МФТИ, 2026/2027

Прикладная статистика

Лекции

1 сентября 2026 г.

Содержание

1 Базовые понятия и асимптотические методы в математической статистике

1.1 Введение

В теории вероятностей предметом исследования являлись модели случайных экспериментов, позволяющие понять, как будут вести себя (как распределены) те или иные характеристики, которые можно наблюдать в результате проведения этих экспериментов. В математической статистике предметом изучения являются обратные задачи теории вероятностей, в которых анализируются экспериментальные данные и требуется сделать вывод о природе (модели) рассматриваемого явления, либо принять некоторое решение на основе этого вывода. Сразу отметим, что под статистическими выводами, или инференцией (от англ.: to infer – делать выводы) обычно подразумевается перенос некоторых характеристик с частного – набора наблюдаемых экспериментальных данных, на общее – все возможные значения результатов рассматриваемого эксперимента, в том числе ненаблюдаемые. Таким образом, математическую статистику можно считать формализованной наукой о познании, заключающемся в выделении закономерностей в случайном.

Исторически, основной практической задачей статистики считались вопросы снижения размерности данных[1] – то есть построение характеристик, максимально полно и точно описывающих определённые аспекты рассматриваемых наблюдений. Позже, новые задачи пришли в статистику из междисциплинарных областей, таких как “Теория принятия решений”, “Машинное обучение”, возникших благодаря развитию вычислительной техники и доступности больших объемов данных, и, как правило, связаных с темой искусственного интеллекта. Эти задачи включают в себя поиск статистических и функциональных зависимостей в данных, построение прогнозов и оценку причинно-следственных связей. Помимо расширения диапазона рассматриваемых задач, накопленный опыт исследований привел к изменению походов к решению классических задач статистики. Одно из существенных изменений связано с развитием концепции робастных оценок, устойчивых к неточностям в предположениях моделей и ошибкам в данных. Устойчивые методы статистических выводов, как правило, используют аппроксимацию с помощью предельных распределений и другие асимптотические методы. Проблемы робастности также имеют место при оценке зависимостей в многомерных данных – поэтому, наряду с классическими подходом, использующим линейные модели, в этом тексте рассматриваются непараметрические и полупараметрические статистические модели, а также особое внимание уделяется проблемам робастной причинно-следственной инференции.

1.2 Генеральная совокупность и выборка

Определение 1.1. (Генеральная совокупность, популяция).

Содержательно генеральную совокупность можно понимать, как множество всех потенциально возможных наблюдений в рамках изучаемой задачи. Формально можно сказать, что генеральная совокупность отождествляется с некоторой случайной величиной $\mathbf{X}$,[2] которая соответствует изучаемому явлению. Так, если $\mathbf{X}$ является доходом наугад выбранного человека из генеральной совокупности, то таким образом мы задаем распределение дохода в ней. Какое распределение выбрать для такого примера – непростой вопрос. Часто для моделирования используют логнормальное, т.е. $\mathbf{X} = e^{\mathrm{Y}}, \mathrm{Y} \sim \mathcal{N}(m, \sigma^2)$.

По определению случайной величины $\mathbf{X}$, это измеримое отображение из некоторого вероятностного про-

[1] Fisher R. (1922). *On the mathematical foundations of theoretical statistics.* Philosophical Transactions of the Royal Society of London, Series A 222, p. 311.

[2] Выделение шрифтом здесь используется исключительно для того, чтобы далее использовать “обычный” символ X преимущественно для обозначения другого объекта в контексте статистических моделей.

странства, порождающее индуцированное пространство

$$\big(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathrm{P}\big),$$

где $\mathcal{B}(\mathbb{R})$ – борелевская сигма-алгебра на вещественной прямой, а P – вероятностная мера на $\mathcal{B}(\mathbb{R})$. Статистическая задача предполагает, что распределение P нам неизвестно, но в то же время обычно делается некоторое изначальное ограничение на его вид. Таким образом, статистическая модель состоит из множества распределений

$$\{\mathrm{P}_\theta,\ \theta \in \Theta\},$$

где θ – параметр этого распределения – то есть, некоторый функционал от P, а Θ – множество его возможных значений. В приведённом выше примере статистической модели для доходов людей в некоторой популяции, параметром является вектор (m, σ^2), а в качестве Θ можно взять $\mathbb{R} \times \mathbb{R}_+$. В более общем случае θ может быть бесконечномерным – например, для той же задачи θ может быть функцией распределения для P, а Θ – множеством ф.р. неотрицательных распределений с фиксированным значенем медианы.

Если в задаче исследуется генеральная совокупность двух или более числовых характеристик, то она отождествляется с распределением некоторого случайного вектора. В ряде случаев в такой ситуации дополнительно накладывается структура зависимостей между переменными в векторе (например, с помощью ориентированного графа). Получающаяся модель включает направленные связи между случайными величинами, а также предположения об их распределениях.[3] В дальнейшем такая модель будет называться процессом, порождающим данные (англ.: data generating process, DGP).

Определение 1.2. Случайная выборка объёма n – это совокупность наблюдений $X_1, X_2, \ldots, X_n$, взятых из генеральной совокупности с помощью процедуры простого случайного выбора, согласно которой любой набор объёма n имеет одинаковый шанс попасть в выборку. Формально, каждый элемент X_i выборки – случайная величина/вектор, имеющий такое же распределение, что и генеральная совокупность $\mathbf{X}$.

Замечание и обозначения. Термин "выборка" имеет два значения:

1. Набор случайных величин $X_1, X_2, \ldots, X_n$ до того, как эксперимент был произведён;

2. Набор полученных значений этих величин $x_1, x_2, \ldots, x_n$ после того, как эксперимент был произведён.

Иными словами, X_1, X_2 – выборка объёма два, $x_1 = 3, x_2 = 10$ – реализация выборки. Поэтому, например, выражение $\mathbb{E}X_1$ имеет содержательный смысл, а $\mathbb{E}x_1$ – нет.

Кроме того, нам иногда требуется обозначать через $x = (x_1, \ldots, x_n)^\top$ точку в $\mathbb{R}^n$. Аналогичной конструкцией до реализации будет $X = (X_1, \ldots, X_n)^\top$ – случайный вектор размера n, также называемый вектором выборки.

Определение 1.3. Статистическая модель $\{\mathrm{P}_\theta, \theta \in \Theta\}$ называется идентифицированной, если теоретически возможно по бесконечно большой выборке с распределением P_{θ_0} установить истинное значение θ_0. Это эквивалентно тому, что разные значения $\theta \in \Theta$ должны давать разные распределения наблюдаемых характеристик.

Пример 1.4. Параметр $\theta \in [0, 2\pi)$ статистической модели $\mathbf{X} \sim \text{Bernoulli}\left(\frac{1}{2}(1 + \sin\theta)\right)$, возникающей в решении одной из задач квантовой механики, не идентифицируем – поскольку почти всем значениям $\frac{1}{2}(1 + \sin\theta)$ соответствуют два различных прообраза θ.

[3] См. например, Pearl, J., M. Glymour, N. Jewell. (2016). *Causal inference in statistics*, New Jersey: Wiley, p. 35.

Пример 1.5. Можно ли идентифицировать параметр $\theta \in (-\infty, 1)$ статистической модели $\mathbf{X} \sim \mathrm{U}[\theta, 1]$, если наблюдаются только значения $\{X_i^2\}_{i=1}^n$?

Определение 1.6. Выборочные статистики, или выборочные характеристики – это некоторые (измеримые) функции выборки. Оценкой (параметра распределения) может называться произвольная выборочная статистика; хорошие оценки могут обладать дополнительными свойствами, такими как состоятельность, несмещенность, асимптотическая нормальность и др.

Некоторые часто используемые выборочные статистики по выборке $X = (X_1, \ldots, X_n)$:

- Выборочное среднее $\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$;
- Выборочная дисперсия $S^2 = \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \bar{X}\right)^2$.
- Эмпирическая (выборочная) функция распределения (э.ф.р.):

$$\hat{F}(x) = \frac{\sum_{i=1}^n \mathbb{1}\{X_i \le x\}}{n}. \tag{1}$$

  Каждому числу $x \in \mathbb{R}$ эта функция сопоставляет долю наблюдений в выборке, оказавшихся слева от этого числа.

Теорема 1.7. *(Гливенко-Кантелли). Эмпирическая ф.р. $\hat{F}(x)$ почти наверное сходится к истинной ф.р. $F_0(x)$ в равномерной метрике:*

$$\sup_{x \in \mathbb{R}} \left|\hat{F}(x) - F_0(x)\right| \xrightarrow[n \to \infty]{\text{п.н.}} 0.$$

1.3 Принцип аналогии

Теорема Гливенко-Кантелли показывает, что основная задача математической статистики (восстановление неизвестного распределения) вполне обоснована: мы можем равномерно по всем точкам на прямой восстановить истинную функцию распределения $\mathbf{X}$.

С помощью теоремы Гливенко-Кантелли можно доказывать сходимость выборочных характеристик к популяционным, записывая их в виде различных функционалов от э.ф.р. Так, среднее по реализовавшейся выборке $\bar{x}$ можно представить в виде интеграла Лебега-Стилтьеса от $\hat{F}(x)$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i =: \int_{\mathbb{R}} x \, d\hat{F}(x),$$

а выборочную медиану – как функционал $T(\hat{F})$, где $T(G) = \inf\{1, x : G(x) \ge 0.5\}$.

Заменяя в этих случаях $\hat{F}(x)$ на истинную функцию распределения $F_0(x)$, можно получить популяционные аналоги рассмотренных величин. Такой подход, известный под названием *принцип аналогии*, даёт универсальный метод нахождения и исследования выборочных конструкций, близких к оцениваемым популяционным параметрам: если интересующий исследователя параметр θ можно выразить в виде функционала T от функции распределения, то есть $\theta_0 = T(F_0)$, то аналогичная конструкция от э.ф.р. $\hat{\theta} := T(\hat{F})$ будет сходиться к θ_0 с ростом размера выборки при условии непрерывности T в точке F_0.[4]

[4] Более того, при достаточно мягких ограничениях на T, главным из которых является дифференцируемость по Адамару в точке F_0 – то есть, возможность хорошего приближения T линейным функционалом на компактных окрестностях этой точки, оказывается, что $T(\hat{F})$ после подходящей нормировки сходится при $n \to \infty$ по распределению к нормальной случайной величине.

1.4 Некоторые предельные теоремы теории вероятностей

В курсе статистики мы часто будем сталкиваться с необходимостью найти распределение случайной величины, являющейся функцией от набора из n других случайных величин. В большинстве случаев такая задача слишком сложна для аналитического решения, поэтому широко используются асимптотические (предельные) методы, в которых неизвестное истинное распределение приближается с помощью более просто устроенного асимптотического (при $n \to \infty$) распределения, которое во многих практически важных случаях можно найти благодаря центральной предельной теореме. По этой причине полезно вспомнить свойства основных видов сходимости случайных величин и векторов.

Определение 1.8. Последовательность случайных величин или векторов X_n сходится к X почти наверное, если

$$\mathrm{P}\big(\omega : X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega)\big) = 1.$$

Определение 1.9. Последовательность $\{X_n\}$ сходится к X по вероятности, если $\forall \varepsilon > 0$

$$\mathrm{P}\big(\|X_n - X\| > \varepsilon\big) \xrightarrow[n\to\infty]{} 0.$$

Определение 1.10. Последовательность $\{X_n\}$ сходится к X в L^2, если выполняется

$$\mathbb{E}(X_n - X)^2 \xrightarrow[n\to\infty]{} 0.$$

По определению дисперсии, это эквивалентно тому, что $\mathbb{E}(X_n - X) \xrightarrow[n\to\infty]{} 0$ и $\mathbb{V}(X_n - X) \xrightarrow[n\to\infty]{} 0$.

Определение 1.11. Последовательность $\{X_n\}$ сходится к X по распределению, если для любой ограниченной непрерывной числовой функции f (соответствующей размерности X_n и X) выполняется

$$\mathbb{E}f(X_n) \xrightarrow[n\to\infty]{} \mathbb{E}f(X).$$

Для сходимости по распределению имеется большое число эквивалентных характеризаций: например, можно показать, что $X_n \xrightarrow{d} X$ тогда и только тогда, когда имеет место поточечная сходимость функций распределения

$$F_{X_n}(x) \xrightarrow[n\to\infty]{} F_X(x),$$

для всех x, являющихся точками непрерывности F_X. Более подробно этот вид сходимости рассматривается в теореме Александрова (см. приложение).

Теорема 1.12. *(Связь различных видов сходимостей случайных величин и векторов).*

$$\textit{из } X_n \xrightarrow{\text{п.н.}} X \textit{ следует } X_n \xrightarrow{p} X,$$

$$\textit{из } X_n \xrightarrow{L^2} X \textit{ следует } X_n \xrightarrow{p} X,$$

$$\textit{из } X_n \xrightarrow{p} X \textit{ следует } X_n \xrightarrow{d} X,$$

причём если $X_n \xrightarrow{d} C$, где C – константа или детерминистический вектор, то выполняется также и $X_n \xrightarrow{p} C$.

Теорема 1.13. *(Центральная предельная теорема в форме Леви).*

Если $\{X_i\}_{i=1}^n$ *– выборка н.о.р. случайных величин, причём* $\mathbb{V}X_1 = \sigma^2 < \infty$*, то*

$$Z_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \xrightarrow{d} \mathcal{N}(0, \sigma^2). \tag{2}$$

Аналогичная теорема верна и для X_i – случайных векторов.

Теорема 1.14. *(Неравенство Берри-Эссеена). Пусть* $F_n(x)$ *– функция распределения случайной величины* Z_n *из* (2)*, и* $\Phi(x)$ *– функция распределения стандартной нормальной величины. Неравенство Берри-Эссеена позволяет оценить (сверху) максимально возможное отклонение между* $F_n(x)$ *и* $\Phi(x)$*:*

$$\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \leq C \frac{\mathbb{E}\,|X_i - \mathbb{E}X_i|^3}{\sigma^3 \sqrt{n}},$$

где C *– некоторая числовая константа, значение которой продолжает уточняться. По последним данным* $\frac{1}{\sqrt{2\pi}} \leq C \leq 0.4784$.

Пример 1.15. Пусть $X \sim \text{Pois}(\lambda)$. Найдите погрешность приближения X нормальной случайной величиной.

Решение. Известно, что суммы независимых пуассоновских случайных величин также имеет распределение Пуассона. Пусть $X_i \sim \text{Pois}(\lambda_i)$, $i \in \{1, \ldots, n\}$ – набор *н.о.р.* случайных величин, таких, что $\lambda_i = \lambda/n$. Тогда $X \overset{d}{=} \sum_{i=1}^{n} X_i$, что даёт основания предположить, что $X \overset{d}{\approx} \mathcal{N}(\lambda, \lambda)$ в силу ЦПТ. С ростом значения параметра λ, нормальное приближение $\tilde{X} \sim \mathcal{N}(\lambda, \lambda)$ к пуассоновской случайной величине $X \sim \text{Pois}(\lambda)$ становится точнее. На рис. 1 представлен график значений ограничений сверху для $\sup_{x \in \mathbb{R}} |F_X(x) - F_{\tilde{X}}(x)|$, рассчитанных с помощью неравенства Берри-Эссеена для разных значений λ.

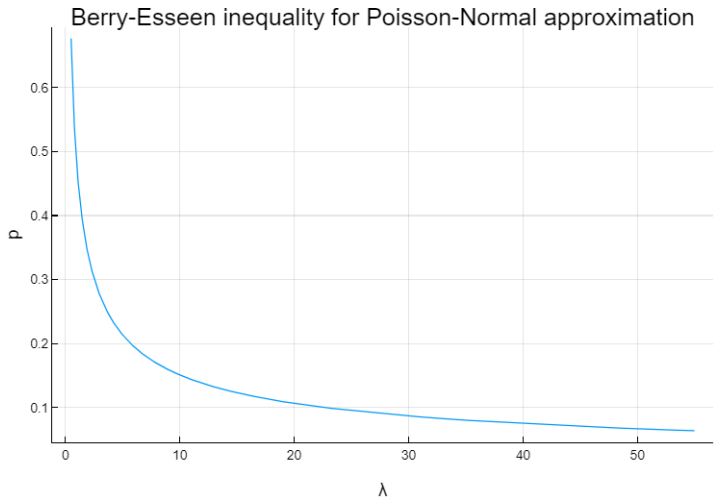


Рис. 1: Значения $C\frac{\mathbb{E}|X_i - \mathbb{E}X_i|^3}{\sigma_i^3 \sqrt{n}}$ для нормального приближения к $X \sim \text{Pois}(\lambda)$, в зависимости от λ.

