Машинное обучение в экономике

Машинное обучение в эконометрике

Потанин Богдан Станиславович

доцент, кандидат экономических наук

2024–2025

- Методы оценивания параметров:
  - Ридж и Лассо регрессии.
  - Пост-Лассо.
  - Двойное машинное обучение.
- Базовые понятия:
  - Регуляризация.
  - Метод моментов.
  - Структурный параметр.
  - Функция шума.
  - Ортогональность по Нейману.
  - Кросс-фиттинг.
  - Эндогенность и неслучайный отбор.

- Машинное обучение, как правило, применяется для прогнозирования с помощью оценок различных характеристик распределения, таких как условные математические ожидания и вероятности.
- Обычно методы машинного обучения дают оценки, обладающие малым смещением и большой дисперсией, поскольку не накладывают структурных предпосылок (например, о линейности) на форму связи между переменными модели.
- В задаче прогнозирования эконометрические методы обычно демонстрируют преимущество на выборках малого и среднего объема, поскольку обладают структурой, позволяющей компенсировать недостаток данных реалистичными предположениями, снижающими дисперсию оценок.

- Основной упор в эконометрическом анализе делается на оценивание параметров моделей, имеющих содержательную экономическую интерпретацию.
- Иногда исследователя интересуют не все, а лишь часть параметров модели, характеризующих связь между основными переменными. В таком случае можно объединить сильные стороны эконометрики (интерпретабельность) и машинного обучения (высокая точность прогнозирования).
- **Основная идея** – часть модели, не представляющая содержательный интерес для исследователя, оценивается методами машинного обучения, а для оценивания структурных параметров применяются эконометрические методы анализа.

Регуляризация

Основная идея

- **Проблема** – машинное обучение позволяет избегать допущения о линейной связи Y_i с X_i, тем самым снижая смещение оценок, но часто серьезно повышает дисперсию на малых выборках.
- **Идея** – для того, чтобы снизить дисперсию оценок и избежать переобучения, пусть и ценой повышения смещения, можно воспользоваться **регуляризацией**.
- Одним из наиболее популярных подходов к регуляризации заключается в накладывании штрафов на параметры модели:

$$\underbrace{\mathrm{L}\left(Y, F(X;\theta)\right)}_{\text{функция потерь}} + \underbrace{\text{penalty}\left(\theta\right)}_{\text{штраф}} \qquad \textbf{минимизируемый функционал}$$

- Функция $\text{penalty}\left(\theta\right)$ накладывает **штраф** (penalty) за определенные, как правило **большие по модулю** значения элементов n_θ-мерного вектора параметров θ модели $F(X;\theta)$.
- **Интуиция** – ограничение $\theta_t = 0$, где $t \in \{1, ..., n_\theta\}$, обычно соответствует исключению параметра θ_t из модели, что приводит к ее упрощению. Регуляризация предлагает в качестве альтернативы накладывать штрафы, приводящие, образно говоря, к естественному отбору среди параметров, когда значительно отличными от 0 оказываются лишь те из них, что оказывают существенное влияние на качество модели.
- В роли параметров θ, например, могут выступать коэффициенты β в обычной линейной или логистической регрессии.

- В большинстве случае функция штрафа задается с помощью Lp-нормы:

$$\text{penalty}(\theta) = ||\theta||_p^p = \sum_{t=1}^{n_\theta} \lambda_t |\theta_t|^p, \text{ где } \lambda_t > 0 \text{ и } p \in \{1, 2, 3, ...\}$$

- Случаи $p = 1$ и $p = 2$ являются наиболее популярными:

$$\text{penalty}(\theta) = \sum_{t=1}^{n_\theta} \lambda_t |\theta_t| \qquad \textbf{Лассо регуляризация}$$

$$\text{penalty}(\theta) = \sum_{t=1}^{n_\theta} \lambda_t \theta_t^2 \qquad \textbf{Ридж регуляризация}$$

- Чем больше значения константа λ_t, тем сильнее накладываемый штраф за большие по абсолютной величине значения параметров θ_t
- Подбор λ_t обычно осуществляется по аналогии с гиперпараметрами, например, с помощью кросс-валидации. Для простоты часто полагают $\lambda_t = \lambda \in R$ для всех t.

