

Ортогонализация

Дмитрий Колодезев
Reliable ML 2025.09.28 v1

ReliableML



Дмитрий
Колодезев



Ирина
Голощапова



Курс
по Causal Inference

https://t.me/reliable_ml

Датафесты и митапы



Reliable ML AB Testing & Causal Inference Meetup



Interpretable & Causal ML



Causal Inference in ML



Reliable ML



плейлист

Все написано до нас

- Causal Inference for The Brave and True
- Introduction to Causal Inference
- CausalML Book
- Causal Artificial Intelligence
- Дружелюбная эконометрика
- Шпаргалка по причинно-следственному анализу

Крупным планом

- Ортогонализация
- Difference in Difference и синтетический контроль
- Матчинг
- Мета-лернеры
- Хороший и плохой контроль
- CATE
- Инструментальные переменные
- Пропущенные переменные
- Causal Discovery
- Обзор библиотек

Исследуем эффект воздействия

- Воздействие (threatment)
- Популяция и выборка
- Конфаундер
- Эндогенность
- Экзогенность
- Систематическая ошибка

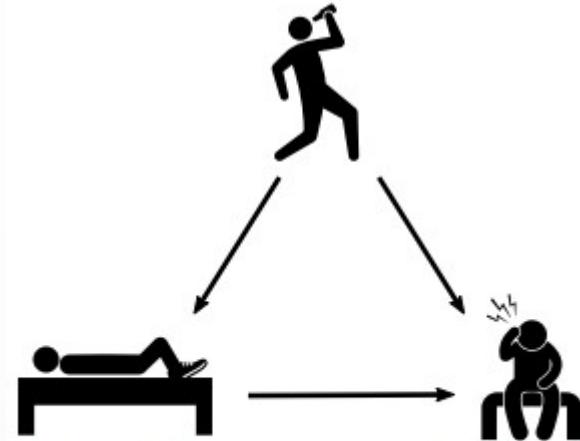
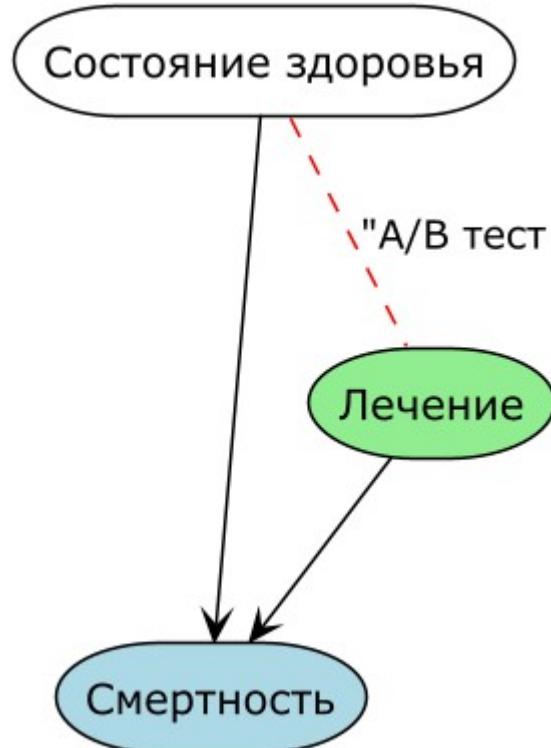
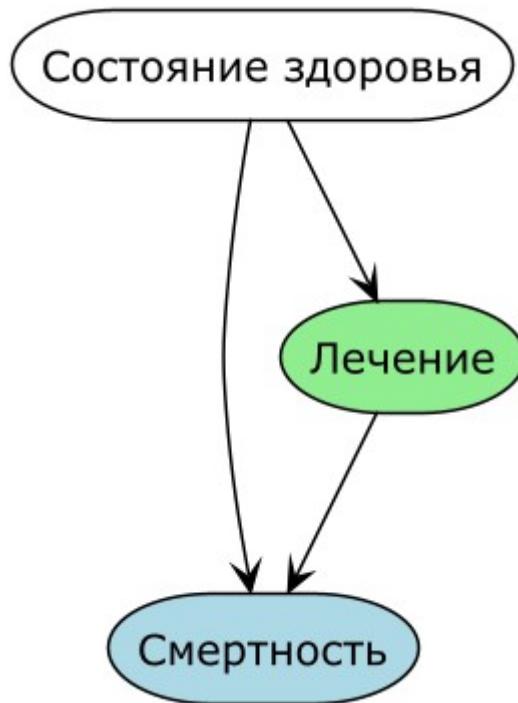


Figure 1.4: Causal structure, where drinking the night before is a common cause of sleeping with shoes on and of waking up with a headaches.

