Машинное обучение в экономике

Эффекты воздействия

Потанин Богдан Станиславович

доцент, научный сотрудник, кандидат экономических наук

2023–2024

Введение

Основные рассматриваемые темы

- Основные понятия:
  - Потенциальные исходы.
  - Эффект воздействия, средний эффект воздействия, локальный средний эффект воздействия и условный средний эффект воздействия.
  - Допущения о независимости, об условной независимости и об экзогенности инструмента.
- Методы оценивания:
  - Оценивание средних эффектов воздействия с помощью условных математических ожиданий, взвешивания на обратные условные вероятности и метода с двойной устойчивостью.
  - Оценивание локального среднего эффекта с помощью двойного машинного обучения.
  - Оценивание условных средних эффектов воздействия с помощью S-learner, T-learner, X-learner и метода трансформации классов.

Повторение теории вероятностей

Некоторые важные результаты

- Рассмотрим случайные векторы Y, X и Z с конечными математическими ожиданиями, а также функцию $g(.)$, такую, что математическое ожидание $g(X)$ существует.
- При введенных предпосылках справедливы следующие формулы.

$$\mathrm{E}\left[\mathrm{E}\left(Y|X\right)\right] = \mathrm{E}\left[Y\right] \qquad \mathrm{E}\left[g(X)Y|X\right] = g(X)\mathrm{E}\left[Y|X\right]$$

$$\mathrm{E}\left[g(X)Y\right] = \mathrm{E}\left[g(X)\mathrm{E}\left(Y|X\right)\right] \qquad \mathrm{E}\left[\mathrm{E}\left(Y|X,Z\right)|X\right] = \mathrm{E}\left[Y|X\right]$$

- Если случайные величины $X_1, ..., X_n$ одинаково распределены и имеют конечное математическое ожидание, то $\mathrm{E}(X_i) = \mathrm{E}(X_j)$ при любых $i, j \in \{1, ..., n\}$, а значит:

$$\mathrm{E}\left(\overline{X}_n\right) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(X_i\right) = \frac{1}{n}\left(\underbrace{\mathrm{E}(X_1) + ... + \mathrm{E}(X_n)}_{n \text{ раз складываем } \mathrm{E}(X_i)}\right) = \mathrm{E}\left(X_i\right)$$

Кроме того, в силу **закона больших чисел** $\overline{X}_n \xrightarrow{p} \mathrm{E}(X_i)$.

- Если $g(.)$ – непрерывная функция и $X_1, ..., X_n$ последовательность случайных векторов такая, что $X_n \xrightarrow{p} x$, где $x \in R^m$, то по **теореме о непрерывном отображении** (CMT) выполняется $g(X_n) \xrightarrow{p} g(x)$.
- В частности, по **теореме Слуцкого** если $X_n \xrightarrow{p} x$ и $Y_n \xrightarrow{p} y$, то:

$$X_n + Y_n \xrightarrow{p} x + y \qquad X_n - Y_n \xrightarrow{p} x - y$$

$$X_n \times Y_n \xrightarrow{p} x \times y \qquad X_n/Y_n \xrightarrow{p} x/y, \text{ где } y \neq 0$$

Эффект воздействия

Определение

- Предположим, что одновременно существуют два гипотетически **режима** (counterfactual states) целевой переменной, обозначаемых Y_{0i} и Y_{1i}. Но в данных мы наблюдаем только один из них, в зависимости от значения бинарной **переменной воздействия** (treatment) T_i.

$$Y_i = \begin{cases} Y_{1i}, \text{ если } T_i = 1 \\ Y_{0i}, \text{ если } T_i = 0 \end{cases} = T_i Y_{1i} + (1 - T_i)\, Y_{0i}$$

- **Эффект воздействия** (treatment effect) T_i на Y_i определяется как:

$$\text{TE}_i = Y_{1i} - Y_{0i}$$

- Например, Y_i может отражать факт соверешения покупки клиентом, а T_i – факт наличия персонального предложения по бесплатной доставке товара на дом.
- **Фундаментальная проблема причинно-следственного анализа** (fundamental problem of casual inference) – на практике мы не можем посчитать эффект воздействия TE_i, поскольку не способны одновременно наблюдать оба потенциальных исхода: Y_{0i} и Y_{1i}.
- Например, мы не знаем, как клиент мог бы повести себя и в случае наличия Y_{1i} и в случае отсутствия Y_{i0} предложения о доставке T_i, поскольку он либо получает это предложение (воздействие) $T_i = 1$, либо нет $T_i = 0$.

