RAboot

Yuto Kitano

2024-08-13

Formatting Data

Formatting Semester Data

```
setwd("/Users/kitanoyuuto/Downloads/warmup training package/01_data/raw/semester_dummy")
semester_1 <- read.csv("semester_data_1.csv", header = TRUE, skip = 1)

semester_2 <- read.csv("semester_data_2.csv")

colnames(semester_2) <- colnames(semester_1)

semester_data <- rbind(semester_1, semester_2) %>%
  select(-Y)

semester_data <- semester_data %>%
  group_by(unitid) %>%
  mutate(semester_start_year = ifelse(any(semester == 1), min(year[semester == 1]), NA))

semester_data <- semester_data %>%
  mutate(semester_flag = ifelse(year >= semester_start_year, 1, 0))
```

Formatting Gradrate Data

```
library(readxl)
library(purrr)
library(stringr)

file_paths <- list.files(path = "/Users/kitanoyuuto/Downloads/warmup training package/01_data/raw/outcor
valid_file_paths <- file_paths[!str_detect(file_paths, "~\\$")]

data_list <- map(valid_file_paths, read_excel)


graduate_data <- bind_rows(data_list) %>%
  filter(year <= 2010) %>%
  mutate(
    tot4yrgrads = as.numeric(tot4yrgrads),
    totcohortsize = as.numeric(totcohortsize),
    women_gradrate_4yr = as.numeric(women_gradrate_4yr),
    m_4yrgrads = as.numeric(m_4yrgrads),
    m_cohortsize = as.numeric(m_cohortsize)
```

```
  ) %>%
  mutate(
    w_grad = round(women_gradrate_4yr * 0.01,3),
    t_grad = round(tot4yrgrads / totcohortsize,3),
    m_grad = round(m_4yrgrads / m_cohortsize,3)
  )
```

Formatting Covariates Data

```
setwd("/Users/kitanoyuuto/Downloads/warmup training package/01_data/raw/covariates")

covariates <- read_xlsx("covariates.xlsx") %>%
  rename("unitid" = "university_id")

covariates <- covariates %>%
  mutate(unitid = str_replace_all(unitid, "aaaa", ""))

cov_wide <- covariates %>%
  pivot_wider(names_from = category, values_from = value,
              values_fill = NA,
              id_cols = c(unitid, year)) %>%
  filter(year >= 1991 & year <= 2010)

id <- unique(semester_data$unitid)

cov_wide <- cov_wide %>%
  filter(cov_wide$unitid  %in% id) %>%
  mutate(year = as.numeric(year)) %>%
   mutate(unitid = as.numeric(unitid))
```

Left_Join

```
data <- left_join(semester_data, graduate_data , by = c("unitid", "year"))

data <- left_join(data, cov_wide, by = c("unitid", "year"))
```

Descriptive Statistics

Table

```
data_nt <- data %>%
  group_by(unitid) %>%
  mutate(mean_quarter = mean(quarter, na.rm = TRUE)) %>%
  filter(mean_quarter == 1) %>%
  ungroup()

data_at <- data %>%
  group_by(unitid) %>%
  mutate(mean_quarter = mean(quarter, na.rm = TRUE)) %>%
  filter(mean_quarter != 1) %>%
  ungroup()
```

```r
summarise <- data %>%
  ungroup() %>%
  summarise(
    totcohortsize = mean(totcohortsize, na.rm = TRUE),
    w_cohortsize = mean(w_cohortsize, na.rm = TRUE),
    m_cohortsize = mean(m_cohortsize, na.rm = TRUE),
    t_grad = mean(t_grad, na.rm = TRUE),
    w_grad = mean(w_grad, na.rm = TRUE),
    m_grad = mean(m_grad, na.rm = TRUE)
  )

summarise_nt <- data_nt %>%
  ungroup() %>%
  summarise(
    totcohortsize = mean(totcohortsize, na.rm = TRUE),
    w_cohortsize = mean(w_cohortsize, na.rm = TRUE),
    m_cohortsize = mean(m_cohortsize, na.rm = TRUE),
    t_grad = mean(t_grad, na.rm = TRUE),
    w_grad = mean(w_grad, na.rm = TRUE),
    m_grad = mean(m_grad, na.rm = TRUE)
  )

summarise_at <- data_at %>%
  ungroup() %>%
  summarise(
    totcohortsize = mean(totcohortsize, na.rm = TRUE),
    w_cohortsize = mean(w_cohortsize, na.rm = TRUE),
    m_cohortsize = mean(m_cohortsize, na.rm = TRUE),
    t_grad = mean(t_grad, na.rm = TRUE),
    w_grad = mean(w_grad, na.rm = TRUE),
    m_grad = mean(m_grad, na.rm = TRUE)
  )

summarise_all <- summarise %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "All_Data")

summarise_nt_long <- summarise_nt %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Control_Group")

summarise_at_long <- summarise_at %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Treatment_Group")

summarise_at_long$Variable
```

```
## [1] "totcohortsize" "w_cohortsize"  "m_cohortsize"  "t_grad"
## [5] "w_grad"        "m_grad"
```

```r
summary_table <- cbind(
  summarise_all$All_Data,
  summarise_nt_long$Control_Group,
  summarise_at_long$Treatment_Group
)
```