Теорема 1.16. *(Теорема Манна-Вальда о непрерывных отображениях).*

Пусть X_n *– последовательность случайных величин или векторов, у которой существует предел* X*. Пусть* g *– непрерывная функция* $\mathbb{R} \to \mathbb{R}$ *или (для векторного случая)* $\mathbb{R}^k \to \mathbb{R}$*.*

$$\textit{Из } X_n \xrightarrow{d} X \textit{ следует } g(X_n) \xrightarrow{d} g(X).$$

Доказательство. Сходимость по распределению эквивалентна тому, что для любой непрерывной ограни-

ченной функции f выполняется

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)], \tag{3}$$

Поскольку $g(\cdot)$ – непрерывная функция, то композиция $f\big(g(\cdot)\big) \equiv f \circ g$ является непрерывной и ограниченной, а значит должно выполняться

$$\mathbb{E}\big[f\big(g(X_n)\big)\big] = \mathbb{E}\big[f\big(g(X)\big)\big],$$

что эквивалентно $g(X_n) \xrightarrow{d} g(X)$. □

Замечание. Условие теоремы Манна-Вальда можно ослабить, допуская разрывность функции g на множестве точек $D_g : \mathrm{P}(X \in D_g) = 0$.

Замечание. Аналогичные утверждения верны и для других видов сходимостей: если g – непрерывная функция, то

$$\text{из } X_n \xrightarrow{p} X \text{ следует } g(X_n) \xrightarrow{p} g(X),$$

$$\text{из } X_n \xrightarrow{L^2} X \text{ следует } g(X_n) \xrightarrow{L^2} g(X),$$

$$\text{из } X_n \xrightarrow{\text{п.н.}} X \text{ следует } g(X_n) \xrightarrow{\text{п.н.}} g(X).$$

Следствие 1.17. *Пусть* $(X_n, Y_n) \xrightarrow{d} (X, Y)$. *Тогда* $X_n + Y_n \xrightarrow{d} X + Y$.

Доказательство. Пусть $f : \mathbb{R} \to \mathbb{R}$ – произвольная непрерывная ограниченная функция. Заметим, что функция $h : \mathbb{R}^2 \to \mathbb{R}$, имеющая вид $h(x, y) = x + y$, является непрерывной, поэтому композиция $f \circ h : \mathbb{R}^2 \to \mathbb{R}$ является непрерывной ограниченной функцией.

Тогда

$$\lim_{n\to\infty} \mathbb{E}f\big(h\,(X_n, Y_n)\big) = \mathbb{E}f\big(h(X, Y)\big)$$

вследствие того, что $(X_n, Y_n) \xrightarrow{d} (X, Y)$.

Это аналогично утверждению $\lim_{n\to\infty} \mathbb{E}[f\,(X_n + Y_n)] = \mathbb{E}[f(X + Y)]$. Поскольку доказанное верно для любой непрерывной ограниченной функции $f : \mathbb{R} \to \mathbb{R}$, то можно сделать вывод, что $X_n + Y_n \xrightarrow{d} X + Y$. □

Замечание. В доказанном следствии сходимость векторов по распределению здесь нельзя заменить на сходимость отдельных случайных величин, поскольку из $X_n \xrightarrow{d} X$ и $Y_n \xrightarrow{d} Y$ не следует $(X_n, Y_n) \xrightarrow{d} (X, Y)$.

Замечание. Покоординатной сходимости по распределению недостаточно для сходимости последовательности векторов. Интересно, что для других видов сходимости у случайных величин, в частности – сходимости по вероятности и почти наверное, это не так: для них покоординатной сходимости компонент вектора достаточно для того, чтобы последовательность случайных векторов также имела предел – то есть, например,

$$\text{из } X_n \xrightarrow{p} X \text{ и } Y_n \xrightarrow{p} Y \text{ следует } (X_n, Y_n) \xrightarrow{p} (X, Y),$$

$$\text{из } X_n \xrightarrow{\text{п.н.}} X \text{ и } Y_n \xrightarrow{\text{п.н.}} Y \text{ следует } (X_n, Y_n) \xrightarrow{\text{п.н.}} (X, Y),$$

$$\text{из } X_n \xrightarrow{L^2} X \text{ и } Y_n \xrightarrow{L^2} Y \text{ следует } (X_n, Y_n) \xrightarrow{L^2} (X, Y).$$

Несмотря на указанное свойство сходимости по распределению, существует важный частный случай – когда одна из координат вектора сходится по распределению к константе.

Пусть $\{X_n\}, \{Y_n\}$ – последовательности случайных величин, причём

$$X_n \xrightarrow{d} X, \quad Y_n \xrightarrow{d} C,$$

где X – случайная величина, а C – константа. Тогда $(X_n, Y_n) \xrightarrow{d} (X, C)$.

Доказательство этого утверждения можно найти в книге Van der Vaart A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. pp. 10-11.

Следствие 1.18. *(лемма Слуцкого).*

Пусть даны две последовательности $\{X_n\}$ и $\{Y_n\}$, для которых выполняется $X_n \xrightarrow{p} C$, где C – константа, и $Y_n \xrightarrow{d} Y$. Тогда

$$X_n + Y_n \xrightarrow{d} C + Y,$$

$$X_n Y_n \xrightarrow{d} CY.$$

Доказательство. Из $X_n \xrightarrow{p} C$ следует $X_n \xrightarrow{d} C$; кроме того, вследствие рассмотренного выше утверждения выполняется $(X_n, Y_n) \xrightarrow{d} (C, Y)$. Тогда утверждения теоремы Слуцкого следуют из теоремы Манна-Вальда и того факта, что отображения $(x, y) \to x + y$ и $(x, y) \to x \cdot y$ являются непрерывными на $\mathbb{R}^2$.

Следствие 1.19. *Асимптотически нормальная оценка является состоятельной.*

Доказательство. Пусть $\hat{\theta}$ – асимптотически нормальная оценка θ, т.е. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V_\theta)$. Обозначим $A_n := \sqrt{n}(\hat{\theta} - \theta)$ и $B_n := \frac{1}{\sqrt{n}}$. Поскольку B_n сходится к нулю, как числовая последовательность, то можно применить теорему Слуцкого, согласно которой $A_n B_n$ имеет предел по распределению, также равный нулю. Но так как $A_n B_n = \hat{\theta} - \theta$, и из сходимости к константе (нулю) по распределению следует сходимость к константе по вероятности, то мы получили $\hat{\theta} - \theta \xrightarrow{p} 0$, откуда следует $\hat{\theta} \xrightarrow{p} \theta$. □

Следствие 1.20. *Несостоятельная оценка не может быть асимптотически нормальной для оцениваемого параметра.*

Пример 1.21. Пусть $\{X_i\}_{i=1}^n$ – случайная выборка из распределения, имеющего конечную дисперсию σ^2 и математическое ожидание μ. Тогда $\bar{X}$ является асимптотически нормальной оценкой μ. Действительно, пусть

$$Z_n := \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}.$$

Тогда $\mathbb{E}Z_n = 0, \mathbb{V}Z_n = 1$, и, поскольку Z_n является суммой независимых одинаково распределённых случайных величин, то условия ЦПТ выполнены, и $Z_n \xrightarrow{d} \mathcal{N}(0, 1)$. Но тогда $\sigma Z_n = \sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, то есть, $\bar{X}$ – асимптотически нормальная оценка μ с асимптотической дисперсией σ^2.

1.5 Дельта-метод

Асимптотически нормальные оценки являются наиболее важным инструментом для статистических выводов (инференции) в практических задачах. Если для некоторого параметра найдена асимптотически нормальная оценка, то существует метод, который позволяет находить асимптотическое распределение у функций от такой оценки; при этом в подавляющем большинстве случаев это распределение тоже будет нормальным.

Лемма 1.22. *Пусть $\hat{\theta}_n$ – асимптотически нормальная оценка (одномерного) параметра θ, построенная по выборке размером n, то есть*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$$

Тогда для любой функции g, такой, что $g'(\theta)$ существует, принимает ненулевые значения и непрерывна в некоторой окрестности θ, выполняется

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\left(0, \big(g'(\theta)\big)^2 \sigma^2\right).$$

Этот результат называется дельта-методом первого порядка, или просто дельта-методом для $\hat{\theta}$ *и* $g(\hat{\theta})$.

Доказательство. По формуле Лагранжа: $g(\hat{\theta}_n) = g(\theta) + g'(\tilde{\theta})(\hat{\theta}_n - \theta)$, где $\tilde{\theta}$ лежит между $\hat{\theta}_n$ и θ.

Поскольку $\hat{\theta}_n \xrightarrow{p} \theta$, то $\tilde{\theta} \xrightarrow{p} \theta$,[5] и поскольку $g'(\theta)$ непрерывна, применение теоремы о непрерывном отображении даёт

$$g'(\tilde{\theta}) \xrightarrow{p} g'(\theta).$$

Умножая выражение в разложении Лагранжа на $\sqrt{n}$, получаем

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) = g'(\tilde{\theta})\sqrt{n}(\hat{\theta}_n - \theta).$$

Так как $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$ по предположению, то правая часть выражения по теореме Слуцкого сходится по распределению к произведению $g'(\theta)$ и нормальной случайной величины $\mathcal{N}(0, \sigma^2)$, что даёт $\mathcal{N}\Big(0, \big(g'(\theta)\big)^2\sigma^2\Big)$, поэтому

$$\sqrt{n}\big(g(\hat{\theta}_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\Big(0, \big(g'(\theta)\sigma^2\big)^2\Big).$$

□

Аналогично доказывается многомерный дельта-метод: если $\theta \in \mathbb{R}^k$, $\hat{\theta}_n$ – асимптотически нормальная оценка, то есть

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right),$$

где Σ – $k \times k$ асимптотическая ковариационная матрица, то если $h : \mathbb{R}^k \to \mathbb{R}^m$ – дифференциируемое отображение, непрерывное в окрестности θ, и $m \times k$ матрица первых производных $J_\theta := \frac{\partial h(x)}{\partial x^\top}\Big|_{x=\theta}$ имеет полный ранг, то

$$\sqrt{n}\big(h(\hat{\theta}_n) - h(\theta)\big) \xrightarrow{d} \mathcal{N}\left(0, J_\theta \Sigma J_\theta^\top\right).$$

Условие полного ранга J_θ аналогично условию $g'(\theta) \neq 0$ для одномерного случая. Если оно не выполнено, то нужно брать компоненту более высокого порядка в разложении функции в ряд Тейлора. Если в одномерном случае $g'(\theta) = 0$ но $g''(\theta) \neq 0$, то $g(\hat{\theta})$ после подходящего преобразования будет иметь асимптотическое распределение $\chi^2(1)$.

[5] Формальное доказательство: поскольку $|\tilde{\theta} - \theta| \leq |\hat{\theta} - \theta|$, то $\forall \varepsilon > 0\ \mathrm{P}(|\tilde{\theta} - \theta| \geq \varepsilon) \leq \mathrm{P}(|\hat{\theta} - \theta| \geq \varepsilon) \xrightarrow[n\to\infty]{} 0$.

2 Оценка причинно-следственных связей

2.1 Введение

Выявление и оценка причинно-следственных зависимостей является важнейшей составляющей большинства задач в области здравоохранения, общественных и поведенческих наук. Эта тема долгое время являлась периферийной для математической статистики и развивалась, преимущественно, экономистами: так, например, в 2021 году Г. Имбенс и Дж. Ангрист получили Нобелевскую премию по экономике за развитие методов причинно-следственного анализа. Такое положение дел начало меняться в конце XX-го века с появлением работ, исследующих теоретические основы причинно-следственной идентификации, в частности, с использованием графических моделей. В настоящее время происходит постепенный синтез подходов к оценке эффектов воздействия, разработанных в различных областях науки – эконометрики, машинного обучения и математической статистики. Количество исследований, связанных с разработкой методологии и статистической оценкой эффектов воздействия с использованием неэкспериментальных данных существенно выросло за последние 10 лет, причём некоторые исследователи подчеркивают важность этой задачи для построения моделей сильного искусственного интеллекта.

Понятию причинно-следственной связи, несмотря на его распространённость и естественность, достаточно сложно дать формальное определение. Основоположник эконометрики Р. Фриш утверждал, что причинно-следственность является феноменом, относящимся к человеческому способу мышления, а не к объективной реальности. Другие авторы связывают причинно-следственность с понятием контрфактуальности – то есть, ответом на вопрос «Что было бы, если?» применительно к некоторой задаче, а также с направленными (односторонними) зависимостями между переменными. Строить статистические оценки направленных зависимостей помогает принцип общей причинности Г. Райхенбаха,[6] который утверждает, что если два случайных события A и B статистически зависимы, то либо A вызывает (является причиной) B, либо B вызывает A, либо у A и B имеется общая причина, – и в таком случае при обуславливании на неё (её удалении), эти события становятся независимыми. Таким образом, фундаментально, статистическая оценка направленных зависимостей отличается от, например, оценки корелляции тем, что исследователю дополнительно нужно знать структуру и направление связей между исследуемыми переменными. При этом во многих практических задачах ситуация осложняется множественностью зависимостей: так, A может являться причиной B, но в то же время оба эти события (частично) могут являться следствием влияния набора дополнительных факторов X, что может привести к существенным искажениям при оценке влияния A на B.

Большинство ранних работ по статистической оценке эффектов воздействия связаны с рандомизированными контролируемыми испытаниями, в которых объекты случайно разделяются на две или более группы, отличающиеся наличием или уровнем воздействия – что позволяет минимизировать влияние посторонних факторов и более точно оценить влияние воздействия на интересующие исследователя характеристики. В то же время, для многих практически важных задач проведение рандомизированных экспериментов недоступно, а любой набор неэкспериментальных данных (англ.: observational data), является неполным – поскольку для каждого исследуемого объекта наблюдаемым является лишь одно состояние, тогда как другие возможные лишь постулируются. Невозможность сравнения контрфактуальных характеристик напрямую приводит к тому, что исследователям требуется дополнять данные посредством теоретических предположений, которые позволяли бы извлечь информацию из имеющихся наблюдений; в работе П. Холланда это заключение было названо фундаментальной проблемой выводов (инференции) о причинно-следственных связях.

[6] Reichenbach, H. (1956). *The Direction of Time*. Edited by Maria Reichenbach. Berkeley and Los Angeles: University of California Press.

2.2 Модель потенциальных исходов Рубина

Исторически, первой из статистических моделей, применимых для идентификации и оценки эффектов воздействия при анализе данных, является модель, названная в честь автора – Д. Рубина, предложившего её в 1969 году; однако, некоторые элементы использованного подхода применительно к случаю рандомизированных контролируемых экспериментов были предложены ещё в 1923 году Е. Нейманом.

В модели Рубина для каждого объекта рассматриваемой популяции постулируется существование спектра различных состояний, описываемых переменной воздействия. Это позволяет говорить о потенциальных исходах зависимой переменной и в этих терминах определять причинно-следственные зависимости. Когда состояний всего два, то их традиционно называют контрольным состоянием и воздействием. Если состояний больше двух, то одно из них назначается контрольным, а остальные расцениваются как различные уровни воздействия, либо как альтернативные варианты воздействий – если сравнивать их интенсивность некорректно.

На практике не всегда существует возможность специфицировать состояния объектов достаточно точно – в силу ограничений, непосредственно связанных с аккуратной формулировкой самих состояний, либо из-за сложности последующего сбора данных о них. В то же время, от этого этапа построения модели зависит интерпретация оценок, а также то, насколько получаемые результаты могут быть распространены с выборки на всю моделируемую популяцию.