Средний эффект воздействия

Определение

- На практике часто рассматривается **средний эффект воздействия**:

$$\text{ATE} = \text{E}(\text{TE}_i) = \text{E}(Y_{1i} - Y_{0i}) = \text{E}(Y_{1i}) - \text{E}(Y_{0i})$$

- Представим, что Y_i отражает зарплату индивида, а T_i - факт наличия высшего образования.
- Тогда Y_{1i} отражает зарплату индивида в случае, когда у него есть высшее образование (экзогенно заданное), а Y_{0i} - в случае, когда у него нет высшего образования.
- У каждого индивида одновременно существуют две гипотетические зарплаты Y_{0i} и Y_{1i}, но мы наблюдаем лишь одну из них Y_i, в зависимости от наличия $T_i = 1$ или отсутствия $T_i = 0$ высшего образования, поэтому мы не можем посчитать эффект воздействия $\text{TE}_i = Y_{1i} - Y_{0i}$.
- Однако, при определенных предпосылках мы можем оценить средний эффект воздействия ATE, который можно интерпретировать как среднюю разницу в заработных платах среди людей в состояниях, когда у них есть высшее образование и в случаях, когда у них нет высшего образования.
- Для простоты дальнейшего изложения будем придерживаться допущения о том, что Y_{ji}, Y_{jt}, T_i и T_t независимы при любых $i \neq t$, где $j \in \{0, 1\}$.

Средний эффект воздействия

Условное распределение потенциальных исходов

- Для того, чтобы корректно оценивать и интерпретировать ATE, крайне важно различать безусловные Y_{1i}, Y_{0i} и условные $(Y_{1i}|T_i = 1)$, $(Y_{0i}|T_i = 0)$ распределения потенциальных исходов.
- Рассмотрим очень простой пример, в котором индивиды без высшего образования всегда зарабатывают две тысячи рублей $Y_{0i} = 2$, а с высшим образованием с равной вероятностью зарабатывают одну или три тысячи рублей $P(Y_{1i} = 1) = P(Y_{1i} = 3) = 0.5$.
- В данном случае легко рассчитать эффекты воздействия и средний эффект воздействия:

$$\text{TE}_i = Y_{1i} - Y_{0i} = Y_{1i} - 2 = \begin{cases} 3-2, \text{ если } Y_{1i} = 3 \\ 1-2, \text{ если } Y_{1i} = 1 \end{cases} = \begin{cases} 1, \text{ если } Y_{1i} = 3 \\ -1, \text{ если } Y_{1i} = 1 \end{cases}$$

$$\text{ATE} = \text{E}(Y_{1i}) - \text{E}(Y_{0i}) = (0.5 \times 1 + 0.5 \times 3) - 2 = 2 - 2 = 0$$

- Предположим, что индивид получает высшее образование $T_i = 1$, только если оно позволяет увеличить доходы $Y_{1i} > Y_{0i}$. Тогда разница в условных на уровень образования ожидаемых зарплатах не будет отражать средний эффект воздействия:

$$\text{E}(Y_{1i}|T_i = 1) - \text{E}(Y_{0i}|T_i = 0) = \text{E}(Y_{1i}|Y_{1i} > Y_{0i}) - \text{E}(Y_{0i}|Y_{1i} < Y_{0i}) = 3 - 2 = 1 \neq \text{ATE}$$

- В данном случае $\text{E}(Y_{1i}|T_i = 1)$ и $\text{E}(Y_{0i}|T_i = 0)$ отражают ожидаемые зарплаты среди людей фактически получившив и не получивших высшее образование соответственно.

Допущение о независимости

Формулировка

- Из предыдущего примера мы заключили, что в общем случае:

$$\text{ATE} = \text{E}(Y_{1i}) - \text{E}(Y_{0i}) \neq \text{E}(Y_{1i}|T_i = 1) - \text{E}(Y_{0i}|T_i = 0)$$

- Однако, на практике в некоторых частных случаях мы можем предположить, что ожидаемый исход не зависит от воздействия, то есть $\text{E}(Y_{1i}|T_i = 1) = \text{E}(Y_{1i})$ и $\text{E}(Y_{0i}|T_i = 0) = \text{E}(Y_{0i})$.
- Например, для измеренеия среднего эффекта воздействия вакцины на излечение от болезни, пациентов случайным образом распределеняют между группой воздействия, получающей лекарство, и контрольной группой, принимающей плацебо.
- В результате то, излечится ли (гипотетически) пациент при принятии вакцины Y_{1i} и без нее Y_{0i} не зависит от того, получит ли он ее фактически $T_i = 1$, поскольку она выдается случайным образом.
- Введем **допущение о независимости** (mean independence):

$$\text{E}(Y_{ji}|T_i = j) = \text{E}(Y_{ji}),\ \text{где } j \in \{0, 1\}$$

- Это допущение обычно соблюдается в рамках контролируемых случайных экспериментов.
- При выполнении допущения о независимости:

$$\text{ATE} = \text{E}(Y_{1i}|T_i = 1) - \text{E}(Y_{0i}|T_i = 0)$$

Допущение о независимости

Оценивание среднего эффекта воздействия

- Предположим, что в данных n_1 наблюдений попали в **группу воздействия** $T_i = 1$, а n_0 наблюдений оказались в **контрольной группе** $T_i = 0$.
- При соблюдении допущения о независимости вследствие закона больших чисел получаем состоятельные оценки:

$$\hat{\mathrm{E}}(Y_{1i}) = \frac{1}{n_1}\sum_{i:T_i=1} Y_{1i} \qquad \hat{\mathrm{E}}(Y_{0i}) = \frac{1}{n_0}\sum_{i:T_i=0} Y_{0i}$$

