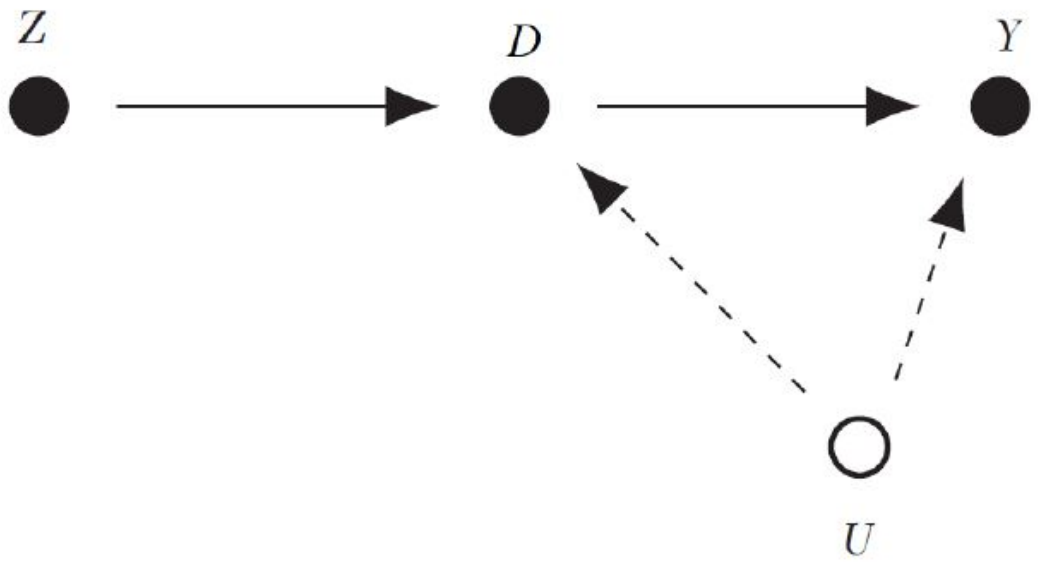
Causal Inference for Policy Evaluation (Spring Semester 2025)

Lab Session 5 - Instrumental Variables (IV)



Key ingredients for an IV design:

- An endogenous treatment.
- A variable that is correlated with this treatment but which does not directly affect the outcome (instrument).

Application

Joshua Angrist and William M. Evans (1998). "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size.'' American Economic Review

Outline for today

1. Introduction to the paper and research question
2. Identification strategy and assumptions
3. Data and descriptive statistics (balancedness table and first stage)
4. Estimation: preliminaries
    - Nonparametric LATE (Wald estimate)
    - Bootstrap
    - Reduced form effect
5. Estimation: Two-stage least squares
6. Extensions
    - Characterizing the compliers
    - Semi-parametric LATE

---

1. Introduction

- What is the research question?
- Why is this question of interest?

- How is the treatment defined, and what are the outcome variables of interest?
- Why would comparing the average labour supply outcomes of women with different numbers of children result in a biased estimate of the effect of fertility on labour supply? What is the endogeneity problem?
- Which instrumental variables do the authors use?

Notation

Follows lecture slides rather than paper.

- $Z_i \in \{0,1\}$ … binary instrument: having a second child of the same gender as the first child (*samesex*)
- $D_i \in \{0,1\}$ … binary treatment status: having more than 2 children yes/no (*morekids*)
- $D^*_{0,i} \in \{0,1\}$ … potential treatment status when $Z_i = 0$
- $D^*_{1,i} \in \{0,1\}$ … potential treatment status when $Z_i = 1$
- Y^*_{dz} … potential outcome under treatment $D = d$ and instrument $Z = z$
- Y_i … observed outcome

What treatment effect do we identify?

Local Average Treatment Effect (LATE), meaning the effect for those who react to the instrument $Z_i = 1$ by having more children (compliers).

- Are there reasons to believe that LATE $\neq$ ATE?

---

2. Identification strategy and assumptions

Discussion of assumptions

What do these assumptions mean in words?

- What could invalidate them? Think of concrete examples or mechanisms.
- Which arguments or evidence can you provide to support that they hold?

(A1) Stable unit treatment value assumption (SUTVA)

$Y_i = D_i Z_i Y^*_{11,i} + D_i(1-Z_i)Y^*_{10,i} + (1-D_i)Z_i Y^*_{01,i} + (1-D_i)(1-Z_i)Y^*_{00,i}$

- No spillovers from treated on non-treated
- Having more than 2 children should not affect the labour supply of women with 2 children.

$D_i = Z_i D^*_{1,i} + (1-Z_i)D^*_{0,i}$

- No spillovers from instrumented on non-instrumented:
- Having two children of the same sex has no effect on the likelihood of having another child for those who have 2 children of mixed sex.

(A2) Exclusion restriction

$Y^*_{d0,i} = Y^*_{d1,i} \equiv Y^*_{d,i}$ for all i and $d \in \{0,1\}$

- No direct effect of the instrument on the potential outcome.

- Is this plausible here?

(A3) Exogeneity

$Y^*_{0,i}, Y^*_{1,i}, D^*_{0,i}, D^*_{1,i} \perp Z_i | X_i$

- No counfounders that determine both Z and Y or D.
- The instrument is randomly assigned (possibly conditional on X).
- Is this plausible here?

