Econometria 1

División de Economía - CIDE

Dr. Francisco Cabrera

**Actividad 6. Especificación, Error de Medición, Experimentación e Identificación.**

1. Let *math10* denote the percentage of students at a Michigan high school receiving a passing score on a standardized math test (see also Example 4.2). We are interested in estimating the effect of per-student spending on math performance. A simple model is



1. The variable *lnchprg* is the percentage of students eligible for the federally funded school lunch program. Why is this a sensible proxy variable for *poverty*?
2. The following table contains OLS estimates, with and without *lnchprg* as an explanatory variable.

Tabla

Descripción generada automáticamente

* 1. Explain why the effect of expenditures on *math10* is lower in column (2) than in column (1). Is the effect in column (2) still statistically greater than zero?
  2. Explain why there is a change in sign for log(enroll) between model in Column (1) and in Column (2)
  3. It appears that pass rates are lower at larger schools, other factors being equal. Explain.
  4. Interpret the coefficient on *lnchprg* in column (2).
  5. What do you make of the substantial increase in *R*2 from column (1) to column (2)?

1. Suppose you conduct this same regression substituting the logarithm of expend and enroll with their levels and squares. How can you test what specification is a better prediction of math10? Describe the steps formally.
2. The following equation explains weekly hours of television viewing by a child in terms of the child’s

age, mother’s education, father’s education, and number of siblings:



We are worried that *tvhours\**  is measured with error in our survey. Let *tvhours* denote the reported

hours of television viewing per week.

* 1. Do you think the CEV assumptions are likely to hold? Explain.
  2. What do the classical errors-in-variables (CEV) assumptions require in this application?
  3. Show formally the bias that the measurement error in *tvhours\** generates.

1. You need to use two data sets for this exercise, JTRAIN2 and JTRAIN3. JTRAIN2 is the outcome of a job training **experiment**. The file JTRAIN3 contains **observational** data, where individuals themselves largely determine whether they participate in job training (self-select). The data sets cover the same time

period.

* 1. In the data set JTRAIN2, what fraction of the men received job training? What is the fraction in

JTRAIN3? Why do you think there is such a big difference?

* 1. Using JTRAIN2, run a simple regression of re78 on train. What is the estimated effect of participating in job training on real earnings?
  2. Now add as controls to the regression in part (ii) the variables re74, re75, educ, age, black, and hisp. Does the estimated effect of job training on re78 change much? why? (Hint: Remember that these are experimental data.)
  3. Do the regressions in parts (b) and (c) using the data in JTRAIN3, reporting only the estimated coefficients on train, along with their t statistics. What is the effect now of controlling for the extra factors, and why?
  4. Define avgre = (re74 + re75)/2. Find the sample averages, standard deviations, and minimum and maximum values in the two data sets. Are these data sets representative of the same populations in 1978?
  5. Almost 96% of men in the data set JTRAIN2 have avgre less than $10,000. Using only these men, run the regression re78 on train, re74, re75, educ, age, black, hisp and report the training estimate and its t statistic. Run the same regression for JTRAIN3, using only men with avgre < 10. For the subsample of low-income men, how do the estimated training effects compare across the experimental and nonexperimental data sets?
  6. Now use each data set to run the simple regression re78 on train, but only for men who were unemployed in 1974 and 1975. How do the training estimates compare now?
  7. Using your findings from the previous regressions, discuss the potential importance of having comparable populations underlying comparisons of experimental and nonexperimental estimates.

1. Does attending a summer school improve test scores? This **fictitious** setting is as follows:

* In the summer break between year 5 and year 6, (roughly corresponding to age 10) there is an optional summer school.
* The summer school could be focusing on the school curriculum, or it could be focused on skills that lead to improved schooling outcomes (for example “grit” as in [Alan et al (2019)](https://academic.oup.com/qje/article-abstract/134/3/1121/5342089?redirectedFrom=fulltext)).
* The summer school is free, but enrollment requires active involvement by parents.
* We are interested in whether participation in summer school improves child outcomes.

We have a dataset to study the research question, including:

Information about person id, school id, an indicator variable that takes the value of 1 if the individual participated in the summer school, information about gender, parental income and parental schooling, and test scores in year 5 (before the treatment) and year 6. The dataset also contains information about whether the individual received a reminder letter.

* 1. Load summercamp.dta into R (data is in the class folder). and include the data in an object called summercamp. Then:

# load tidyr package

library("tidyr")

# make data tidy (make long)

school\_data<-summercamp%>%

pivot\_longer(

cols = starts\_with("test\_year"),

names\_to = "year",

names\_prefix = "test\_year\_",

names\_transform = list(year = as.integer),

values\_to = "test\_score",

)

# Load skimr

library("skimr")

# Use skim() to skim the data

skim(school\_data)

* 1. Correlate missing values for parental\_schooling with parental income (hint: create a dummy for missing values). Is there evidence that missing values are not random?
  2. Assume all “missing values at random”. Hence drop NA rows.
  3. Why do we want to run the code below?

