Problem Set 1. データ分析プログラミングの実践

AUTHOR
Ito Hiroki

PUBLISHED
May 22, 2021

- 中間報告？

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。

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

- 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"Semesters or Quarters? The Effect of the Academic Calendar on Postsecondary Student Outcomes"(Bostwick et al.)[1]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。

背景

アメリカの大学の多くは2学期制を採用しており、近年は4学期制から2学期制への移行をしている大学が増えている。主な理由は、学力を向上させ、夏季インターンシップの機会を増やすためだと主張しているが、学事歴の効果に明確なエビデンスはない。
また、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半数以下であり、このような修了率の低さは、学生に直接的・間接的な負担を強いることから、卒業率は政策担当者の関心事項となっている。 4学期制と2学期制の違いについては、以下を参照すること。

本課題の問い：

4学期制から2学期制への制度変更が、学生の卒業率に影響を及ぼすのか

参考資料

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	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	科目ごとの期間が長いのでより深くまで学べる。 夏季インターンシップの機会が多い	多くの科目を選択できる。 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	試験勉強の先延ばしをしてしまう	インターンの参加や留学時期が合わせにくい。

＊：約半数が専攻を変更する。

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
  - 各種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
  - 各大学のsemesterとquarterのダミーデータが格納されているファイル。

2. 提出課題

評価基準

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行目に格納されているデータを列名としてください。

3. 2つのデータを縦に結合してください。
    - ヒント:型に注意

(b) gradrate_tidy [難易度3]

1. 生データを読み込み、全て縦に結合してください。
    - ヒント：データは各年に分かれているため、'for'や 'purrr::map'を用いると良い

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てください。

(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'の最後についている"aaaa"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)
    - ヒント：pivot_wider

(d) gradrate_ready [難易度1]

1. 4年卒業率、男女合計と男子学生を計算し、有効数字3桁表示にしてください。
    - ヒント：型に注意

2. gradrateが欠損値の値を削除してください。

(e) covariates_ready [難易度2]

1.

(f) master [難易度2]

1. キーとなる変数を考え、結合してください。semester_dummy_tidy, covariates_tidy, gradrate_readyを結合しなさい。

データの分析

(a) 記述統計の作成

1. 問題背景などを知る上で役に立つ記述統計を表として作成しなさい

図の作成

1. 卒業率の平均推移をプロットしてください。

    1. ヒント：summarize関数を使い、.byの引数にyearを入れる

2. semester導入率の推移をプロットしてください。

3. 各種変数と4年卒業率の散布図を作成してください。（Advanced）作成する際に、unitidで平均を算出した数値を使用してください。

    1. ヒント：ggplot2::facet_wrap等を参照。

(b)回帰分析

1. 以下の式を推定してください。 回帰式の段階から、学生に考えてもらう。

$$Y_{it} = \beta_0 + \beta_1 Semester_{it} + \varepsilon_{it}$$

1. Control変数も加えた状態で推定しなさい。また、結果について上記の式と比較して議論しなさい。

$$Y_{it} = \beta_0 + \beta_1 Semester_{it} + \gamma X_{it} + \varepsilon_{it}$$

(adv.) Difference-in-Difference

$$Y_{it} = \beta_0 + \beta_1 Treatment_{it} + \beta_2 After_{it} + \beta_3 Treatment_{it} \times After_{it} + \varepsilon_{it}$$

$$Y_{it} = \gamma X_{it} + \beta_1 D_{it} + unit_i + year_t + \varepsilon_{it}$$

Footnotes

1. American Economic Journal:Economic Policy 2022年2月号掲載↩