

中級ミクロデータサイエンス期末課題

Problem Set 2

横浜国立大学経済学部 3 年
学籍番号 2125178
廣江友哉

2024 年 2 月 5 日

(a) 記述統計

a-1, 問題背景などを知る上で役に立つ記述統計を作成し、内容について議論しなさい

Table 1—Institution-Level Summary Statistics				
Variable	N	Overall, N = 13,889 [†]	Never Switcher, N = 12,825 [†]	Switcher, N = 1,064 [†]
Cohort Size	13,889	1,099 (1,183)	1,086 (1,170)	1,258 (1,317)
Women's Cohort Size	13,889	599 (629)	593 (620)	675 (722)
Men's Cohort Size	13,889	500 (571)	493 (566)	583 (619)
4-Year Graduates	13,889	211 (93, 448)	216 (94, 457)	169 (78, 332)
Women's 4-Year Graduates	13,889	135 (60, 277)	138 (61, 282)	105 (50, 217)
Men's 4-Year Graduates	13,889	73 (29, 173)	75 (30, 178)	60 (23, 124)
4-Year Graduation Rate	13,889	0.37 (0.23)	0.38 (0.23)	0.27 (0.18)
Women's 4-Year Graduation Rate	13,865	0.41 (0.23)	0.42 (0.23)	0.32 (0.19)
Men's 4-Year Graduation Rate	13,824	0.32 (0.23)	0.33 (0.23)	0.23 (0.17)
Women's Cohort Size(%)	13,889	0.56 (0.51, 0.61)	0.56 (0.51, 0.61)	0.56 (0.51, 0.60)
White Cohort Size(%)	13,889	0.79 (0.64, 0.88)	0.79 (0.64, 0.88)	0.79 (0.67, 0.87)
In-State Tuition	13,889	8,562 (3,140, 17,300)	9,186 (3,252, 17,600)	3,875 (2,414, 11,194)
faculty	13,889	188 (96, 437)	185 (96, 430)	216 (101, 492)
costs	13,889	64 (31, 156)	64 (31, 153)	67 (32, 193)
[†] Mean (SD); Median (IQR)				

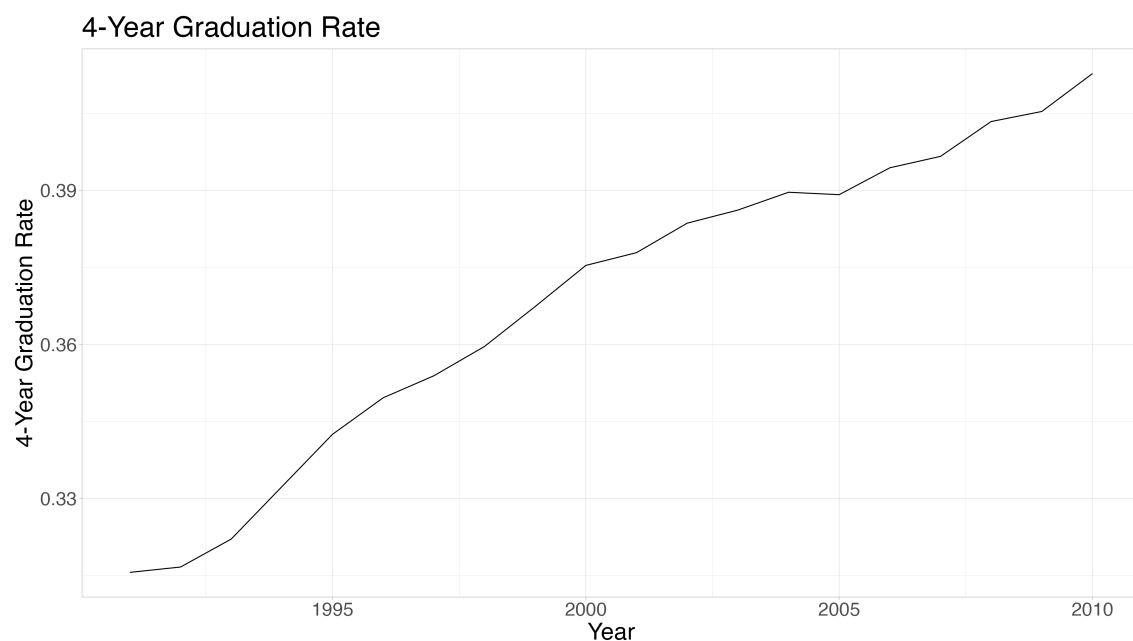
クォーター制からセメスター制への移行を 1991 年から 2005 年までの間に行った大学は調査対象の大学 731 校中 56 校だった。これは元論文の "Switcher"*¹ 76 校と比較してだいぶ少ない数となっているため、R で書いたコードの中に誤りが含まれる、もしくは、そもそもの計算方法に誤りがある