- Тогда по теореме Слуцкого состоятельная оценка среднего эффекта воздействия может быть получена как:

$$\widehat{\mathrm{ATE}} = \hat{\mathrm{E}}(Y_{1i}) - \hat{\mathrm{E}}(Y_{0i}) = \frac{1}{n_1}\sum_{i:T_i=1} Y_{1i} - \frac{1}{n_0}\sum_{i:T_i=0} Y_{0i}$$

- Например, если среди $n = 100$ пациентов $n_0 = 60$ получили плацебо и $n_1 = 40$ получили вакцину, причем среди получивших плацебо излечились 15, а среди получивших вакцину 20, то:

$$\widehat{\mathrm{ATE}} = (1/40) \times 20 - (1/60) \times 15 = 0.5 - 0.25 = 0.25$$

- **Вывод** – согласно полученной оценке в среднем вакцина повышает вероятность исцеления на 0.25.

Допущение о независимости

Технический комментарий о числе наблюдений в группе воздействия и контрольной группе

- При оценивании $\mathrm{E}(Y_{1i})$ и $\mathrm{E}(Y_{0i})$ для простоты изложения ранее и далее **предполагается**, что n_1 и n_0 являются константами.
- Однако, за пределами контролируемых экспериментов размеры группы воздействия и контрольной группы, как правило, являются случайными величинами:

$$n_1 = \sum_{i=1}^{n} T_i \qquad n_0 = \sum_{i=1}^{n} 1 - T_i$$

- Воспользуемся законом больших чисел (в числителе и знаменателе) и теоремой Слуцкого:

$$\hat{\mathrm{E}}(Y_{1i}) = \frac{1}{n_1}\sum_{i:T_i=1} Y_{1i} = \frac{\frac{1}{n}\sum\limits_{i=1}^{n} T_i Y_{1i}}{\frac{1}{n}\sum\limits_{i=1}^{n} T_i} \xrightarrow{p} \frac{\mathrm{E}(T_i Y_{1i})}{\mathrm{E}(T_i)} =$$

$$= \frac{\mathrm{E}(Y_{1i}|T_i=1)P(T_i=1) + 0 \times P(T_i=0)}{P(T_i=1)} = \mathrm{E}(Y_{1i}|T_i=1)$$

- По аналогии нетрудно показать, что $\hat{\mathrm{E}}(Y_{0i}) \xrightarrow{p} \mathrm{E}(Y_{0i}|T_i=0)$, откуда при допущении о независимости следует состоятельность введенной ранее оценки:

$$\widehat{\mathrm{ATE}} = \hat{\mathrm{E}}(Y_{1i}) - \hat{\mathrm{E}}(Y_{0i}) \xrightarrow{p} \mathrm{E}(Y_{1i}|T_i=1) - \mathrm{E}(Y_{0i}|T_i=0) = \underbrace{\mathrm{E}(Y_{1i}) - \mathrm{E}(Y_{0i})}_{\text{допущение о независимости}} = \mathrm{ATE}$$

Допущение о независимости

AB-тестирование

- Часто под **AB-тестированием** понимается проверка гипотезы $\text{H}_0 : \text{ATE} = 0$ при допущении о независимости.
- Например, представим, что клиентская база продавца составляет 1000 человек. Из них он случайным образом отобрал 100 и предоставил им специальное предложение, согласно которому при покупке телефона они получат в подарок наушники.
- Из 100 человек, получивших предложение, покупку совершили 50, а из оставшихся 900 покупку осуществили 360 человек.
- Оценим средний эффект воздействия, то есть насколько, всреднем, возросла вероятность покупки благодаря предоставлению предложения:

$$\widehat{\text{ATE}} = 50/100 - 360/900 = 0.1$$

- Протестируем гипотезу $\text{H}_0 : \text{ATE} = 0$ против альтернативы $\text{H}_1 : \text{ATE} > 0$ с помощью теста о разнице долей, тестовая статистика которого, при верной нулевой гипотезе, в асимптотике имеет стандартное нормальное распределение:

$$Z = \frac{0.1}{\sqrt{0.410 \times (1 - 0.410)(1/100 + 1/900)}} \approx 1.93 \qquad \text{p-value} = 1 - \Phi(Z) \approx 0.03$$

Допущение об условной независимости

Формулировка

- **Проблема** – если допущение о независимости не соблюдается, то введенная ранее оценка $\widehat{\text{ATE}}$ окажется несостоятельной, поскольку:

$$\widehat{\text{ATE}} \xrightarrow{p} \text{E}(Y_{1i}|T_i = 1) - \text{E}(Y_{0i}|T_i = 0) \neq \underbrace{\text{E}(Y_{1i}) - \text{E}(Y_{0i})}_{\text{ATE}}$$

