Econometría 1

División de Economía - CIDE

Dr. Francisco Cabrera

**Actividad 7. Variables Instrumentales y Variable Dependiente Dicotómica.**

1. **Consider a simple model to estimate the effect of personal computer (PC) ownership on college grade point average for graduating seniors at a large public university:**



where *PC* is a binary variable indicating PC ownership.

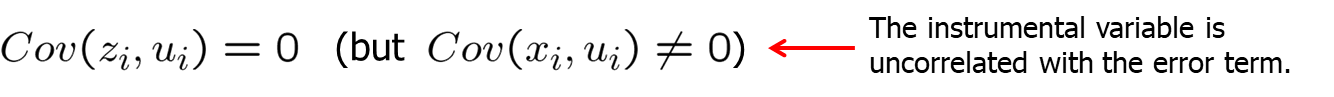
1. Why might PC ownership be correlated with *u*?
2. Explain why *PC* is likely to be related to parents’ annual income. Does this mean parental

income is a good IV for *PC*? Why or why not?

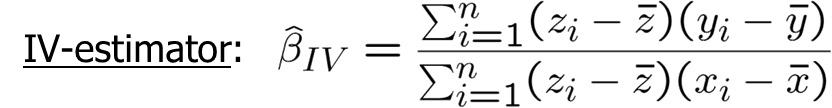
1. Suppose that, four years ago, the university gave grants to buy computers to roughly one half

of the incoming students, and the students who received grants were randomly chosen. Carefully explain how you would use this information to construct an instrumental variable for *PC*.

1. **Consider the following:**



1. Show that



1. Now consider the simple regression model

**

And let *z* be a *binary* instrumental variable for x. Use to show that the IV estimator of b1 can be written as: 

Where  and ** are the sample averages of *yi* and *xi* over the part of the sample with *zi* = 0, and where  and are the sample averages of *yi* and *xi* over the part of the sample with *zi* = 1. This estimator, known as a *grouping estimator*, was first suggested by Wald (1940).

*\*Remember We used this estimator en class to evaluate the effect of a random grant to incentivize unemployed people to take a training.*

1. **Upload the data file *card.dta* in the class folder “assignments”. Card (1993, 1995) uses wage and education data for a sample of men in 1976 to estimate the return to education. He uses a dummy variable for whether someone grew up near a four-year college (*nearc4*) as an IV for education.**

Other controls we might use in the log(wage)-education regression are:

* + - experience
    - a black dummy variable
    - dummy for living in an SMSA (city indicator - Standard Metropolitan Statistical Area)
    - Residence in 1966.

1. Is nearc4 a *credible* IV for education? Explain why or why not.
2. Now, obtain the IV estimate of the education on wages. How does this estimate compare with the OLS estimate?
3. To probe the credibility of this instrument, we will relate it to IQ. For a subsample of men in the data, an IQ score is available. Regress IQ on *nearc4* to check whether average IQ scores vary by whether the man grew up near a four-year college. What do you conclude?
4. Now regress IQ on nearc4, smsa66, and the 1966 regional dummy variables reg662,...,reg669. Are IQ and nearc4 related?
5. From parts c and d what do you conclude about the importance of controlling for smsa66 and the 1966 regional dummies in the log(wage) equation?
6. **Create your own data for IV simulation:**

#Clear the environment

rm(list = ls())

#Load packages

library(pacman)

p\_load(tidyverse, fixest, modelsummary)

#create your data

library(tidyverse)

set.seed(666)

df <- tibble(z = rnorm(1000),

u = rnorm(1000),

# x1 is endogenous since it correlates with u1 by construction

x = 0.5\*z + 4\*u + rnorm(1000),

y = 3\*x + 5\*u) #true value is 3

#regress

library(fixest)

ols\_estimator <- lm(y ~ x, df)

iv\_estimator <- feols(y ~ 1 | x ~ z, df, se = 'hetero')

summary(ols\_estimator)

summary(iv\_estimator)

library(modelsummary)

msummary(list(ols\_estimator, iv\_estimator), stars = TRUE, gof\_omit = '^(?!Num)', coef\_omit = "(Intercept)")