- Как правило величины коэффициентов θ тесно связаны с единицами измерения признаков X.
- Например, в линейной регрессии если коэффициент при весе в килограммах равняется $\theta_k = 100$, то этот же коэффициент при весе в граммах будет равняться $\theta_k^* = 100/1000 = 0.1$.
- **Проблема** – если на все коэффициенты накладывается один и тот же штраф, например, λ при использовании Lp-нормы, то его сила будет зависеть от единиц измерения признаков.
- **Решение** – привести признаки к сопоставимым единицам измерения, например, за счет стандартизации.
- Кроме того, часто стандартизация снижает сложность оптимизационной задачи (через снижение погрешностей, связанных с операциями над числами с плавающей точкой), тем самым повышая скорость нахождение минимума методами численной оптимизации.

- Даже сохраняя линейную форму связи $\mathrm{E}(Y_i|X_i) = X_i\beta$, линейная регрессия может аппроксимировать очень сложные зависимости, за счет того, что X_i могут быть разнообразными функциями (например, полиномы и сплайны) от исходных данных.
- Чем больше функций от исходных данных включает исследователь, тем, как правило, ниже смещение, но выше дисперсия оценок параметров и прогнозов.
- **Проблема** – при включении большого числа функций от исходных данных число оцениваемых коэффициентов β_i может оказаться чрезвычайно велико, что приведет к крайне большой дисперсии оценок.
- **Решение** – воспользоваться, например, Лассо регуляризацией, минимизируя:

$$\sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \sum_{t=1}^{n_\beta} \lambda_t |\beta_t|$$

- **Полезное свойство Лассо регуляризации** – часто оценки коэффициентов при наименее значимых (с точки зрения вклада в прогностическое качество модели) регрессорах обнуляются $\hat{\beta}_i = 0$, что эквивалентно их исключению из модели.

- Преимущество Ридж регуляризации в линейной регрессии заключается в возможности получения аналитических оценок коэффициентов и их характеристик:

$$\hat{\beta} = \left(X^T X + \Lambda\right)^{-1} X^T Y, \text{ где } \Lambda = \text{diag}\left(\lambda, ..., \lambda\right)$$

$$\mathsf{E}\left(\hat{\beta}|X\right) = \beta - \underbrace{\lambda \left(X^T X + \Lambda\right)^{-1} \beta}_{\text{смещение}}$$

$$\text{Cov}\left(\hat{\beta}|X\right) = \left(X^T X + \Lambda\right)^{-1} X^T \text{Cov}\left(\varepsilon|X\right) X \left(X^T X + \Lambda\right)^{-1}$$

- Можно показать, что смещение увеличивается по мере роста штрафа λ.
- Производная $\text{Cov}\left(\hat{\beta}|X\right)$ по λ является отрицательно определенной матрицей, поэтому увеличение штрафа приводит к уменьшению дисперсии оценок.
- Если случайные ошибки ε_i гетероскедастичны, то существует такая константа c, что при $\lambda \in (0, c)$ оценки Ридж регрессии более эффективны, чем МНК.

- Если в модели используется лишь один регрессор (без константы) и $\beta \neq 0$, то легко показать, что смещение возрастает вместе со штрафом λ:

$$\partial \text{bias}\left(\hat{\beta}|X\right)/\partial\lambda = \partial\left|\lambda\beta/\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)\right|/\partial\lambda = \left|\beta\sum_{i=1}^{n} X_i^2/\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)^2\right| > 0$$