- Обычно оно нарушается в неконтролируемых экспериментах, например, когда имеется самоотбор в число тех, кто решил принять участие в новой программе лояльности магазина или вследствие конкурсного отбора в университеты.
- **Решение** – ввести менее сильное допущение, чем допущение о независимости.
- Введем **допущение об условной независимости**, согласно которому:

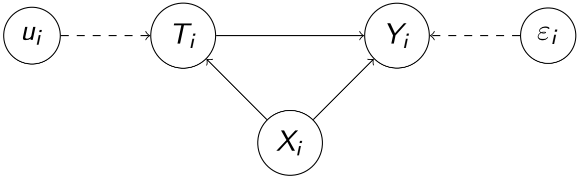
$$\text{E}(Y_{1i}|X_i, T_i = 1) = \text{E}(Y_{1i}|X_i) \qquad \text{E}(Y_{0i}|X_i, T_i = 0) = \text{E}(Y_{0i}|X_i)$$

- Например, мы можем предположить, что при прочем равном возрасте и интеллекте X_i информация о полученном уровне образования T_i не влияет на наши ожидания по поводу того, сколько бы индивид зарабатывал в случаях, когда у него есть Y_{1i} и когда у него нет Y_{0i} высшего образования.
- В приведенной формулировке допущение об условной независимости также подразумевает **предпосылку об общем носителе**, согласно которой $0 < P(T_i = 1|X_i) < 1$. В противном случае X_i может принимать такие значения, при которых события $T_i = 1$ и $T_i = 0$ невозможны.

Допущение об условной независимости

Интерпретация

- Обычно допущение об условной независимости соблюдается, когда X_i отражает все факторы, которые могут быть статистически связаны и с T_i, и с Y_{ji}.



- Предполагается, что связь между T_i и Y_i обусловлена наблюдаемыми в данных переменными X_i, именуемыми **смешивающими** (confounders).
- Прерывистыми линиями отображены связи T_i и Y_i с агрегированными ненаблюдаемыми переменными u_i и ε_i.

Допущение об условной независимости
Пример

- Представим, что мы оцениваем эффект воздействия высшего образования T_i на зарплату Y_i.
- В состоянии без высшего образования индивид зарабатывает $Y_{0i} = 1$.
- Зарплата при наличии высшего образования описывается как:

$$Y_{1i} = 1 + X_i \implies \text{TE}_i = 1 + X_i - 1 = X_i \implies \text{ATE} = \text{E}(X_i) = 0.5$$

  Где X_i отражает любовь индивида к обучению, причем $P(X_i = 0) = P(X_i = 1) = 0.5$.
- Вероятность получения образования зависит **только** от любви к обучению:

$$P(T_i = 1|X_i = 1) = 0.75 \qquad P(T_i = 1|X_i = 0) = 0.5$$

- Допущение о независимости не соблюдается, так как $\text{E}(Y_{1i})$ и $\text{E}(Y_{1i}|T_i = 1)$ не совпадают:

$$\text{E}(Y_{1i}) = 1 + \text{E}(X_i) = 1 + 0.5 = 1.5$$
$$\text{E}(Y_{1i}|T_i = 1) = 1 + 1 \times P(X_i = 1|T_i = 1) + 0 \times P(X_i = 0|T_i = 1) =$$
$$1 + \frac{P(T_i = 1|X_i = 1)P(X_i = 1)}{P(T_i = 1|X_i = 1)P(X_i = 1) + P(T_i = 1|X_i = 0)P(X_i = 0)} = 1 + \frac{0.75 \times 0.5}{0.75 \times 0.5 + 0.5 \times 0.5} = 1.6$$

- Однако, соблюдается допущение об условной независимости, поскольк Y_{0i} константа, а также:

$$\text{E}(Y_{1i}|X_i, T_i = 1) = 1 + \text{E}(X_i|X_i, T_i = 1) = 1 + X_i = \text{E}(Y_{1i}|X_i)$$

Допущение об условной независимости

Оценивание с помощью условных математических ожиданий

- При соблюдении допущения об условной независимости:

$$
\begin{aligned}
\text{ATE} = \text{E}(Y_{1i}) - \text{E}(Y_{0i}) = \text{E}\left(\text{E}(Y_{1i}|X_i) - \text{E}(Y_{0i}|X_i)\right) = \\
= \text{E}\left(\text{E}(Y_{1i}|X_i, T_i = 1) - \text{E}(Y_{0i}|X_i, T_i = 0)\right) = \\
= \text{E}\left(\text{E}(T_iY_{1i} + (1 - T_i)Y_{0i}|X_i, T_i = 1) - \text{E}(T_iY_{1i} + (1 - T_i)Y_{0i}|X_i, T_i = 0)\right) = \\
= \text{E}\left(\text{E}(Y_i|X_i, T_i = 1) - \text{E}(Y_i|X_i, T_i = 0)\right) \implies \frac{1}{n}\sum_{i=1}^{n}\text{E}(Y_i|X_i, T_i = 1) - \text{E}(Y_i|X_i, T_i = 0) \xrightarrow{p} \text{ATE}
\end{aligned}
$$

