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

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

背景

アメリカの大学の多くは2学期制を採用しており、近年では4学期制から2学期制への移行を進める大学が増加している。主な移行の理由としては、学業成績の向上と夏季インターンシップの機会増加が主張されている。しかし、学事歴の変更が、それらの指標にどの程度の影響を及ぼすのかついては、明確なエビデンスがほとんどない。

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50%以下、6年以内卒業者の割合は約60%という実態がある。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、学生に直接的・間接的な負担を強いる原因となっている。よって、学事歴の変更が卒業率や学業成績に与える影響は、政策担当者にとっても関心事項である。

4学期制と2学期制の違いについては、Appendixと論文の第2セクション「Background」を参考にすること。

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。

2. 提出課題

1. データ整形と記述統計に基づくレポートを提出すること
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

評価基準

レポートについて

- 変数の欠損値などの問題についてどう対応したかを議論しているか
- 変数に異常値などがないかを議論しているか
- 図表が見やすいか

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

(d) gradrate_ready [難易度1]

1. 男女合計と男子学生の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. '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

(f) master [難易度2]

1. キーとなる変数を考え、semester_dummy_tidy, covariates_ready, gradrate_readyを適切に結合しなさい。

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
    - ヒント：summrize, ggplot2, kableExtra

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

Footnotes

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