- Поскольку $\text{Cov}(\varepsilon|X)$ положительно определена, то дисперсия падает с ростом λ:

$$\partial \text{Var}\left(\hat{\beta}|X\right)/\partial\lambda = \partial\left(X^T \text{Cov}(\varepsilon|X)\, X/\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)^2\right)\partial\lambda =$$

$$= \underbrace{\left(-2/\left(\sum_{i=1}^{n} X_i^2 + \lambda\right)^3\right)}_{<0} \underbrace{X^T \text{Cov}(\varepsilon|X)\, X}_{>0} < 0$$

Регуляризация в линейном регрессионном анализе

Пост-Лассо

- Напомним, что при Лассо регуляризации в линейных регрессионных моделях некоторые из коэффициентов β могут обращаться в 0.
- **Проблема** – включение большого числа регрессоров с нулевыми коэффициентами может привести к снижению эффективности оценок вследствие серьезного смещения.
- **Решение** – применить двухшаговую процедуру, на первом шаге которой оценивается Лассо регрессия, а на втором – обычная МНК регрессия, в которой в качестве объясняющих переменных используются лишь те, при которых коэффициенты оказались отличными от нуля в Лассо регрессии.
- Поскольку МНК регрессия используется после Лассо, описанный метод именуется **пост-Лассо**.
- **Примечание** – эффективность оценок пост-Лассо может быть ниже, чем у обычной Лассо регрессии.

Двойное машинное обучение (DML)

Частично линейная регрессия

- Рассмотрим **частично линейную модель** (partially linear model):

$$Y_i = \alpha T_i + g(X_i) + \varepsilon_i, \text{ где } \mathrm{E}(\varepsilon_i | T_i, X_i) = 0 \text{ и } (T_i, X_i, \varepsilon_i) \text{ i.i.d.}$$

- В качестве основного параметра интереса для исследователя выступает $\alpha \in R$.
- Например, Y_i может отражать прибыль фирмы, T_i – долю акций, принадлежищих государству, α – влияние государственного участия на прибыль при прочих равных значениях контрольных переменных X_i (размер, возраст, объем долга и т.д.).
- **Проблема** – неизвестная функция $g(X_i)$ может оказаться нелинейной и тогда МНК оценки могут оказаться несостоятельными.
- **Наивное решение** – применить методы машинного обучения, например, Ридж или Лассо регрессию с большим числом функций от X_i (полиномы и сплайны).
- **Проблема** – методы машинного обучения могут дать достаточно точные прогнозы $\hat{Y}_i$, но полученная с их помощью оценка $\hat{\alpha}$ может оказаться неэффективной.

Двойное машинное обучение (DML)

Классический метод оценивания частично линейной регрессии

- Вычтем из обеих частей регрессионного уравнения условное математическое ожидание, что позволит нам избавиться от $g(X_i)$:

$$Y_i - \mathrm{E}(Y_i|X_i) = \alpha(T_i - \mathrm{E}(T_i|X_i)) + (g(X_i) - \mathrm{E}(g(X_i)|X_i)) + (\varepsilon_i - \mathrm{E}(\varepsilon_i|X_i)) =$$

$$= \alpha(T_i - \mathrm{E}(T_i|X_i)) + \varepsilon_i - \mathrm{E}\left(\underbrace{\mathrm{E}(\varepsilon_i|X_i, T_i)}_{0}|T_i\right) = \alpha(T_i - \mathrm{E}(T_i|X_i)) + \varepsilon_i$$

- Случайная ошибка полученного уравнения имеет нулевое условное математическое ожидание:

$$\mathrm{E}(\varepsilon_i|T_i - \mathrm{E}(T_i|X_i)) = \mathrm{E}\left(\underbrace{\mathrm{E}[\varepsilon_i|T_i - \mathrm{E}(T_i|X_i), X_i, T_i]}_{0}|T_i - \mathrm{E}(T_i|X_i)\right) = 0$$

- Следовательно, для того, чтобы получить состоятельную оценку параметра α, достаточно с помощью МНК оценить регрессию без константы $Y_i - \mathrm{E}(Y_i|X_i)$ на $T_i - \mathrm{E}(T_i|X_i)$.
- **Проблема** – нам неизвестны условные математические ожидания $\mathrm{E}(Y_i|X_i)$ и $\mathrm{E}(T_i|X_i)$.
- **Решение** – их можно оценить с помощью методов непараметрической статистики, в частности, машинным обучением, например, регрессионными деревьями.