If we observe the confounders X that determine both Z and Y, we can obtain a valid IV under conditional exogeneity and common support (see lecture slides).

(A4) Monotonicity

$D^*_{1,i} \geq D^*_{0,i}$ for all i and $D^*_{1,i} > D^*_{0,i}$ for some i

- There exist compliers but no defiers.
- The instrument moves the endogenous variable in one direction, i.e. the instrument is *relevant*.
- How can we provide supportive evidence for it?

---

3. Data and descriptive statistics

- Census Public Use Micro Samples (PUMS) 1980
- We use a random sample covering one third of the observations.
- What is the unit of observation? – 1 line = 1 household, separate variables for mothers and fathers.
- What is the time dimension? – One cross-section.
- Preliminary data prep done by the authors:
  - Children are matched to female household head or the spouse of a male household head.
  - Mothers for whom the number of children did not match the reported number were deleted from the data.
- Here, focus on women with two and more children. The paper also analyses the married sample.

Variable name	Description
Treatment variables	
morekids	had more than 2 kids
kidcount	count of kids in household
Instruments	
samesex	first two kids are of same sex
multi2nd	second birth twins
Outcome variables	
weeksm	weeks worked per year, mom
hourswm	hours worked per week, mom
weeksd	weeks worked per year, dad

Variable name	Description
hourswd	hours worked per week, dad
workedm	worked for pay, mom
workedd	worked for pay, dad
incomem	moms labour income
incomed	dads labour income
faminc	family income
lfaminc	log family income
nonmomi	income not generated by mom
lnonmomi	log income not generated by mom
Characteristics of the children	
ageqk	age in quarters, first born
ageq2nd	age in qtrs second kid
ageq3rd	age in qtrs of 3rd kid
boy1st	first birth boy
boy2nd	2nd birth boy
boys2	first two births boys
girls2	first two births girls
Characteristics of the mother	
agem	age in years of mom
agefstm	age of mom when kid first born
blackm	=1 if mom black
hispm	=1 if mom hispanic
othracem	=1 if mom other race (white is ref)
educm	moms education
Characteristics of the father (married sample)	
msample	married sample
agefstd	age of dad when kid first born
aged	age of dad
blackd	=1 if dad black
hispd	=1 if dad hispanic
othraced	=1 if dad other race (white is ref)

Load Packages

```
# Empty working space
rm(list=ls())

#Load Packages
# Define packages that you need,
packages_vector <- c( "haven", "dplyr",  "sandwich",  "jtools", "fBasics",
                      "xtable",  "stargazer", "data.table", "tidyverse", "ggplot2",
```

```
                      "AER", # AER package for ivreg command
                      "causalweight") # for semiparametric LATE
lapply(packages_vector, require, character.only = TRUE)


# Set working directory
getwd()
work_dir <- "Q:/Arbeitsmarktökonomie/Data-Lehre/Methods/2025 CIPE/Lab Sessions/Session 4 - IV/Mario"
setwd(work_dir)
```

Read in the data

```
# Random sample of 1980 PUMS data (1/3 of observations)
data <-read_dta("AngristEvans1980_sample.dta")

# Inspect
head(data)
```

```
## # A tibble: 6 x 35
##   morekids kidcount samesex multi2nd ageqk ageq2nd ageq3rd boy1st boy2nd boys2
##      <dbl>    <dbl>   <dbl>    <dbl> <dbl>   <dbl>   <dbl>  <dbl>  <dbl> <dbl>
## 1        1        5       1        0    57      51      47      1      1     1
## 2        0        1       0        1     5      NA       0      0      0     0
## 3        0        2       1        0    38      13       0      0      0     0
## 4        0        2       0        0    51      38       0      0      1     0
## 5        1        3       1        0    41      37       8      1      1     1
## 6        0        2       0        0    18       6       0      1      0     0
## # i 25 more variables: girls2 <dbl>, agem <dbl>, agefstm <dbl>, agefstd <dbl>,
## #   aged <dbl>, blackm <dbl>, hispm <dbl>, othracem <dbl>, blackd <dbl>,
## #   hispd <dbl>, othraced <dbl>, educm <dbl>, faminc <dbl>, weeksm <dbl>,
## #   hourswm <dbl>, weeksd <dbl>, hourswd <dbl>, workedm <dbl>, workedd <dbl>,
## #   incomem <dbl>, incomed <dbl>, lfaminc <dbl>, nonmomi <dbl>, lnonmomi <dbl>,
## #   msample <dbl>
```

Sample selection criteria for the main analysis

```
# Number of observations before sample selection
print("Sample size before sample restrictions:")
```

```
## [1] "Sample size before sample restrictions:"
```

```
nrow(data)
```

```
## [1] 440623
```

```r
# only keep women aged 21-35
data <- dplyr::filter(data, agem >= 21 & agem <= 35)

# who were older than 15 at first birth
data <- dplyr::filter(data, agefstm >= 15)

# who have 2 or more children
data <- dplyr::filter(data, kidcount >= 2)

# second child older than 4 quarters (1 year)
data <- dplyr::filter(data, ageq2nd > 4)

# Number of observations after sample selection
print("Sample size after sample restrictions:")
```

```
## [1] "Sample size after sample restrictions:"
```

```r
nrow(data)
```

```
## [1] 197071
```

```r
# save final data set
save(data, file="AngristEvans1980_reduced.RData")
```

- The census does not allow to track children across households.
- What do the sample selection criteria ensure?
- What do they imply for the representativeness of the estimates? I.e., is this a selective sample?

