Actividad 3

Ejercicio 1. Propiedades Asintóticas.

1. Suponga que $\{x_N\}$ es una secuencia de vectores aleatorios de dimensión K tal que $x_N \xrightarrow{p} \beta$ y que $\sqrt{N}(x_N - \beta) \xrightarrow{d} z$. Suponga que $a(\cdot) : \mathbb{R}^K \to \mathbb{R}^r$ puede diferenciarse una vez, siendo $A(\beta) \equiv \frac{\partial a(\beta)}{\partial \beta'}$ la matriz de primeras derivadas de dimensión $r \times K$ evaluadas en β.

   (a) Pruebe que $\sqrt{N}(a(x_N) - a(\beta)) \xrightarrow{d} A(\beta)z$.

   (b) Asuma además que $\sqrt{N}(x_N - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. ¿A qué es igual z en este caso particular?

2. Sea z_n definido por:

$$z_n = \begin{cases} 0 & \text{con probabilidad } \frac{n-1}{n}, \\ n^2 & \text{con probabilidad } \frac{1}{n}. \end{cases}$$

   Demuestre que $\text{plim}_{n\to\infty} z_n = 0$ pero $\lim_{n\to\infty} E(z_n) = \infty$.

3. Demuestre la Ley Débil de los Grandes Números (LLN) de Chebyshev, mostrando que $\bar{z}_n \to \mu$.

   Hint:
   $(\bar{z}_n - \mu) = (\bar{z_n} - E(\bar{z}_n)) + (E(\bar{z}_n) - \mu)$.

4. Let

$$Z_n \xrightarrow{d} Z \quad \text{where} \quad g \text{ is continuous} \implies g(Z_n) \xrightarrow{d} g(Z)$$

   (a) Mention the theorem we refer to, and explain intuitively what it implies.

   (b) Now using Chat GPT, and your own experience, follow the next steps in R:

      i. Define the function $g(u) = u^2$: This is a continuous function.

      ii. Generate a sequence of random variables Z_n**: We generate ‘n‘ observations from a standard normal distribution $N(0, 1)$.

      iii. Apply the function g to Z_n: We calculate $g(Z_n)$.

      iv. Theoretical distribution: Since Z is $N(0, 1)$, $g(Z) = Z^2$ follows a chi-square distribution with 1 degree of freedom.

      v. Plot the distributions: Plot the empirical distribution of $g(Z_n)$ and compare it with the theoretical chi-square distribution.

   (c) ¿Why is this important for OLS estimation of the Least Squares estimator? show this formally (see B.H, section 7.2)

5. Realice tres simulaciones en R para ilustrar la eficiencia, consistencia y la normalidad asintótica de las estimaciones de OLS. Considere el siguiente DGP:

$$y_i = 1 + 2x_{i1} + \epsilon_i,$$

   donde $\epsilon \sim N(0, 1)$.

   En específico, muestre cómo afecta a) $E[X_1 e] \neq 0$; y b) la presencia de alta multicolinearidad, tanto a la consistencia como a la eficiencia de la estimación asimptótica.

c) Suponga ahora que $E(X_1 e) = 0$; $E(X_2 e) \neq 0$; y que $corr(X_1, X_2) \to 1$. Muestre el proceso $\hat{\beta}_1 \to_p \beta_1$ ¿Cuál es su conclusión?

Ejercicio 2. Topicos Aplicados

1) 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:

$$tvhours^* = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 mothereduc + \beta_5 siblings + u$$

We are worried that $tvhours^*$ is measured with error in our survey. Let tvhours denote the reported hours of television viewing per week.

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

Ejercicio 3. Outcomes Potenciales

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").
- 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 data set also contains information about whether the individual received a *reminder letter*.

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

```r
# 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",
       )
```

b.

```r
# Load skimr
library("skimr")
# Use skim() to skim the data
skim(school_data)
```

c. 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?
d. Assume all "missing values at random". Hence drop NA rows.
e. Why do we want to run the code below?

```
# Standardize test score
# Group analysis data 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)))
```

f. Create the bar chart of pre-summer school test scores (in SD) and summer school attendance as below ¿Is there evidence of a selection bias? Explain.

```
# 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")
```

g. 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.

h. 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")
```

i. 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"?

j. Run an OLS estimation relating letter and summer camp attendance, including a set of sensible controls.

k. 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?

l. 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)? why?

2) 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()
```

a. Discuss the data depicted in the graph plot and draw a Direct Acyclical Graph (check: https://mixtape.scunning.com/03-directed_acyclical_graphs) on the relationship between gifts, return rates, and revenues.
b. Run a regression of gifts on return rates. Does the gift make customers return to the store?
c. What is the regression a well-trained economist would run after the experiment to know the effect of the gifts?
d. Describe formally, this is with the use of the potential outcomes notation, why controlling for "return rates" in the experimental regression creates a bias.
e. 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.