Удивительные открытия

- Те, кого лечат, умирают чаще (к сожалению)
- Повышение цен ведет к росту продаж (иногда)
- Чем больше пожарных, тем больше ущерб
- Преступность больше там где патрулируют
- Мокрая земля вызывает дождь (в каком-то смысле)
- Алкоголь в малых дозах полезен (нет)
- Курильщики чаще доживают до 100 (**на бумаге**)

Давайте лечить не всегда

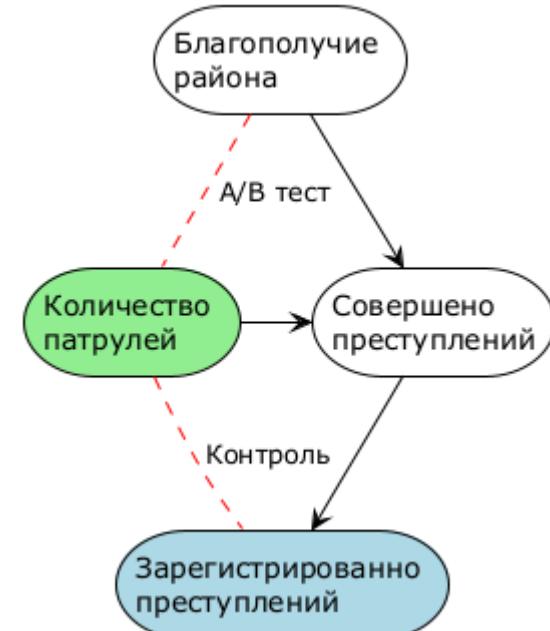
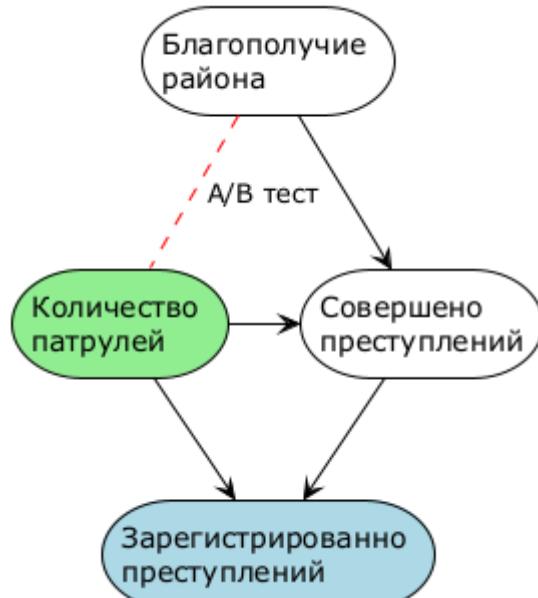
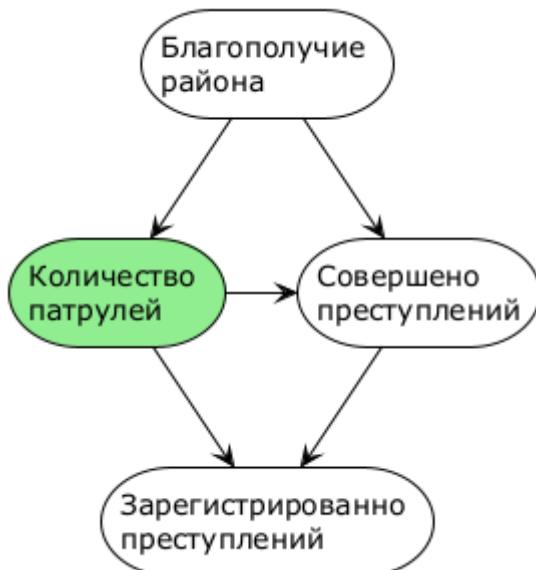


Ученые рекомендуют

Департаменту полиции следует определять количество полицейских в ряде районов случайным образом, то есть независимо от наблюдаемого там уровня преступности. В этом случае обратная причинно-следственная связь в указанных районах пропадет, что позволит состоятельно оценить коэффициент β_2 при помощи обычного МНК.