- **Проблема –** условные математические ожидания $\text{E}(Y_i|X_i, T_i)$ неизвестны.
- **Решение –** поскольку Y_i всегде наблюдается в данных, то можно получить состоятельные оценки $\hat{\text{E}}(Y_i|X_i, T_i)$, например, **методами машинного обучения**.
- В итоге средний эффект воздействия оценивается как:

$$
\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\text{E}}(Y_i|X_i, T_i = 1) - \hat{\text{E}}(Y_i|X_i, T_i = 0)
$$

Допущение об условной независимости

Оценивание с помощью взвешивания на обратные условные вероятности

- Обратим внимание, что при соблюдении допущения об условной независимости:

$$\mathrm{E}\left(T_i Y_i / P(T_i = 1|X_i)|X_i\right) = \underbrace{\mathrm{E}\left(Y_{1i}/P(T_i = 1|X_i)|X_i, T_i = 1\right) P(T_i = 1|X_i)}_{\text{формула полной вероятности, второе слагаемое обнуляется}} = \mathrm{E}\left(Y_{1i}|X_i\right) \implies$$

$$\implies \mathrm{E}\left(T_i Y_i / P(T_i = 1|X_i)\right) = \mathrm{E}\left(\mathrm{E}\left(T_i Y_i / P(T_i = 1|X_i)|X_i\right)\right) = \mathrm{E}\left(\mathrm{E}\left(Y_{1i}|X_i\right)\right) = \mathrm{E}(Y_{1i})$$

- По аналогии можно показать, что при допущении о независимости:

$$\mathrm{E}\left(\left(1 - T_i\right) Y_i / \left(1 - P(T_i = 1|X_i)\right)\right) = \mathrm{E}\left(Y_{0i}\right)$$

- В итоге получаем:

$$\mathrm{ATE} = \mathrm{E}\left(T_i Y_i / P(T_i = 1|X_i)\right) - \mathrm{E}\left(\left(1 - T_i\right) Y_i / \left(1 - P(T_i = 1|X_i)\right)\right)$$

- Из полученны результатов следует альтернативный способ оценивания ATE, именуемый **оценкой с помощью взвешивания на обратные условные вероятности** (inverse probability weighting):

$$\widehat{\mathrm{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{\hat{P}(T_i = 1|X_i)} - \frac{\left(1 - T_i\right) Y_i}{1 - \hat{P}(T_i = 1|X_i)}$$

- **Преимущество** – достаточно с помощью **методов машинного обучения** оценить $\hat{P}(T_i = 1|X_i)$.

Допущение об условной независимости

Двойная устойчивость

- Мы рассмотрели два способа оценивания ATE при допущении об условной независимости, первый из которых опирается на оценки $\mathrm{E}\left(Y_i|X_i, T_i\right)$, а второй – на оценки $P\left(T_i = 1|X_i\right)$.
- **Проблема** – точность оценок каждого из этих способов зависит от точности оценок соответствующих условных математических ожиданий или вероятностей. Если они оценены неточно, то и итоговая оценка ATE также будет неточной.
- **Решение** – обеспечить **двойную устойчивость**, то есть совместить оба способа, чтобы оценка ATE оказывалась состоятельной, если по крайней мере один из них дает состоятельную оценку.

$$\widehat{\mathrm{ATE}} =$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{1i} - \hat{g}_{0i}}_{\text{стремится к ATE если } \hat{\mathrm{E}}\left(Y_i|X_i,\, T_i\right) \text{ состоятельна}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{T_i(Y_i - \hat{g}_{1i})}{\hat{g}_{Ti}} - \frac{(1-T_i)(Y_i - \hat{g}_{0i})}{1-\hat{g}_{Ti}}}_{\text{стремится к 0 если } \hat{\mathrm{E}}\left(Y_i|X_i,\, T_i\right) \text{ состоятельна}} =$$

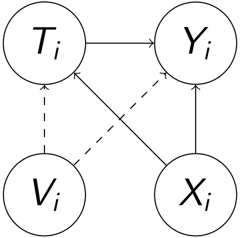
$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{T_iY_i}{\hat{g}_{Ti}} - \frac{(1-T_i)Y_i}{1-\hat{g}_{Ti}}}_{\text{стремится к ATE если } \hat{P}\left(T_i = 1|X_i\right) \text{ состоятельна}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{\left(T_i - \hat{g}_{Ti}\right)\hat{g}_{0i}}{1-\hat{g}_{Ti}} - \frac{\left(T_i - \hat{g}_{Ti}\right)\hat{g}_{1i}}{\hat{g}_{Ti}}}_{\text{стремится к 0 если } \hat{P}\left(T_i = 1|X_i\right) \text{ состоятельна}}$$