*¹ Switcher ... クォーター制からセメスター制に変更があった大学

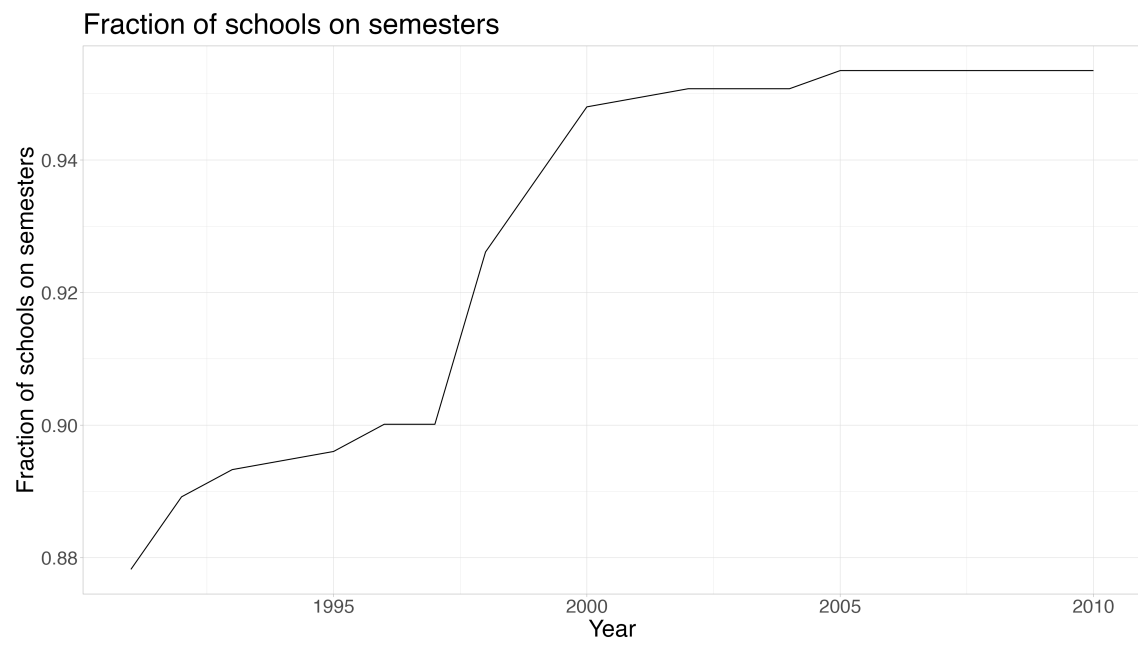
可能性がある。女性の4年卒業率は男性の4年卒業率と比較して高い傾向にあり、これは "Switcher" "Never Switcher" に関わらず共通している。一方で、卒業率や卒業者数を確認すると、4年卒業率は "Switcher" で $27\% \pm 18\%$ 、"Never Switcher" で $38\% \pm 23\%$ となっている。男性についても4年卒業率を確認すると、"Switcher" で $23\% \pm 17\%$ 、"Never Switcher" で $33\% \pm 23\%$ となっており、女性は、"Switcher" で $32\% \pm 19\%$ 、"Never Switcher" で $42\% \pm 23\%$ となっている。従ってセメスター制への移行を決断する大学は、4年卒業率が低い傾向にあると言える。

