Warm-up Problemset

PUBLISHED
July 19, 2024

この事前課題では、一つのプロジェクト例を通じて、データ分析に取り組むことで、RAbootcampに必要なスキルを身につけることを目的としている。**必ずしも、全ての問題に取り組む必要はない。** 15時間程度を目安に取り組み、途中でも構わないのでそこまでを提出すること。

1. 問題設定

本課題は以下の論文の一部をレプリケーション（リプロダクション）するものである。
BOSTWICK, Valerie, Stefanie Fischer, and Matthew Lang. (2022). “Semesters or Quarters? The Effect of the Academic Calendar on Postsecondary Student Outcomes.” *American Economic Journal: Economic Policy*

https://www.aeaweb.org/articles?id=10.1257/pol.20190589

背景

アメリカでは、2学期制と4学期制を採択する大学があり、どちらの学期制度が適切か長く議論が続いている。近年は、4学期制から2学期制へ移行する大学が多い。2学期制の方が、(1)学業成績が向上し、(2)夏季インターンシップなどの機会がより充実すると考えられているためだ。しかしながら、このような学期制の変更が学生のアウトカムにどのような影響を及ぼすのか、明確なエビデンスがほとんどない。(4学期制と2学期制の違いについては、宿題Appendix (a)や論文の第2セクション「Background」を参考にすること。)

この議論の背景として、アメリカの大学では入学後4年以内に卒業する学生の割合は50%以下、6年以内卒業者の割合は約60%という実態がある[1]。このような修了率の低さと学位取得までの期間が長いことは、学生に金銭的・時間的な負担を強いている。その意味でも、適切な学期制を検討することが政策課題として重要である。

参考資料として、論文に掲載されている2つの図を示す。Figure.1は、2学期制を採用している大学の割合と4年卒業率の推移である。そして、Figure.2は、4学期制から2学期制への移行が卒業率に与える効果を表してる。

Figure 1: Fraction of Schools on Semesters and Four-Year Graduation Rates

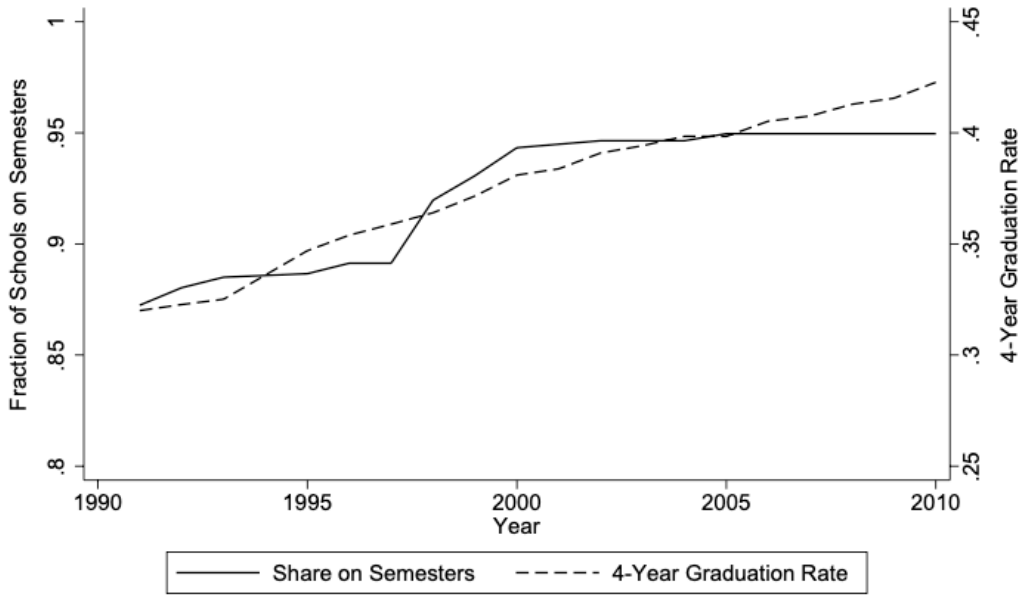
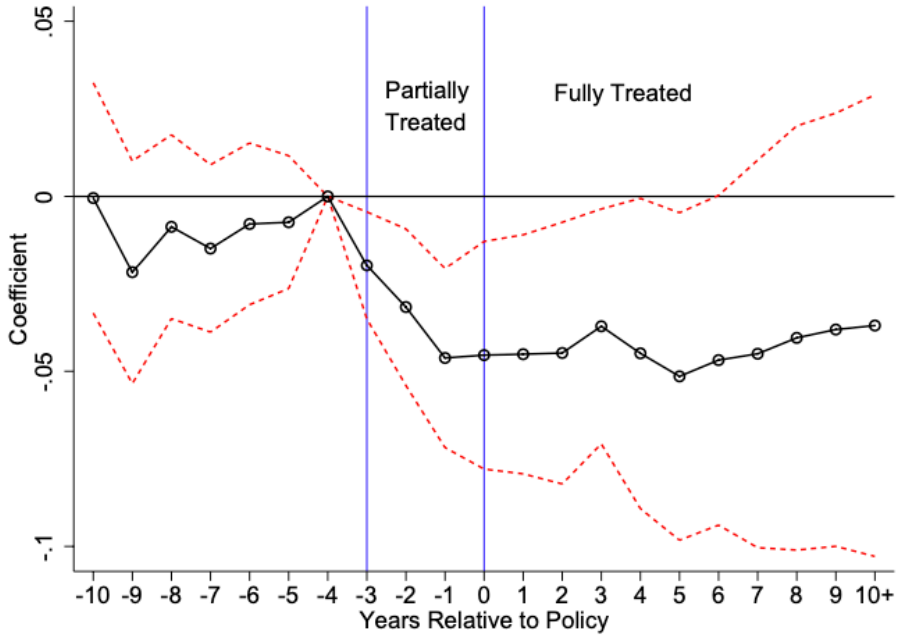