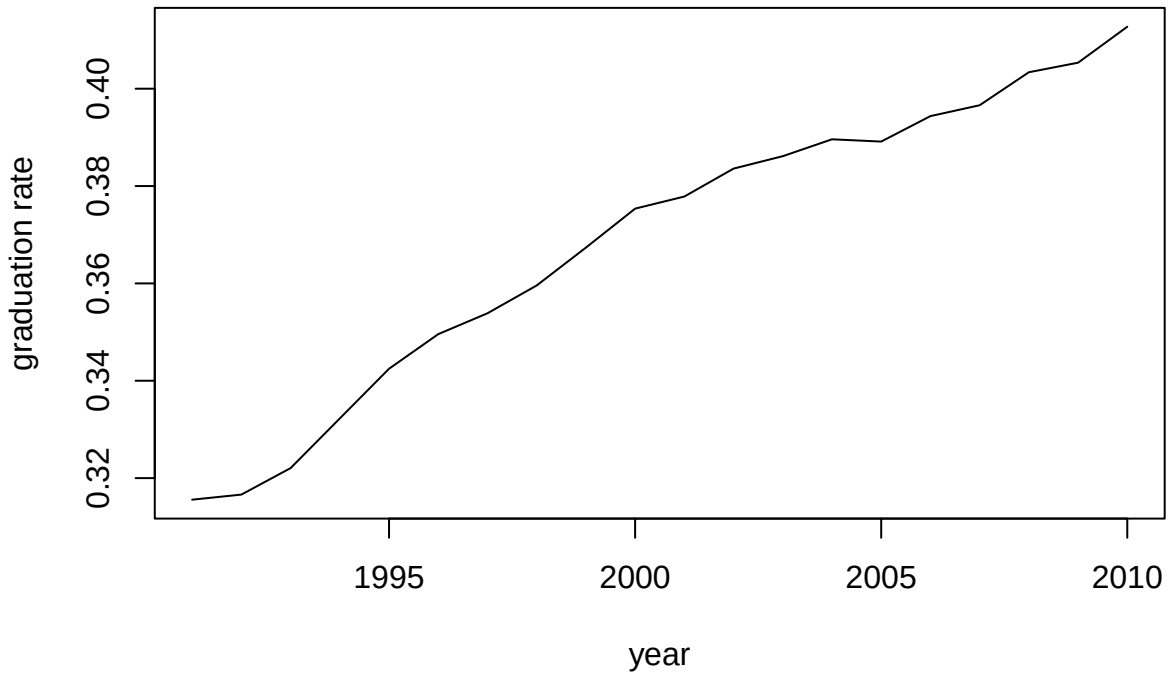
```r
colnames(summary_table) <- c("All_Data", "Control_Group", "Treatment_Group")

rownames(summary_table) <- c("totcohortsize", "w_cohortsize", "m_cohortsize", "t_grad", "w_grad", "m_gra

summary_table <- as.data.frame(summary_table)

library(xtable)
```

	All_Data	Control_Group	Treatment_Group
totcohortsize	1099.45	1695.19	1070.39
w_cohortsize	599.50	880.66	585.78
m_cohortsize	499.95	814.53	484.60
t_grad	0.37	0.38	0.37
w_grad	0.41	0.42	0.41
m_grad	0.32	0.33	0.32

