Домашнее задание №2

Практическая эконометрика

11 октября 2018 г.

1 Общие требования

Задание выполняется 1-2 студентами, сдаётся на онэкон в виде архива с 3 файлами - оформленного связного текста с ответами на вопросы, таблицами и графиками (если они необходимы по заданию) в формате pdf, файла с кодом в R и файла с именами студентов в формате txt. Они должны называться HA2text, HA2code и authors. Часть вопросов факультативные, ответы на них оцениваются бонусными баллами (которые выходят за пределы 60 баллов на домашки). На on.econ вывешен текст статьи, папка с данными и шаблон, указывающий, в каком формате вам надо сдать код. Пожалуйста, пользуйтесь этим шаблоном. В случае несоответствия результата запрошенному формату пункты можем не зачитывать (иначе они протребуют отдельной ручной проверки)

2 срока сдчи задания: 14 октября 23:59 - промежуточный дедлайн, 21 октября 23:59 - окончатльный. К промежуточному дедлайну нужно сдать по крайней мере 2 блока вопросов из 4. После промежуточного дедлайна можно будет досдать не больше 2 разделов

Если в процессе у вас возникли сложности, не стесняйтесь спросить помощи у товарища (или у товарища семинариста). Вообще, помощи у товарища не надо стесняться просить. Но если вы будете злоупотреблять этой помощью, у нас есть способы об этом узнать (способы по очевидным причинам не разглашаются).

2 Задание

В рамках этой и следующей домашней работы мы попробуем целиком реплицировать результаты статьи Vincent Pons "Will a Five-Minute Discussion Change Your Mind? A Countrywide Experiment on Voter Choice in France"American Economic Review 2018, 108(6): 1322–1363

На Гугл диске лежат данные. Данные содержат непосредственно данные (Data) и Stata код (Do-files). Мы будет реплицировать с помощью R, но Stata код полезно почитать, если вы застряли.

Единственные 2 файла с данными, которые вам понадобятся: analysis и intermediate/ base_randomization_v2

Установите в R пакет `readstata13`, чтобы прочитать набор данных. В некоторых пунктах указаны другие рекомендованные пакеты. Испольуйте их или любые другие пакеты R на ваше усмотрение.

В рамках этой домашней работы вы можете целиком пропустить подпункты B, C, D раздела II. Также можно пропускать все Panel B в табличках с результатами.

Вопросы этой работы разбиты на 4 логических блока, которые содержат 2 типа вопросов: поработать с данными или ответить на содержательный вопрос. Напротив кажого пункта есть пояснения о том, как он будет оцениваться

2.1 Общий обзор статьи

(2 балла + 1 бонус)

1. Какой исследовательский вопрос интересует авторов? Что именно они хотят измерить? (1 балл за 1-2 предложения)
2. Почему этот вопрос интересен с точки зрения авторов? А именно: какой вклад он вносит в своей области (политические предпочтения)? С какими другими исследовательскими вопросами связано это исследование? (1 балл за описание по пунктам)
3. Знаете ли вы, почему ещё этот вопрос может быть интересен (со ссылками на литературу)? (1 бонусный балл в случае интересной и релевантной ссылки на литературу)

2.2 Рандомизация

(5 баллов + 2 бонуса)

В этом пункте мы попытаемся реплицировать рандомизацию из статьи

1. С помощью функции `read.dta13` импортируйте файл `base_randomization_v2.dta`

2. Сделайте функцию `generate_stratum`, генерируюющую на данных номер `stratum`. Для этого внутри групп с общими `territory` и `department_code` отсортируйте данные по убыванию переменной с называнием `PO_name` (для избирательных участков это будет `prop_leftabstention`, а для муниципалитетов `prop_leftabstention_mun`). Присвойте каждому блоку из 5 строчек одно и то же целое число Вам может помочь пакет `dplyr` и какие-то из следующих функций: `order`, `seq`, `ave`, `seq_along`, `group_by`, `mutate`. Документация: `https://dplyr.tidyverse.org/` (1 балл. Условие пункта изменилось. К сдаче принимается старый вариант, если вы его уже сделали)

3. Сделайте функцию `generate_treatment`, которая для заданной колонки `stratum` создает колонку `treatment` таким образом, как это описано в статье на страницах 1335-1336. Если вы чувствуете, что не справляетесь с этим пунктом, просто назначайте treatment = 1 с вероятностью 80% (1 балл в случае просто 80% + 1 бонусный балл в случае «честной» репликации)

4. Сделайте функцию `allocate_canvassers`, которая имея колонки `stratum` и `treatment`, отталкиваясь от количества зарегистрированных граждан на этом избирательном участке (`nb_registered_prim`) и целевых чисел по объему эксперимента (`target_ter`), возвращает `data.frame` с колонками (1 бонусный балл)