Под потенциальными исходами в модели Рубина понимаются числовые характеристики, возможно случайные – причём по одной величине приписывается каждой паре из объекта наблюдения и его состояния. В простейшем случае, когда состояний всего два, для i-го объекта популяции их можно обозначить за Y_{i1} и Y_{i0}. Иногда представляется целесообразным, особенно для рассмотрения общих понятий, сопоставить потенциальные исходы со случайным процессом $Y = Y(\omega, d)$, где $d \in \{0, 1\}$ соответствует отсутствию/наличию воздействия, а $\omega \in \Omega$ – состояние мира, отражающее случайность в выборе конкретного объекта, и случайность, связанную с любыми другими возможными причинами. Если потенциальных исходов всего два, то можно ввести индикатор воздействия $D_i : \Omega \to d_i$, который определяет, в каком из состояний находится интересующий нас объект наблюдения. Таким образом получается набор случайных величин $Y(D_i)$ – потенциальные исходы для i-го объекта при случайном уровне воздействия D_i. Индивидуальный эффект воздействия можно определить, как разность между потенциальными исходами

$$\delta_i = Y_{i1} - Y_{i0}, \tag{4}$$

и, поскольку δ_i в такой постановке может быть случайной величиной, то практически всегда рассматривается лишь её усреднённое значение

$$\delta_i^E = \mathbb{E}Y_{i1} - \mathbb{E}Y_{i0}. \tag{5}$$

Заметим, что нет оснований определять эффект воздействия исключительно как разность двух величин. Например, иногда исследователя может интересовать процентное различие между двумя величинами, и тогда можно рассматривать эффект воздействия как частное $\left(\frac{Y_{i1}}{Y_{i0}} - 1\right) \times 100\%$ (в таких случаях часто используется термин *аплифт* от англ.: uplift – подъём); также возможны и другие функциональные формы.

Если ввести переменную Y_i, обозначающую фактически наблюдаемый потенциальный исход для i-го объекта, то для контрольной группы данная переменная будет принимать одно из возможных значений Y_{i0}, а для экспериментальной – одно из возможных значений Y_{i1}. Формально это можно записать в виде уравнения (6):

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0} \tag{6}$$

Одним из возможных решений фундаментальной проблемы причинно-следственной инференции является агрегация эффектов воздействия для отдельных элементов популяции в один средний эффект, при этом можно попытаться использовать информацию об объектах из экспериментальной группы для выводов о $\mathbb{E}Y_{i1}$ и, аналогично, объекты из контрольной группы можно использовать для оценки популяционного значения $\mathbb{E}Y_{i0}$.

Рубин в своей работе ввёл идентификационное предположение об устойчивости индивидуальной величины воздействия – набор требований, достаточных для того, чтобы считать среднее по подвыборкам в группе воздействия и в контрольной группе несмещёнными оценками соответствующих популяционных значений.

Устойчивость индивидуальной величины воздействия:

1. Эффект воздействия на один элемент популяции не зависит от значений воздействия у других элементов популяции;

2. Эффект воздействия на один элемент популяции не зависит от механизма получения этого воздействия, то есть, наблюдаемое значение случайной величины Y_i соответствует потенциальному исходу $Y(D_i)$.

Первую часть этого предположения можно интерпретировать как отсутствие экстерналий (внешних эффектов). Вторая компонента предположения постулирует ключевое свойство контрфактуальных состояний, необходимое для разрешения фундаментальной проблемы инференции – благодаря ему средние по подвыборкам Y_i с разными уровнями воздействия становятся состоятельными оценками соответствующих $\mathbb{E}Y(d)$. В некоторых работах это свойство называется состоятельностью (англ.: consistency), а в эконометрической литературе распространено близкое по смыслу понятие экзогенности. Оптимизационное поведение исследуемых объектов нередко ведёт к нарушению данного предположения в ситуациях, когда они самостоятельно отбираются для того, чтобы подвергнуться воздействию. Такие же проблемы можно наблюдать в случаях, когда воздействие осуществляется на основе тех или иных характеристик объектов изучения.

Предложенный подход можно проиллюстрировать следующим примером из книги Ангриста и Пишке.[7]

Пример 2.1. Исследуйте влияние госпитализации на здоровье людей, используя данные опроса населения США The National Health Interview Survey, в котором включались вопросы, “Был ли госпитализирован опрашиваемый в течение последних 12 месяцев”, и “Как вы оцениваете своё здоровье по шкале от 1 (отличное) до 5 (плохое)”.

Таблица 1: Описательная статистика показателя самооценки уровня здоровья

Группа	Размер выборки	Среднее значение	Стандартная ошибка
Воздействие	7774	2.79	0.014
Контрольная	90049	2.07	0.003

Из данных в таблице следует, что опрошенные после госпитализации имеют в среднем существенно худшие показатели здоровья. Однако, с другой стороны, наверняка такие люди изначально обладали существенно худшим здоровьем, чем те, кто не был госпитализирован.

Обозначая за Y_{i0} уровень здоровья человека, не подвергнутого госпитализации, а Y_{i1} – уровень здоровья того же человека после госпитализации, можно получить следующее представление для среднего эффекта воздействия:

$$\underbrace{\mathbb{E}\left(Y_i \mid D_i = 1\right) - \mathbb{E}\left(Y_i \mid D_i = 0\right)}_{\text{наблюдаемая разность}} = \underbrace{\mathbb{E}\left(Y_{i1} \mid D_i = 1\right) - \mathbb{E}\left(Y_{i0} \mid D_i = 1\right)}_{\text{ATT}} +$$

[7] Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

$$+\underbrace{\mathbb{E}\left(Y_{i0} \mid D_i = 1\right) - \mathbb{E}\left(Y_{i0} \mid D_i = 0\right)}_{\text{смещение выборки}}$$

Нас интересует первый член в правой части равенства – ATT (от англ. average treatment at treated – средний эффект воздействия в части популяции подвергшейся воздействию) – поскольку он является наиболее подходящим доступным приближением среднего эффекта воздействия по всей популяции ATE (от англ. average treatment effect). При этом, наблюдаемая разность также содержит компоненту, интерпретируемую как смещение выборки, – то есть среднее различие здоровья между теми, кто был, и кто не был госпитализирован. В случаях, когда воздействие назначается случайно (включая рандомизированные контролируемые испытания), переменная D_i независима от всех других переменных модели (экзогенна); в частности, независимость с потенциальными исходами можно обозначить следующим способом:

$$D_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \text{ для всех } i.$$

Это условие, называемое безусловной независимостью (или игнорируемостью), гарантирует, что сравниваемые группы сопоставимы по всем наблюдаемым и ненаблюдаемым характеристикам, кроме самого воздействия. Тогда наблюдаемые средние в каждой группе являются несмещёнными оценками соответствующих математических ожиданий потенциальных исходов; кроме того

$$\mathbb{E}(Y_{i0} \mid D_i = 1) - \mathbb{E}(Y_{i0} \mid D_i = 0) = \mathbb{E}(Y_{i0}) - \mathbb{E}(Y_{i0}) = 0,$$

поэтому смещение выборки равно нулю – и разность средних значений имеет причинно-следственную интерпретацию. Такой подход широко используется на практике, часто под названием A/B-тестирования (сплит-тестирования). В рассмотренном же примере фундаментальные отличия между экспериментальной и контрольной группами означают, что условия устойчивости индивидуальной величины воздействия нарушены, и без дополнительной информации средний эффект воздействия не идентифицирован.

2.3 A/B тесты

A/B-тестирование – один из наиболее распространённых методов сравнения двух вариантов воздействия на некоторую целевую аудиторию, основанный на проведении рандомизированного контролируемого испытания. Название метода связано с случайным разделением испытуемых на две группы – контрольную (A) и экспериментальную (B), – причем к группе B применяется новое воздействие (изменённый дизайн вебстраницы, новый алгоритм, лекарство) и сравниваются наблюдаемые отклики с группой A, где воздействие остаётся стандартным (или отсутствует). Целью является выяснить, приводит ли нововведение к статистически значимому изменению некоторой целевой метрики (конверсия, время на сайте, доход, выживаемость и т. п.). Как уже было показано ранее, в случае A/B тестирования разность выборочных средних является несмещённой оценкой эффекта воздействия. Поэтому задача сводится к стандартной проблеме статистического вывода: проверить гипотезу $H_0 : \text{ATE} = 0$, или построить доверительный интервал для этого параметра. Метод тестирования может отличаться в зависимости от типа отклика (непрерывный, бинарный, счётный), а также от используемой метрики. В простом случае можно использовать нормальное приближение:

$$\bar{Y}_1 - \bar{Y}_0 \overset{d}{\approx} \mathcal{N}\left(\text{ATE}, \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right) \Rightarrow CI_{1-\alpha}(\text{ATE}) = \bar{Y}_1 - \bar{Y}_0 \pm q_{\alpha/2}\left(\frac{\hat{\sigma}_0^2}{n_0} + \frac{\hat{\sigma}_1^2}{n_1}\right), \tag{7}$$

где $\bar{Y}_1$ и $\bar{Y}_0$ – выборочные средние по тестовой и контрольной выборкам соответственно; n_1 и n_0 – размеры этих выборок, а σ_i^2 и $\hat{\sigma}_i^2, i \in \{0, 1\}$ – значения популяционной дисперсии и их (состоятельные) оценки. Кроме

того, используются стандартные обозначения: $1-\alpha$ – уровень доверительного интервала, $q_{\alpha/2}$ – $\frac{\alpha}{2}$-квантиль для используемого распределения (стандартного нормального в данном случае).

Пример 2.2. Найдите минимальный размер сбалансированных выборок ($n_0 = n_1 = n$) для проведения A/B-теста, при котором тест, основанный на доверительном интервале (7), имеет значимость α и мощность $1-\beta$ по отношению к альтернативе $H_a : \text{ATE} \geq \delta$, где $\delta > 0$ – так называемый минимально детектируемый эффект. Предполагается, что дисперсия в обоих выборках равна σ^2.

Использование ковариатов (контрольных переменных) в A/B тестах: хотя случайное назначение воздействия само по себе обеспечивает несмещённость оценки, включение контрольных переменных в анализ может уменьшить её дисперсию и повысить мощность тестов. Это соответствует использованию регрессионных моделей или методов сопоставления, однако в A/B-тестах часто применяют стратифицированную рандомизацию: разбивают популяцию на слои по важным признакам и внутри каждого слоя проводят отдельную рандомизацию. Это гарантирует баланс ковариат в малых выборках и снижает вариативность оценок; оценка ATE в такой ситуации берется, как взвешенное среднее по слоям.

3 Метод бутстрапа

3.1 Бутстрап для коррекции смещения оценок

Помимо использования приближений, основанных на ЦПТ, существует ещё один универсальный метод нахождения (приближенных) распределений сложных функций от выборок – так называемый бутстрап. В основе этого метода лежит принцип "матрёшки": наблюдаемая реализация выборки $x = (x_1, \dots, x_n)$ является частью исследуемой популяции с неизвестной ф.р. F_X, поэтому если сделать новую популяцию, состоящую только из точек вектора x, то выборки из неё будут отличаться от x похожим образом на то, как сам x отличается от полной популяции. Идея того, что выборочные значения данных можно использовать в роли всей популяции для оценки параметров популяционного распределения, нашла отражение в названии метода, которое происходит от англ. boot strap – петли на обуви, ухватившись за которые, барон Мюнхгаузен, по слухам, вытащил себя из болота.[8]

Рассмотрим типичную ситуацию, когда мы вычисляем оценку $\hat{\theta}_n(X)$ параметра θ, и нам требуется найти какие-то характеристики её распределения.

Предположим, что известно, что $\hat{\theta}_n(X)$ является смещённой, но состоятельной оценкой параметра θ, и пусть

$$\text{bias}_0 = \mathbb{E}\hat{\theta}_n(X) - \theta.$$

Если величина смещения известна, то скомпенсировать его не представляет труда – поэтому рассмотрим ситуацию, когда такой аналитический расчет невозможен. В этом случае можно проделать следующее рассуждение: состоятельность оценки равносильна тому, что при расчете оценки по элементам всей популяции, мы будем получать точное значение θ, а значит смещение возникает из-за пропуска в выборке некоторых возможных значений $\mathbf{X}$ (и, соответственно, избыточного веса включенных значений). Тогда, если взять подвыборку X^* из рассматриваемой выборки, то оценка, посчитанная по ней, будет тоже смещена относительно значения "по всей популяции", в роли которой теперь выступают точки исходной выборки $(x_1, \dots, x_n)$. Естественным в такой ситуации представляется усреднить наблюдаемое значение смещения по разным возможным подвыборкам из исходной выборки (которые обычно делаются с помощью процедуры случайного выбора с возвращением, имеют такой же размер n и называются бутстрап-выборками), и вычесть его из исходного значения оценки. В этом случае оценка с поправленным смещением будет иметь вид

$$\hat{\theta}_n(x) - \big(\bar{\theta}_n(X^*) - \hat{\theta}_n(x)\big) = 2\hat{\theta}_n(x) - \bar{\theta}_n(X^*),$$

где $\bar{\theta}_n(X^*)$ – среднее значение статистики на бутстрап-выборках.

Схожий алгоритм можно применить для поправки доверительных интервалов. В базовом варианте, достаточно применить коррекцию к статистикам $a(x), b(x)$, задающим края доверительного интервала. Другие методы используют оценки э.ф.р. и рассматриваются далее.

3.2 Непараметрический бутстрап для построения доверительных интервалов

В предыдущих разделах мы видели, что для построения доверительных интервалов нужно

- Подобрать статистику S, являющуюся хорошей точечной оценкой исследуемого параметра θ;
- Найти характеристики (например, квантили) распределения этой статистики.

[8] Anatolyev S. (2007). *The basics of bootstrapping.* Quantile, №3, pp. 1–12.

Большинство используемых на практике статистических моделей достаточно сложны, поэтому как-правило, вторая часть такой задачи не имеет точного аналитического решения. Редкие исключения включают в себя случаи, когда вся рассматриваемая выборка имеет нормальное распределение, – но обычно такое предположение является слишком нереалистичным. Из-за недоступности точных решений, распространено использование асимптотических распределений статистик в качестве приближения для истинных распределений в малых (конечных) выборках. Однако, такие приближения могут быть сильно неточными, или сложными для вычисления.

Любые вероятностные характеристики S можно выразить, если известна её функция распределения

$$F_S(u, F_X) := \mathrm{P}\big(S(X_1, \ldots, X_n) \leq u\big), \tag{8}$$

которая зависит от размера выборки n, популяционной функции распределения F_X и, косвенно, от значения параметра θ. При известной F_X, даже в ситуации, когда аналитическое нахождение $F_S(u, F_X)$ является слишком сложным, её можно восстановить со сколь угодно высокой точностью с помощью метода Монте-Карло. Для этого нужно повторить достаточно много раз эксперимент с генерацией из F_X случайных выборок размером n и расчётом на них значений S. В результате на полученных значениях статистики можно построить э.ф.р. $\hat{F}_S(u, F_X)$, которая по теореме Гливенко-Кантелли является хорошей оценкой $F_S(u, F_X)$, а значит, пригодна для вычисления квантилей для доверительных интервалов и других характеристик распределения.

На практике F_X обычно неизвестна, однако доступна её оценка – э.ф.р. $\hat{F}_X$. Учитывая, что $\hat{F}_X$ соответствует равномерному дискретному распределению с носителем – точками исходной выборки $\{x_i\}_{i=1}^n$ (в предположении, что все x_i разные), то процедура Монте-Карло генерации заключается в выборе случайным образом с возвращением n элементов из исходной выборки (*ресэмплинг*), что соответствует бутстрап-выборкам, рассмотренным ранее.

Таким образом мы можем построить бутстрап-э.ф.р. $\hat{F}_S(u, \hat{F}_X)$, и поскольку $\hat{F}_X$ равномерно сходится к F_X с ростом n, то можно ожидать (в предположении непрерывности соответствующего отображения), что $\hat{F}_S(u, \hat{F}_X)$ будет сходиться по n к $F_S(u, F_X)$, – а значит при достаточно больших n есть основания считать, что они близки – поэтому можно использовать $\hat{F}_S(u, \hat{F}_X)$ для нахождения вероятностных характеристик S.

3.3 Полупараметрический и параметрический бутстрап

Рассмотренная выше процедура построения бутстрап-выборок с помощью ресэмплинга не предполагает никаких дополнительных знаний о виде распределения генеральной совокупности, и называется непараметрическим бутстрапом. Очевидным недостатком этого метода является то, что он может воспроизводить только те элементы, которые были в исходной выборке, что ограничивает число возможных значений статистики, распределение которой оценивается с помощью бутстрапа. Тем не менее, существуют подходы, помогающие решить эту проблему.

В методе полупараметрического бутстрапа ресэмплинг делается из сглаженной версии выборочного распределения. Это позволяет получать элементы, похожие на наблюдаемую выборку, но в то же время не повторяющие её в точности. Оказывается, что это можно очень просто сделать, сначала сделав ресэмплинг, как непараметрическом бутстрапе, а затем добавляя к полученным наблюдениям случайный шум. Такой метод часто применяют для выборок из дискретных распределений, в которых наблюдения могут повторяться и иметь низкую вариативность.