Еще пол можно случайно менять, для точности

Патрулируй случайно

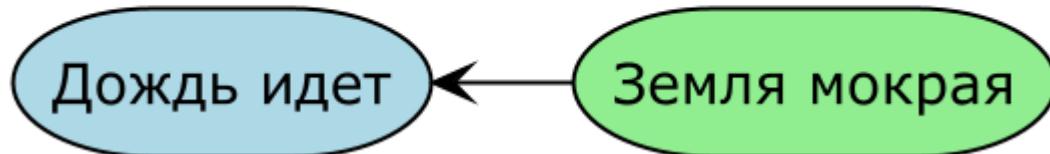
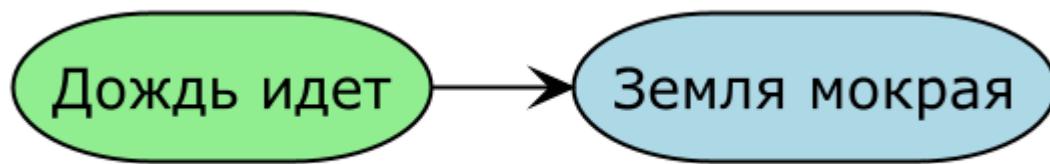


Но стрелка не в ту сторону!

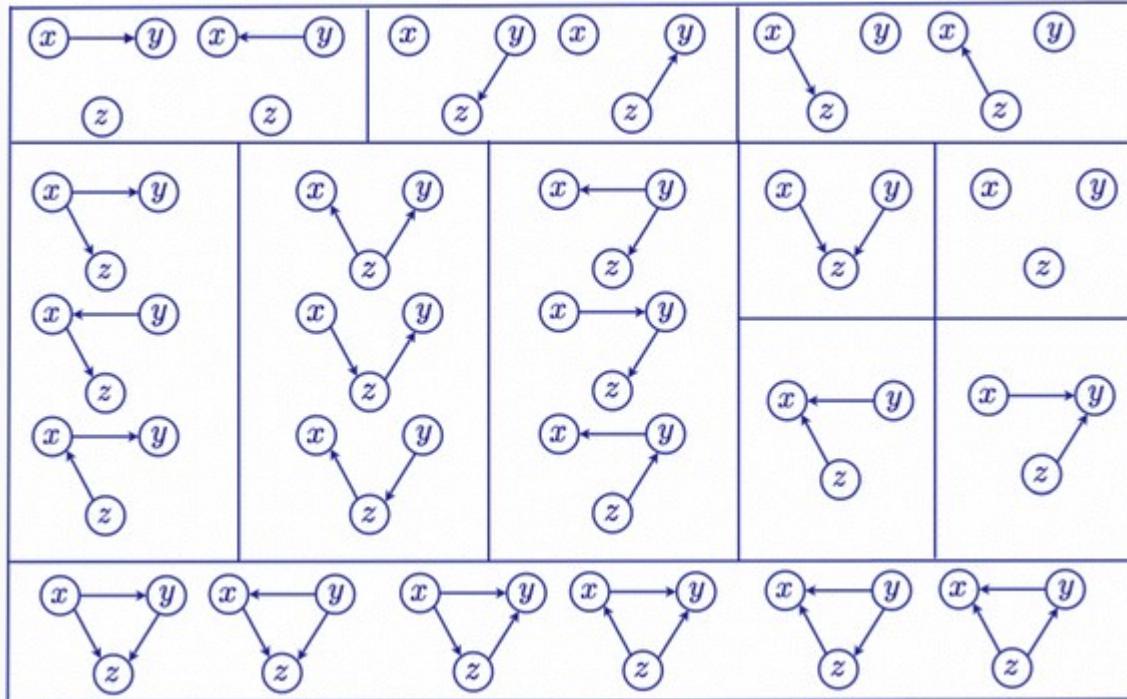
- Как мы узнаем о причинах?
 - Априорное знание механизма
 - Наблюдение через действие (вмешательство)
 - Наблюдение связи (корреляция, да)
 - Последовательность