Figure 2: Event Study: Institution-Level Analysis

(a) 4-year Graduation Rates



本課題の問い

4学期制から2学期制への移行は卒業率に影響を及ぼすのか

用いるデータ

- gradrate_data(1991.csv - 2016.csv)
  - 卒業者数などの結果変数に関連するデータが格納されているファイル
  - 各年に分かれている
- covariates.xlsx

    - その他の変数データが格納されているファイル
- semester_data_1.csv, semester_data_2.csv
    - 大学ごとの2学期制と4学期制の分類データが格納されているファイル

論文に使われている元データに関心がある学生はAppendix (b)を参照すること。また、クリーニング後のデータについても配布している。(`clean_covariates`, `clean_outcome`, `clean_semester_dummy`, `master`)

2. 提出について

提出期限

- 2024年8月18日 23:59

提出の流れ

1. コードとアウトプットが含まれるフォルダを、Githubにアップロードする(**public repository**にすること)
2. 提出用google formに(1)氏名と(2)レポジトリURLを記入し送信する(準備中)

- Githubの使い方については、別紙「」を参照すること
- 原則、Github以外の提出は認めない
- レポート形式にまとめる必要はない

提出フォルダ構造

提出フォルダに、`cleaning`, `analysis`の3つのフォルダを作成し、クリーニングや分析をする際には、適切な場所にコードを保存すること。また、Rを用いていれば、コードファイルの形式は問わない(.r, .rmd, .qmdなど)。.Rmdや.qmd形式で書く場合は、必ずしもcleaningとanalysisを分ける必要はない。

Example

- 01_cleaning
    - code_file.r
- 02_analysis
    - code_file.r

3. 課題

データ整理と変換

通常、データは、そのまま分析できる形式で保存されていない。ここでは、様々な問題を抱える生データをどのように整理・変換すればよいかを学ぶ。プログラミングに慣れていない参加者は、参考資料やオフィスアワーを活用してほしい。また、コードの相談やgithubの使い方がわからない場合も、オフィスアワーで質問すること。

- 基本的にtidyverseRを使用して書くことを推奨する
- 可読性や拡張性を意識すること
    - マジック・ナンバーを使わない、適切な変数名をつける、など

(a) Semester Dataの整形

1. 生データを読み込みなさい (semester_dummy_1.csv, semester_dummy_2.csv)

2. semester_dummy_1.csvについては、1行目を列名としなさい

3. 2つのデータを適切に結合しなさい
    - ヒント：型に注意

4. 'Y'列を削除しなさい

5. semester制が導入された年の列を作成しなさい。

6. 5.を用いてsemester制導入後を示すダミー変数を作成しなさい
    - 2001年にsemester制が導入された場合、1991~2000年は0, 2001年以降は1となる変数

