7316 - INTRODUCTION TO R

Final Assignment

Uyen Vu ([uyen.vu@phdstudent.hhs.se](mailto:uyen.vu@phdstudent.hhs.se))

5/8/2022

Table of Contents

# LOAD LIBRARIES

library(tidyverse)  
library(dplyr)  
library(tidyr)  
library(rio)  
library(knitr)  
library(ggplot2)  
library(ggthemes)

# EXERCISE 1: REPLICATE FIGURE 4

## Load the dataset

data\_original <- rio::import("data/pub\_pvt\_scatters.dta")  
uniid <- rio::import("data/univ\_names.xlsx")

## Merge data

* Steps:
  + When we merge 2 data frames we only keep obs of common university ID
  + We reorder the columns for better view
* Command:

df <- merge(data\_original, uniid, by="unitid", all = FALSE)

df <- dplyr::relocate(df, name, .after = unitid)

## Create a common deflator

library(dplyr)  
df <- df %>%  
 group\_by(year) %>%  
 mutate(cpi\_all = mean(cpi, na.rm = TRUE))

## Create the real value of the appropriation

df$real\_approp <- df$nominal\_approp / df$cpi\_all

## Differentiate public and private university

df$Private <- as.factor(df$Private)

class(df$Private)

## Reorder data by unit and by year

df <- df %>% dplyr::arrange(unitid, year)

## Create log of variable

df <- df %>%  
 mutate(log\_real\_approp = log(real\_approp))  
df <- df %>%  
 mutate(log\_foreign\_freshmen = log(ENROLL\_FRESH\_NON\_RES\_ALIEN\_DEG))

## Create difference in log values by university between 2005 and 2012

* We first create 2 temporary data frames, one for log\_real\_approp and another for log\_foreign\_freshmen

tmp1 <- df %>%   
 select(name, year, log\_real\_approp) %>%  
 filter(year == 2005 | year == 2012)  
tmp2 <- df %>%  
 select(name, year, log\_foreign\_freshmen) %>%  
 filter(year == 2005 | year == 2012)

* Calculate the difference between 2005 and 2012 for log\_real\_approp

tmp1 <- tmp1 %>%  
 tidyr::spread(key = "year", value = "log\_real\_approp")  
tmp1 <- as.data.frame(tmp1)  
tmp1$change\_logreal <- tmp1$`2012` - tmp1$`2005`  
tmp1 <- tmp1 %>%  
 select(name, change\_logreal)

* Calculate the difference between 2005 and 2012 for log\_foreign\_freshmen

tmp2 <- tmp2 %>%  
 tidyr::spread(key = "year", value = "log\_foreign\_freshmen")  
tmp2 <- as.data.frame(tmp2)  
tmp2$change\_logforeign <- tmp2$`2012`-tmp2$`2005`  
tmp2 <- tmp2 %>%  
 select(name, change\_logforeign)

* Create a data frame for the school type (whether private or public)

tmp3 <- df %>%  
 as.data.frame() %>%  
 select(name, Private) %>%  
 distinct()

* Create data frame for figure 4

df\_fig4 <- merge(x = tmp1, y = tmp2, by = "name")  
df\_fig4 <- merge(x = df\_fig4, y = tmp3, by = "name")  
df\_fig4$Private <- as.integer(df\_fig4$Private)  
df\_fig4$Private[df\_fig4$Private == 1] <- "Public"  
df\_fig4$Private[df\_fig4$Private == 2] <- "Private"