Направление стрелок

- В данных они произошли одновременно
- Неразличимы только по данным



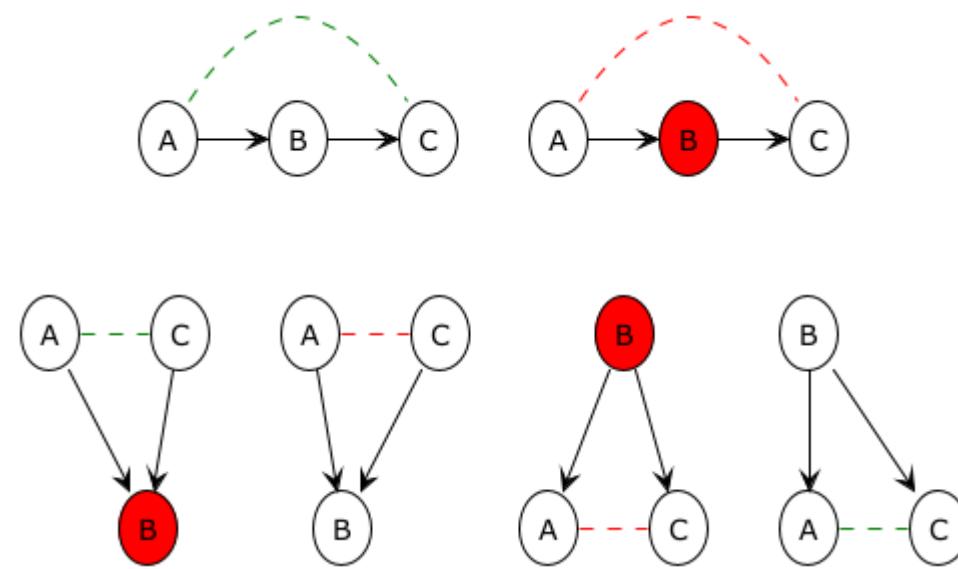
Классы эквивалентности



Introduction to the foundations of causal discovery

Как убрать зависимость?

- Действовать случайно — RCT
- Проконтролировать (учесть) переменную



Принцип Райенбаха

- Из корреляции между А и Б следует:
 - Связь А → Б
 - Связь Б → А
 - Общая причина $A \leftarrow W \rightarrow B$
- Так что корреляция, конечно, не подразумевает каузации, но кое-что она подразумевает
- Ассоциация, не корреляция — см **Григорий Чернов**

Какое лечение лучше?

- Эдвард Симпсон, 1951 год
- https://ru.wikipedia.org/wiki/Парадокс_Симпсона
- Что еще не учли?

	Treatment A	Treatment B
Small Stones	93%(81/87)	87%(234/270)
Large Stones	73%(192/263)	69%(55/80)
Both	78%(273/350)	83%(289/350)

Что можно сделать

- Те, кого лечат, умирают чаще.
- Назначенное лечение зависит от здоровья
 - Хотим, чтобы не зависело — как будто случайно
- Смертность зависит от здоровья
 - Нас интересует зависимость от лечения, а не от здоровья

Эндогенность и смещение

В параграфе 6.2 мы доказали, что из предпосылки $E(\varepsilon_i | x_i) = 0$ следует равенство нулю соответствующей ковариации: $cov(x_i, \varepsilon_i) = 0$. Поэтому в случае, когда данная предпосылка об экзогенности регрессора выполнена, получаем:

$$\widehat{\beta}_2 \xrightarrow{p} \beta_2 + \frac{cov(x_i, \varepsilon_i)}{var(x_i)} = \beta_2 + \frac{0}{var(x_i)} = \beta_2$$

..

Отклонение от «обычного»

- Те, кого лечат, умирают чаще.
- Каков объем лечения сверх обычного?
 - Убираем выборочное смещение $rT = T - (T \sim X)$
- Какова смертность сверх обычного
 - Объясняем часть дисперсии $rY = Y - (Y \sim X)$
- Строим модель $rY \sim rT$