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.

本課題のデータ分析の問い:

用いるデータ:

- gradrate_data.xlsx
  - 卒業率等が格納されているデータ
- covariates.xlsx
  - 各種変数が格納されているデータ
- semester_data_1.csv, semester_data_2.csv
  - 各大学のsemesterとquarterのダミーが格納されているデータ

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。

(b) gradrate_tidy [難易度2]

1. 生データを読み込みなさい (grad_rate.xlsx)
2. 女子学生の4年卒業率に0.01をかけて、0から1のスケールにし変えてください。
3. long型のデータに変更してください。
    1. ヒント：（tidyr::pivot_longer）

(c) covariates_tidy [難易度2]

1. 生データを読み込みなさい (covariates.xlsx) い
2. 'university_id'の最後についている"000"を削除しなさい。
    1. ヒント：stringr
3. 'university_id'という列名を'unitid'に変更しなさい

(d) gradrate_ready [難易度2]

1. 4年卒業率、男女合計と男子学生を計算しなさい。
    1. ヒント：(型に注意)
2. 1991 - 2010のみを抽出しなさい

(e) master [難易度2]

1. unitidをキーとして,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やtable()を参照。

(b)回帰分析

1. 以下の式を推定してください。

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

2. 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

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

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