Практическая эконометрика. Лекция 4. Снижение дисперсии оценки: CUPED, контрольные переменные, пре- и постстратификация

преподаватели 2023: Ольга Сучкова, Алексей Замниус, Анна Ставнийчук. При поддержке Георгия Калашнова

28 сентября 2023 г.

Table of Contents

Зачем нужны контрольные переменные?

чтобы

- Снизить дисперсию. Почему так происходит?
- Потому что **ковариаты** «съедают» часть дисперсии

Что еще могут контрольные переменные:

- Получить **гетерогенные эффекты**. **Предикторы**
- Попробовать проанализировать неэкспериментальные данные **Смесители (confounders)**
- Разобраться в эффекте **Медиаторы**
- Испортить несмещенность **Bad Control**

Сегодня мы обо всем этом коротко поговорим, про большую часть эффектов поговорим позже

Получить гетерогенные эффекты

- Помните исследование про переселение в Чикаго?

Плохой контроль[1]

Примеры:

- Контроль на место работы при исследовании влияния образования на доходы

схема

Интуиция:

- Люди, которые получили white collar job без образования сами по себе крутые – sample bias

[1] Angrist и Pischke 2008, Глава 3.2.3

Плохой контроль

Примеры:

- Контролировать на явку в эксперименте с выборами
- Контроль на место работы при исследовании влияния образования на доходы

Формально:

$$(Y_1, Y_0, X) \perp T \text{ — не выполнено}$$

$$(Y_1, Y_0, X_1, X_0) \perp T \text{ — предположим это}$$

$$E(Y|X = 1T = 1) - E(Y|X = 1T = 0) =$$

$$\mathrm{E}(\mathrm{Y}_1|X_1 = 1) - E(Y_0|X_0 = 1T = 0) =$$

$$\mathrm{E}(\mathrm{Y}_1|X_1 = 1) - E(Y_0|X_0 = 1) =$$

$$\mathrm{E}(\mathrm{Y}_1 - Y_0|X_1 = 1) + (E(Y_0|X_1 = 1) - E(Y_0|X_0 = 1)) =$$

LATE + Смещение выборки

Backdoor-критерий (J. Pearl)

Для упорядоченной пары переменных (T, Y) в ориентированном ациклическом графе G, набор переменных X удовлетворяет критерию backdoor относительно пары (T, Y), если ни один узел из X не является «потомком» от T, и X блокирует все пути между T и Y, которые содержат стрелку, входящую в узел T.

Backdoor-критерий (J. Pearl)

- Post-treatment не удовлетворяет этому критерию. Это «плохие контрольные переменные»
- Pre-treatment - под вопросом: зависит от структуры графа G
- Вывод: перед оценкой регрессий нарисуйте схему взаимодействия между показателями

Table of Contents

Стратификация эксперимента на Французских выборах (Pons 2018)

Precinct ID	# reg. citizens	PO
1	1033	0.103
2	918	0.083
3	1175	0.093
4	1184	0.103
5	1158	0.119
6	854	0.082
7	963	0.092
8	876	0.097
9	962	0.117
10	961	0.121
11	997	0.067
12	907	0.087
13	971	0.067
14	1246	0.120
15	1098	0.096
16	1021	0.104
17	1218	0.076

→

Precinct ID	# reg. citizens	PO	**Stratum**
10	961	0.121	**1**
14	1246	0.120	**1**
5	1158	0.119	**1**
9	962	0.117	**1**
16	1021	0.104	**1**
1	1033	0.103	**2**
4	1184	0.103	**2**
8	876	0.097	**2**
15	1098	0.096	**2**
3	1175	0.093	**2**
7	963	0.092	**3**
12	907	0.087	**3**
2	918	0.083	**3**
6	854	0.082	**3**
17	1218	0.076	**3**
13	971	0.067	**4**
11	997	0.067	**4**

groups.

* 4 precincts are assigned to Treatment
* 1 precinct is assigned to Control.

Precinct ID	# reg. citizens	PO	Stratum	**Treatment**
10	961	0.121	1	**1**
14	1246	0.120	1	**0**
5	1158	0.119	1	**1**
9	962	0.117	1	**1**
16	1021	0.104	1	**1**

Second stratum an

Престратификация treatment переменной

Почему бы заранее не убедиться в том, что по всем переменным у нас баланс

1. Разбить пространство на «бины» (bins), например, по возрастанию признака X
2. В каждом «бине» по 2 наблюдения. Случайным образом 1 из них направляем в тритмент-группу, второе в контрольную
3. Внутри каждого «бина» идеальный баланс по X
4. Итог: по сравнению с обычной рандомизацией дисперсия оценки ниже, плюс гарантируется баланс по переменной X

Контролирование и CUPED

- Можно «просто» включить в регрессию контрольные переменные.
- Аналитики пользуются CUPED - Controlled Experiments Using Pre-Existing Data (сотрудники Microsoft, 2013), реально это Residualized Outcome Regressions (в основе - теорема Фриша-Ву-Ловелла)
- См пример применения Netflix, 2016
- Вспомним формулу MDE. Можно достичь нужной точности при меньшем объёме выборки, т.е.ускорить АБ (см Демешев, 2021).

Обозначения (статья о Нетфликс)

- Y - метрика (зависимая переменная, напр. часы просмотра Нетфликс в месяц)
- $\mu = E(Y)$, $\sigma^2 = Var(Y)$
- Пользователей можно разбить на K страт по переменной X. Среднее в страте μ_k, дисперсия σ_k^2, численность страты в выборке n_k, так что $\sum_{k=1}^{K} n_k = N$
- p_k - доля людей из k-той страты в генеральной совокупности
- Y_{kj} - метрика j-го человека из k-той страты
- $\overline{Y} = \frac{1}{N} \sum_{k=1}^{K} \sum_{j=1}^{n_k} Y_{kj}$ - обычное среднее
- $\widehat{Y}_{strat} = \sum_{k=1}^{K} p_k \overline{Y_k}$ - среднее при стратификации, где $\overline{Y_k} = \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{kj}$

Покажем, что дисперсия при стратификации снижается

Содержательные результаты

- Дисперсия обычной оценки - это сумма внутригрупповой и межгрупповой дисперсии. Пре-стратификация убирает межгрупповую дисперсию.
- О пре-стратификации надо думать заранее, до АБ-теста.
- Пост-стратификация снижает дисперсию, так как «исправляет» выборку

Что надо запомнить

- Метрику (outcome variable) лучше выбирать заранее
- Можно продумать дизайн эксперимента заранее и сделать престратификацию
- Надо проверять баланс контрольных переменных в тритмент- и контрольной группе (balance on covariates)
- В качестве контрольных переменных нельзя брать post-treatment