Параметрический бутстрап предполагает, что данные происходят из известного распределения с неизвестными параметрами. В этом методе сначала оцениваются параметры распределения по имеющимся данным, а затем оцененные распределения используются для симуляции выборок. Параметрический бутстрап позво-

ляет учитывать известную структуру данных в популяции и использовать эту информацию для генерации бутстрап-выборок. Однако важно помнить, что правильный выбор теоретического распределения и точные оценки его параметров имеют решающее значение для точности и надежности результатов применения такого метода.

3.4 Построение доверительных интервалов с помощью бутстрапа

Один из самых первых бутстрап-методов был предложен Эфроном[9] и предполагал следующий алгоритм действий для посторения доверительного интервала для параметра θ:

1. Сгенерировать N бутстрап-выборок $\{X_i^*\}_{i=1}^n$ на основе исходной выборки $\{x_i\}_{i=1}^n$ (непараметрический подход), либо из оценённого распределения (параметрический подход);
2. Вычислить значение $\hat{\theta}^* := \hat{\theta}(X_1^*, \ldots, X_n^*)$ для каждой из N бутстрап-выборок;
3. В полученной выборке $\hat{\theta}^*$ найти выборочные квантили уровня $q_{\alpha/2}(\hat{\theta}^*)$ и $q_{1-\alpha/2}(\hat{\theta}^*)$, которые и будут границами доверительного интервала для θ. Здесь $1-\alpha$, как обычно, требуемый уровень доверия.

Несмотря на простоту метода Эфрона, он не имеет под собой твёрдого математического обоснования, и большее распространение на практике получил метод Холла.[10] В нём бутстрап-выборки используются для вычисления аналогов конструкции $S = \hat{\theta} - \theta$, выражающей различие между оценкой по выборке и её популяционным аналогом. Важно заметить, что S содержит в своём выражении популяционный параметр – то есть не является выборочной статистикой. Тем не менее, аналог S на бутстрап-выборках X^* можно вычислить, поскольку нам доступны все наблюдения новой уменьшенной популяции. В этом случае аналогом $\hat{\theta}$ станут значения $\hat{\theta}^*$, вычисленные по X^*, а вместо θ популяционным параметром станет $\hat{\theta}$, поэтому

$$S(X^*) = \hat{\theta}^* - \hat{\theta}.$$

Поскольку $S(X^*)$ вычислимы, то можно построить бутстрап-э.ф.р. $\hat{F}_S(u, \hat{F}_X)$. Если этот функционал непрерывен в окрестности F_X, то в силу теоремы Гливенко-Кантелли, должно выполняться

$$\hat{F}_S(u, \hat{F}_X) \xrightarrow[n\to\infty]{\text{п.н.}} F_S(u, F_X), \tag{9}$$

что эквивалентно сходимости $S(X^*)$ по распределению к своему пределу $S = \hat{\theta} - \theta$. Далее, сходимость заменяется на приближенное равенство распределений, что позволяет выразить границы доверительного интервала. Пусть $q^*_{\alpha/2}$ и $q^*_{1-\alpha/2}$ – квантили бутстрап-распределения статистики $\hat{\theta}^*$. Тогда

$$\hat{\theta}^* - \hat{\theta} \overset{d}{\approx} \hat{\theta} - \theta \Rightarrow \hat{\theta}^* \overset{d}{\approx} 2\hat{\theta} - \theta \Rightarrow \mathrm{P}\big(q^*_{\alpha/2} \le 2\hat{\theta} - \theta \le q^*_{1-\alpha/2}\big) \approx 1 - \alpha \Rightarrow$$

$$\Rightarrow CI^*_{1-\alpha}(\theta) = \big(2\hat{\theta} - q^*_{1-\alpha/2}, 2\hat{\theta} - q^*_{\alpha/2}\big).$$

Аналогичный подход можно применять для построения бустстрап-статистик вида $\frac{\hat{\theta}^* - \hat{\theta}}{s.e.(\hat{\theta}^*)}$ и $\left|\frac{\hat{\theta}^* - \hat{\theta}}{s.e.(\hat{\theta}^*)}\right|$, которые в предположении непрерывности должны сходиться по распределению к $\frac{\hat{\theta} - \theta}{s.e.(\hat{\theta})}$ и $\left|\frac{\hat{\theta} - \theta}{s.e.(\hat{\theta})}\right|$ соответственно. В некоторых случаях применение статистик подобного вида помогает добиться того, что бутстрап-аналог сходится к истинному распределению очень быстро: модуль разности между функциями распределения может

[9] Efron, B. (1979). *Bootstrap methods: Another look at the jackknife*. The Annals of Statistics. 7: 1-26.
[10] Hall, P. (1986). *On the bootstrap and confidence intervals*. The Annals of Statistics. 14: 1431–1452.

иметь асимптотику $n^{-3/2}$ и выше вместо типичного $n^{-\frac{1}{2}}$. Такой случай в бутстрапе называют асимптотической рафинацией.

```r
#example on bootstrap confidence interval calculation for sample mean
data(mtcars)
head(mtcars)
bootstrap <- function(data, n) {
  resampled_data <- lapply(1:n, function(i) {
    resample <- sample(data, replace = TRUE)
    mean(resample)
    # Perform desired operations on the resampled data, e.g., compute a statistic
    # and return the mean
  })
  return(resampled_data)
}
bootstrapped_samples <- bootstrap(mtcars$hp, 500)
#bootstrapped_samples
quantile(unlist(bootstrapped_samples), probs = c(0.025, 0.25, 0.5, 0.75, 0.975))

#Hall's bootstrap confidence interval
print(c(2*mean(mtcars$hp) - quantile(unlist(bootstrapped_samples), probs = 0.975),
       2*mean(mtcars$hp) - quantile(unlist(bootstrapped_samples), probs = 0.025)))
```

3.5 Использование бутстрапа при A/B тестировании

Рассмотрим пример, в котором требуется изучить влияние нового алгоритма рекомендаций на сумму покупки (англ.: average revenue per user, ARPU), причём ключевой метрикой будет средний чек (или общая выручка на пользователя).

Сложность анализа здесь заключается в том, что распределение исследуемой переменной имеет тяжелые хвосты: например 90% пользователей могут не покупать ничего (сумм покупок – 0 рублей), 9% – покупать на 500-2000 рублей, а 1% – это ''киты", которые покупают на 50 000 рублей. Среднее значение сильно зависит от этих 1% "китов". Даже с не очень маленькими выборками, если в контрольную попал один супер-кит, а в тест нет, оценка на основе доверительного интервала (7), скорее всего, приведёт к неверному выводу. Также вероятно, что классический z-тест в таком примере скажет, что дисперсии огромны, и различий нет – даже если новый алгоритм реально поднял средний чек у среднего сегмента.

Ключевая особенность бутстрапа, обеспечивающая его удобство в таких ситуациях – это возможность использования для сложно устроенных статистик. В приведённом выше примере, можно бутстрапить не среднее, а медиану или урезанное среднее – ведь для бизнеса часто важнее, что стало лучше у 80% пользователей, а не что один кит купил на 10к больше. В другом случае, если метрикой является отношение (например, клики/показы), то часто бутстрап остаётся единственным доступным вариантом, потому что числитель и знаменатель коррелируют. Дельта-метод в такой ситуации может быть очень сложным, а бутстрап дает корректный доверительный интервал автоматически. Кроме того, с точки зрения бизнеса, интерпретация результата, полученного с помощью бутстрапа, наподобие "в 99% наших симуляций новая модель показала рост" гораздо интуитивнее классического тестирования.

Обратной стороной простоты применения и универсальности бутстрапа являются случаи, когда теоретические основания применимости метода не выполнены: в таком случае фактический уровень доверительных

интервалов и значимость тестов, будут отличаться от расчётных. Применение бутстрапа может привести к проблемам в следующих случаях:

- При малых размерах выборок – из-за большой погрешности приближения F_X с помощью $\hat{F}_X$ и, как следствие, неточности у бутстрап-оценки распределения тестовой статистики;

- При невыполнении условия непрерывности статистического функционала $T(F)$ в точке F_X. Важным примером такого случая являются статистики максимума и минимума по выборке. На рис. 2 изображены ф.р. для $\mathcal{N}(0,1)$ и э.ф.р. по выборке из этого распределения размером 100. Максимальное значение для $\mathcal{N}(0,1)$ не ограничено, и соответствующий супремум равен $+\infty$. В то же время, максимум по выборке всегда конечен. Поэтому функционал $T(F)$, вычисляющий супремум распределения, резко меняет значение при сдвиге аргумента от истинной к выборочной ф.р. даже несмотря на то, что расстояние (в равномерной метрике) между ними может быть сколь угодно мало.

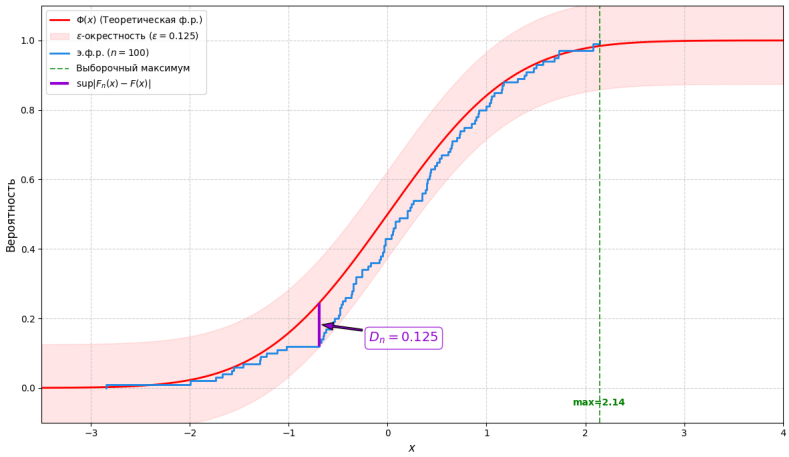


Рис. 2: Популяционная и выборочная ф.р., а также ε-окрестность точки F_X функционального пространства.

- При нарушении предположения о случайности выборки – например, при обработке данных онлайн-сервисов, где пользователи могут совершать много действий внутри одной сессии, и эти действия, как правило, коррелируют друг с другом. В таких ситуациях можно применять блочный бутстрап, отбирая в новые выборки целые целые кластеры данных, например по userID или sessionID.

```
import numpy as np
import pandas as pd
from sklearn.utils import resample

# 1. Симулируем данные (контроль А и тест Б)
# В тесте Б средний чек реально выше на 300 руб., но есть выбросы
np.random.seed(42)

# Группа А: много нулей и немного покупок
group_a = np.concatenate([
    np.zeros(800),        # 800 не купили
    np.random.exponential(500, 150), # 150 покупок со средним 500
    np.random.normal(50000, 10000, 5) # 5 супер-китов
])
```

```
# Группа Б: алгоритм лучше активирует средний сегмент
group_b = np.concatenate([
    np.zeros(750),
    np.random.exponential(800, 200), # 200 покупок со средним 800 (выше!)
    np.random.normal(50000, 10000, 5)
])

# Вывод: Среднее А:  2000, Среднее Б:  2300 (разница есть, но из-за выбросов t-test даст p=0.4)

n_iterations = 10000
diffs = []

for _ in range(n_iterations):
    # Извлекаем выборки того же размера С ВОЗВРАЩЕНИЕМ
    sample_a = resample(group_a, n_samples=len(group_a), replace=True)
    sample_b = resample(group_b, n_samples=len(group_b), replace=True)

    # Считаем разницу средних (можно заменить на медиану: для этого замените в коде sample_b.mean() на
        np.median(sample_b)
    # - и вы получите доверительный интервал для разницы медиан, что гораздо устойчивее к «китам»).

    diff = sample_b.mean() - sample_a.mean()
    diffs.append(diff)

# Превращаем в массив
diffs = np.array(diffs)

# 2. Доверительный интервал (95%)
ci_lower = np.percentile(diffs, 2.5)
ci_upper = np.percentile(diffs, 97.5)

# 3. Принимаем решение: если 0 не входит в интервал -> разница статистически значима
```

3.6 Разложение Эджворта функции распределения стандартизованной суммы

В этом разделе представлена техника разложения функций распределения асимптотически нормальных статистик в полиномиальные ряды, основанная на идеях Ф. Эджворта.[11] Она позволяет количественно оценить ошибку аппроксимации при использовании ЦПТ и бутстрапа, а также служит строгой основой для доказательства существования асимптотически рафинированных статистик – то есть, обладающих аномально высокой скоростью сходимости к истинному распределению и, как следствие, пригодных для использования бутстрапа в выборках меньшего объёма.

Пусть $\{X_i\}_{i=1}^n$ – случайная выборка из распределения с $\mathbb{E}X_1 = \mu$, $\mathbb{V}X_1 = \sigma^2$ и конечными моментами до порядка $s \geq 4$. Рассмотрим стандартизованную сумму:

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i, \qquad Y_i := \frac{X_i - \mu}{\sigma}.$$

[11] Cramér, H. (1937). Sur un nouveau théorème-limite de la théorie des probabilités. Actualités Scientifiques et Industrielles, 736, 5-23.

Характеристическая и кумулянтная функции для Z_n:

$$\varphi_{Z_n}(t) = \mathbb{E}e^{itZ_n} = \left(\mathbb{E}e^{i\frac{t}{\sqrt{n}}Y_1}\right)^n = \varphi_Y^n(t/\sqrt{n}) \Rightarrow \ln\varphi_{Z_n}(t) = n\ln\varphi_Y(t/\sqrt{n}), \tag{10}$$

причём производящая функция кумулянтов $\ln\varphi_{Z_n}(t)$ корректно определена в окрестности нуля при всех достаточно больших n, поскольку Z_n сходится по распределению к $\mathcal{N}(0,1)$, а следовательно $\varphi_{Z_n}(t)$ должна сходиться к строго положительной функции $e^{-\frac{t^2}{2}}$.

Разложим $\ln\varphi_Y(u)$ в ряд Тейлора около $u=0$:

$$\ln\varphi_Y(u) = \sum_{j=1}^{s}\frac{1}{j!}\frac{\partial^j \ln\varphi(u)}{\partial u^j}\Big|_{u=0}u^j + o(u^s) = \sum_{j=1}^{s}\frac{\kappa_j}{j!}(iu)^j + o(u^s),$$

где $\kappa_j = (-i)^j\frac{\partial^j\ln\varphi_Y(u)}{\partial u^j}\Big|_{u=0}$ называется j-м кумулянтом распределения Y; их значения можно выразить через моменты распределения:

$$\kappa_1 = -i\frac{\varphi_Y'(0)}{\varphi_Y(0)} = -i^2\frac{\mathbb{E}Y}{1} = 0,$$

$$\kappa_2 = (-i)^2\frac{\varphi_Y''(0)\varphi_Y(0) - \big(\varphi_Y'(0)\big)^2}{\varphi_Y^2(0)} = 1,$$

$$\kappa_3 = (-i)^3\frac{\varphi_Y'''(0)\varphi_Y^2(0) + 2\big(\varphi_Y'(0)\big)^3 - 3\varphi_Y(0)\varphi_Y'(0)\varphi_Y''(0)}{\varphi_Y^3(0)} = \mathbb{E}Y^3,$$

$$\kappa_4 = \frac{\varphi_Y^{(4)}(0)}{\varphi_Y(0)} - \frac{3\big(\varphi_Y''(0)\big)^2}{\varphi_Y^2(0)} - \frac{6\big(\varphi_Y'(0)\big)^4}{\varphi_Y^4(0)} - \frac{4\varphi_Y^{(3)}(0)\varphi_Y'(0)}{\varphi_Y^2(0)} + \frac{12\big(\varphi_Y'(0)\big)^2\varphi_Y''(0)}{\varphi_Y^3(0)} = \mathbb{E}Y^4 - 3.$$

Подставляем $u = \frac{t}{\sqrt{n}}$ и из (10) получаем представление для $\ln\varphi_{Z_n}(t)$ в виде ряда:

$$\ln\varphi_{Z_n}(t) = n\left[\sum_{j=1}^{4}\frac{\kappa_j}{j!}\left(\frac{it}{\sqrt{n}}\right)^j + O(n^{-\frac{5}{2}})\right] = -\frac{t^2}{2} + \frac{\kappa_3}{6}\frac{(it)^3}{\sqrt{n}} + \frac{\kappa_4}{24}\frac{(it)^4}{n} + O(n^{-\frac{3}{2}}) \Rightarrow$$

$$\varphi_{Z_n}(t) = e^{-\frac{t^2}{2}}\exp\left(\frac{\kappa_3}{6}\frac{(it)^3}{\sqrt{n}} + \frac{\kappa_4}{24}\frac{(it)^4}{n} + O(n^{-\frac{3}{2}})\right).$$

Первый член $-\frac{t^2}{2}$ соответствует характеристической функции стандартного нормального распределения, остальные члены – это поправки, убывающие с ростом n. Разлагая экспоненту в ряд $e^x = 1 + x + \frac{x^2}{2} + o(x^2)$, получаем

$$\varphi_{Z_n}(t) = e^{-\frac{t^2}{2}}\left[1 + \frac{\kappa_3}{6}\frac{(it)^3}{\sqrt{n}} + \frac{1}{n}\left(\frac{\kappa_4}{24}(it)^4 + \frac{\kappa_3^2}{72}(it)^6\right) + O(n^{-\frac{3}{2}})\right]. \tag{11}$$

Чтобы получить плотность $f_{Z_n}(x)$, применяем обратное преобразование Фурье. Ключевое наблюдение: умножение характеристической функции на $(it)^j$ соответствует j-й производной плотности:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}(it)^j e^{-\frac{t^2}{2}}\,dt = (-1)^j\frac{d^j}{dx^j}\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-itx}e^{-\frac{t^2}{2}}\,dt\right) = (-1)^j\frac{d^j}{dx^j}\phi(x) = H_j(x)\phi(x),$$

где $\phi(x)$ – плотность $\mathcal{N}(0,1)$; $H_j(x)$ – полиномы Эрмита: $H_0(x)=1$, $H_1(x)=x$, $H_2(x)=x^2-1$, $H_3(x)=x^3-3x$, $H_4(x)=x^4-6x^2+3$, $H_5(x)=x^5-10x^3+15x$, $H_6(x)=x^6-15x^4+45x^2-15$.