a-2, 4年卒業率の平均推移を計算し、図で示しなさい

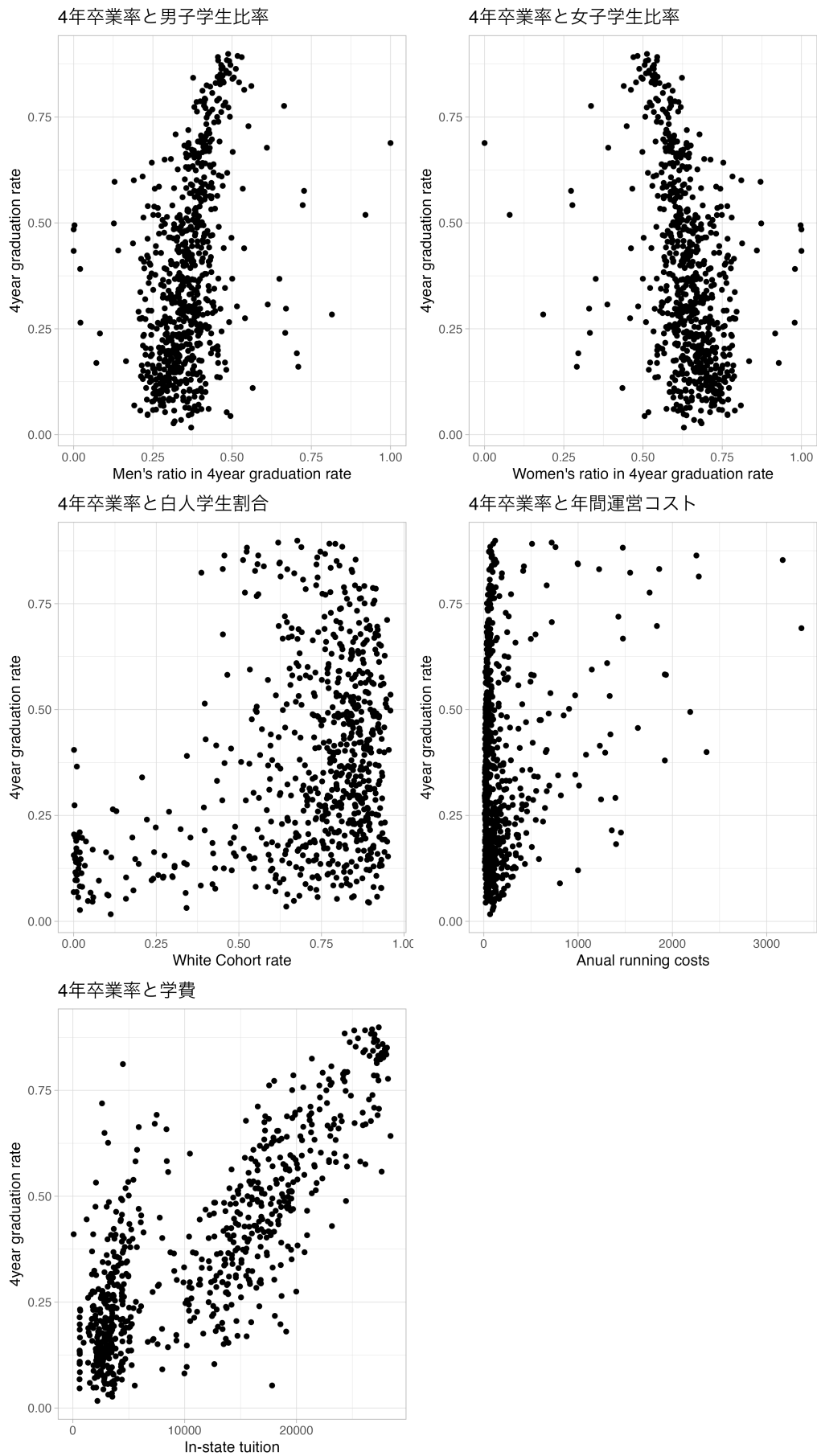
各年ごとにすべての大学の4年卒業率の平均を計算し、図で示すと以下のようになる。



a-3, semester 導入率を計算し、図で示しなさい



a-4, 変数に処理を加えた上で、以下の散布図を作成しなさい。また、重要だと考える結果について議論しなさい



五つの散布図から読み取れるとこで特徴的なのは、学費と4年卒業率に強い相関関係があることだ。これは、学費が高い大学ほどできるだけ早く卒業して費用を抑えたいというインセンティブが学生に対して働くからかもしれないし、学費が高い大学ほど一人一人に対して学業の支援が行き届いているため卒業率が高くなっている可能性もある。

(b) 回帰分析

b-1, yearofsem, yearstosem, treated を作成しなさい

以下のようなコードで、3つの変数を作成した。

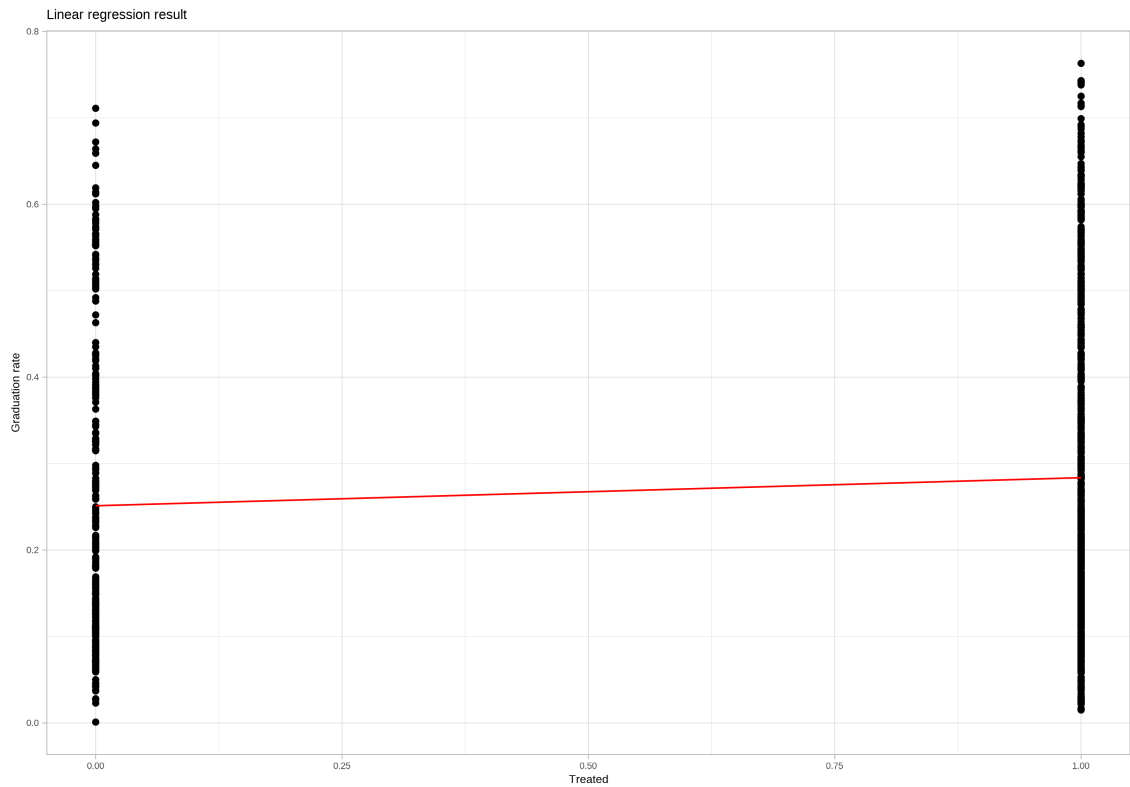
```
1 # 相対パスを取得してデータを読み込む
2 data <- readr::read_csv(basics$get_absolute_path("src/build/table_1/output/intermediate.csv"))
3
4 # Switcher に該当する大学名をリスト化
5 switcher_instnms_list <- data %>%
6   dplyr::distinct(across(-c("instnm", "...1", "year", "totcohortsize",
7     "w_cohortsize", "m_cohortsize", "tot4yrgrads", "m_4yrgrads",
8     "w_4yrgrads", "women_gradrate_4yr", "gradrate4yr", "men_gradrate_4yr",
9     "instatetuition", "costs", "faculty", "white_cohortsize",