(b) Gradrate Dataの整形

1. 生データを読み込み、適切に結合しなさい
    - ヒント：'for'や'purrr::map'を参照

2. 女子学生の4年卒業率に0.01をかけて、0から1のスケールに変更しなさい

3. 男女合計の4年卒業率と男子学生の4年卒業率を計算し、新たな列として追加しなさい
    - ヒント：型に注意

4. 計算した卒業率を有効数字3桁に調整しなさい

5. 1991年から2010年までのデータフレームに変形しなさい

(c) Covariates Dataの整形

1. 生データを読み込みなさい (covariates.xlsx)

2. 'university_id'という列名を'unitid'に変更しなさい

3. 'unitid'に含まれる"aaaa"という文字を削除しなさい

- ヒント：stringr

4. ‘category’列に含まれる’instatetuition’, ‘costs’, ’faculty’, ’white_cohortsize’を別の列として追加しなさい(wide型に変更しなさい)
    - ヒント：pivot_wider
5. outcomeやsemester_dummyに含まれる年を調べ、covariatesデータの期間を他のデータに揃えなさい
6. outcome_dataに含まれるunitidを特定し、covariatesに含まれるunitidをoutcomeデータに揃えなさい

(d) Master Dataの作成

1. 結合に用いる変数を考え、semester_data, covariates_data, gradrate_dataを適切に結合しなさい
    - ヒント：left_join

分析

(a) 記述統計

1. 「(d) Master Dataの作成」で作成したデータの、各列に含まれるNAの数を数えなさい。
2. 問題背景などを知る上で役に立つ記述統計を作成しなさい

- 参考：
  論文Table 1

3. 4年卒業率の平均推移を計算し、図で示しなさい

- 参考：
  論文Figure 1

4. semester導入率を計算し、図で示しなさい

- 参考：
  論文Figure 1

注意：問2, 問3のプロットは別々の図として作成することを推奨する

5. 以下の3つの変数を横軸、「4年卒業率」を縦軸にとった、散布図を作成しなさい。
    1. 女子学生比率
    2. 白人学生割合
    3. 学費(instatetuition)

- 作成の際には関数を作成することを強く推奨する
- ヒント:
  “rlang” package (enquo関数, sym関数)
  使い方については、以下のHPが参考になる
  https://www.tidyverse.org/blog/2018/07/ggplot2-tidy-evaluation/#tidy-facets-with-vars
  https://ggplot2.tidyverse.org/reference/tidyeval.html

(b) 回帰分析

1. 以下の式を推定し、表にまとめなさい。

- s：大学、t：年、Y_{st}：大学s, 年tの4年卒業率、$After_{st}$：大学s, 年tの時の、semester制導入のダミー変数

$$Y_{st} = \beta_0 + \beta_1 \text{After}_{st} + \varepsilon_{st} \tag{1}$$

4. Appendix

(a) 2学期制（Semester）と4学期制(Quarter)の違い

- 詳細は論文の第2セクション 「Background」を参照

	2学期制	**4学期制**
学事歴	8月下旬 - 5月上旬	9月下旬 - 6月下旬
授業期間	約15週間	約10週間
履修科目数（1学期）	5科目程度	3科目 - 4科目
メリット	授業期間が長いのでより難しい内容まで学べる。 夏季インターンシップの機会が多い	多くの科目を履修できる。 学期が細かく分かれているため、専攻を変えやすい。*
デメリット	試験勉強を先延ばしにする	インターンの参加や留学時期が合わせにくい。

* 約半数の学生が専攻を変更する背景がある

(b) 論文に使用されたデータソース

本課題に取り組む際はLMSのデータを使用すること
ダウンロードには会員登録が必要である。

URL

https://www.openicpsr.org/openicpsr/project/124861/version/V1/view

5. 参考資料

R

私たちのR
R for Data Science (2e)

Rで計量政治学入門

Advanced R(上級者向け)

cheat sheet

Posit Cheatsheets

レポート関連

Quarto[2]
Overleaf

因果推論

Causal Inference The Mixtape

Footnotes

1. 日本は約89% (令和4年度学校基本調査：最低修業年数卒業者 / 卒業者計)↩
2. 課題資料はQuartoで作成している↩