Подставляя (11) в обратное преобразование Фурье, получаем разложение для плотности:

$$f_{Z_n}(x) = \phi(x)\left[1 + \frac{\kappa_3}{6\sqrt{n}}H_3(x) + \frac{1}{n}\left(\frac{\kappa_4}{24}H_4(x) + \frac{\kappa_3^2}{72}H_6(x)\right) + O(n^{-\frac{3}{2}})\right].^{12} \tag{12}$$

Далее, интегрируя плотность, переходим к функции распределения $F_{Z_n}(x)$. Используем свойство полиномов Эрмита:

$$\int_{-\infty}^{x} H_j(u)\phi(u)du = \int_{-\infty}^{x}(-1)^j\frac{d^j}{du^j}\phi(u)\,du = (-1)^j\frac{d^{j-1}\phi(u)}{du^{j-1}}\Big|_{-\infty}^{x} = -H_{j-1}(x)\phi(x).$$

Следовательно:

$$F_{Z_n}(x) = \Phi(x) - \phi(x)\left[\frac{\kappa_3}{6\sqrt{n}}H_2(x) + \frac{1}{n}\left(\frac{\kappa_4}{24}H_3(x) + \frac{\kappa_3^2}{72}H_5(x)\right)\right] + O(n^{-\frac{3}{2}})$$

Подставляя явные выражения для H_2, H_3, H_5, получаем классическое разложение Эджворта второго порядка:

$$F_{Z_n}(x) = \Phi(x) - \phi(x)\left[\frac{\kappa_3}{6\sqrt{n}}(x^2-1) + \frac{1}{n}\left(\frac{\kappa_4}{24}(x^3-3x) + \frac{\kappa_3^2}{72}(x^5-10x^3+15x)\right)\right] + O(n^{-\frac{3}{2}})$$

В более компактной форме:

$$F_{Z_n}(x) = \Phi(x) + \frac{1}{\sqrt{n}}p_1(x)\phi(x) + \frac{1}{n}p_2(x)\phi(x) + O(n^{-\frac{3}{2}})$$

где полиномы $p_k(x)$ зависят от кумулянтов распределения.

Интерпретация получившегося результата:

- Компонента, имеющая порядок $n^{-\frac{1}{2}}$, пропорциональна скошенности распределения $\kappa_3 = \mathbb{E}Y^3$;
- Компонента, имеющая порядок n^{-1}, учитывает эксцесс $\kappa_4 = \mathbb{E}Y^4 - 3$ и квадрат скошенности κ_3^2;
- Для симметричного распределения ($\kappa_3 = 0$) первый поправочный член исчезает, и ошибка нормального приближения по ЦПТ становится $O(n^{-1})$.

3.7 Разложение Эджворта пивотальных статистик

Анализируя процедуру, представленную в предыдущем разделе, можно заметить, что её ключевые этапы можно применить и для асимптотически нормальной статистики с произвольной структурой. Особенно важный случай – когда такая статистика является асимптотически пивотальной (поворотной), то есть её предельное распределение не зависит от параметров. Пусть

$$S_n = \frac{\hat{\theta}_n - \theta}{\sqrt{\hat{\sigma}^2}},$$

[12] Нужно отметить, что порядок $O(n^{-\frac{3}{2}})$ у остатка в разложении может нарушиться при этом переходе. Например, если Z_n имеет дискретное распределение, то очевидно, что его плотность или ф.р. нельзя хорошо приблизить конечным числом компонент абсолютно непрерывного распределения. Для корректности (12) достаточно потребовать, чтобы распределение Z_n имело абсолютно непрерывную компоненту, что эквивалентно условию $\limsup_{|t|\to\infty}|\varphi_{Z_n}(t)| < 1$, см. Hall P. (1992). *The bootstrap and Edgeworth expansion*. Springer-Verlag, pp. 65-67.

где θ – некоторый параметр анализируемого абсолютно непрерывного распределения, $\hat{\theta}_n$ – его оценка по случайной выборке размером n, а $\hat{\sigma}^2$ – состоятельная оценка дисперсии $\hat{\theta}_n$. Разложение в ряд около $t = 0$ производящей функции кумулянтов для S_n имеет вид

$$\ln \varphi_{S_n}(t) = \ln\bigl(\mathbb{E}e^{itS_n}\bigr) = \kappa_1^{(n)}(it) + \frac{1}{2}\kappa_2^{(n)}(it)^2 + \ldots + \frac{1}{j!}\kappa_j^{(n)}(it)^j + \ldots, \tag{13}$$

где $\kappa_j^{(n)} = (-i)^j \frac{\partial^j \ln \varphi_{S_n}(u)}{\partial u^j}\Big|_{u=0}$ – кумулянты распределения S_n.

Предположим, что каждый такой кумулянт имеет порядок $n^{-\frac{j-2}{2}}$ и может быть разложен в степенной ряд по n^{-1}:

$$\kappa_j^{(n)} = n^{-\frac{j-2}{2}}\left(\kappa_{j,1} + n^{-1}\kappa_{j,2} + n^{-2}\kappa_{j,3} + \ldots\right), \qquad j \geq 1, \tag{14}$$

где $\kappa_{1,1} = 0$ и $\kappa_{2,1} = 1$. Последние два соотношения отражают тот факт, что S_n центрировано и масштабировано так, чтобы $\kappa_1^{(n)} = \mathbb{E}S_n \to 0$ и $\kappa_2^{(n)} = \mathbb{V}S_n \to 1$. Такое разложение $\kappa_j^{(n)}$ возможно, например, если S_n на $o_p(1)$ отличается от стандартизованной суммы *н.о.р.* величин.

Объединяя (13) и (14), получаем

$$\begin{aligned}
\ln \varphi_{S_n}(t) &= -\frac{1}{2}t^2 + n^{-\frac{1}{2}}\left\{\kappa_{1,2}(it) + \frac{1}{3!}\kappa_{3,1}(it)^3\right\} + n^{-1}\left\{\frac{1}{2}\kappa_{2,2}(it)^2 + \frac{1}{4!}\kappa_{4,1}(it)^4\right\} + \ldots, \\
\varphi_{S_n}(t) &= e^{-\frac{t^2}{2}}\left[1 + n^{-\frac{1}{2}}\Bigl(\kappa_{1,2}(it) + \frac{1}{6}\kappa_{3,1}(it)^3\Bigr) + \right. \\
&\quad \left. + n^{-1}\Bigl(\frac{1}{2}\kappa_{2,2}(it)^2 + \frac{1}{24}\kappa_{4,1}(it)^4 + \frac{1}{2}\Bigl(\kappa_{1,2}(it) + \frac{1}{6}\kappa_{3,1}(it)^3\Bigr)^2\Bigr) + O(n^{-\frac{3}{2}})\right].
\end{aligned}$$

Повторяя с получаенной характеристической функцией операции аналогично предыдущему разделу, получаем

$$F_{S_n}(x) = \Phi(x) + \frac{1}{\sqrt{n}}p_1(x)\phi(x) + \frac{1}{n}p_2(x)\phi(x) + O(n^{-\frac{3}{2}}), \tag{15}$$

где

$$p_1(x) = -\Bigl(\kappa_{1,2} + \frac{1}{6}\kappa_{3,1}H_2(x)\Bigr) = -\Bigl(\kappa_{1,2} + \frac{1}{6}\kappa_{3,1}(x^2 - 1)\Bigr),$$

$$\begin{aligned}
p_2(x) &= -\Bigl(\frac{1}{2}(\kappa_{2,2} + \kappa_{1,2}^2)H_1(x) + \frac{1}{24}(\kappa_{4,1} + 4\kappa_{1,2}\kappa_{3,1})H_3(x) + \frac{1}{72}\kappa_{3,1}^2 H_5(x)\Bigr) \\
&= -x\Bigl\{\frac{1}{2}(\kappa_{2,2} + \kappa_{1,2}^2) + \frac{1}{24}(\kappa_{4,1} + 4\kappa_{1,2}\kappa_{3,1})(x^2 - 3) + \frac{1}{72}\kappa_{3,1}^2(x^4 - 10x^2 + 15)\Bigr\}.
\end{aligned}$$

Важное наблюдение: вне зависимости от значений компонент кумулянтов, полином $p_1(x)$ является чётным, а $p_2(x)$ – нечётным.

Аналогичное построение можно сделать и для нестандартизованных статистик вида $\sqrt{n}(\hat{\theta}_n - \theta)$; в таком случае $\kappa_{2,1} = \sigma^2$ – асимптотическая дисперсия, и первый член разложения будет функцией распределения $\mathcal{N}(0, \sigma^2)$.

3.8 Асимптотическая рафинация бутстрапа

Теперь рассмотрим бутстрап-аналог пивотальной статистики $S_n^* = \frac{\hat{\theta}_n^* - \hat{\theta}_n}{\sqrt{\hat{\sigma}_*^2}}$, где $\hat{\theta}_n^*$ и $\hat{\sigma}_*^2$ вычислены по бутстрап-выборке.

Условное распределение S_n^* (при фиксированном векторе исходной выборки X) также допускает разло-

жение Эджворта:

$$F_{S_n^*|X}(x) = \Phi(x) + \frac{1}{\sqrt{n}}\hat{p}_1(x)\phi(x) + \frac{1}{n}\hat{p}_2(x)\phi(x) + O_p(n^{-\frac{3}{2}}), \tag{16}$$

где $\hat{p}_k(x)$ – те полиномы, аналогичные (15), но вычисленные с использованием выборочных кумулянтов $\hat{\kappa}_j^{(n)}$ вместо истинных $\kappa_j^{(n)}$.

Суть рафинации заключается в сравнении двух разложений: ошибка аппроксимации базового бутстрапа равна:

$$\mathrm{P}(S_n \le x) - \mathrm{P}(S_n^* \le x \mid X) = \frac{1}{\sqrt{n}}\big(p_1(x) - \hat{p}_1(x)\big)\phi(x) + O_p(n^{-1})$$

Поскольку выборочная скошенность $\hat{\kappa}_3^{(n)}$ сходится к истинной $\hat{\kappa}_3^{(n)}$ со скоростью $O_p(n^{-\frac{1}{2}})$, разность $\big(p_1(x) - \hat{p}_1(x)\big)$ имеет порядок $O_p(n^{-\frac{1}{2}})$, следовательно, общая ошибка бутстрапа для пивотальной статистики составляет:

$$\text{Ошибка} = O_p(n^{-\frac{1}{2}}) \times \frac{1}{\sqrt{n}} = O_p(n^{-1}).$$

Это фундаментальный результат: бутстрап пивотальной статистики автоматически корректирует скошенность, обеспечивая ошибку порядка $O(n^{-1})$, в то время как нормальная аппроксимация (ЦПТ) дает ошибку порядка $O(n^{-\frac{1}{2}})$. В случае применения бутстрапа для статистик, не обладающих свойством пивотальности, ошибка приближения будет иметь порядок $O(n^{-\frac{1}{2}})$, поскольку первый член разложения будет зависеть от бутстрап-оценки неизвестного параметра.

Оказывается, используя уже полученные результаты, можно довольно легко добиться дальнейшей рафинации – от $O(n^{-1})$ к $O(n^{-\frac{3}{2}})$. Для этого рассмотрим симметризованную t-статистику

$$T_n = \frac{|\hat{\theta}_n - \theta|}{\sqrt{\hat{\sigma}^2}} \xrightarrow{d} |\mathcal{N}(0,1)|.$$

Используя (15) и (16), получаем выражения для функций распределения T_n и её бутстрап-аналога T_n^*:

$$F_{T_n}(x) = \Pr\left\{-x \le \frac{|\hat{\theta}_n - \theta|}{\sqrt{\hat{\sigma}^2}} \le x\right\} = 2\Phi(x) - 1 + \frac{2}{n}p_2(x)\phi(x) + O(n^{-\frac{3}{2}}),$$

$$F_{T_n^*}(x) = 2\Phi(x) - 1 + \frac{2}{n}\hat{p}_2(x)\phi(x) + O(n^{-\frac{3}{2}})$$

Таким образом, ошибки аппроксимации для асимптотики и бутстрапа имеют порядки

$$F_{T_n^*}(x) - F_{T_n}(x) = \frac{2\phi(x)}{n}\big(\hat{p}_2(x) - p_2(x)\big) + O(n^{-\frac{3}{2}}) = O(n^{-\frac{3}{2}}),$$

$$2\Phi(x) - 1 - F_{T_n}(x) = O(n^{-1}).$$

4 Линейные модели и метод наименьших квадратов

4.1 Модель простой (парной) регрессии. Две задачи анализа данных

В качестве оцениваемого статистического параметра в многомерном распределении может рассматриваться функциональная зависимость, например, среднего значения одной переменной от значения другой (функция регрессии). Термин “регрессия” появился в статистическом анализе благодаря работе Ф. Гальтона “Регресс к среднему при наследовании” (1886 г.), в которой исследовалась зависимость между ростом родителей и их детей. Гальтон обнаружил закономерность, что у высоких родителей дети, как правило, тоже имеют рост выше среднего – однако меньший, чем у их родителей; и наоборот. Таким образом, отклонения роста в следующем поколении сглаживаются – происходит возвращение назад (регресс) к среднему.

Обозначая популяционное среднее за m, рост родителей в популяции за $\mathbf{X}$, а рост детей за $\mathbf{Y}$, найденную в работе Гальтона зависимость можно выразить уравнением

$$\mathbf{Y} - m \approx \frac{2}{3}(\mathbf{X} - m). \tag{17}$$

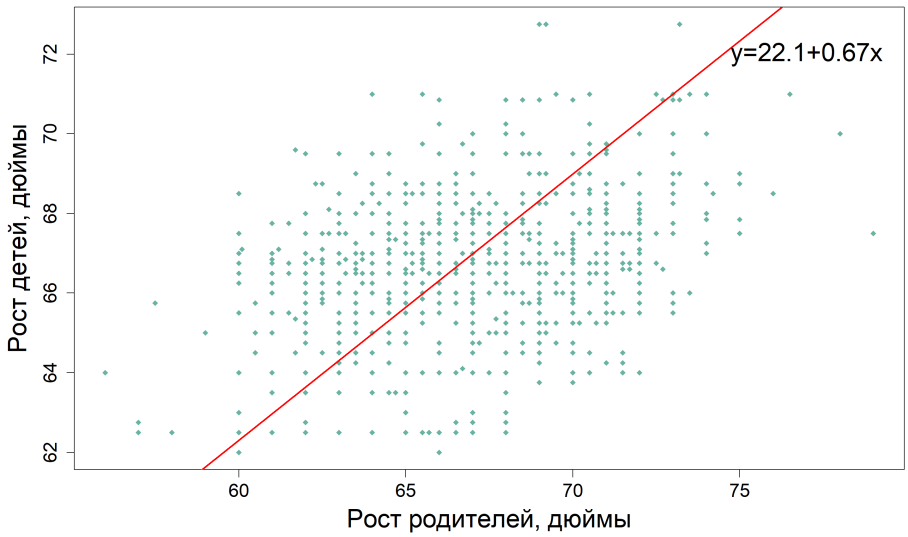


Рис. 3: Зависимость между ростом родителей и детей, оценённая с использованием данных из оригинальной работы Гальтона.