10     "per_white_cohort", "per_women_cohort")), .keep_all = TRUE) %>%
11   dplyr::group_by(unitid) %>%
12   dplyr::filter(dplyr::n() > 1) %>% # unitidが重複している行だけを抽出
13   dplyr::ungroup() %>%
14   dplyr::select(instnm) %>%
15   dplyr::distinct()
16
17 # Switcher に該当する大学のデータを抽出
18 switcher_data <- data %>%
19   dplyr::mutate(is_switcher = dplyr::if_else( # is_switcher列を追加
20     # 条件(大学名がswitcher_instnms_listに含まれるか)
21     as.character(instnm) %in% switcher_instnms_list$instnm,
22     TRUE, # 条件が真の場合
23     FALSE)) %>% # 条件が偽の場合
24   dplyr::filter(is_switcher == TRUE) # is_switcher が TRUE の行だけを抽出
25
26 # yearofsemとunitidを対応させたデータを作成
27 yearofsem_unitid <- switcher_data %>%
28   dplyr::filter(semester == 1) %>%
29   dplyr::group_by(unitid) %>%
30   dplyr::summarize(yearofsem = min(year)) %>%
31   dplyr::ungroup()
32
33 # yearofsem, yearstosem, treated を追加
34 switcher_data <- switcher_data %>%
