Problem Set 1. データ分析プログラミングの実践(1)

PUBLISHED
November 21, 2023

プログラミングを学ぶために、実際に手を動かすことが欠かせない。この課題は、一つのプロジェク ト例を通じて、以下の一連の作業に実際に取り組むことを目標としている。
また、本課題は2部構成であり、1部はデータ整形、2部は分析/レポート執筆となっている。

1. 生データの読み込み
2. 変数型の確認と変換
3. 中間データの保存と読み込み
4. 変数の欠損値の確認と対応
5. 変数の値の変換
6. 複数のデータセットの結合
7. 回帰分析
8. 図表の作成と保存
9. 文書の作成

- この課題のために必要なデータとコードの型は、以下のGitHub remote repositoryにある
  link : 未設定
- また、この課題を完成させるために、コードの例として、以下のプロジェクトを参考にしてほしい
  link : https://github.com/Chishio318/Peanuts-Data-Project

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*

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

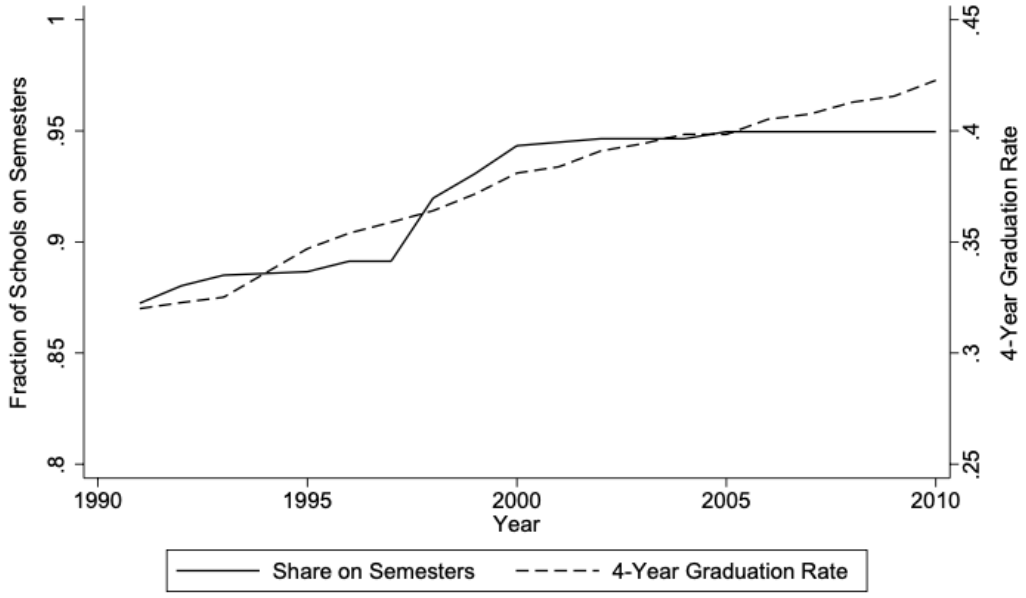
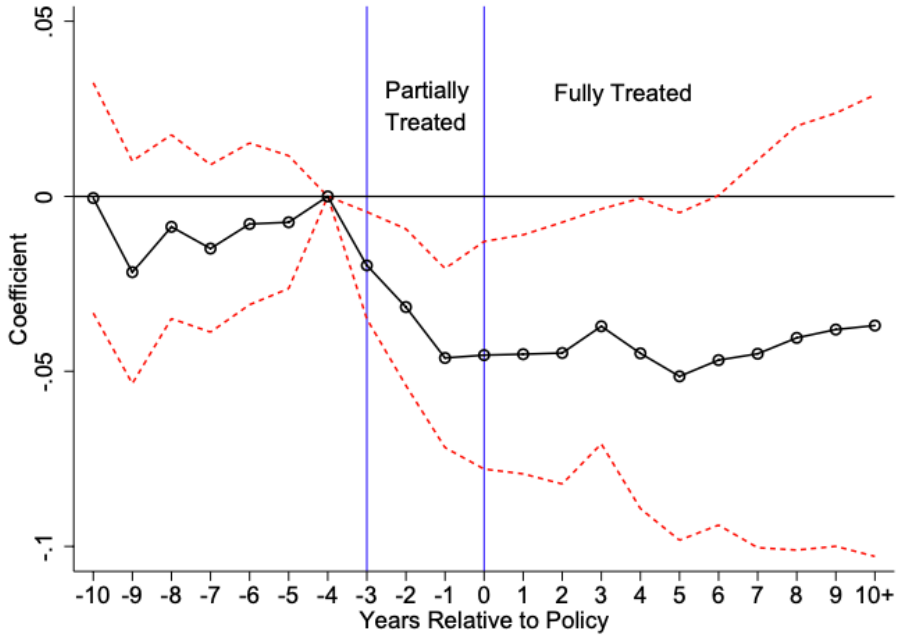


Figure 2: Event Study: Institution-Level Analysis

(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。

2. 提出課題

1. データ整形過程のコードとアウトプットを提出すること
    - Giithubレポジトリの共有を推奨
2. 分析過程のコードを書き、合同・個別オフィスアワーの時間を活用して、古川（伊藤、吹原）から添削・フィードバックを受けること
    - 実際にコード全体がスムーズに実行され、結果を出すことを示すこと

提出期限

- 2023年12月19日（火曜）

評価基準

コードについて

- tidyverseRを使用して書いているか
- エラーなく実行できるか
- 読みやすいか、拡張しやすいか
    - マジック・ナンバーなどを使わず、説明がされているか
    - 適切な変数名をつけているか
    - わかりやすいコメントが書かれているか（任意）

3. 分析課題

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。なお、以下のステップは「指針」であり、コード については、参考資料やPeanuts Data Projectなどを参考にすること。

(a) semester_dummy_tidy[難易度2]

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

(b) gradrate_tidy [難易度3]

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

(c) covariates_tidy [難易度3]

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'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

(d) gradrate_ready [難易度1]

1. 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

2. 計算した卒業率を有効数字3桁に調整しなさい

3. 卒業率に欠損値が含まれている行を削除しなさい

(e) covariates_ready [難易度2]

1. 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

2. 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

(f) master [難易度2]

1. 結合に用いる変数を考え、semester_dummy_tidy, covariates_ready, gradrate_readyを適切に結合しなさい
    - ヒント：left_joinなど

取り組むにあたって気をつけてほしいこと

- コードを調べるときは、英語のリソースを探すことを強く勧める
- プログラミングの課題に一緒に取り組む仲間も、大切なリソースである。まず自力で取り組み、その後、お互いに助言をし合うことを勧める

5. 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

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

6. 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

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↩