1$reg669)
x1$age2 <- x1$age76^2

Z1 <- cbind(x1$nearc4, x1$age76, x1$age2, x1$black,
x1$reg76r,x1$smsa76r, x1$smsa66r, x1$reg662,
x1$reg663,x1$reg664, x1$reg665, x1$reg666,
x1$reg667, x1$reg668, x1$reg669)

res <- lm_iv(y,X,Z1, Reps=1000)
rownames(res) <- c("intercept","ed76","exp", "exp2",
"black", "reg76r","smsa76r", "smsa66r",
"reg662","reg663","reg664", "reg665",
"reg666","reg667", "reg668", "reg669")
res
```

Instrumental Variables

Concerns with Distance to College

The results suggest the OLS estimate is biased down. It is unclear why this would be.

1) The first concern is that distance to college is not an instrumental variable.

```
tab_cols <- c("Near College", "Not Near College")
tab_rows <- c("ed76","exp","black","south66",
"smsa66r","reg76r","smsa76r")
table_dist <- matrix(NA,7,2)
# Creating mean of each variable for each type
for (i in 1:7) {
        table_dist[i,1] <-
        mean(x1[x1$nearc4==1,colnames(x1)==tab_rows[i]])
        table_dist[i,2] <-
        mean(x1[x1$nearc4==0,colnames(x1)==tab_rows[i]])
}
colnames(table_dist) <- tab_cols
rownames(table_dist) <- tab_rows
table_dist
```

Instrumental Variables

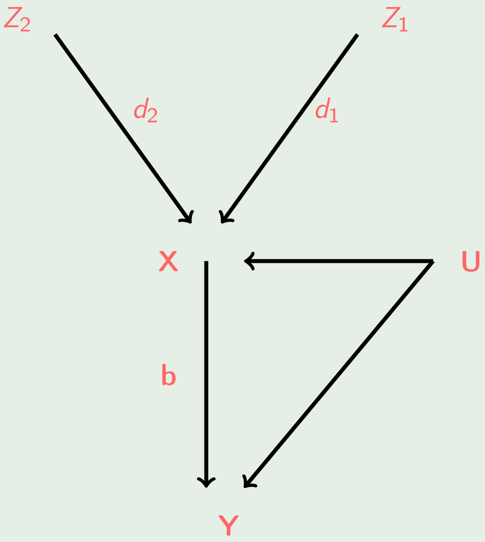
Concerns with Distance to College

The results suggest the OLS estimate is biased down. It is unclear why this would be.

2) The second concern is the additivity assumption.
   The assumption allows the unobserved characteristics of the student and their distance to college to determine the amount of schooling they receive.
   It does not allow the two effects to interact, **students from families with less means, living near college may have a big effect on their propensity to go to college. However, for students from wealthy families, it has little or no effect.**

Instrumental Variables

Test of Instrument Validity



How do we know if an instrument satisfies the assumptions of the model? Often we don't.

Two different estimates of b:

- $c1 = d_1 \cdot b$.
- $c_2 = d_2 \cdot b$

These two sets of regressions give us an **over-identification test**:

$$\left| \frac{c_1}{d_1} - \frac{c_2}{d_2} \right|$$

Instrumental Variables

Test of Instrument Validity

Assume that we have two valid instruments (unobserved characteristics do not affect these two measures):

- Distance to a 4 year college.
- Whether both parents were at home when the young man was 14.

```
Z2 <- cbind(x1$momdad14, x1$age76, x1$age2, x1$black,
        x1$reg76r, x1$smsa76r, x1$smsa66r, x1$reg662,
        x1$reg663, x1$reg664, x1$reg665, x1$reg666,
        x1$reg667, x1$reg668, x1$reg669)
```

Instrumental Variables

Test of Instrument Validity