Из рис. 3 очевидно, что в наблюдаемой выборке рост детей не полностью определяется ростом родителей в соответствии с (17), но можно предположить, что отклонения от линейной зависимости для каждой семьи являются случайными и имеют нулевое среднее. Таким образом, найденной закономерности соответствует модель

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{18}$$

где β_0, β_1 – оцениваемые числовые коэффициенты, а ε_i – значения реализаций случайной величины, называемой ошибкой регрессии. Исследователю известны только значения выборки $\{(y_i, x_i)\}_{i=1}^n$, а значения ε_i являются ненаблюдаемыми (за исключением случаев, когда данные для подобной модели генерируются искуственно), поэтому далее не будем делать специального различия в обозначениях между ошибками – случайными величинами и их реализациями. Как правило, предполагается, что ε_i включают в себя все переменные, влияющие на y_i, но по какой-то причине не включенные в модель (например, недоступные ис-

следователю).

Свойства коэффициентов и ошибок в модели (18) связаны друг с другом. Например, наличие константы β_0 в модели позволяет всегда считать, что $\mathbb{E}\varepsilon_i = 0$ (иначе, если $\mathbb{E}\varepsilon = c \neq 0$, то мы можем переназначить $\beta_0' := \beta_0 + c$ и $\varepsilon' := \varepsilon - c$).

В модели (18) случайность в y связана только с случайностью ошибки регрессии ε. Концептуально это означает возможность считать значения x полностью контролируемыми (*детерминистические регрессоры*), и в таком случае коэффициент β_1 имеет смысл предельного эффекта переменной x на среднее значение y; в таком контексте x называется *независимой* или *объясняющей*, а y – *зависимой* переменной, хотя эти названия часто некорректно отражают истинное положение дел.

Модели с детерминистическими регрессорами длительное время использовались в практике прикладных исследований, однако они совершенно неприемлемы для ситуаций, когда $\{x_i\}_{i=1}^n$ являются неэкспериментальными (англ.: observational data), и их формирование просходит без вмешательства исследователя. В дальнейшем будем всегда считать это предположение выполненным.

Проблему детерминистических регрессоров можно проиллюстрировать с помощью двух возможных интерпретаций результата, полученного Гальтоном:

1. Прогнозная интерпретация. Если в дополнение к проанализированной выборке рассмотреть несколько других семей из той же популяции, то можно ожидать, что отклонение роста от популяционного среднего у детей в этих семьях будет близко к $\frac{2}{3}$ отклонения от среднего у их родителей.

2. Причинно-следственная интерпретация. Если над семейной парой из той же популяции поставить эксперимент, который должен каким-то образом повлиять на рост участников эксперимента (например, длительный комплекс гимнастических упражнений и соответствующим образом подобранное питание), то это повлияет также и на рост их детей.

Вторая интерпретация полученного результата является гораздо менее правдоподобной: она предполагает наличие причинно-следственной связи от x к y и, в частности, гипотетическую возможность исследователя изменять значение x_i так, что ненаблюдаемая переменная ε_i при этом не подвергается никакому влиянию. Кроме того, такая интерпретация очевидно не может быть верной для всех линейных моделей, поскольку те же самые данные Гальтона свидетельствуют о положительной связи и в обратной модели – в которой зависимой переменной является рост родителей, а объясняющей – рост детей. Абсурдность предположения о том, что рост детей может повлиять на рост родителей привела Гальтона в конечном итоге к отказу от исследования понятия причинно-следственных связей в пользу ненаправленного понятия корреляции (от англ. co-relation – взаимосвязь).[13]

Модификация модели (18), позволяющая отказаться от обязательной причинно-следственной импликации, но сохраняющая прогнозные свойства, заключается в том, что мы считаем значения роста родителей x_i реализациями некоторой случайной величины X_i (так называемые *стохастические регрессоры*); при этом между случайными величинами существует линейная связь, которую мы можем оценить по выборке реализовавшихся значений.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{19}$$

В такой модели уже не всегда можно интерпретировать β_1, как среднее изменение $\mathbf{Y}$ в ответ на вмешательство, приводящее к единичному изменению в $\mathbf{X}$ – поскольку эта переменная может быть связана с ε, и тогда такое вмешательство приведёт к непредсказуемому изменению у ошибки модели.

[13] Pearl J., Mackenzie D. (2018). *The book of why. The new science of cause and effect.* Basic books. p. 402.

Определение 4.1. Модель (19) называется простой или парной регрессией с константой.

Позднее регрессией стали называть функциональную связь между произвольными случайными величинами. Напомним формальное определение: пусть $\mathbf{Y}$ – случайная величина с конечным математическим ожиданием, а $\mathbf{X}$ – случайный вектор. Функцией регрессии $\mathbf{Y}$ на $\mathbf{X}$ называется функция условного среднего $\varphi(x) = \mathbb{E}(\mathbf{Y} \mid \mathbf{X} = x)$, где $x \in \mathbb{R}^k$ – значение вектора $\mathbf{X}$.

Задача оценки (иногда используется термин "восстановления") регрессии заключается в нахождении оценки неизвестной функциональной зависимости $\mathbf{Y}$ от $\mathbf{X}$ по выборочным наблюдениям и является частным случаем задачи обучения модели по прецедентам (задачей машинного/статистического обучения). Общий случай такой задачи формулируется следующим образом:

- $\mathbf{X}$ – множество объектов (их информационных описаний),
- $\mathbf{Y}$ – множество ответов (оценок, предсказаний или прогнозов, зависимых переменных),
- $y(\mathbf{X}, \varepsilon) = \mathbf{Y}$ – неизвестная (истинная, популяционная) зависимость, ε – ненаблюдаемые переменные ("ошибки").

Дано:

- $\{x_1, \ldots, x_n\}$ – подмножество значений $\mathcal{X}$ в обучающей выборке,
- $y_i = y(x_i, \varepsilon_i), \quad i = 1, \ldots, n$ – известные ответы.

Найти: оценку (т.е. функцию от данных), приближающую функцию $y(\cdot)$ на всём $\mathcal{X}$.

Задача обучения по прецедентам заключается в конкретизации того, как задаются объекты и какими могут быть ответы; в каком смысле $\hat{y}(x)$ или $a(x)$ приближает $y(x, \varepsilon)$, и как строить функцию $\hat{y}$ или алгоритм a.

В зависимости от рассматриваемой задачи, оценённая модель может использоваться для предсказания (то есть, вычисления значений функции $\hat{y}$ в точках x, не входивших в обучающую выборку), либо для анализа структуры оценённой модели.

4.2 Метод наименьших квадратов для оценки парной регрессии

В модели парной регрессии $\mathbf{X}$ и $\mathbf{Y}$ – это случайные величины, характеризующие популяционные значения некоторых числовых характеристик. Обычно предполагается, что исследователю доступна *н.о.р.* выборка реализаций их значений $\{(y_i, x_i)\}_{i=1}^n$, в которой для каждого из наблюдений i зависимость между Y_i и X_i имеет вид

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{20}$$

Параметры модели (20) как правило оцениваются методом наименьших квадратов (МНК, англ. – ordinary least squares, OLS), идея которого заключается в минимизации суммы квадратов остатков $\hat{\varepsilon}_i$ (то есть, оценённых значений ошибок регрессии):

$$\hat{\beta}_0, \hat{\beta}_1 := \operatorname{argmin}_{\beta_0, \beta_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Для нахождения $\hat{\beta}_0$ и $\hat{\beta}_1$ вычислим условия первого порядка:

$$\frac{\partial}{\partial \beta_0} \left[\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \right] = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \Rightarrow \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \Rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

$$\frac{\partial}{\partial\beta_1}\left[\sum_{i=1}^{n}(Y_i-\hat\beta_0-\hat\beta_1X_i)^2\right]=-2\sum_{i=1}^{n}X_i(Y_i-\hat\beta_0-\hat\beta_1X_i)=0\Rightarrow\sum_{i=1}^{n}X_i(Y_i-\bar Y)=\hat\beta_1\sum_{i=1}^{n}X_i(X_i-\bar X)\Rightarrow$$

$$\Rightarrow\hat\beta_1=\frac{\sum_{i=1}^{n}X_i(Y_i-\bar Y)}{\sum_{i=1}^{n}X_i(X_i-\bar X)}. \tag{21}$$

Для проверки условий второго порядка, запишем матрицу гессиана и проверим её знакоопределённость.

$$H=\begin{pmatrix}2n & 2\sum_{i=1}^{n}X_i\\ 2\sum_{i=1}^{n}X_i & 2\sum_{i=1}^{n}X_i^2\end{pmatrix}.$$

Определитель H положителен в силу неравенства Коши-Буняковского-Шварца (кроме случая, когда все x_i одинаковы – тогда определитель равен нулю), поэтому по критерию Сильвестра матрица H – положительно-определённая, – а значит найденные значения $\hat\beta_0$ и $\hat\beta_1$ минимизируют сумму квадратов остатков.

Пример данных с оценённой линейной регрессией представлен на рис. 4.

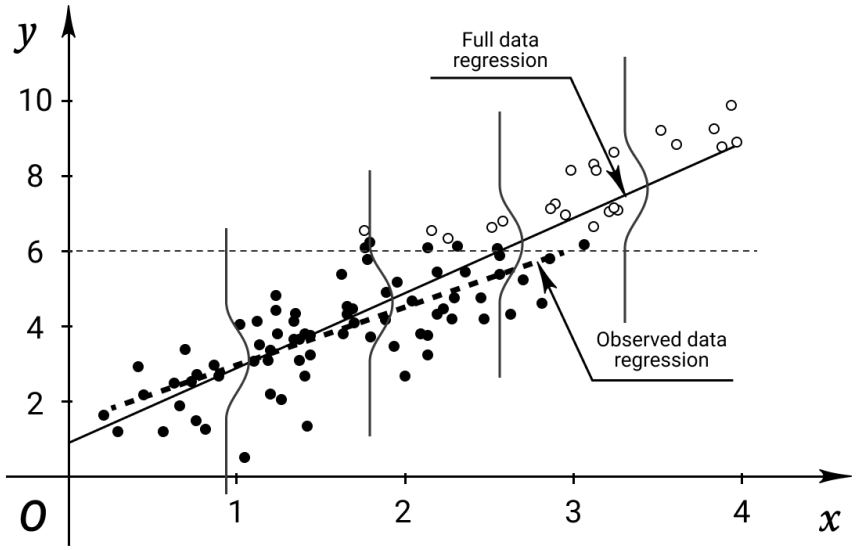


Рис. 4: Выборка (чёрные точки) и полная популяция (все точки) в модели парной регрессии.

Оценку коэффициента наклона $\hat\beta_1$ можно выразить с помощью выборочной корреляции и выборочных дисперсий:

$$\hat\beta_1=\frac{\sum_{i=1}^{n}X_i(Y_i-\bar Y)}{\sum_{i=1}^{n}X_i(X_i-\bar X)}=\frac{\sum_{i=1}^{n}(X_i-\bar X)(Y_i-\bar Y)}{\sum_{i=1}^{n}(X_i-\bar X)^2}=$$

$$=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar X)(Y_i-\bar Y)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar X)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar Y)^2}}\cdot\sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar Y)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar X)^2}}=:\rho_{xy}\sqrt{\frac{S_y^2}{S_x^2}}. \tag{22}$$

Выборочная корреляция ρ_{xy} оценивает меру линейной зависимости между случайными величинами, но не говорит о направлении (причинно-следственной) связи между ними. Аналогично, оценка коэффициента β_1 в линейной модели без дополнительных предположений не может быть интерпретирована, как мера причинно-следственного эффекта $\mathbf{X}$ на $\mathbf{Y}$.

Аналогичные (21) оценки β_0,β_1 в модели (20) можно получить методом максимального правдоподобия в случае, когда ошибки ε имеют нормальное распределение. Общее название методов оценивания, в которых семейство распределений не специфицируется, но получаемая оценка является ОМП для одной из возможных

параметризаций – методы *квази-максимального правдоподобия*.

4.3 Множественная регрессия

Естественным обобщением модели парной регрессии является множественная линейная регрессия с k объясняющими переменными $(X_1, \ldots, X_k)$; при этом популяционная зависимость $\mathbf{Y}(\mathbf{X}, \varepsilon)$ является линейной по векторному параметру $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^\top$.

Для типичной задачи, решаемой с использованием множественной регрессии, предполагается, что известна случайная выборка $\{(y_i, x_i)\}_{i=1}^n$, где $x_i = (x_{i1}, \ldots x_{ik})$ – реализация случайного вектора $(X_1, \ldots, X_k)$, и для каждого наблюдения зависимость имеет вид

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \varepsilon_i,$$

или в векторном (матричном) виде

$$Y = X\beta + \varepsilon, \tag{23}$$

где

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad - n \times 1 \text{ вектор значений зависимой переменной,}$$

$$X = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{nk} \end{pmatrix} \quad - n \times (k+1) \text{ матрица (значений) регрессоров,}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad - (k+1) \times 1 \text{ вектор коэффициентов регрессии,}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad - n \times 1 \text{ вектор ошибок регрессии.}$$

Оценённая модель (23) имеет вид

$$\hat{Y}_i(X_{i1}, \ldots, X_{ik}) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_k X_{ik},$$

или, эквивалентно,

$$\hat{Y} = X\hat{\theta}, \quad \text{где } \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^\top - k+1\text{-мерный вектор оценок параметров модели.}$$

Уравнению (23) могут соответствовать различные процессы порождения данных (англ.: data generating process, DGP). От выбора конкретного вида DGP зависит ответ, можно ли какой-то из коэффициентов β_j интерпретировать как причинно-следственный эффект на Y от изменения X, то есть, верно ли, что если в i-м наблюдении каким-то образом изменить значение x_{ij} на единицу, то, в среднем, y_i должен измениться на

β_j. Также от вида DGP и метода оценивания зависят статистические свойства получаемых оценок.

Основным методом оценки множественных регрессий также является метод наименьших квадратов:

$$\hat{\beta} := \operatorname{argmin}_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_k X_{ik})^2, \tag{24}$$

или в векторном виде, выражая сумму квадратов остатков с помощью скалярного произведения

$$\hat{\beta} := \operatorname{argmin}_{\beta} (Y - X\beta)^\top (Y - X\beta). \tag{25}$$

Хорошо известно решение этой оптимизационной задачи $\hat{\beta} = (X^\top X)^{-1} X^\top Y$ (в предположении, что матрица $X^\top X$ обратима, которое эквивалентно линейной независимости столбцов матрицы X). В оценённой модели $\hat{Y} = X\hat{\beta}$, что можно записать как

$$\begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \hat{\beta}_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \hat{\beta}_1 \begin{pmatrix} X_{11} \\ \vdots \\ X_{n1} \end{pmatrix} + \ldots + \hat{\beta}_k \begin{pmatrix} X_{1k} \\ \vdots \\ X_{nk} \end{pmatrix},$$

– то есть вектор прогнозных (оценённых) значений $\hat{Y}$, в отличие от настоящего Y, лежит в $k+1$-мерном линейном подпространстве $\mathbb{R}^n$, порождённом векторами-столбцами матрицы X. Условия первого порядка для (25) можно переписать в виде

$$X^\top (Y - X\hat{\beta}) = 0_{k+1},$$

где 0_{k+1} – вектор-столбец из $k+1$ нулей. Замечая, что $Y - X\hat{\beta} = \hat{\varepsilon}$, получаем условие $X^\top \hat{\varepsilon} = 0_{k+1}$, означающее, что каждый из столбцов матрицы X ортогонален вектору остатков регрессии $\hat{\varepsilon}$. Обозначая линейное подпространство, порождённое столбцами матрицы X за $\mathcal{L}(X)$, видим, что минимизация суммы квадратов остатков приводит к тому, что вектор Y раскладывается в сумму $X\hat{\beta} \in \mathcal{L}(X)$ и $\hat{\varepsilon} \perp \mathcal{L}(X)$, то есть $\hat{Y} = X\hat{\beta}$ является ортогональной проекцией вектора Y на $\mathcal{L}(X)$. Этот результат становится очевидным, если заметить, что минимизация $\hat{\varepsilon}^\top \hat{\varepsilon}$ – это минимизация (квадрата) длины вектора, равного разности между Y и произвольным вектором из $\mathcal{L}(X)$, рис. 5.