Двойное машинное обучение (DML)

Линейный метод наименьших квадратов как частный случай метода моментов

- Метод наименьших квадратов (МНК) предполагает минимизацию квадратов отлокнений:

$$\beta = \underset{\tilde{\beta}}{\text{argmin}}\ \mathrm{E}\left(\left(Y_i - X_i\tilde{\beta}\right)^2\right)$$

- Условия первого порядка данной оптимизационной задачи:

$$\mathrm{E}\left(\left(Y_i - X_i\beta\right)X_i\right) = \mathrm{E}\left(\varepsilon_i X_i\right) = (0, ..., 0)$$

- Решая соответствующее равенство получаем:

$$\beta = \mathrm{E}\left(\left(X_i^T X_i\right)^{-1}\right)\mathrm{E}\left(X_i^T Y_i\right)$$

- Линейный МНК можно помыслить как метод моментов (ММ), в котором моментные тождества задаются условием первого порядка, а значит для оценивания коэффициентов достаточно заменить теоретические моменты их выборочными аналогами:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

Двойное машинное обучение (DML)

Классический подход через призму метода моментов

- Напомним, что МНК оценка параметров линеной регрессии является оценкой метода моментов, опирающейся на следующее моментное тождество:

$$\mathrm{E}\left(\varepsilon_i X_i\right) = \mathrm{E}\left(\left(Y_i - X_i\beta\right) X_i\right) = (0, ..., 0)$$

- По аналогии можно показать, что в рассматриваемой ранее регрессии без константы $Y_i - \mathrm{E}\left(Y_i|X_i\right)$ на $T_i - \mathrm{E}\left(T_i|X_i\right)$ параметр α является единственным решением моментного тождества:

$$\mathrm{E}\left(\left[Y_i - \mathrm{E}\left(Y_i|X_i\right) - \alpha\left(T_i - \mathrm{E}\left(T_i|X_i\right)\right)\right]\left[T_i - \mathrm{E}\left(T_i|X_i\right)\right]\right) = 0$$

- Для краткости обозначим $g_Y(X_i) = \mathrm{E}\left(Y_i|X_i\right)$ и $g_T(X_i) = \mathrm{E}\left(T_i|X_i\right)$.
- Выражая α из тождества получаем:

$$\alpha = \frac{\mathrm{E}\left(\left(Y_i - g_Y(X_i)\right)\left(T_i - g_T(X_i)\right)\right)}{\mathrm{E}\left(\left(T_i - g_T(X_i)\right)^2\right)}$$

Двойное машинное обучение (DML)

Основная идея метода

- **Проблема** – исследователю неизвестны не только истинные математические ожидания, через которые выражается параметр α, но и входящие в них условные математические ожидания $g_Y(X_i)$ и $g_T(X_i)$.
- **Решение** – оценить неизвестные условные математические ожидания с помощью классических методов непараметрической статистики или машинного обучения.
- В результате получаем двухшаговую процедуру, на **первом** шаге которой с помощью машинного обучения оцениваются **функции шума**:

$$\hat{g}_Y(x) = \hat{\mathrm{E}}(Y_i|X_i = x) \qquad \hat{g}_T(x) = \hat{\mathrm{E}}(T_i|X_i = x)$$

- На **втором** шаге теоретические моменты заменяются на выборочные, в которых вместо истинных условных математических ожиданий используются оцененные на первом шаге функции шума:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} (Y_i - \hat{g}_Y(X_i))(T_i - \hat{g}_T(X_i))}{\sum_{i=1}^{n} (T_i - \hat{g}_T(X_i))^2}$$