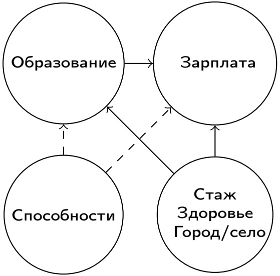
$$\hat{g}_{1i} = \hat{\mathrm{E}}(Y_i|X_i, T_i = 1) \qquad \hat{g}_{0i} = \hat{\mathrm{E}}(Y_i|X_i, T_i = 0) \qquad \hat{g}_{Ti} = \hat{\mathrm{P}}(T_i = 1|X_i)$$

Допущение об условной независимости
Двойное машинное обучение

- Средний эффект воздействия можно также оценить с помощью DML метода, рассмотрев уравнение:
$$Y_i = g(T_i, X_i) + \varepsilon_i \qquad \text{ATE} = \text{E}\left(g(1, X_i) - g(0, X_i)\right)$$
- Допущение об условной независимости можно сформулировать как $\text{E}(\varepsilon_i | X_i, T_i) = 0$.
- Рассмотрим вклад, удовлетворяющий условию ортогональности по Нейману:
$$\psi = \frac{T_i(Y_i - g_1(X_i))}{g_T(X_i)} - \frac{(1 - T_i)(Y_i - g_0(X_i))}{1 - g_T(X_i)} + g_1(X_i) - g_0(X_i) - \text{ATE}$$
$$g_1(X_i) = E(Y_i | X = x_i, T_i = 1), \quad g_0(X_i) = \text{E}(Y_i | X = x_i, T_i = 0), \quad g_T(X_i) = P(T_i = 1 | X_i)$$
- Решая $E(\psi) = 0$ для ATE, а также подставляя оценки (полученные с помощью машинного обучения) неизвестных функций и применяя кросс-фиттинг, получаем оценку, обладающую свойством двойной устойчивости:
$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n} \frac{T_i(Y_i - \hat{g}_1^{(q_i)}(X_i))}{\hat{g}_T^{(q_i)}(X_i)} - \frac{(1 - T_i)(Y_i - \hat{g}_0^{(q_i)}(X_i))}{1 - \hat{g}_T^{(q_i)}(X_i)} + \hat{g}_1^{(q_i)}(X_i) - \hat{g}_0^{(q_i)}(X_i)$$
- Функции $\hat{g}_1^{(k)}$, $\hat{g}_0^{(k)}$ и $\hat{g}_T^{(k)}$ оцениваются с помощью машинного обучения на данных, **не** вошедших в k-ю из K выборок. Также, $q_i = k$, если наблюдение i **не** вошло в k-ю выборку.

- **Проблема –** на практике допущение об условной независимости часто нарушается вследствие **эндогенности**, из-за чего описанные ранее методы обычно дают несостоятельные оценки эффектов воздействия.





- Эндогенность обычно возникает из-за наличия ненаблюдаемых характеристик V_i, одновремено влияющих и на целевую переменную Y_i, и на вероятность воздействия $T_i = 1$.
- **Решение –** адаптировать метод инструментальных переменных.

Допущение об экзогенности инструментальной переменной

Локальный средний эффект воздействия

- Рассмотрим случай, когда эндогенный регрессор T_i и инструментальная переменная Z_i являются бинарными переменными, отражающими, например, факт наличия высшего образования у индивида и его родителей соответственно. Кроме того, временно проигнорируем все остальные признаки X_i.
- По аналогии с Y_{1i} и Y_{0i} введем гипотетические состояния переменной воздействия T_{1i} и T_{0i}.

$$T_i = \begin{cases} T_{1i}, \text{ если } Z_i = 1 \\ T_{0i}, \text{ если } Z_i = 0 \end{cases} = Z_i T_{1i} + (1 - Z_i) T_{0i}$$

- Допустим **экзогенность (валидность) инструмента** - вектор $(Y_{1i}, Y_{0i}, T_{1i}, T_{0i})$ и инструмент Z_i независимы.
- Выделим четыре **группы** индивидов:

	$T_{1i} = 1$	$T_{1i} = 0$
$T_{0i} = 1$	Всегда согласные (always takers)	Отрицатели (deniers)
$T_{0i} = 0$	Соблюдатели (compliers)	Всегда несогласные (never takers)

- Без введения дополнительных строгих допущений, например, о том что эффект воздействия T_i на Y_i является одинаковым для всех индивидов, в общем случае оценить ATE не получится, но можно оценить **локальный ссредний эффект воздействия** LATE, отражающий ATE среди соблюдателей.

$$\text{LATE} = \text{E}(Y_{1i} - Y_{0i} | \underbrace{T_{1i} > T_{0i}}_{\text{соблюдатели}}) = \underbrace{\frac{\text{E}(Y_i|Z_i = 1) - \text{E}(Y_i|Z_i = 0)}{P(T_i = 1|Z_i = 1) - P(T_i = 1|Z_i = 0)}}_{\text{если нет отрицателей, есть соблюдатели и } Z_i \text{ экзогенна}}$$

- **Вывод** – при отсутствии отрицателей, наличи соблюдателей и экзогенном инструменте Z_i для оценивания LATE достаточно оценить $\text{E}(Y_i|Z_i)$ и $P(T_i|Z_i)$, для чего достаточно посчитать соответствующие доли.