Define key variables

```r
# Endogenous Variable
data$d <- data$morekids # has more than 2 kids

# Instrument
data$z <- data$samesex # first two kids are of same sex

# Store each variable in own R object
attach(data)

# Labour market outcomes (mother)
#y_names<- c("workedm", "weeksm", "hourswm", "incomem", "lfaminc")
```

Descriptive Statistics

We replicate the first column of Table 2 (page 445) for the 1980 data using all women in the sample.

Slightly simplified sample selection criteria, so that exact figures might not match paper.

```
x_desc_names<-c("kidcount", "morekids", "boy1st", "boy2nd", "boys2",
                "girls2", "samesex", "multi2nd",
                "agem", "agefstm",
                "workedm", "weeksm", "hourswm", "incomem", "faminc")

desc <- fBasics::basicStats(data[x_desc_names]) %>%
                t() %>%
                as.data.frame() %>%
                dplyr::select(Mean, Stdev, nobs)

print("Descriptive statistics")
```

```
## [1] "Descriptive statistics"
```

```
print(round(desc, digits=3))
```

```
##                Mean      Stdev   nobs
## kidcount      2.552      0.807 197071
## morekids      0.402      0.490 197071
## boy1st        0.511      0.500 197071
## boy2nd        0.510      0.500 197071
## boys2         0.263      0.440 197071
## girls2        0.242      0.428 197071
## samesex       0.505      0.500 197071
## multi2nd      0.009      0.093 197071
## agem         30.129      3.507 197071
## agefstm      20.142      2.953 197071
## workedm       0.566      0.496 197071
## weeksm       20.811     22.271 197071
## hourswm      18.804     18.917 197071
## incomem   7162.109  10829.388 197071
## faminc   42368.075  26635.494 197071
```

Check balancedness of covariates across $Z = 0$ and $Z = 1$

Supporting evidence for the **exogeneity** of the instrument.

Compare the average characteristics of mothers with first two children of the same gender vs. different genders.

See Table 4, page 459, column 1980 PUMS.

```
# Define a vector of covariates
x_diff <- cbind(agem, agefstm, blackm, hispm, othracem, educm)
x_names <- colnames(x_diff)

# Define a function estimating the differences across samesex
```

```r
balance_check.model <- function(x){

    # Conditional means
    mean_z0 <- mean(x[z==0])
    mean_z1 <- mean(x[z==1])

    # Difference in means
    diff_z <- lm(x ~ z)
    cov <- vcovHC(diff_z, type = "HC")
    robust.se <- sqrt(diag(cov))

    list(mean_z0 = mean_z0,
         mean_z1 = mean_z1,
        diff = diff_z$coefficients[2],
        robust.se = robust.se[2],
        pval = 2*pnorm(-abs(diff_z$coefficients[2]/robust.se[2])) )
}

# Run function and bind to number of observations
diff_output <- apply(x_diff, 2, balance_check.model)
diff_output <- as.data.frame(rbindlist(diff_output))

obs <- c(nrow(data[z==0,]),
           nrow(data[z==1,]),
           NA, NA, NA)
diff_output <- rbind(diff_output, obs)

# Format # Display in desired format
rownames(diff_output)<- c(x_names, "Observations")
colnames(diff_output)<- c("E(X|Z=0)", "E(X|Z=1)", "Difference", "s.e.", "p-value")
print("Difference in means for demographic variables by same sex")
```

```
## [1] "Difference in means for demographic variables by same sex"
```

```r
print(round(diff_output, digits=3))
```

```
##              E(X|Z=0)  E(X|Z=1) Difference  s.e. p-value
## agem           30.136    30.122     -0.013 0.016   0.405
## agefstm        20.135    20.150      0.015 0.013   0.253
## blackm          0.119     0.120      0.001 0.001   0.558
## hispm           0.030     0.030      0.001 0.001   0.469
## othracem        0.029     0.028     -0.001 0.001   0.167
## educm          12.135    12.127     -0.008 0.011   0.449
## Observations 97516.000 99555.000         NA    NA      NA
```

- What do we conclude from this?

First-stage effect of Z on D

Check for **instrument relevance**: first stage regression. See upper panel of Table 6, columns (1) and (2), page 462 in the paper.

First, without control variables

$D_i = \alpha_1 + \pi_1 Z_i + \varepsilon_{1,i}$

```
ols.m.morekids.1 <- lm(d ~ samesex )
summ(ols.m.morekids.1, robust = "HC1")
```

```
## MODEL INFO:
## Observations: 197071
## Dependent Variable: d
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,197069) = 677.95, p = 0.00
## R² = 0.00
## Adj. R² = 0.00
##
## Standard errors: Robust, type = HC1
## ------------------------------------------------
##                      Est.   S.E.   t val.      p
## ----------------- ------ ------ -------- ------
## (Intercept)          0.37   0.00   240.97   0.00
## samesex              0.06   0.00    26.04   0.00
## ------------------------------------------------
```

- How do we interpret this first stage?
- F-statistic should be > 10 for a strong instrument.
- With weak instrument, complier share is very small such that estimate becomes highly sensitive to the denominator.

```
cov <- vcovHC(ols.m.morekids.1, type = "HC")
robust.se.morekids.1 <- sqrt(diag(cov))
```