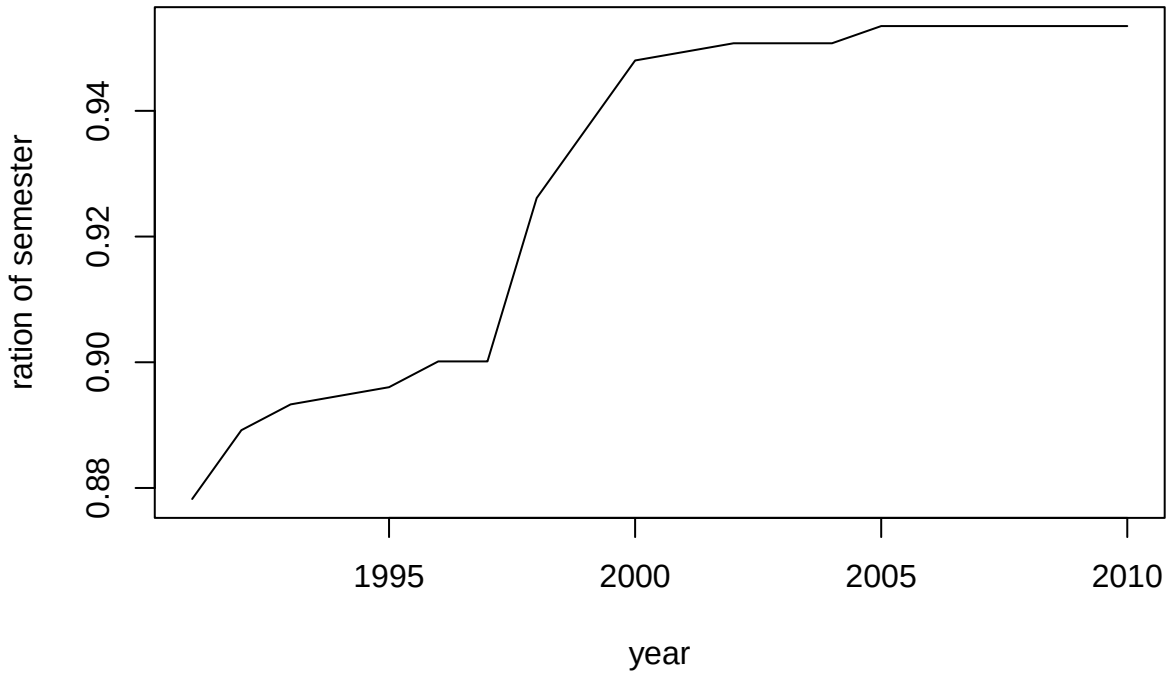
Table 1: Summary Table

Figure_grad

```r
data_fig <- data %>%
  group_by(year) %>%
  summarise(t_grad = mean(t_grad, na.rm = TRUE))

data_fig_2 <- data %>%
  group_by(year) %>%
  summarise(semester = mean(semester, na.rm = TRUE))

plot(x = data_fig$year, y = data_fig$t_grad, type = "l", xlab = "year",
     ylab = "graduation rate")
```

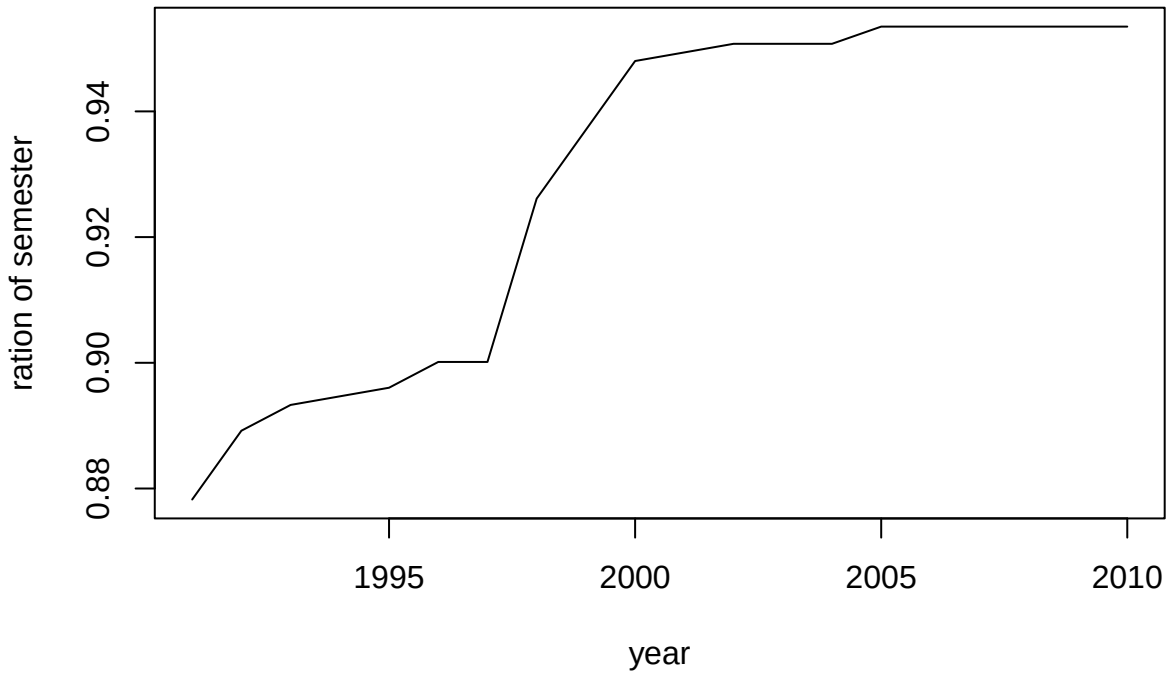


```r
plot(x = data_fig_2$year, y = data_fig_2$semester, type = "l", xlab = "year",
     ylab = "ration of semester")
```