Допущение об экзогенности инструментальной переменной

Двойное машинное обучение

- **Проблема** – без использования дополнительных наблюдаемых признаков X_i оценка LATE может иметь достаточно большую дисперсию. Кроме того, инструмент Z_i может коррелировать с некоторыми признаками X_i, что часто приводит к нарушению допущения об экзогенности.
- **Решение** – воспользоваться двойным машинным обучением, позволяющим использовать инструментальную переменую Z_i вместе с признаками X_i:

$$Y_i = g_Y(Z_i, X_i) + \varepsilon_i^{(Y)} \qquad T_i = g_T(Z_i, X_i) + \varepsilon_i^{(T)} \qquad Z_i = g_Z(X_i) + \varepsilon_i^{(Z)}$$

$$g_Y(Z_i, X_i) = \mathrm{E}\left(Y_i|Z_i, X_i\right) \qquad g_T(Z_i, X_i) = \mathrm{E}\left(T_i|Z_i, X_i\right) \qquad g_Z(X_i) = \mathrm{E}\left(Z_i|X_i\right)$$

$$\mathrm{E}(\varepsilon_i^{(Y)}|Z_i, X_i) = 0 \qquad \mathrm{E}(\varepsilon_i^{(T)}|Z_i, X_i) = 0 \qquad \mathrm{E}(\varepsilon_i^{(Z)}|X_i) = 0$$

- Функции g_Y, g_T и g_Z оцениваются с помощью машинного обучения. При выполнении некоторых дополнительных технических условий двойное машинное обучение позволяет получить состоятельную и асимптотически нормальную оценку LATE (выражение оценки опущено для краткости).

$$\text{LATE} = \frac{\mathrm{E}(g_Y(1, X_i)) - \mathrm{E}(g_Y(0, X_i))}{\mathrm{E}(g_T(1, X_i)) - \mathrm{E}(g_T(0, X_i))}$$

- **Преимущество 1** – благодаря использованию инструментальных переменных решает проблему эндогенности, являющуюся серьезным препятствием к состоятельному оцениванию эффектов воздействия с использованием обсуждавшихся ранее методов.
- **Преимущество 2** – в отличие от классического линейного метода инструментальных переменных не опирается на допущения о линейной связи между переменными.

- Средний эффект воздействия не всегда достаточно информативен для принятия конкретных решений. Поэтому в качестве альтернативы часто оценивают **условный средний эффект воздействия**, что в маркетинге именуется **uplift** моделированием.

$$\text{CATE}_i = \text{E}\left(Y_{1i}|X_i\right) - \text{E}\left(Y_{0i}|X_i\right) = \underbrace{\text{E}(Y_i|X_i, T_i = 1) - \text{E}(Y_i|X_i, T_i = 0)}_{\text{допущение об условной независимости}}$$

- При **допущении об условной независимости** можно получить состоятельную оценку условного среднего эффекта воздействия:

$$\widehat{\text{CATE}}_i = \hat{\text{E}}(Y_i|X_i, T_i = 1) - \hat{\text{E}}(Y_i|X_i, T_i = 0)$$

- Можно либо оценить $\hat{\text{E}}(Y_i|X_i, T_i)$ по всей выборке (**Single-learner/S-learner**), либо отдельно $\hat{\text{E}}(Y_i|X_i, T_i = 1)$ и $\hat{E}(Y_i|X_i, T_i = 0)$ по группе воздействия и контрольной группе соответственно (**Two-learner/T-learner**).
- Однако существуют и иные, менее очевидные подходы к оцениванию CATE_i.

Условный средний эффект воздействия

Метод трансформации класса

- Скронструируем **псевдоисход**:

$$Y_i^* = Y_i\left(\frac{T_i}{P(T_i = 1|X_i)} + \frac{1 - T_i}{1 - P(T_i = 1|X_i)}\right)$$

- Обозначим через $\hat{Y}_i^*$ величину Y_i^*, посчитанную с использованием $\hat{P}(T_i = 1|X_i)$ вместо $P(T_i = 1|X_i)$.
- При **соблюдении допущения об условной независимости** можно показать, по аналогии с тем, как это было сделано с оцениванием ATE с помощью взвешивания на обратные условные вероятности, что:

$$\mathrm{E}(Y_i^*|X_i) = \mathrm{CATE}_i$$

- Следовательно, для оценивания CATE_i можно воспользоваться двухшаговой процедурой, часто именуемой **методом трансформации классов**.
- **Первый шаг** – вычислить $\hat{P}(T_i = 1|X_i)$ и $\hat{Y}_i^*$.
- **Второй шаг** – оценить $\mathrm{E}(\hat{Y}_i^*|X_i)$ и получить $\widehat{\mathrm{CATE}}_i = \hat{\mathrm{E}}(\hat{Y}_i^*|X_i)$.