Second, controlling for the gender mix of the first two children and mother's demographic characteristics

$D_i = \alpha_1 + \pi_1 Z_i + \delta_1 Boy1st + \delta_2 Boy2nd + X_i'\beta_1 + \varepsilon_{1,i}$

- Why do we control for the gender of the first two children?
- Why do we include additional control variables?

```
x <- cbind(boy1st, boy2nd, agem, agefstm, blackm, hispm, othracem)
```

```r
ols.m.morekids.2 <- lm(d ~ samesex + x)
cov <- vcovHC(ols.m.morekids.2, type = "HC")
robust.se.morekids.2 <- sqrt(diag(cov))

# Output Coefficients
stargazer(ols.m.morekids.1,ols.m.morekids.2,
          se=list(robust.se.morekids.1, robust.se.morekids.2),
          type="text",
          keep=c("boy1st", "boy2nd", "samesex"),
          keep.stat = c("n", "rsq", "f"),
          align=TRUE, dep.var.labels = c("More than two children"),
          dep.var.labels.include = TRUE)
```

```
##
## ===============================================================================
##                                  Dependent variable:
##              ---------------------------------------------------------
##                                More than two children
##                            (1)                            (2)
## -------------------------------------------------------------------------------
## samesex                0.057***                       0.059***
##                         (0.002)                        (0.002)
##
## xboy1st                                               -0.010***
##                                                        (0.002)
##
## xboy2nd                                               -0.009***
##                                                        (0.002)
##
## -------------------------------------------------------------------------------
## Observations           197,071                        197,071
## R2                      0.003                          0.084
## F Statistic  677.951*** (df = 1; 197069) 2,254.119*** (df = 8; 197062)
## ===============================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

---

4. Estimation: Preliminaries

Nonparametric LATE (Wald estimate)

$$LATE = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}$$

We estimate the LATE for *workedm*, based on the full sample.

See Table 5 (first 2 columns).

```
    # Conditional outcomes (for participation decision)
    E_workedm_1 = mean(workedm[z==1])
    E_workedm_0 = mean(workedm[z==0])

    # Conditional treatment
    E_d_1 = mean(d[z==1])
    E_d_0 = mean(d[z==0])

    # Difference in conditional outcomes
    diff_workedm = E_workedm_1 - E_workedm_0

    # Difference in conditional treatment = FIRST STAGE
    diff_d = E_d_1 - E_d_0

    #  Wald Estimate / LATE #
    wald_workedm = diff_workedm/diff_d

    # Present results in simple table
    tab_wald <- rbind(cbind(E_workedm_1, E_workedm_0, diff_workedm),
                      cbind(E_d_1, E_d_0, diff_d),
                      cbind(NA, NA, wald_workedm))
    colnames(tab_wald) <- c("Z=1", "Z=0", "Difference")
    rownames(tab_wald) <- c("E(Y|Z)", "E(D|Z)", "Wald estimate")
    print(round(tab_wald, digits=3))
```

```
##                 Z=1   Z=0 Difference
## E(Y|Z)        0.562 0.570     -0.007
## E(D|Z)        0.431 0.373      0.057
## Wald estimate    NA    NA     -0.128
```

- How do we interpret this effect?
- What about inference? Bootstrap.

```
# define a function which we can later use for inference (bootstrap)
est_LATE <-function(y,d,z){

    E_y_1 = mean(y[z==1])
    E_y_0 = mean(y[z==0])

    # conditional treatment
    E_d_1 = mean(d[z==1])
    E_d_0 = mean(d[z==0])
```

```
    # difference in conditional outcomes
    diff_y = E_y_1 - E_y_0

    # difference in conditional treatment = FIRST STAGE
    diff_d = E_d_1 - E_d_0

    # LATE
    late= diff_y/diff_d

    list(late=late,
         E_y_1= E_y_1,
         E_y_0= E_y_0,
         E_d_1= E_d_1,
         E_d_0= E_d_0,
         diff_y=diff_y,
         diff_d=diff_d
         )
}

# estimate the LATE on labor force participation
LATE <- est_LATE(y=workedm, d=d, z=z)
print("est_LATE output")
```

```
## [1] "est_LATE output"
```

```
print(LATE)
```

```
## $late
## [1] -0.1283451
##
## $E_y_1
## [1] 0.5623926
##
## $E_y_0
## [1] 0.5697629
##
## $E_d_1
## [1] 0.4306564
##
## $E_d_0
## [1] 0.3732311
##
## $diff_y
## [1] -0.007370263
```

```
##
## $diff_d
## [1] 0.05742536
```

Bootstrap

```
# Define a function for the bootstrap
bootstrap.late<-function(y,d,z,boot){

  obs<-length(y) # store the number of observations
  mat=c() # empty matrix for storing effect estimates
  temp=c() # empty vector to count bootstrap replications

  # The bootstrap loop starts here:
  while(length(temp)<boot){

    # draw a bootstrap sample
    sboot<-sample(x=1:obs, # observations that are drawn from y
                  size=obs, # number of obs as in original data
                  replace=TRUE) # with replacement (one obs allowed to appear more than once)