## Replicate Figure 4

library(ggplot2)  
png(filename = "figure4.png", unit = "cm", width = 12, height = 12, res = 800)  
  
ggplot(data = df\_fig4, aes(x = change\_logreal, y = change\_logforeign, color = Private, shape = Private)) +  
 geom\_point(size = 2) +  
 scale\_color\_manual(values = c("darkred","#0567B9")) +  
 scale\_shape\_manual(values = c(0,16)) +  
 geom\_text(label = df\_fig4$name, size = 2, vjust = 0, nudge\_y = 0.1, show.legend = FALSE) +  
 geom\_smooth(method = "lm", formula = "y ~ x", se = FALSE, show.legend = FALSE, size=0.5) +  
 ylim (-0.4,3) +  
 labs(x = "Change in log(total appropriations in state)",  
 y = "Change in log(foreign freshmen)") +  
 theme\_classic(base\_size = 14) +  
 ggtitle("AAU (2005-2012)") +  
 theme(plot.title = element\_text(hjust = 0.5, size = 14)) +  
 theme(axis.title = element\_text(size = 12)) +  
 theme(legend.text = element\_text(size = 8), legend.title = element\_blank(), legend.position = c(0.89,0.9), legend.box.background = element\_rect(fill = "white", color = "black"))  
  
dev.off()