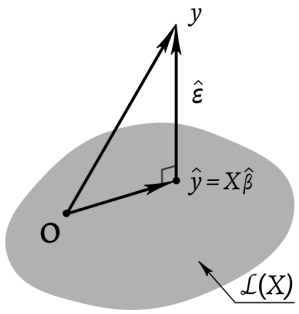


Рис. 5: Вектор $\hat{Y} = X\hat{\beta}$ по построению лежит в линейном подпространстве векторов из $\mathbb{R}^n$, порождённых столбцами матрицы X. Метод наименьших квадратов заключается в минимизации (квадрата) длины $\hat{\varepsilon}$

Пример 4.2. Докажите следующие утверждения для простой линейной регрессии с константой $Y = \beta_0 + \beta_1 X + u$, оценённой с помощью МНК.

a) $\sum_{i=1}^n \hat{u}_i = 0$, где $\hat{u}$ – МНК-остатки;

b) $\sum_{i=1}^n \hat{u}_i X_i = 0$;

c) $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$;

d) $\bar{Y} = \overline{\hat{Y}}$;

e) $\sum_{i=1}^n \hat{Y}_i \hat{u}_i = 0$;

Решение. Столбцы матрицы регрессоров имеют вид $(1, \ldots, 1)^\top$ и $(X_1, \ldots, X_n)^\top$. Утверждения a), b) и e) следуют из ортогональности МНК-остатков к столбцам матрицы регрессоров. При этом используется, что $\hat{Y}$ лежит в линейном пространстве, порождённом этими столбцами. Утверждения c) и d) можно доказать следующим образом:

$$Y = \hat{Y} + \hat{u} \Rightarrow \bar{Y} = \overline{\hat{Y}} + \overline{\hat{u}} \Rightarrow \bar{Y} = \overline{\hat{Y}}.$$

$$\overline{\hat{Y}} = \frac{1}{n} \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}.$$

4.4 Теорема Фриша-Ву-Ловелла

Теорема Фриша, Ву и Ловелла (далее – FWL) предлагает метод исключения отдельных переменных из множественной линейной регрессии и является ключевым результатом для метода наименьших квадратов. Для её доказательства нам потребуется вспомогательный результат из линейной алгебры, характеризующий вид матриц операторов ортогональной проекции в конечномерных пространствах.

Теорема 4.3. *(О матрице оператора ортогональной проекции в $\mathbb{R}^n$; доказательство в приложении).*

Оператор $\mathcal{P}$ задаёт ортогональную проекцию в $\mathbb{R}^n$ тогда и только тогда, когда (в произвольном базисе) его матрица P имеет следующие свойства:

- $P^\top = P$ *(симметричность);*
- $P^2 = P$ *(идемпотентность).*

Теорема 4.4. *(Фриш, Ву, Ловелл). Рассмотрим модель множественной регрессии в матричной форме*

$$Y = X\beta + Z\gamma + \varepsilon,$$

где X и Z – $n \times k$ и $n \times \ell$ матрицы регрессоров, столбцы которых (совместно) линейно независимы. Оценённая методом МНК модель имеет вид

$$Y = X\hat{\beta} + Z\hat{\gamma} + \hat{\varepsilon}. \tag{26}$$

Утверждение: если сначала оценить регрессию Y на Z, потом регрессию X на Z (каждый столбец в X по отдельности), обозначая получающиеся остатки как $\hat{\varepsilon}_Y$ и $\hat{\varepsilon}_X$ (последний является $n \times k$ матрицей), то МНК-оценка в регрессии $\hat{\varepsilon}_Y$ на $\hat{\varepsilon}_X$ в точности совпадает с $\hat{\beta}$ в (26).

Доказательство. По свойствам метода наименьших квадратов, остатки регрессии в "большой" модели $\hat{\varepsilon}$ ортогональны всем столбцам X и Z. Обозначим $P_Z := Z(Z^\top Z)^{-1} Z^\top$ и $Q_Z := I_n - P_Z$. Эти матрицы симметричны, поскольку

$$\big(Z(Z^\top Z)^{-1} Z^\top\big)^\top = Z\big(Z(Z^\top Z)^{-1}\big)^\top = Z\big(Z(Z^\top Z)^{-1}\big)^\top = Z\big((Z^\top Z)^{-1}\big)^\top Z^\top = Z(Z^\top Z)^{-1} Z^\top.$$

Здесь использовалось свойство $(AB)^\top = B^\top A^\top$, и синим цветом в выражении помечена матрица A.

Кроме того, матрицы P_Z и Q_Z идемпотентны:

$$P_Z^2 = Z(Z^\top Z)^{-1}Z^\top Z(Z^\top Z)^{-1}Z^\top = Z(Z^\top Z)^{-1}Z^\top = P_Z,$$

$$Q_Z^2 = (I_n - P_Z)(I_n - P_Z) = I_n^2 - P_Z I_n - I_n P_Z + P_Z^2 = I_n - 2P_Z + P_Z = I - P_Z = Q_Z.$$

Из доказанного следует, что эти матрицы задают операторы ортогональной проекции, причём их действие на произвольный вектор $Z\theta$, лежащий в пространстве $\mathcal{L}(Z)$, порождённом векторами-столбцами матрицы Z, происходит следующим образом:

$$P_Z Z\theta = Z(Z^\top Z)^{-1}Z^\top Z\theta = Z\theta, \quad Q_Z Z\theta = (I_n - P_Z)Z\theta = Z\theta - Z\theta = 0_n.$$

Таким образом, P_Z задаёт проектор на $\mathcal{L}(Z)$, а Q_Z – проектор на $\mathcal{L}(Z)^\perp$.

Применяя Q_Z к (26), получаем

$$Q_Z Y = Q_Z X\hat\beta + Q_Z Z\hat\gamma + Q_Z\hat\varepsilon \Rightarrow Q_Z Y = Q_Z X\hat\beta + \hat\varepsilon, \tag{27}$$

поскольку $\hat\varepsilon \in \mathcal{L}(Z)^\perp$.

Умножая (27) на $X^\top$ и учитывая, что $X^\top\hat\varepsilon = 0_k$ в силу ортогональности МНК-остатков, получаем

$$X^\top Q_Z Y = X^\top Q_Z X\hat\beta + X^\top\hat\varepsilon \Leftrightarrow X^\top Q_Z Y = X^\top Q_Z X\hat\beta \Rightarrow \hat\beta = (X^\top Q_Z X)^{-1}X^\top Q_Z Y. \tag{28}$$

Теперь можно сравнить $\hat\beta$ из (28) с МНК-оценками регрессии остатков $\hat\varepsilon_Y$ на $\hat\varepsilon_X$. Во-первых, легко заметить, что МНК-остатки в регрессии Y или X на Z – это $Q_Z Y$ и $Q_Z X$ соответственно; например

$$Y = Z\theta + \varepsilon_Y, \quad \hat\theta = (Z^\top Z)^{-1}Z^\top Y, \quad \hat Y = Z\hat\theta, \quad \hat\varepsilon_Y = Y - \hat Y = \big(I_n - Z(Z^\top Z)^{-1}Z^\top\big)Y = Q_Z Y.$$

Регрессию остатков можно записать следующим образом:

$$Q_Z Y = Q_Z X\delta + u, \tag{29}$$

где δ – коэффициент в регрессии, а u – ошибка. Тогда МНК-оценка (29) – это

$$\hat\delta = \big((Q_Z X)^\top Q_Z X\big)^{-1}(Q_Z X)^\top Q_Z Y = (X^\top Q_Z^\top Q_Z X)^{-1}X^\top Q_Z^\top Q_Z Y = (X^\top Q_Z X)^{-1}X^\top Q_Z Y = \hat\beta,$$

где мы использовали симметрию и идемпотентность проектора Q_Z.

```
# A numeric example on FWL theorem
# draw data according to a DGP Y = 1 + 2x_1 + 3x_2 + 4z_1 + 5z_2 + ε
# with x, z, ε ~ i.i.d. N(0, 1).

n = 10
X <- matrix(rnorm(2 * n, 0, 1), ncol = 2)
X <- cbind(rep(1, n), X)
Z <- matrix(rnorm(2 * n, 0, 1), ncol = 2)
epsilon <- rnorm(n, 0, 1)
Y <- X %*% c(1, 2, 3) + Z %*% c(4, 5) + epsilon

# obtain OLS estimates of Y = Xβ + Zγ + ε
```

```
ols_full <- lm(Y ~ X + Z - 1)
summary(ols_full)

# perform FWL procedure and show OLS estimates in the regression of residuals
Qz <- diag(x = 1, ncol = n, nrow = n) - Z %*% solve(t(Z) %*% Z) %*% t(Z)
eps_y <- Qz %*% Y
eps_x <- Qz %*% X
ols_fwl <- lm(eps_y ~ eps_x - 1)
summary(ols_fwl)
```

4.5 Идентификация параметров регрессионной модели

Рассмотрим стандартную модель множественной регрессии на основе случайной выборки $\{(Y_i, X_i)\}_{i=1}^n$, где X_i – k-мерный случайный вектор; также будем считать, что $n \gg k$, и матрица X регрессоров имеет полный ранг с вероятностью 1. Соотношение между переменными в популяции можно записать так:

$$\mathbf{Y} = \mathbf{X}^\top \beta + \varepsilon. \tag{30}$$

С одной стороны, оценка β методом наименьших квадратов в такой ситуации однозначно определена и уникальна; с другой – модель (30) позволяет считать вектором параметров β совершенно произвольный вектор нужной размерности. Действительно, заменяя фиксированный β на некоторый новый вектор γ, мы можем записать

$$\mathbf{Y} = \mathbf{X}^\top \gamma + \underbrace{\mathbf{X}^\top(\beta - \gamma) + \varepsilon}_{u},$$

где u – новая (ненаблюдаемая) ошибка регрессионной модели. Свойства вектора коэффициентов и ошибки модели таким образом являются взаимосвязаными, поэтому существуют два разных подхода к определению того, чем является вектор параметров β, который мы оцениваем с помощью линейных моделей.

1. Определение β на основе свойств ошибки модели. Обычно в таком случае β предполагается равным коэффициентам "популяционной регрессии" – то есть, предельным значением МНК-оценки при использовании всех наблюдений в популяции. Умножая условия первого порядка для МНК на $\frac{1}{n}$ и переходя к пределу, можно получить следующие моментные условия, идентифицирующие β:

$$\frac{1}{n}\sum_{i=1}^{n} -2X_{ij}(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_k X_{ik}) = 0 \xrightarrow[n\to\infty]{} \mathbb{E}(\mathbf{X}_j \varepsilon) = 0 \quad \forall j \in \{1, \ldots, k\}.$$

   Поскольку практически всегда предполагается $\mathbb{E}\varepsilon = 0$, то полученные условия можно переписать в виде $\operatorname{cov}(\mathbf{X}, \varepsilon) = 0$. Близкий по смыслу случай – когда предполагается, что $\mathbf{X}^\top\beta$ является функцией регрессии $\mathbf{Y}$, то есть $\mathbb{E}(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}^\top\beta$, и тогда $\mathbb{E}(\varepsilon \mid \mathbf{X}) = 0$, откуда следует некоррелированность $\mathbf{X}$ с ошибкой модели.

2. Определение свойств ошибки модели на основе β, заданного из каких-то дополнительных условий. Чаще всего такая ситуация возникает, когда весь β, или какие-то его компоненты определяются, как эффекты воздействия на $\mathbf{Y}$ соответствующих компонент $\mathbf{X}$ в гипотетическом рандомизированном эксперименте. Для отличия этого случая от предыдущего, будем использовать обозначения с *do*-оператором Перла: например, $\mathbb{E}\big(\varepsilon \mid do(\mathbf{X} = x)\big) = 0$ подразумевает, что параметр β идентифицирован из условия ортогональности/некоррелированности ε ко всем компонентам $\mathbf{X}$ в модифицированном процессе порождения

данных, для которого исследователь имеет возможность назначать произвольные значения $\mathbf{X}$ – и в таком случае уравнение (30) соответствует модели потенциальных исходов Рубина для $\mathbf{Y}$.

Более подробно отличие между этими двумя случаями будет рассмотрено далее; формальное определение *do*-оператора приводится в 7.1.

4.6 Регрессия по всей популяции

Как было показано ранее, суть метода наименьших квадратов заключается в ортогональной проекции вектора Y на линейную оболочку столбцов матрицы регрессоров X, являющуюся подпространством в $\mathbb{R}^n$. Кроме того, поскольку $\hat{Y} = X\hat{\beta}$, то оценку $\hat{\beta}$ можно интерпретировать, как координаты вектора проекции $\hat{Y}$ в базисе столбцов X. Рассмотрим более подробно связь между методом наименьших квадратов в конечных выборках и аналогичной процедурой в полной популяции. Напомним, что элементы популяции можно рассматривать, как вектора в гильбертовом пространстве L^2, геометрия которого определяется скалярным произведением вида $\langle \xi, \zeta \rangle = \operatorname{cov}(\xi, \zeta)$; $\xi, \zeta \in L^2$. Таким образом, по свойствам скалярного произведения, каждой случайной величине ξ в L^2 сопоставляется длина/норма $\|\xi\|_2 = \sqrt{\langle \xi, \xi \rangle} = \sqrt{\mathbb{V}\xi}$.

Рассмотрим задачу минимизации среднеквадратичного отклонения между $\mathbf{Y}$ и линейными функциями от случайного вектора $\mathbf{X} = (1, \mathbf{X}_1, \ldots, \mathbf{X}_k)^\top$, то есть $\mathbb{E}(\mathbf{Y} - \mathbf{X}^\top b)^2 \to \min_{b \in \mathbb{R}^{k+1}}$.

$$\beta = \arg\min_b \mathbb{E}(\mathbf{Y} - \mathbf{X}^\top b)^2 \xrightarrow{\frac{\partial}{\partial b}} -2\mathbb{E}[\mathbf{X}(\mathbf{Y} - \mathbf{X}^\top b)] = 0 \Rightarrow \beta = \big(\mathbb{E}(\mathbf{X}\mathbf{X}^\top)\big)^{-1}\mathbb{E}(\mathbf{X}\mathbf{Y}), \tag{31}$$

в предположении, что матрица $\mathbb{E}(\mathbf{X}\mathbf{X}^\top)$ является обратимой. Заметим, что в оптимуме $\mathbb{E}(\mathbf{Y} - \mathbf{X}^\top b)^2 = \mathbb{V}(\mathbf{Y} - \mathbf{X}^\top b) = \|\mathbf{Y} - \mathbf{X}^\top b, \mathbf{Y} - \mathbf{X}^\top b\|_2^2$, поэтому геометрический смысл задачи – поиск в линейном подпространстве L^2 вида $\mathbf{X}^\top b, b \in \mathbb{R}^{k+1}$, точки, наиболее близкой к $\mathbf{Y}$, а решением является ортогональная проекция $\mathbf{Y}$ на это линейное подпространство.

Выражение $\big(\mathbb{E}(\mathbf{X}\mathbf{X}^\top)\big)^{-1}\mathbb{E}(\mathbf{X}\mathbf{Y})$ является пределом по вероятности для МНК-оценки по случайной выборке: достаточно заметить, что реализации векторов $\mathbf{X}$ соответствуют строкам матрицы регрессоров X, и тогда

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right) \xrightarrow{p} \big(\mathbb{E}(\mathbf{X}\mathbf{X}^\top)\big)^{-1}\mathbb{E}(\mathbf{X}\mathbf{Y}),$$

в силу ЗБЧ и непрерывности операции обращения матрицы. Если же $\det \mathbb{E}(\mathbf{X}\mathbf{X}^\top) = 0$, то предельный переход не позволяет однозначно восстановить вектор коэффициентов, и параметр β не идентифицирован.

4.7 Наилучший линейный предиктор и ошибка спецификации модели

Найденное в (31) значение β позволяет интерпретировать $\mathbf{X}^\top\beta$ как наилучший линейный предиктор для $\mathbf{Y}$ на основе вектора $\mathbf{X}$. В то же время, свойства L^2 помогают ответить и на более общий вопрос: "какая функция от $\mathbf{X}$ наилучшим образом приближает $\mathbf{Y}$?".