- Введем отдельное обозначение для **моментного тождества** (score):

$$\psi\left(\alpha, g_T(X_i), g_Y(X_i)\right) = \mathsf{E}\left(\left[Y_i - g_X(X_i) - \alpha\left(T_i - g_T(X_i)\right)\right]\left[T_i - g_T(X_i)\right]\right) = 0$$

- **Проблема** – вместо $\psi = \psi\left(\alpha, g_T(X_i), g_Y(X_i)\right)$ используется $\hat{\psi} = \psi\left(\alpha, \hat{g}_T(X_i), \hat{g}_Y(X_i)\right)$. Однако, как правило $\mathsf{E}(\hat{\psi}) \neq 0$, поскольку оценки $\hat{g}_T(X_i)$ и $\hat{g}_Y(X_i)$ могут иметь достаточно сильное смещение, в частности, из-за регуляризации (**regularization bias)**.

- **Решение** – частично данная проблема смягчается за счет формы функции ψ, удовлетворяющей условию **ортогональности по Нейману**:

$$\partial \mathsf{E}\left(\psi\left(\alpha; g_Y(X_i) + q\underbrace{\left(\hat{g}_Y(X_i) - g_Y(X_i)\right)}_{\text{смещение}}, g_T(X_i) + q\underbrace{\left(\hat{g}_T(X_i) - g_T(X_i)\right)}_{\text{смещение}}\right)\right) / \partial q|_{q=0} = 0$$

Интуиция – благодаря ортогональности по Нейману при малом смещении $\hat{g}_T$ и $\hat{g}_Y$ можно ожидать, что $\mathsf{E}(\hat{\psi}) \approx 0$. Это оправдывает то, что мы выражаем $\hat{\alpha}$ из равенства $\mathsf{E}(\hat{\psi}) = 0$.

Двойное машинное обучение (DML)

Проблема переобучения

- **Проблема** – даже несмотря на регуляризацию, многие методы машинного обучения склонны к переобучению (**overfitting bias**), из-за чего по крайней мере внутривыборочные оценки $Y_i - \hat{g}_Y(X_i)$ и $T_i - \hat{g}_T(X_i)$ могут существенно отклоняться от $Y_i - g_Y(X_i)$ и $T_i - g_T(X_i)$, тем самым снижая точность оценок второго шага.
- **Решение** – применить разбиение выборки (**sample splitting**) на две части – первая часть выборки используется на первом шаге, то есть для оценивания g_Y и g_T, а вторая – на втором шаге для оценивания α с использованием полученных на первом шаге оценок $\hat{g}_Y$ и $\hat{g}_T$.
- **Проблема** – мы используем лишь по половине выборки для каждого из шагов, что может снижать эффективность наших оценок.
- **Решение** – использовать различные части выборки для оценивания и прогнозирования.

- Обозначим через $\hat{g}_Y^{(1)}$, $\hat{g}_T^{(1)}$ и $\hat{g}_Y^{(2)}$, $\hat{g}_T^{(2)}$ оценки функций g_Y и g_T, полученные на первой и второй половинах выборки соответственно. То есть обе половины выборки поочередно используются на первом шаге.
- Введем вспомогательную переменную q_i, такую, что $q_i = 1$ если наблюдение i **не** вошло в первую половину выборки, и $q_i = 2$ – в противном случае.
- Оценим $\hat{\alpha}$ таким образом, чтобы для каждого наблюдения i на втором шаге использовались оценки функций g_Y и g_T, которые были получены без использования i-го наблюдения:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{g}_Y^{(q_i)}(X_i) \right) \left(T_i - \hat{g}_T^{(q_i)}(X_i) \right)}{\sum_{i=1}^{n} \left(T_i - \hat{g}_T^{(q_i)}(X_i) \right)^2}$$

- **Проблема** – использование лишь половины выборки может существенно снизить эффективность оценок функций g_Y и g_T.
- **Решение** – реализовать кросс-фиттинг по аналогии с кросс-валидацией, разбив выборку на K (примерно) равных частей, где $\hat{g}_Y^{(k)}$ и $\hat{g}_T^{(k)}$ оцениваются на данных, **не** вошедших в k-ю из этих выборок (обычно $K = 5$):