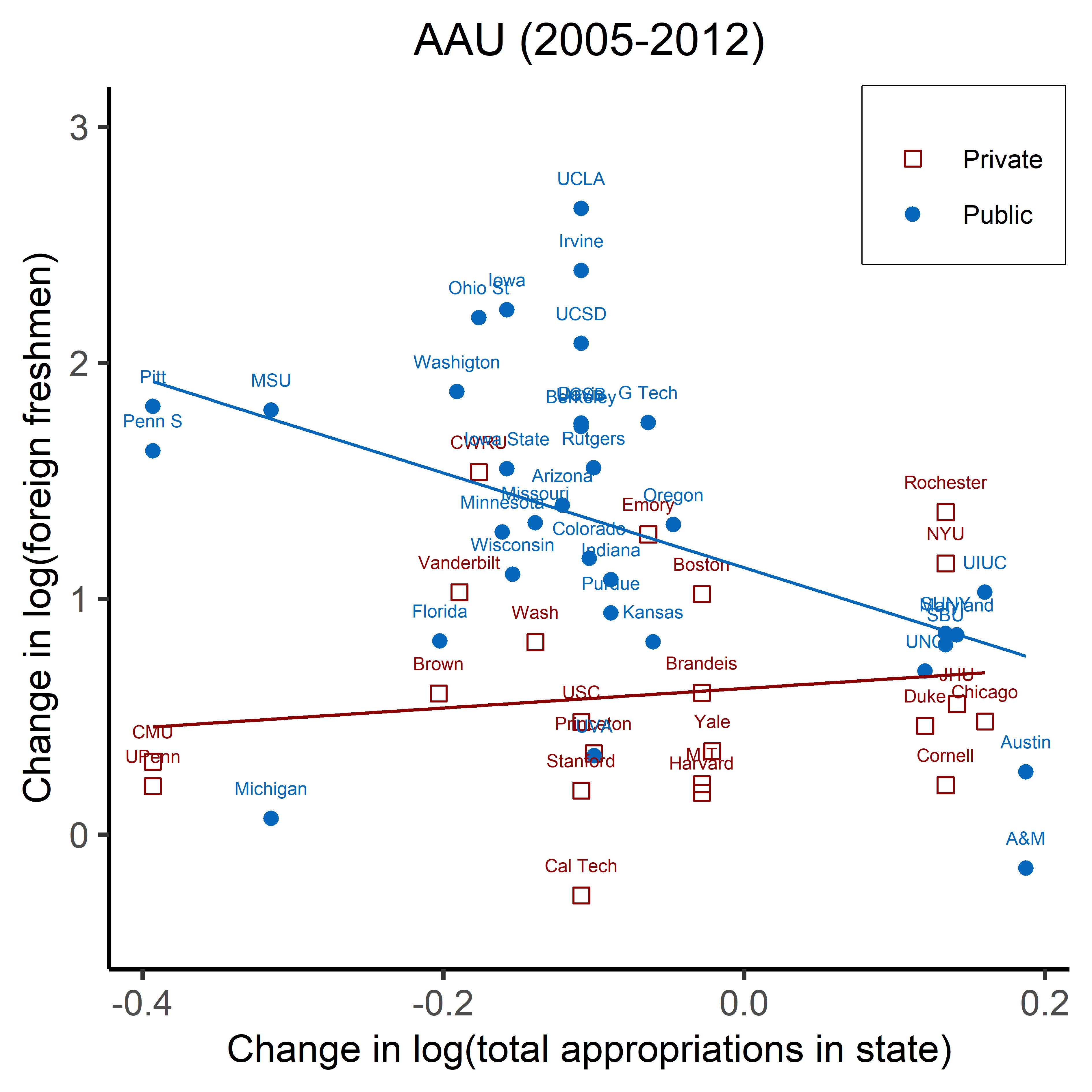


Figure 4. FOREIGN ENROLLMENT AND STATE LEVEL APPROPRIATION TO HIGHER EDUCATION: AAU UNIVERSITY, 2005-2012

# EXERCISE 2: REPLICATE TABLE 2

## Load the dataset

unidata <- rio::import("data/univ\_data.dta")

## Create variable for total state appropriation of all other universities within the same state

unidata$total\_state\_ap <- unidata$nominal\_approp - unidata$state\_ap\*100000  
unidata$l\_total\_state\_ap <- log(unidata$total\_state\_ap)

## Find the balanced sample

library(tidyr)  
balanced <- unidata %>%   
 drop\_na(ENROLL\_FRESH\_NON\_RES\_ALIEN\_DEG, AMERICAN\_OOS, IN\_STATE\_FRESHMEN)

## Run the regression from table 2

* Load library **lfe** and **stargazer**

library(lfe)  
library(stargazer)

* Create data frame for each group: research, AAU and nonResearch

research <- balanced %>% filter(Research == 1)  
nonResearch <- balanced %>% filter(nonResearch == 1)  
AAU <- balanced %>% filter(AAU == 1)

* Run **OLS** regression

research\_ols <- felm(l\_ENROLL\_FRESH\_NON\_RES\_ALIEN\_DEG ~ l\_state\_ap + l\_population | unitid + year | 0 | unitid, data = research, weights = research$weight)  
  
nonResearch\_ols <- felm(l\_ENROLL\_FRESH\_NON\_RES\_ALIEN\_DEG ~ l\_state\_ap + l\_population | unitid + year | 0 | unitid, data = nonResearch, weights = nonResearch$weight)  
  
AAU\_ols <- felm(l\_ENROLL\_FRESH\_NON\_RES\_ALIEN\_DEG ~ l\_state\_ap + l\_population | unitid + year | 0 | unitid, data = AAU, weights = AAU$weight)

* Run **1st stage** regression

research\_1st <- felm(l\_state\_ap ~ l\_total\_state\_ap | unitid + year | 0 | state\_of\_college, data = research)  
  
nonResearch\_1st <- felm(l\_state\_ap ~ l\_total\_state\_ap | unitid + year | 0 | state\_of\_college, data = nonResearch, weights = nonResearch$weight)  
  
AAU\_1st <- felm(l\_state\_ap ~ l\_total\_state\_ap | unitid + year | 0 | state\_of\_college, data = AAU, weights = AAU$weight)

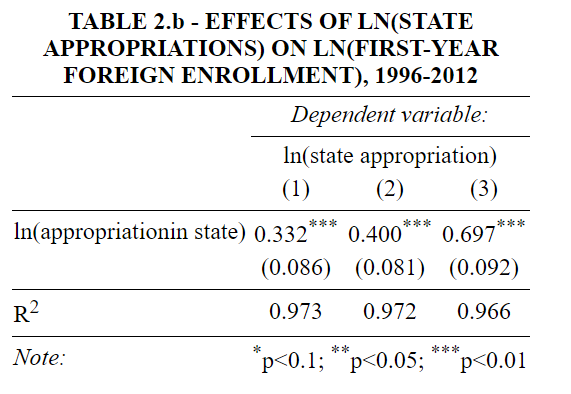
* Run **2nd stage** regression

research\_iv <- felm(l\_ENROLL\_FRESH\_NON\_RES\_ALIEN\_DEG ~ l\_population | unitid + year | (l\_state\_ap ~ l\_total\_state\_ap) | state\_of\_college, data = research, weights = research$weight)  
  