35   dplyr::left_join(yearofsem_unitid, by = "unitid") %>% # yearofsemを追加
36   dplyr::mutate(yearstosem = year - yearofsem) %>% # yearstosemを追加
37   dplyr::mutate(treated = dplyr::if_else(yearstosem < 0, 0, 1)) # treatedを追加
38
```

b-2, 以下の式を推定し、結果について議論しなさい

s は大学, k は相対年数 (yearstosem)

$$Y_{sk} = \beta_0 + \beta_1 \text{treated}_{sk} + \epsilon_{sk}$$

Characteristic	N = 2 ¹
term	
(Intercept)	1 (50%)
treated	1 (50%)
estimate	
0.0324515101437764	1 (50%)
0.251244372990354	1 (50%)
std.error	
0.010312317701766	1 (50%)
0.0122582820608957	1 (50%)
statistic	
2.64731305598667	1 (50%)
24.3635213980392	1 (50%)
p.value	
1.66034091603913e-104	1 (50%)
0.0082335511838119	1 (50%)
¹ n (%)	



上記の結果から以下のような回帰式が得られた。

$$Y_{sk} = 0.2512 + 0.0325\text{treated}_{sk} + \epsilon_{sk}$$

したがって、セメスター制に移行すると、4年卒業率が 3.25%(SD 1.23%) 上昇するという結果が得られた。

b-3, 数式の問題点を指摘しなさい

まず、説明変数が一つと少ないため、その他の変数が目的変数へ与える効果を考慮できていない。さらに、treated がダミー変数であるためセメスター制へ移行した後の効果の推移を考慮できていない。