    # redefine y, d, z from the bootstrap sample  (no covariates here)
    yb<-y[sboot]
    db<-d[sboot]
    zb<-z[sboot]

    # estimate the LATE within the bootstrap sample
    est<-c(est_LATE(y=yb, d=db, z=zb))

    # add the estimates as an additional row in the effects matrix
    # one column per effect (as in output of estimator function)
    if (sum(is.na(est))==0) mat<-rbind(mat, est)

    # increase length by 1 (ran 1 more bootstrap repetition)
    temp<-c(temp,1)

  }

  # store standard deviations of the estimated effect
  list(se_diff_y=sd(as.numeric(mat[,6])), # column 6 stores the numerator difference
       se_diff_d=sd(as.numeric(mat[,7])), # column 7 stores the denominator difference
       se_late=sd(as.numeric(mat[,1]))) #column 1 stores the LATE
  }
```

```
  #-------------------------------------------------------

  # Set seed for replicability - right before estimation!
  set.seed(12345)

  # Run the bootstrap on the original data
  LATE <- est_LATE(y=workedm, d=d, z=z)

  # estimate SE separately using the bootstrap
  inf.LATE<- bootstrap.late(y=workedm,d=d,z=z,boot=99)
```

```
results<- cbind(tab_wald, inf.LATE)
results
```

```
##                Z=1       Z=0       Difference   inf.LATE
## E(Y|Z)         0.5623926 0.5697629 -0.007370263 0.002140361
## E(D|Z)         0.4306564 0.3732311 0.05742536   0.002003153
## Wald estimate NA        NA        -0.1283451   0.03689578
```

```
print("P-value")
```

```
## [1] "P-value"
```

```
print(round(2*pnorm(-abs(LATE$late/inf.LATE$se_late)), digits=3))
```

```
## [1] 0.001
```

Reduced form effect of Z on Y

Reduced-form regression of Y on Z:

$Y_i = \alpha_{RF} + \pi_{RF} Z_i + X_i' \beta_{RF} + \varepsilon_{RF,i}$

- Because the IV is (conditionally) random, the reduced form gives an unbiased estimate of the effect of the instrument on the outcomes (that operates via the endogenous treatment only!).
- Numerator of the Wald estimate.
- We still assume the instrument has no direct effect on the outcome.
- More informative when the IV is, e.g. a policy intervention. Then the reduced-form measures the intention to treat effect (ITT), which includes that some instrumented observations do not actually take up the treatment.

```
x <- cbind(boy1st, boy2nd, agem, agefstm, blackm, hispm, othracem)
y_mat = cbind(workedm, weeksm, hourswm, incomem, lfaminc)
y_names = colnames(y_mat)

# Define function for several outcomes
itt.model <- function(y){
        itt.m <- lm(y ~ z + x)
```

```
        cov <- vcovHC(itt.m, type = "HC")
        robust.se <- sqrt(diag(cov))

    list(itt.coeff = itt.m$coefficients[2],
         robust.se = robust.se[2],
         pval = 2*pnorm(-abs(itt.m$coefficients[2]/robust.se[2])) )
}

itt_output <- apply(y_mat, 2, itt.model)
itt_output <- as.data.frame(rbindlist(itt_output))
obs <- c(nrow(data), NA, NA)
itt_output <- rbind(itt_output, obs)
rownames(itt_output)<- c(y_names, "Observations")
print("Reduced form estimates of labour supply models")
```

```
## [1] "Reduced form estimates of labour supply models"
```

```
print(round(itt_output, digits=3))
```

```
##               itt.coeff robust.se  pval
## workedm          -0.007     0.002 0.002
## weeksm           -0.241     0.098 0.014
## hourswm          -0.226     0.084 0.007
## incomem         -91.480    47.897 0.056
## lfaminc           0.003     0.006 0.575
## Observations 197071.000        NA    NA
```

---

5. Estimation: Two-stage least squares (2SLS)

Reminders

- The 2SLS estimation is the parametric version of the IV estimation.
- With control variables, it imposes a specific functional form on the outcome and treatment equations.
- Controlling for covariates to increase precision (under A3), and/or account for conditional exogeneity of IV (under A3', e.g. gender of first two children).
- Imposes homogeneous treatment effect.

Implementation

2SLS first extracts the exogenous variation from D (first stage from before):

$$D_i = \alpha_1 + \pi_1 Z_i + X_i'\beta_1 + \varepsilon_{1,i}$$

Then, replace endogenous treatment D in the outcome equation by the predicted $\hat{D}_i$ obtained from this first stage:

$$Y_i = \alpha_{IV} + \pi_{IV}\hat{D}_i + X_i'\beta_{IV} + \varepsilon_{IV,i}$$

The command `ivreg` from the `AER` package directly integrates these two steps and gives you standard errors corrected for the first stage estimation.

- Note that it does not allow you to show the results of the first stage estimation.
- Shows some model diagnostics, e.g. F-test of the first stage ('weak instruments').

We first estimate the effect on the labor force participation of women to demonstrate the application of the package:

```
iv.m <- ivreg(workedm ~ d + x  | z + x)
summary(iv.m, vcov = sandwich, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = workedm ~ d + x | z + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9272 -0.5178  0.3002  0.4313  0.7288
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4612828  0.0180108  25.612  < 2e-16 ***
## d           -0.1150971  0.0370254  -3.109  0.00188 **
## xboy1st     -0.0001912  0.0022003  -0.087  0.93076
## xboy2nd     -0.0042123  0.0021967  -1.918  0.05517 .
## xagem        0.0222252  0.0011698  19.000  < 2e-16 ***
## xagefstm    -0.0262497  0.0017176 -15.283  < 2e-16 ***
## xblackm      0.1041820  0.0042507  24.509  < 2e-16 ***
## xhispm      -0.0391096  0.0086700  -4.511 6.46e-06 ***
## xothracem    0.0401819  0.0070451   5.704 1.18e-08 ***
##
## Diagnostic tests:
##                      df1    df2 statistic p-value
## Weak instruments       1 197062   773.301  <2e-16 ***
## Wu-Hausman             1 197061     2.471   0.116
## Sargan                 0     NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4832 on 197062 degrees of freedom
## Multiple R-Squared: 0.04961, Adjusted R-squared: 0.04957
## Wald test: 687.1 on 8 and 197062 DF,  p-value: < 2.2e-16
```

```
# Important for coding: Endogenous variables (d)
# can only appear before the vertical line;
# instruments (z) can only appear after the vertical line;
# exogenous regressors that are not instruments (x)
# must appear both before and after the vertical line.
```

Main results

We now replicate Table 7, columns 1-2, page 465, for the outcomes of interest.

This table compares the 2SLS estimates π_{IV} with the 'naive' OLS estimates π_{OLS} from a regression

$Y_i = \alpha_{OLS} + \pi_{OLS} D_i + X_i' \beta_{OLS} + \varepsilon_{OLS,i}$

```
# Define OLS function for several outcomes
ols.model <- function(y){

    ols.m <- lm(y ~ d + x)
    cov <- vcovHC(ols.m, type = "HC")
    robust.se <- sqrt(diag(cov))

    list(ols.coeff = ols.m$coefficients[2],
         robust.se = robust.se[2],
         pval = 2*pnorm(-abs(ols.m$coefficients[2]/robust.se[2])))
}

ols_output <- apply(y_mat, 2, ols.model)
ols_output<-as.data.frame(rbindlist(ols_output))
rownames(ols_output)<- y_names
```

```
# Define 2sls function for several outcomes
tsls.model <- function(y){
        iv.m <- ivreg(y ~ d + x  | z + x)
        iv_sum <-summary(iv.m, vcov = sandwich)

        list( iv.coeff = iv_sum$coefficients[2,1],
              robust.se = iv_sum$coefficients[2,2],
              pval = 2*pnorm(-abs(iv_sum$coefficients[2,1]/iv_sum$coefficients[2,2])))
 }

iv_output <- apply(y_mat, 2, tsls.model)
iv_output<-as.data.frame(rbindlist(iv_output))


output<-cbind(ols_output,iv_output)
```

```r
rownames(output)<- y_names
colnames(output) <- c("OLS", "se", "p", "2SLS", "se", "p")
print(round(output,digits=3))
```

```
##               OLS     se p      2SLS      se     p
## workedm    -0.173  0.002 0    -0.115   0.037 0.002
## weeksm     -8.861  0.100 0    -4.101   1.649 0.013
## hourswm    -6.520  0.087 0    -3.847   1.405 0.006
## incomem -3718.223 48.955 0 -1555.290 807.002 0.054
## lfaminc    -0.134  0.006 0     0.056   0.100 0.576
```

- What do we learn from comparing OLS to IV estimates?
- Can the LATE answer the research question?
- Why is the effect on family income so small/non-significant?

---

6. Extensions

Characterizing the compliers

- We cannot individually identify compliers.
- But we can quantify the size of the complier group.
- And we can describe the distribution of complier characteristics.

Estimate the effect of Z on D (first stage) in subsamples defined by characteristics X to assess who is under or over-represented among the compliers:

$$D_i = \alpha_1 + \pi_1 Zi + X_i'\beta_1 + \epsilon_{1,i}$$

Let's try for subsample of mothers who * have years of schooling above the median * are married.

```r
# define a variable being one if education is above the median
educm_h <- ifelse(educm > median(educm), 1, 0)

# Estimate first stages (controlling for other characteristics)

# full sample
ols.m.morekids.full <- lm(d ~ samesex + x)
cov <- vcovHC(ols.m.morekids.full, type = "HC")
robust.se.morekids.full <- sqrt(diag(cov))

# highly educated
ols.m.morekids.educh <- lm(d[educm_h == 1] ~ samesex[educm_h == 1] + x[educm_h == 1,])
cov <- vcovHC(ols.m.morekids.educh, type = "HC")
robust.se.morekids.educh <- sqrt(diag(cov))
```

```r
# married
ols.m.morekids.married <- lm(d[msample == 1] ~ samesex[msample == 1] + x[msample == 1,])
cov <- vcovHC(ols.m.morekids.married, type = "HC")
robust.se.morekids.married <- sqrt(diag(cov))


# Output Coefficients
stargazer(ols.m.morekids.full, ols.m.morekids.educh, ols.m.morekids.married,
          se=list(robust.se.morekids.full, robust.se.morekids.educh,robust.se.morekids.married),
          type="text",
          keep=c("samesex"),
          keep.stat = c("n", "rsq", "f"),
          align=TRUE, dep.var.labels = c("More than one child", "More than one child", "More than one ch
          dep.var.labels.include = TRUE)
```