nonResearch\_iv <- felm(l\_ENROLL\_FRESH\_NON\_RES\_ALIEN\_DEG ~ l\_population | unitid + year | (l\_state\_ap ~ l\_total\_state\_ap) | state\_of\_college, data = nonResearch, weights = nonResearch$weight)  
  
AAU\_iv <- felm(l\_ENROLL\_FRESH\_NON\_RES\_ALIEN\_DEG ~ l\_population | unitid + year | (l\_state\_ap ~ l\_total\_state\_ap) | state\_of\_college, data = AAU, weights = AAU$weight)

## Arrange regression result onto the table

* Table for **1st stage** regression

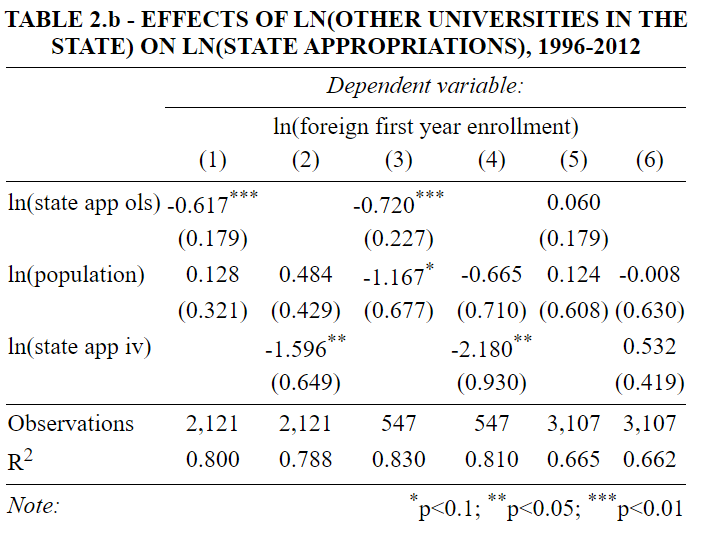
stargazer(research\_1st, AAU\_1st, nonResearch\_1st,   
 title = "TABLE 2.b - EFFECTS OF LN(STATE APPROPRIATIONS) ON LN(FIRST-YEAR FOREIGN ENROLLMENT), 1996-2012",   
 dep.var.labels = c("ln(state appropriation)"),  
 covariate.labels = c("ln(appropriation\_other in state)"),  
 omit.stat = c("LL", "f", "ser", "adj.rsq","n"),  
 no.space = TRUE,  
 align = TRUE,   
 type = 'html',  
 out = "1ststage.html")



Panel B - First Stage

* Table for **OLS** and **2nd stage** regression

stargazer(research\_1st, AAU\_1st, nonResearch\_1st,   
 title = "TABLE 2.b - EFFECTS OF LN(STATE APPROPRIATIONS) ON LN(FIRST-YEAR FOREIGN ENROLLMENT), 1996-2012",   
 dep.var.labels = c("ln(state appropriation)"),  
 covariate.labels = c("ln(appropriation\_other in state)"),  
 omit.stat = c("LL", "f", "ser", "adj.rsq","n"),  
 no.space = TRUE,  
 align = TRUE,   
 type = 'html',  
 out = "1ststage.html")



Panel A - OLS & Second Stage

# KEY NOTES

I don’t get the same result for everything and have not figured out the reasons.

* I got a slightly different number of observations for non-research group, hence the estimation for this group comes differently too.
* My estimation for 1st stage also has problem
* All the R2 come different from that in the article