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{g}_Y^{(q_i)}(X_i) \right) \left(T_i - \hat{g}_T^{(q_i)}(X_i) \right)}{\sum_{i=1}^{n} \left(T_i - \hat{g}_T^{(q_i)}(X_i) \right)^2}$$

  Где $q_i = k$, если наблюдение i **не** вошло в k-ю выборку.
- **Проблема** – результаты оценивания могут быть чувствительны к конкретному разбиению на K частей.
- **Решение** – повторить кросс-фиттинг m раз и либо усреднить все полученные оценки, либо взять ту из них, что является выборочной медианой.

$$\text{Зарплата}_i = \alpha \times \text{Образование}_i + g(\text{Возраст}_i) + \varepsilon_i$$

Для оценивания $g_Y^{(q_i)}(X_i)$ и $g_Y^{(q_i)}(X_i)$ используется метод ближайших соседей с одним соседом.

Возраст_i (X_i)	20	30	40	50	60	24	37	44	47	90
Образование_i (T_i)	1	0	1	0	1	0	1	0	1	0
Зарплата_i (Y_i)	1	2	3	4	5	6	7	8	9	10
Разбиение выборки	Первая часть					Вторая часть				
$\hat{g}_Y^{(q_i)}(\text{Возраст}_i) = \hat{\text{E}}(\text{Зарплата}_i \mid \text{Возраст}_i)$	6	6	7	9	9	1	3	3	4	5
$\hat{g}_T^{(q_i)}(\text{Возраст}_i) = \hat{\text{E}}(\text{Образование}_i \mid \text{Возраст}_i)$	0	0	1	1	1	1	1	1	0	1

- Нетрудно показать, что $\hat{\alpha} = -10/6$, поскольку:

$$\sum_{i=1}^{n} \left(Y_i - \hat{g}_Y^{(q_i)}(X_i)\right)\left(T_i - \hat{g}_T^{(q_i)}(X_i)\right) = (1-6)(1-0) + \ldots + (10-5)(0-1) = -10$$

$$\sum_{i=1}^{n} \left(T_i - \hat{g}_T^{(q_i)}(X_i)\right)^2 = (1-0)^2 + \ldots + (0-1)^2 = 6$$

Двойное машинное обучение (DML)

Резюме

- Описанный метод именуется **двойным машинным обучением** (DML), поскольку предполагает применение методов машинного обучения при оценивании функций $\hat{g}_Y$ и $\hat{g}_T$, а также кросс-фиттинга.
- При достаточно слабых допущениях DML метод дает состоятельную и асимптотически нормальную оценку $\hat{\alpha}$.
- Идейно DML опирается на метод моментов, поскольку выражение, используемое для оценивания α, выводится из равенства $E(\psi) = 0$.
- **Проблема** – использование оценок $\hat{g}_Y$ и $\hat{g}_T$ вместо истинных значений g_Y и g_T может приводить к неточностям в оценивании $\hat{\alpha}$.
- **Решение** – кросс-фиттинг и подбор функции ψ, удовлетворяющей ортогональности по Нейману.
  - Ортогональность по Нейману позволяет сгладить смещение вследствие регуляризации.
  - Кросс-фиттинг помогает снизить смещение, обусловленное переобучением.
- Иногда кросс-фиттинг реализуется упрощенным образом – параметр α оценивается на каждой из K подвыборок и полученный результат усредняется. Такой подход называется DML, а рассмотренный ранее – DML2.
- Авторы метода рекомендуют применять DML2, особенно на малых выборках.
- В рамках курса, если не сказано иного, предполагается использование DML2.