```
##
## ==========================================================================================================
##                                                  Dependent variable:
##                        ---------------------------------------------------------------------------------
##                             More than one child        More than one child         More than one chi
##                                     (1)                        (2)                         (3)
## ----------------------------------------------------------------------------------------------------------
## samesex                           0.059***
##                                    (0.002)
##
## samesex[educm_h == 1]                                      0.050***
##                                                            (0.004)
##
## samesex[msample == 1]                                                                  0.065***
##                                                                                         (0.003)
##
## ----------------------------------------------------------------------------------------------------------
## Observations                        197,071                    58,245                      125,725
## R2                                   0.084                      0.053                       0.077
## F Statistic                2,254.119*** (df = 8; 197062) 404.020*** (df = 8; 58236) 1,307.318*** (df = 8;
## ==========================================================================================================
## Note:                                                                        *p<0.1; **p<0.05; **
```

- Are highly educated and married women under or overrepresented among the compliers?
- In paper: 2SLS estimation in subsample of college-educated women indicates smaller labor supply effects.
- Authors conclude that childbearing has stronger negative effects in groups with low socioeconomic status.

Semi-parametric LATE

Used when

- We observe the confounders X that determine both Z and Y conditional on D (i.e. under the conditional exogeneity assumption A3' and common support A5).
- X is multidimensional (i.e. many cells defined by X), complicates nonparametric estimation.
- Remember: If instrument is (known to be) randomly assigned, do not need any control variables.

Roadmap based on Frölich (2007) using inverse probability weighting as an estimator 1. Estimate the model for $p(X_i) \equiv Pr(Z_i = 1|X_i = x)$ using probit, and calculate predicted probabilities $\hat{p}(X_i)$.

2. Calculate the LATE by reweighting observations by the inverse of their conditional instrument probabilities.

$$\text{LATE} = \frac{E\left[\frac{Y_i Z_i}{\hat{p}(X_i)} - \frac{Y_i(1-Z_i)}{(1-\hat{p}(X_i))}\right]}{E\left[\frac{D_i Z_i}{\hat{p}(X_i)} - \frac{D_i(1-Z_i)}{(1-\hat{p}(X_i))}\right]}$$

3. Bootstrap everything for inference.

1. Manually estimate the p-scores

```
# estimate pscores manually
pscore.model <- glm(z ~ x, family = binomial(link = "probit"))
summ(pscore.model, , robust = "HC1")
```

```
## MODEL INFO:
## Observations: 197071
## Dependent Variable: z
## Type: Generalized linear model
##   Family: binomial
##   Link function: probit
##
## MODEL FIT:
##  ²(7) = 177.95, p = 0.00
## Pseudo-R² (Cragg-Uhler) = 0.00
## Pseudo-R² (McFadden) = 0.00
## AIC = 273015.37, BIC = 273096.90
##
## Standard errors: Robust, type = HC1
## ------------------------------------------------
##                      Est.   S.E.   z val.      p
## ----------------- ------- ------ -------- ------
## (Intercept)         -0.04   0.03    -1.50   0.13
## xboy1st              0.05   0.01     8.86   0.00
## xboy2nd              0.05   0.01     9.57   0.00
```

```
## xagem               -0.00   0.00    -1.40   0.16
## xagefstm             0.00   0.00     1.76   0.08
## xblackm              0.01   0.01     0.84   0.40
## xhispm               0.01   0.02     0.73   0.46
## xothracem           -0.02   0.02    -1.38   0.17
## ------------------------------------------------
```

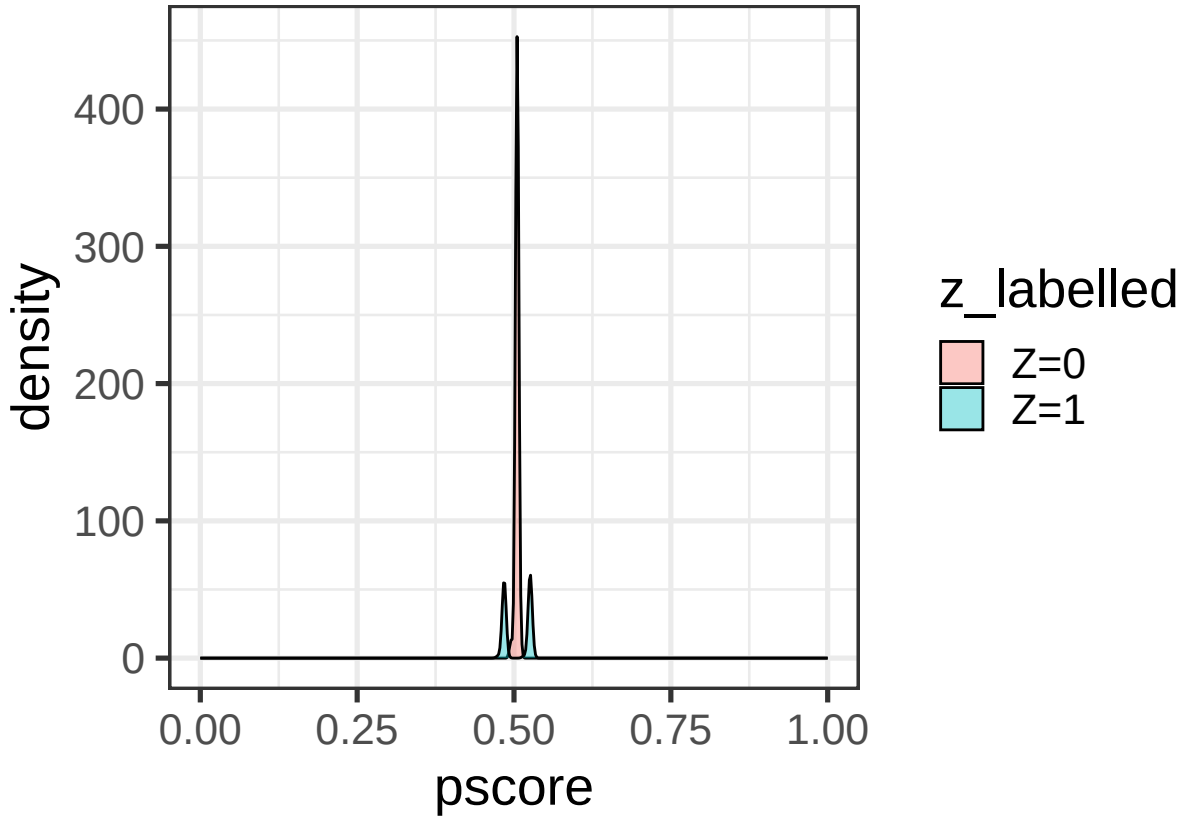
- Indicates if you should be worried about any confounders.
- Coefficients on gender of first child are significant because slightly higher probability of having boys.

2. Check common support