```
# Bootstrap
set.seed(123456789)
bs_diff <- matrix(NA,1000,1)
N <- length(y)
for (i in 1:1000) {
        index_bs <- round(runif(N, min = 1, max = N))
        y_bs <- y[index_bs]
        X_bs <- X[index_bs,]
        Z1_bs <- Z1[index_bs,]
        Z2_bs <- Z2[index_bs,]
        bs_diff[i,] <-
        (solve(t(Z1_bs)%*%X_bs)%*%t(Z1_bs)%*%y_bs)[2,1] -
        (solve(t(Z2_bs)%*%X_bs)%*%t(Z2_bs)%*%y_bs)[2,1]
        # note the parentheses around the beta estimates.
        # print(i)
}
summary(bs_diff)
quantile(bs_diff,c(0.05,0.95))
```

Instrumental Variables

Test of Instrument Validity

- The mean difference is about 0.68 and the 90% confidence interval includes zero.
- We cannot rule out that distance to college and having both parents at home are both valid instruments.
- The results tell us clearly that the instrument may be valid or may be invalid.

Better LATE than Nothing

The IV independence of the unobserved characteristic can be dropped and we can simply reinterpret the result.

Heterogeneous Effects

It is not reasonable to assume that the policy has the same effect on everyone. Some people get more out of attending college than others.

- Treatment effect positive.
- Treatment effect negative.

If the treatment effect is heterogeneous, the instrumental variable approach is not valid. **Average treatment effect can be a solution.**

- We cannot measure the average treatment effect if the U and Z interact in affecting X

For a subset of the population ([Card, 2001]): **Local Average Treatment Effect** or **LATE**.

Better LATE than Nothing

There are four groups of people. These four groups are characterized by the probability that they accept the treatment corresponding to the instrument.

4 groups

1. Compliers: $Pr(X = 1|Z = 1, C) = Pr(X = 0|Z = 0, C) = 1$
2. Always Takers: $Pr(X = 1|Z = 1, A) = Pr(X = 1|Z = 0, A) = 1$
3. Never Takers: $Pr(X = 0|Z = 1, N) = Pr(X = 0|Z = 0, N) = 1$
4. Defiers: $Pr(X = 0|Z = 1, D) = Pr(X = 1|Z = 0, D) = 1$

There is no expectation that these groups are immutable.

*Importantly, we do not observe which group a particular person is in.

Better LATE than Nothing

We can write down the **intent to treat effect**

Law of Total Expectation: states that if X is a random variable whose expected value $E(X)$ is defined, and Y is any random variable on the same probability space, then

$$E(X) = E(E(X|Y))$$

We can always write out probability of an event as a weighted sum of all the conditional probabilities of the event. That is,

$$Pr(A) = Pr(A|B)Pr(B) + Pr(A|C)Pr(C)$$

where $Pr(B) + Pr(C) = 1$.

Better LATE than Nothing

We can write down the **intent to treat effect**

$$E(Y|Z=1) - E(Y|Z=0)$$

$$=$$

$$\sum_{T\in\{C,A,N,D\}} (E(Y|Z=1,T) - E(y|Z=0,T))Pr(T)$$

where $T = \{C, A, N, D\}$. We can write ou expected outcome conditional on the instrument,

$$\begin{aligned} E(Y|Z=1,T) = {} & E(Y|Z=1, X=1, T)Pr(X=1|Z=1,T) \\ & + E(Y|Z=1, X=0, T)Pr(X=0|Z=1,T) \end{aligned} \tag{1}$$

The **expected income conditional on the instrument** is an average of the expected income conditional on both the instrument and the treatment allocation, weighted by the probability of receiving the treatment allocation conditional on the instrument.

Better LATE than Nothing

We can write down the **intent to treat effect**

The effect of Z on Y is only through X, so

$$E(Y|X=1,Z=1)=E(Y|X=1,Z=0)=E(Y|X=1)$$

This implies the following for our intent to treat estimates for each group.

$$E(Y|Z=1,C)-E(Y|Z=0,C)=E(Y|X=1,C)-E(Y|X=0,C)$$

$$E(Y|Z=1,A)-E(Y|Z=0,A)=E(Y|X=1,A)-E(Y|X=0,A)=0$$
$$E(Y|Z=1,N)-E(Y|Z=0,N)=E(Y|X=1,N)-E(Y|X=0,N)=0$$
$$E(Y|Z=1,D)-E(Y|Z=0,D)=E(Y|X=1,D)-E(Y|X=0,D)$$

Better LATE than Nothing

We can write down the **intent to treat effect**

Given the additional assumption that there are no defiers ($Pr(D) = 0$) (**monotonicity assumption**), the fraction of compliers is:

$$E(Y|X = 1, C) - E(Y|X = 0, C) = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{Pr(X = 1|Z = 1) - Pr(X = 1|Z = 0)}$$

Note the value on the bottom of the fraction is the percent of compliers.

This fraction is the discrete version of the IV estimate presented above. The LATE estimate is an alternative interpretation of the original estimate.

Better LATE than Nothing

LATE Estimator

The intent to treat (empirical analog) divided by the effect of Z on X is,

$$\hat{\mu}_{y1} = \frac{\sum_{i=1}^{N} y_i \mathbb{1}(z_i = 1)}{\sum_{i=1}^{N} \mathbb{1}(z_i = 1)}$$

and

$$\hat{\mu}_{y0} = \frac{\sum_{i=1}^{N} y_i \mathbb{1}(z_i = 1)}{\sum_{i=1}^{N} \mathbb{1}(z_i = 1)}$$

where $\mathbb{1}()$ is an indicator function. This function is 1 if the value inside the parenthesis is true, 0 if it is false.

Better LATE than Nothing

LATE Estimator

We can also write out the analog estimators for the two bottom probabilities.

$$\hat{p}_{11} = \frac{\sum_{i=1}^{N} \mathbb{1}(x_1 = 1 \& z_i = 1)}{\sum_{i=1}^{N} \mathbb{1}(z_i = 1)}$$

and

$$\hat{p}_{10} = \frac{\sum_{i=1}^{N} \mathbb{1}(x_1 = 1 \& z_i = 0)}{\sum_{i=1}^{N} \mathbb{1}(z_i = 0)}$$

Putting all this together, we have the LATE estimator.

$$\hat{\mu}_{LATE} = \frac{\hat{\mu}_{y1} - \hat{\mu}_{y0}}{\hat{p}_{11} - \hat{p}_{10}}$$

Better LATE than Nothing

LATE Estimates of Returns to Schooling

```
X2 <- X[,1] > 12 # college indicator
# using college proximity as an instrument.
mu_y1 <- mean(y[Z1[,1]==1])
mu_y0 <- mean(y[Z1[,1]==0])
p_11 <- mean(X2[Z1[,1]==1])
p_10 <- mean(X2[Z1[,1]==0])
# LATE, divide by 4 to get the per-year effect
((mu_y1 - mu_y0)/(p_11 - p_10))/4

# this allows comparison with the OLS estimates.
# using living with both parents as an instrument.
mu_y1 <- mean(y[Z2[,1]==1])
mu_y0 <- mean(y[Z2[,1]==0])
p_11 <- mean(X2[Z2[,1]==1])
p_10 <- mean(X2[Z2[,1]==0])
((mu_y1 - mu_y0)/(p_11 - p_10))/4
```

References

Adams, C. P. (2020).
Learning Microeconometrics with R.
Chapman and Hall/CRC.

Card, D. (1993).
Using geographic variation in college proximity to estimate the return to schooling.
NBER working paper, (w4483).

Card, D. (2001).
Estimating the return to schooling: Progress on some persistent econometric problems.
Econometrica, 69(5):1127–1160.