- **Проблема** – если T_i является эндогенной переменной, то $\mathrm{E}(\varepsilon_i | T_i, X_i) \neq 0$, откуда $\mathrm{E}(\psi) \neq 0$, что не позволяет оценить α описанным ранее способом.
- **Решение** – найти **инструментальную переменную** Z_i (случай с несколькими инструментами рассматривается по аналогии), то есть такую, что $\mathrm{E}(\varepsilon_i | X_i, Z_i) = 0$ и $\mathrm{E}(\mathrm{Cov}(T_i, Z_i | X_i)) \neq 0$. После этого рассмотреть такую ψ, что $\mathrm{E}(\psi) = 0$ и соблюдается ортогональность по Нейману, например:

$$\psi = (Y_i - g_Y(X_i) - \alpha (T_i - g_T(X_i))) (Z_i - g_Z(X_i)), \text{ где } g_Z(X_i) = \mathrm{E}(Z_i | X_i)$$

- По аналогии с предыдущим примером применив кросс-фиттинг получаем:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{g}_Y^{(q_i)}(X_i) \right) \left(Z_i - \hat{g}_Z^{(q_i)}(X_i) \right)}{\sum_{i=1}^{n} \left(T_i - \hat{g}_T^{(q_i)}(X_i) \right) \left(Z_i - \hat{g}_Z^{(q_i)}(X_i) \right)}$$

Двойное машинное обучение (DML)

Неслучайный отбор

- Наблюдаемость Y_i может зависеть от некоторого правила, например, запрлата Y_i наблюдается лишь для работающих $Z_i = 1$ индивидов и ненаблюдается для безработных $Z_i = 0$:

$$\underbrace{Y_i^* = \alpha T_i + g(X_i) + \varepsilon_i}_{\text{целевое уравнение}} \qquad \underbrace{Z_i^* = r(W_i) + u_i}_{\text{уравнение отбора}}$$

$$Y_i = \begin{cases} Y_i^*, \text{ если } Z_i = 1 \\ \text{ненаблюдаем, в противном случае} \end{cases} \qquad Z_i = \begin{cases} 1, \text{ если } Z_i^* \geq 0 \\ 0, \text{ в противном случае} \end{cases}$$

- Поскольку в данных мы наблюдаем лишь $(Y_i^* | Z_i = 1)$, а не Y_i^*, то нарушается допущение о нулевом условном математическом ожидании случайной ошибки:

$$\mathrm{E}(\varepsilon_i | Z_i = 1) = \mathrm{E}(\varepsilon_i | u_i \geq -r(W_i)) = h(W_i) \implies \mathrm{E}(Y_i^* | X_i, T_i, W_i, Z_i = 1) = \alpha T_i + g(X_i) + h(W_i)$$

- **Проблема** – если ε_i и u_i зависимы, то функция $h(W_i) \neq 0$ является пропущенной переменной, что приведет к несостоятельности DML оценки $\hat{\alpha}$.
- **Решение** – если T_i не входит в W_i, то можно объединить переменные X_i и W_i, получив регрессионное уравнение, в котором α можно оценить DML методом:

$$Y_i = \alpha T_i + g^*(X_i^*) + v_i, \text{ где } g^*(X_i^*) = g(X_i) + h(W_i) \text{ и } X_i^* = (X_i, W_i)$$

Двойное машинное обучение (DML)

Несколько структурных параметров

- **Проблема** – иногда исследователю необходимо оценить не один, а сразу несколько структурных параметров a_j, где $j \in \{1, ..., d\}$.

$$Y_i = \alpha_1 T_{1i} + ... + \alpha_d T_{di} + g(X_i) + \varepsilon_i$$

- Например, параметры α_j могут отражать отдачу от различных уровней образования: базовый, бакалавриат и магистратура.
- **Решение** – оценить каждый из параметров α_j поочередно, используя DML метод для следующего уравнения:

$$Y_i = \alpha_j T_{ji} + g_j(X_i, T_{1i}, ..., T_{(j-1)i}, T_{(j+1)i}, ..., T_{di}) + \varepsilon_i$$

- Для тестирования гипотез о связи между параметрами α_j можно применить бутстрап.