```
# add pscore to data frame
data$pscore <- pscore.model$fitted.values

# generate a labelled factor IV
data$z_labelled <- factor(z,
                          levels = c(0,1),
                          label = c("Z=0", "Z=1"))

# Check for common support in propensity score
ggplot(data, aes(x = pscore, fill = z_labelled)) +
                        geom_density(alpha=0.4)  +
                        theme_bw(base_size = 20) +
                                        xlim(0, 1)
```



- Poor common support.

3. Estimate the semiparametric LATE

Can use the `lateweight` command from the** `causalweight` package which implements all of the above mentionned steps.

- But as we have seen in Session 2, a bit of a black box, so check common support manually.
- Only works with binary treatment and instrument.
- Can specify trimming.

For more details, see https://cran.r-project.org/web/packages/causalweight/causalweight.pdf

```
# Exemplarily for the labour supply decision of mothers
late_workedm <- lateweight(y=workedm, # outcome
                           d=d, # binary endogenous treatment
                           z=z, # binary instrument
                           x=x, # observed confounders
                           LATT=FALSE, # for LATE
                           logit=FALSE, # probit p-score model
                           boot=2) # number of bootstrap replications
# increase boot to 199 or higher and go for a walk

# Display results
```

```r
print("LATE: ")
```

```
## [1] "LATE: "
```

```r
round(c(late_workedm$effect),3)
```

```
## [1] -0.115
```

```r
print("standard error: ")
```

```
## [1] "standard error: "
```

```r
round(c(late_workedm$se.effect),3)
```

```
## [1] 0.015
```

```r
print("p-value: ")
```

```
## [1] "p-value: "
```

```r
late_workedm$pval.effect
```

```
## [1] 8.299563e-14
```

```r
# For all outcomes
sp.model <- function(y){
      sp.m <- lateweight(y=y, # outcome
                         d=d, # binary endogenous treatment
                         z=z, # binary instrument
                         x=x, # observed confounders
                         LATT=FALSE, # for LATE
                         logit=FALSE, # probit p-score model
                         boot=9) # number of bootstrap replications
      list( sp.iv.coeff = sp.m$effect,
            boot.se = sp.m$se.effect,
            pval = sp.m$pval.effect)
}

# Apply to all outcomes of interest
sp.iv_output <- apply(y_mat, 2, sp.model)
sp.iv_output<-as.matrix(rbindlist(sp.iv_output))

# Bind OLS and IV results
output_all<-cbind(output, sp.iv_output,c(tab_wald[3,3],NA,NA,NA,NA),c(inf.LATE$se_late,NA,NA,NA,NA)
)
rownames(output_all)<- y_names
colnames(output_all)<- c("OLS", "se", "p", "2SLS", "se", "p", "Semi IV", "se", "p", "Wald", "se")

# Display table
```

```r
print("OLS, 2SLS and semiparametric LATE estimates of labour supply models")
```

```
## [1] "OLS, 2SLS and semiparametric LATE estimates of labour supply models"
```

```r
print(round(output_all,digits=3))
```

```
##               OLS     se p      2SLS      se     p   Semi IV      se     p
## workedm    -0.173  0.002 0    -0.115   0.037 0.002    -0.115   0.024 0.000
## weeksm     -8.861  0.100 0    -4.101   1.649 0.013    -4.102   0.964 0.000
## hourswm    -6.520  0.087 0    -3.847   1.405 0.006    -3.848   1.399 0.006
## incomem -3718.223 48.955 0 -1555.290 807.002 0.054 -1555.335 742.227 0.036
## lfaminc    -0.134  0.006 0     0.056   0.100 0.576     0.056   0.114 0.623
##           Wald    se
## workedm -0.128 0.037
## weeksm      NA    NA
## hourswm     NA    NA
## incomem     NA    NA
## lfaminc     NA    NA
```

References

Frölich M (2007). Nonparametric IV Estimation of Local Average Treatment Effects with Covariates. *Economics Letters*, 139, 35-75.