* 1. Is z a relevant and exogenous instrument”? Is it a “weak” instrument?
  2. Now change in the code above the x relation with z to 0.2 instead of 0.5. what happens? Is there a reason to think that the IV estimator is now biased? In what direction? show this formally with the corresponding formula.
  3. What happens with the SE now that the relation of x and z is weaker? Show this using the corresponding formula.
  4. Now, imagine there is some additional explanatory variable V which is unobserved but partially explains the instrument…

set.seed(666)

df <- tibble(v = rnorm(1000),

z2 = -v + rnorm(1000),

u2 = 0.1\*v + rnorm(1000),

x2 = 0.2\*z2 + 4\*u2 + rnorm(1000), # all coefficients stay the same here

y2 = 3\*x2 + 5\*u2) # the true effect is 3

ols\_est2 <- lm(y2 ~ x2, df)

iv\_est2 <- feols(y2 ~ 1 | x2 ~ z2, data = df, se = 'hetero')

msummary(list(ols\_est2, iv\_est2), stars = TRUE, gof\_omit = '^(?!Num)', coef\_omit = "(Intercept)")

Formally describe, using the corresponding formula the direction and reason for the inconsistency of bIV. What estimator is better OLS or IV?

* 1. Now suppose that the relevance of the instrument improves in the following way:

set.seed(7)

df <- tibble(v = rnorm(1000),

z2 = -v + rnorm(1000),

u2 = 0.1\*v + rnorm(1000),

#we just change x3

x3 = 3\*z2 + 4\*u2 + rnorm(1000), # all coefficients stay the same here

y2 = 3\*x3 + 5\*u2) # the true effect is 3

ols\_est3 <- lm(y2 ~ x3, df)

iv\_est3 <- feols(y2 ~ 1 | x3 ~ z2, data = df, se = 'hetero')

msummary(list(ols\_est3, iv\_est3), stars = TRUE, gof\_omit = '^(?!Num)', coef\_omit = "(Intercept)")

What is the “rule of thumb” to say that z is a relevant instrument in terms of t?

* 1. What is the reason why the IV instrument is now consistent despite not being purely “exogenous” Show this with the formula used above.
  2. Now run: thef <- fitstat(iv\_est3, 'ivf')[["ivf1::x3"]][["stat"]] Discuss the result.

1. Let *grad* be a dummy variable for whether a student-athlete at a large university graduate in five years. Let *hsGPA* and *SAT* be high school grade point average and SAT score, respectively. Let *study* be the number of hours spent per week in an organized study hall. Suppose that, using data on 420 student-athletes, the following logit model is obtained:



where  is the logit function.

1. Holding *hsGPA* fixed at 3.0 and *SAT* fixed at 1,200, compute the estimated difference in the graduation probability for someone who spent 10 hours per week in study hall and someone who spent 5 hours per week.
2. Remember that:  such that the Probit coefficient β1 is the change in a “latent variable” associated with a one unit change in X. The association between x and the “latent variable” is linear but the relation between X and Y is non-linear because it is mediated by f(z), which is nonlinear. Since the dependent variable is a nonlinear function of the regressors, the coefficient on X has no simple interpretation.

HMDA data in R provides data that relates to mortgage applications filed in Boston in the year of 1990. The variable we are interested in modelling is deny, an indicator for whether an applicant’s mortgage application has been accepted (deny = no) or denied (deny = yes). A regressor that ought to have power in explaining whether a mortgage application has been denied is *pirat*, the size of the anticipated total monthly loan payments relative to the applicant’s income.

* 1. Estimate the linear probability model for: A picture containing clock, watch, gauge

     Description automatically generatedand report your results below:
  2. Estimate the probit model and report your results below:

denyprobit <- glm(deny ~ pirat,

family = binomial(link = "probit"),

data = HMDA

coeftest(denyprobit, vcov. = vcovHC, type = "HC1")

#note that this is an heteroskedastic robust z-test, not a t-test due to normality in probit.

* 1. Write the estimated equation for the probit model:
  2. Compute the prediction for change in the P/I ratio from 0.3 to 0.4.
  3. Include the variable *single* and compute the estimated differences in probability between single and non-single people, or the AME of being single.
  4. Write the estimated equation in e. (hint: this is not directly interpretable as AME, but the signs would tell you if the probability of being rejected is higher or lower for higher P/I ratios and for black people.)
  5. Estimate the logit including the *P/I ratio* and *single* variables, by changing in the chunk of code above:

family = binomial(link = "logit"),

* 1. Estimate the pseudo R-squared and the LR test.