- **Проблема** – иногда группа воздействия может включать малое число наблюдений, что осложняет оценивание $\hat{\mathrm{E}}(Y_i|X_i, T_i = 1)$ по данным группы воздействия.
- **Решение** – рассмотрим вспомогательную переменную:

$$D_{1i} = Y_{1i} - \mathrm{E}(Y_i|X_i, T_i = 0)$$

- Обратим внимание, что при соблюдении **допущения об условной независимости**:

$$\mathrm{E}(D_{1i}|X_i) = \mathrm{E}(Y_{1i}|X_i) - \mathrm{E}(Y_i|X_i, T_i = 0) = \mathrm{E}(Y_{1i}|X_i) - \mathrm{E}(Y_{0i}|X_i) = \mathrm{CATE}_i$$

- Следовательно, $\hat{\mathrm{E}}(D_{1i}|X)$ можно рассматривать как оценку CATE_i, полученную с помощью Y_{1i} и $\hat{\mathrm{E}}(Y_i|X_i, T_i = 0)$, то есть без использования неэффективно оцениваемого по малому числу наблюдений группы воздействия $\mathrm{E}(Y_i|X_i, T_i = 1)$.
- **Проблема** – в данных отсутствуют наблюдения по D_{1i}.
- **Решение** – рассмотреть $\hat{D}_{1i} = Y_{1i} - \hat{\mathrm{E}}(Y_i|X_i, T_i = 0)$ и оценить $\widehat{\mathrm{CATE}}_i = \hat{\mathrm{E}}\left(\hat{D}_{1i}|X_i\right)$.

Условный средний эффект воздействия

Алгоритм X-learner

- **Первый шаг** – по аналогии с T-learner оцениваются условные математические ожидания $\text{E}(Y_i|X_i, T_i = 1)$ и $\text{E}(Y_i|X_i, T_i = 0)$.
- **Второгой шаг** – рассчитываются вспомогательные переменные:

$$\hat{D}_{1i} = Y_{1i} - \hat{\text{E}}(Y_i|X_i, T_i = 0) \qquad \hat{D}_{0i} = \hat{\text{E}}(Y_i|X_i, T_i = 1) - Y_{0i}$$

- **Третий шаг** – оцениваются $\text{E}\left(\hat{D}_{1i}|X_i\right)$ и $\text{E}\left(\hat{D}_{0i}|X_i\right)$.
- **Четвертый шаг** – оцениваются условные вероятности попадания в группу воздействия $P(T_i = 1|X_i)$.
- **Пятый шаг** – с помощью взвешивания оценивается условный эффект воздействия:

$$\widehat{\text{CATE}}_i = \left(1 - \hat{P}(T_i = 1|X_i)\right)\hat{\text{E}}\left(\hat{D}_{1i}|X_i\right) + \hat{P}(T_i = 1|X_i)\hat{\text{E}}\left(\hat{D}_{0i}|X_i\right)$$

- **Интуиция взвешивания** – чем меньше наблюдений в группе воздействия, тем больший вес присваивается $\hat{\text{E}}\left(\hat{D}_{1i}|X_i\right)$, который не зависит от $\hat{\text{E}}\left(Y_i|X_i, T_i = 1\right)$.

Качество оценивания условных средних эффектов воздействия

Постановка проблемы

- Для измерения качества прогнозирования условных эффектов воздействия хотелось бы использовать метрику:

$$\text{MSE}_0 = \frac{1}{n}\sum_{i=1}^{n}\left(\text{CATE}_i - \widehat{\text{CATE}}_i\right)^2$$

- **Проблема** – на практике CATE_i неизвестно.
- **Решение** – в качестве косвенной метрики качества прогнорзов условных эффектов воздействия использовать метрику качества точности прогнозов исхода:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\text{E}}\left(Y_i|X_i, T_i\right)\right)^2$$

- **Проблема** – модель, точнее оценивающая Y_i, не обязательно будет точнее оценивать CATE_i. Например, даже если оценки $\hat{\text{E}}\left(Y_i|X_i, T_i = 1\right)$ и $\hat{\text{E}}\left(Y_i|X_i, T_i = 0\right)$ обладают очень большим, но примерно одинаковым смещением, то их разница, то есть $\widehat{\text{CATE}}_i$, может оказаться весьма точной оценкой, поскольку смещения сократятся.

- Рассмотрим псевдоисход Y_i^*, такой, что $\mathrm{E}(Y_i^*|X_i) = \mathrm{CATE}_i$ и соблюдены некоторые дополнительные, технические условия. Например, можно использовать псевдоисход метода трансформации классов:

$$Y_i^* = Y_i\left(\frac{T_i}{P(T_i = 1|X_i)} + \frac{1 - T_i}{1 - P(T_i = 1|X_i)}\right)$$

- Рассмотрим следующую метрику качества:

$$\mathrm{MSE}^* = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i^* - \widehat{\mathrm{CATE}}_i\right)^2$$

- Можно показать, что при определенных условиях, в частности, больших объемах выборки, MSE^* является достаточно точной аппроксимацией MSE_0.
- Существует множество иных подходов к аппроксимации MSE_0 с помощью метрик, зависящих от различных псевдоисходов.