# Standardize test score

# Group analysisdata by year

analysisdata<-group\_by(summercamp,year)

# Create a new variable with mutate

analysisdata<-mutate(analysisdata, test\_score=(test\_score-mean(test\_score))/sd(test\_score))

# show mean of test\_score

print(paste("Mean of test score:",mean(analysisdata$test\_score)))

#show sd of test\_score

print(paste("SD of test score:",sd(analysisdata$test\_score)))

* 1. Create a bar chart of pre-summer school test scores (in SD) and summer school attendance ¿Is there evidence of a selection bias?

# Load patchwork

library("patchwork")

# Create raw chart element

rawchart<-ggplot(analysisdata%>%filter(year==5),x=as.factor(fill))+

theme\_classic()

p2<-rawchart+

geom\_bar(aes(x=as.factor(summercamp),y=test\_score),

stat="summary",fun="mean")+

labs(y="Test Score Year 5", x="Attended Summer School")

* 1. Denote, formally (i.e. using expected values and potential outcomes notation), how the selection bias arises in the case of a naive comparison between those who attend summer school and those who do not.
  2. What can we conclude regarding selection bias after the table generated by:

# Load libraries

library(modelsummary)

library(estimatr)

# Filter and modify data

testdata<-filter(analysisdata,year==5)

testdata<-ungroup(testdata)

testdata<-mutate(testdata,Treated=ifelse(summercamp==1,"Summer Camp","No Summer Camp"))

testdata<-select(testdata,female,parental\_schooling,parental\_lincome,test\_score,Treated)

testdata<-rename(testdata,`Female`=female,

`Parental schooling (years)`=parental\_schooling,

`Parental income (log)`=parental\_lincome,

`Test Score`=test\_score)

# Table with balancing test

datasummary\_balance(~Treated,

data = testdata,

title = "Balance of pre-treatment variables",

notes = "Notes: Goya, Goya, Universidad!",

fmt= '%.5f',

dinm\_statistic = "p.value")

* 1. Reproduce the same table for students receiving a reminding letter and those who do not. What can you conclude about the letter “assignation” does it appear to be “as good as random”?
  2. Run an OLS estimation relating letter and summer camp attendance, including a set of sensible controls.
  3. Run a regression without controls that allows you to accomplish Gauss-Markov assumptions *with standardized test scores after summer camp (i.e. year 6)* as a dependent variable ¿What is the interpretation of this effect? ¿Is it an ATE or an ATT?
  4. Run a regression **with good controls** that allow you to accomplish Gauss-Markov assumptions *with standardized test scores after summer camp (i.e. year 6)* as a dependent variable ¿is the result of the estimator of interest different from (k)?

1. Imagine you own a snickers store. An employee suggests giving “little gifts” (i.e. key rings) to clients to **make them return** to the store. You have 200 stores; an economist thus creates an experiment to evaluate the effect of the “little gifts” on **revenues**.

The data of this little experiment is given by:

library(pacman)

p\_load(tidyverse, glue)

#Clear the environment

rm(list = ls())

# experimental data at the store level dataset

set.seed(22)

n\_stores <- 200

true\_gift\_effect <- 100

noise <- 50

data\_downstream <- tibble(store\_id = 1:n\_stores) %>%

mutate( #mutate allows to create new\_columns in data frame.

# treatment (random in this case)

gives\_gift=rbinom(n\_stores, 1, prob = 0.5),

# return rate increased by 20% if given gifts

return\_rate=rnorm(n\_stores, mean = 0.5, sd=0.1) + gives\_gift\*0.1,

# outcome (influenced by return rate)

# gifs impact revenue through return rate

revenue= 50 + true\_gift\_effect\*10\*return\_rate + rnorm(n\_stores, mean=0, sd=noise)

)

# plot to visualize the relationship

data\_downstream %>%

mutate(treatment=ifelse(gives\_gift==1, "gift", "no gift")) %>%

ggplot(aes(return\_rate, revenue, color=treatment)) +

geom\_point() +

labs(title=glue(" ")) + geom\_rug()

1. Discuss the data depicted in the graph plot and draw a DAG on the relationship between gifts, return rates, and revenues.
2. Run a regression of gifts on return rates. Does the gift make customers return to the store?
3. What is the regression a well-trained economist would run after the experiment to know the effect of the *gifts*?
4. Describe formally, this is with the use of the potential outcomes notation, why controlling for “return rates” in the experimental regression creates a bias.
5. Show how controlling for *return rates* biases our coefficient in a regression setting (i.e., run a regression controlling for *return rates*). Discuss your result and show formally, with the use of potential outcomes notation, why the regressor *gifts* is downward biased.