    - `stratum` в неизменном виде
    - `treatment` заменяет на `NA` если эта stratum не должна войти в эксперимент
    - создает колонку `allocated`, которая равна 1, если по правилам эксперимента туда следует отправить агитаторов

5. Разберитесь, как, воспользовавшись этими функциями и колонкой `level_randomization` воспроизвести рандомизацию. В качестве результата вы должны получить функцию, которая берет на вход исходный `data.frame` и возвращает его же с дополнительными колонками: `treatment`, `allocated`. Выбросите строчки с пропущенным `treatment`. (у вас может получиться разный результат в зависимости от того, целиком вы сделали предыдущие пункты или нет. Вы получите 1 балл за любой результат, который даст хороший balance on covariates и использует правильным способом переменную `level_randomization`).

6. Почему авторы делают такую сложную рандомизацию, а не попросту назначают в treatment группу избирательный участок с фиксированной вероятностью? Выпишите, какую цель они преследуют и отметье все детали в этом способе рандомизировать, которые помогают достичь этой цели (по пунктам). (1 балл)

7. Что необычного отмечают сами авторы статьи в своём способе рандомизировать? (выпишите в виде пунктов) (1 балл)

2.3 Balance on covariates

(3 балла + 1 бонус)

В этом пункте вы будете реплицировать таблицу Table 2—Summary Statistics.

1. С помощью функции `read.dta13` импортируйте `analysis.dta`
2. Постройте таблицу с описательной статистикой. Сгруппируйте данные по `treatment` и внутри группы посчитайте средние. В качестве результата вы должны иметь функцию `summary_table`, которая принимает на вход исходный `data.frame` и производит `data.frame` с колонками `treatment`, `control` и названиями переменных в `row.names`. Вам может пригодиться пакет `dplyr` (функции `group_by`, `t`, `summarise` or `summarise_all`). Вам поможет этот материал: `https://dplyr.tidyverse.org/` (2 балла)
3. Посчитайте тем же способом стандартные отклонения и проведите t тест. В качестве результата вы должны иметь функцию `balance_on_covariates`, которая принимает на вход исходный `data.frame` и производит `data.frame` с колонкой `t_test` с посчитанной t статистикой и названиями переменных в `row.names`. (1 бонусный балл)
4. С какой целью авторы приводят эту таблицу? (1 балл)

2.4 Результаты

(5 баллов + 3 бонуса)

1. С помощью функции `read.dta13` импортируйте `analysis.dta`
2. Отфильтруйте данные по `territory_in == 1`.
3. Зачем нужна эта фильтрация? (1 бонусный балл)
4. Создайте функцию `make_models`, которая принимает на вход название переменной y, название переменной, отвечающей за предыдущие результаты, вектор названий прочих контрольных переменных и данные. Возвращает 3 оцененные модели. Вам могут пригодиться функции update, reformulate. (1 балл)
5. С помощью пакета stargazer и функции `make_models` реплицируйте таблицы с 4 по 10 (panel A). Вы можете использовать параметр `type='html'` для экспорта в word или `type='latex'` для экспорта в Latex. Переменная `out` содержит имя файла, в который необходимо записать результат. Со остальными параметрами разберитесь

сами по мере необходимости. Ваш код должен в точности воспроизводить те таблицы, которые содержатся в вашей работе (править руками их нельзя). (2 балла если из таблиц можно получить коэффициент при treatment для каждой модификации модели + 1 бонусный балл в случае удачного оформления: в таблице нет ничего лишнего, строчки имеют человекочитаемые названия и из каждой колонки понятно, о какой модели идет речь)

6. Какой из полученных вами результатов (какой коэффициент в какой таблице) отвечает на главный исследовательский вопрос? Проинтерпретируйте его. (0.5 баллов)

7. Объясните, почему несмотря на то, что авторы провели эксперимент, они включают контрольные переменные? Сравните оценки коэффициентов и их стандарные ошибки.Почему авторы называют оценку within estimator? Как бы они оценивали не within estimator? Зачем делать именно within estimator? (0.5 баллов)

8. В регрессиях в качестве контрольных переменных используются резльтаты прошлых выборов. Является ли это "плохим контролем"? (0.5 баллов)

9. Зачем нужен плацебо тест, если авторы уже проверяли balance on covariates? (0.5 баллов)

10. Можете ли вы сами задать ещё какие-нибудь содержательные вопросы к результатам? Может авторы в каком-то месте неверно интерпретируют коэффициент. Может на их данных можно еще что-то важное посчитать? (1 бонусный балл в случае интересного вопроса)

11. Ваши стандартные ошибки отличаются от тех, которые были в статье. Почему? Сделайте так, чтобы стандартные ошибки совпадали? (Бонусный вопрос. Не оценивается вовсе)