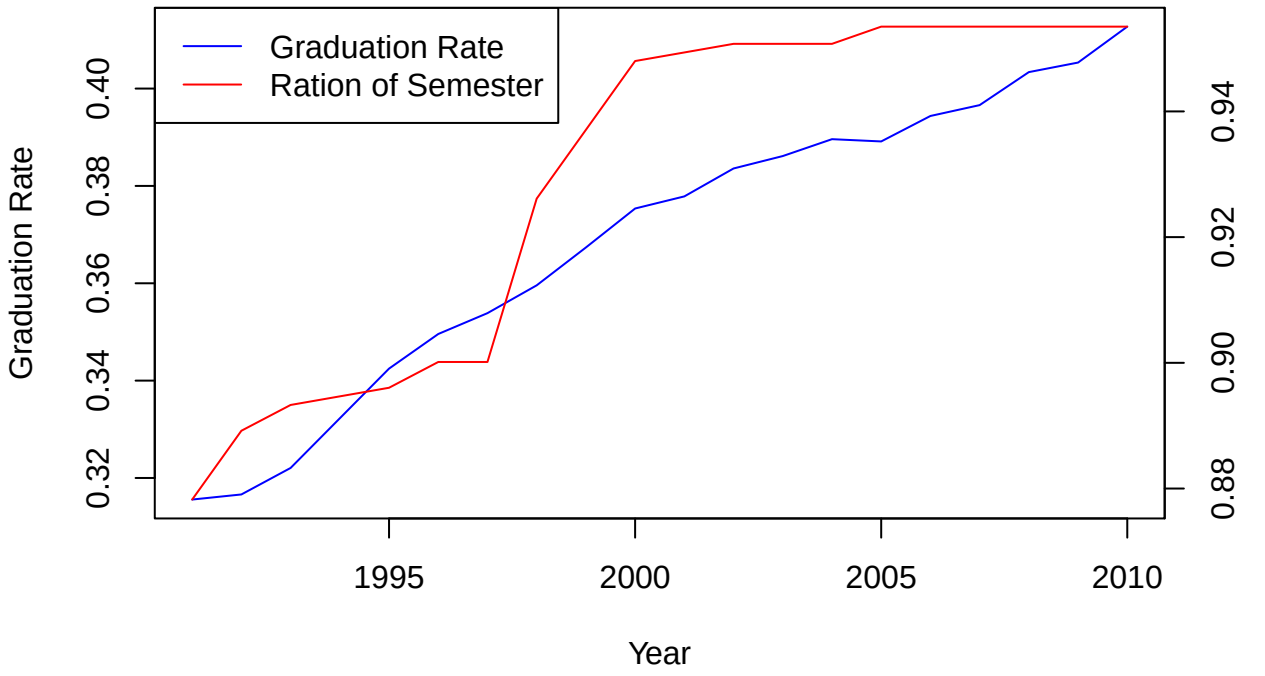


Figure_semester

```r
plot(x = data_fig_2$year, y = data_fig_2$semester, type = "l", xlab = "year",
     ylab = "ration of semester")
```



```r
plot(x = data_fig$year, y = data_fig$t_grad, type = "l", col = "blue",
     xlab = "Year", ylab = "Graduation Rate")

par(new = TRUE)

plot(x = data_fig_2$year, y = data_fig_2$semester, type = "l", col = "red",
     axes = FALSE, xlab = "", ylab = "")

axis(side = 4, at = pretty(range(data_fig_2$semester)))

mtext("Ration of Semester", side = 4, line = 3)

legend("topleft", legend = c("Graduation Rate", "Ration of Semester"),
       col = c("blue", "red"), lty = 1)
```



Inference

```r
library(stargazer)
lm_model <- lm(t_grad ~ data$semester_flag, data = data)
```

Table 2

	Dependent variable:
	t_grad
semester_flag	0.122***
	(0.013)
Constant	0.251***
	(0.013)
Observations	13,243
R^2	0.007
Adjusted R^2	0.007
Residual Std. Error	0.224 (df = 13241)
F Statistic	89.741*** (df = 1; 13241)
Note:	*p<0.1; **p<0.05; ***p<0.01