Теорема 4.5. *(Оптимизационное свойство условного математического ожидания). Пусть в пространстве L^2 над некоторым вероятностным пространством заданы $\mathbf{Y}$ – скалярная случайная величина и $\mathbf{X} = (1, \mathbf{X}_1, \ldots, \mathbf{X}_k)^\top$ – случайный вектор. Тогда среди всех измеримых функций $f : \mathbb{R}^{k+1} \to \mathbb{R}$ минимум расстояния между $\mathbf{Y}$ и $f(\mathbf{X})$ достигается при f – функции регрессии Y на $\mathbf{X}$; т.е. функция $\varphi(\mathbf{X}) = \mathbb{E}(\mathbf{Y} \mid \mathbf{X})$ есть решение задачи*

$$\mathbb{E}\big(\mathbf{Y} - f(\mathbf{X})\big)^2 \to \min_f.$$

Доказательство. Покажем, что для любой f выполнено неравенство

$$\mathbb{E}\big(\mathbf{Y} - f(\mathbf{X})\big)^2 \geq \mathbb{E}\big(\mathbf{Y} - \varphi(\mathbf{X})\big)^2.$$

Имеем

$$\mathbb{E}\big(\mathbf{Y} - f(\mathbf{X})\big)^2 = \mathbb{E}\big(\mathbf{Y} - \varphi(\mathbf{X}) + \varphi(\mathbf{X}) - f(\mathbf{X})\big)^2 =$$

$$= \mathbb{E}\big(\mathbf{Y} - \varphi(\mathbf{X})\big)^2 + 2\mathbb{E}\left[\big(\mathbf{Y} - \varphi(\mathbf{X})\big)\big(\varphi(\mathbf{X}) - f(\mathbf{X})\big)\right] + \mathbb{E}\big(\varphi(\mathbf{X}) - f(\mathbf{X})\big)^2. \tag{32}$$

Используя закон повторных математических ожиданий можно показать, что второе слагаемое в (32) всегда равно нулю:

$$\mathbb{E}\left[\big(\mathbf{Y} - \varphi(\mathbf{X})\big)\big(\varphi(\mathbf{X}) - f(\mathbf{X})\big)\right] = \mathbb{E}\Big[\mathbb{E}\left[\big(\mathbf{Y} - \varphi(\mathbf{X})\big)\big(\varphi(\mathbf{X}) - f(\mathbf{X})\big) \mid \mathbf{X}\right]\Big] =$$

$$= \mathbb{E}\Big[\big(\varphi(\mathbf{X}) - f(\mathbf{X})\big)\mathbb{E}\left[\big(\mathbf{Y} - \varphi(\mathbf{X})\big) \mid \mathbf{X}\right]\Big] = 0,$$

поскольку $\mathbb{E}\left[\big(\mathbf{Y} - \varphi(\mathbf{X})\big) \mid \mathbf{X}\right] = \mathbb{E}(\mathbf{Y} \mid \mathbf{X}) - \mathbb{E}\left[\varphi(\mathbf{X}) \mid \mathbf{X}\right]$ и $\mathbb{E}\left[\varphi(\mathbf{X}) \mid \mathbf{X}\right] = \varphi(\mathbf{X})$, так как $\varphi(\mathbf{X})$ не содержит случайных компонент при фиксированном $\mathbf{X}$.

Третье слагаемое в (32) всегда неотрицательное, и требуемое неравенство доказано. □

Таким образом, зная совместное распределение $\mathbf{Y}$ и $\mathbf{X}$, можно построить оптимальную (в среднеквадратичном смысле) модель, предсказывающую значения $\mathbf{Y}$ по значениям $\mathbf{X}$, причём в большинстве случаев эта функция нелинейна по $\mathbf{X}$. Таким образом, в популяционном пространстве L^2 возникает критически важное разделение: линейная проекция $\mathbf{X}^\top\beta$ может не совпадать с оптимальной моделью $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$. Расстояние $\|\mathbb{E}[\mathbf{Y} \mid \mathbf{X}] - \mathbf{X}^\top\beta\|_2$ и есть строгая математическая мера ошибки спецификации линейной модели.

В выборочном пространстве $\mathbb{R}^n$ концепция "ошибки спецификации" отсутствует в явном виде: вектор Y всегда можно разложить на проекцию $\hat{Y}$ и ортогональный остаток $\hat{\varepsilon}$, независимо от того, как были сгенерированы данные. Тем не менее, если в популяции, из которой сгенерирована выборка, функции $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$ и $\mathbf{X}^\top\beta$ существенно отличаются, то стандартная модель множественной регрессии, оценённая по выборке, будет неоптимальной – и, скорее всего, будет уступать по качеству прогнозов моделям, позволяющим воспроизводить более гибкую зависимость Y от X. Большие проблемы ошибочная спецификация модели приносит и в задачу оценки эффектов воздействия.

Даже при точной спецификации популяционной зависимости, оценка модели на выборке может оказаться несостоятельной, если выборка является смещённой, то есть, систематически отличается от популяции по какому-то признаку. Пример такой ситуации изображён на рис. 6.

Ниже представлена сводная таблица 2, формализующая принцип аналогии между свойствами линейной модели на конечной выборке и в генеральной совокупности.

Таблица 2: Дуальность выборочных и популяционных свойств линейной модели

Концепция	**Конечная выборка (геометрия $\mathbb{R}^n$)**	**Генеральная совокупность (геометрия L^2)**
Базовое пространство	Евклидово пространство $\mathbb{R}^n$ со стандартным скалярным произведением $\langle u, v\rangle = u^\top v$.	Гильбертово пространство $L^2(\Omega, \mathcal{F}, P)$ случайных величин с конечным вторым моментом и скалярным произведением $\langle U, V\rangle = \operatorname{cov}(U, V)$.
Исходные данные	Вектор $y \in \mathbb{R}^n$ и матрица регрессоров $X \in \mathbb{R}^{n\times(k+1)}$.	Случайная величина $\mathbf{Y}$ и случайный вектор $\mathbf{X} = (1, \mathbf{X}_1, \ldots, \mathbf{X}_k)^\top$.
Целевое подпространство	Линейная оболочка столбцов матрицы X: $\mathcal{L}(X) = \{Xb : b \in \mathbb{R}^{k+1}\} \subset \mathbb{R}^n$.	Линейная оболочка регрессоров: $\mathcal{L}(\mathbf{X}) = \{\mathbf{X}^\top b : b \in \mathbb{R}^{k+1}\} \subset L^2$.
Критерий оптимальности	Минимизация суммы квадратов остатков (SSR): $\min_b \|y - Xb\|^2$	Минимизация среднеквадратичной ошибки (MSE): $\min_b \mathbb{E}(\mathbf{Y} - \mathbf{X}^\top b)^2 = \min_b \|\mathbf{Y} - \mathbf{X}^\top b\|_2^2$
Оценка / Параметр	Оценка МНК $\hat\beta = (X^\top X)^{-1} X^\top y$	Идентифицированный параметр линейной модели $\beta = (\mathbb{E}[\mathbf{X}\mathbf{X}^\top])^{-1}\mathbb{E}[\mathbf{X}\mathbf{Y}]$
Геометрическая операция	Ортогональная проекция y на $\mathcal{L}(X)$; $\hat y = P_X y$; матрица проектора $P_X = X(X^\top X)^{-1}X^\top$	Ортогональная проекция $\mathbf{Y}$ на $\mathcal{L}(\mathbf{X})$: $\hat{\mathbf{Y}} = \mathbf{X}^\top\beta$
Условия ортогональности	Вектор остатков ортогонален столбцам X: $X^\top(y - X\hat\beta) = 0_{k+1}$	Популяционная ошибка ортогональна регрессорам: $\operatorname{cov}(\mathbf{X}, \mathbf{Y} - \mathbf{X}^\top\beta) = 0_{k+1}$
Наилучшая линейная модель (предиктор)	Эмпирическая линия регрессии (подогнанные значения): $\hat y_i = X_i^\top \hat\beta$	BLP-функция: $\hat{\mathbf{Y}} = \mathbf{X}^\top\beta$
Наилучшая модель (предиктор)	*Не определена для конечной выборки.*	Проекция $\mathbf{Y}$ на подпространство всех функций от $\mathbf{X}$ в L^2 (УМО): $\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]$.
Связь (асимптотика): По ЗБЧ, $\frac{1}{n}X^\top X \xrightarrow{p} \mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ и $\frac{1}{n}X^\top y \xrightarrow{p} \mathbb{E}[\mathbf{X}\mathbf{Y}]$, следовательно, $\hat\beta \xrightarrow{p} \beta$.		

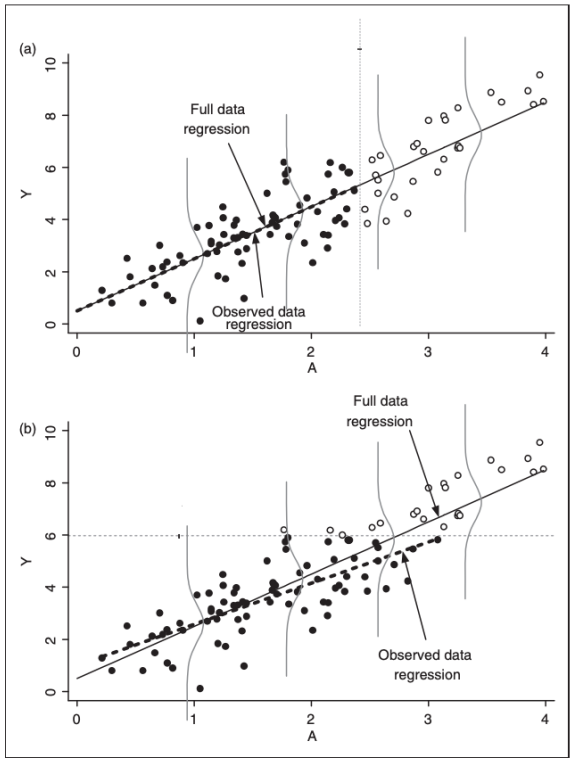


Рис. 6: Диаграмма рассеяния для парной регрессии на *смещённых выборках*; в одном случае все $X_i < 2.5$, а в другом случае $Y_i < 6$. Также изображена условная плотность ошибок регрессии в предположении экзогенности $\mathbb{E}(\varepsilon_i \mid X) = 0$

5 Инференция в линейных моделях

5.1 Характеристики оценённой линейной модели

Оценённая линейная модель позволяет разделить вектор Y на две компоненты: часть, "объяснённую" с помощью X (предсказанные/восстановленные значения Y), и необъяснённые остатки,

$$Y = \hat{Y} + \hat{\varepsilon}. \tag{33}$$

Обозначим суммы квадратов отклонений от среднего у координат этих векторов следующим образом:

$$SST := \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \quad SSE := \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})^2, \quad SSR := \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2,$$

Названия этих конструкций связаны с аббревиатурами: SST – sum of squares total (общая), SSE – sum of squares explained (объяснённая), SSR – sum of squares residual (остатков). В литературе встречаются и иные расшифровки.

Если регрессионная модель содержит константу, то $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$, и, применяя усреднение по выборке к обеим сторонам равенства (33), получаем $\bar{Y} = \bar{\hat{Y}}$,

Утверждение 5.1. Для множественной регрессии с константой выполняется

$$SST = SSE + SSR.$$

Доказательство. Мы хотим доказать

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{\hat{Y}}\right)^2+\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2. \tag{34}$$

Заметим, что $Y-\bar{Y}1_n=(\hat{Y}-\bar{\hat{Y}}1_n)+(Y-\hat{Y})$, поскольку $\bar{\hat{Y}}=\bar{Y}$ для регрессий с константой, и SST, SSE и SSR являются квадратами норм (длин) соответствующих векторов. Поскольку $(\hat{Y}-\bar{\hat{Y}})\in\mathcal{L}(X)$ и $(Y-\hat{Y})\in\mathcal{L}^{\perp}(X)$ по свойствам МНК-метода, то требуемое равенство (34) верно в силу теоремы Пифагора. □

Существует несколько стандартных характеристик оценённой модели, которые позволяют понять, насколько хорошо она соответствует зависимости между данными в выборке:

- Коэффициент детерминации R^2 в моделях с константой;
- Скорректированный коэффициент детерминации R^2_{adj};
- Стандартная ошибка регрессии – оценка стандатного отклонения ошибки модели;
- F-статистика значимости коэффициентов регрессии.

Первая из этих характеристик является отношением выборочных дисперсий $\hat{Y}$ и Y, или, что эквивалентное, отношением сумм квадратов

$$R^2:=\frac{SSE}{SST}.$$

Для анализа свойств R^2 удобно перейти от исходной модели к модели в отклонениях от средних значений (центрирование). Для этого можно применить теорему FWL к стандартной линейной модели, осуществляя регрессию каждой из компонент на вектор-константу. Пусть исходная модель, оценённая МНК, имеет вид

$$Y=X\hat{\theta}+\hat{\varepsilon}, \tag{35}$$

где X содержит константу (столбец из единиц). Умножая обе части равенства (92) на матрицу $Q_1:=I_n-\frac{1}{n}1_n1_n^{\top}$, проецирующую на подпространство $\mathbb{R}^n$ векторов, ортогональных константе, получаем

$$\tilde{Y}=\tilde{X}\hat{\theta}+\hat{\varepsilon}, \tag{36}$$

где $\tilde{Y}=Y-\bar{Y}$ – вектор отклонений от среднего значения зависимой переменной. Аналогично, столбцы $\tilde{X}$ являются столбцами X, из которых вычли соответствующие средние значения (столбец единиц при этом полностью зануляется). Вектор $\hat{\varepsilon}$ под действием проектора Q_1 не изменяется, так как он изначально был ортогонален вектору-константе.

В уравнении (36) остатки $\hat{\varepsilon}$ по-прежнему ортогональны столбцам матрицы $\tilde{X}$, поэтому это уравнение задаёт разложение $\tilde{Y}$ на $\tilde{X}\hat{\theta}\in\mathcal{L}(\tilde{X})$ и $\hat{\varepsilon}\in\mathcal{L}^{\perp}(\tilde{X})$.

Лемма 5.2. *В множественной регрессии с константой выполняется* $R^2=\widehat{\mathrm{corr}}^2(\hat{Y},Y)$. *Для парной регрессии также верно* $Y_i=\beta_0+\beta_1X_i+\varepsilon_i$ *является* $R^2=\widehat{\mathrm{corr}}^2(X,Y)$, *где* X *– вектор значений* X_i.

Доказательство. Рассмотрим модель в отклонениях от средних значений (36). Поскольку $\widehat{\tilde{Y}}:=\tilde{X}\hat{\theta}$ является ортогональной проекцией $\tilde{Y}$ на пространство $\mathcal{L}(\tilde{X})$, косинус угла α между $\tilde{Y}$ и $\widehat{\tilde{Y}}$ – это

$$\cos(\alpha)=\frac{\|\widehat{\tilde{Y}}\|}{\|\tilde{Y}\|},$$

где $\|\cdot\|$ – евклидова норм в $\mathbb{R}^n$. Также, по определению скалярного произведения

$$\langle \widehat{\tilde{Y}}, \tilde{Y} \rangle = \|\widehat{\tilde{Y}}\| \cdot \|\tilde{Y}\| \cdot \cos(\alpha) \Rightarrow \cos(\alpha) = \frac{\langle \widehat{\tilde{Y}}, \tilde{Y} \rangle}{\|\widehat{\tilde{Y}}\| \cdot \|\tilde{Y}\|} \Rightarrow \tag{37}$$

$$\frac{\|\widehat{\tilde{Y}}\|^2}{\|\tilde{Y}\|^2} = \frac{\left(\langle \widehat{\tilde{Y}}, \tilde{Y} \rangle\right)^2}{\|\widehat{\tilde{Y}}\|^2 \cdot \|\tilde{Y}\|^2}.$$

Заметим, что $\|\widehat{\tilde{Y}}\|^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = SSE$, $\|\tilde{Y}\|^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 = SST$, и $\langle \widehat{\tilde{Y}}, \tilde{Y} \rangle = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})$, поэтому

$$R^2 = \frac{SSE}{SST} = \frac{\frac{1}{(n-1)^2}\left(\sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})\right)^2}{\frac{1}{n-1}\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \cdot \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\widehat{\mathrm{cov}}^2(\hat{Y}, Y)}{\hat{\mathbb{V}}(\hat{Y}) \cdot \hat{\mathbb{V}}(Y)} = \widehat{\mathrm{corr}}^2(\hat{Y}, Y).$$

□

Связь $R^2 = \cos^2(\alpha)$ можно использовать для интерпретации двух экстремальных случаев $R^2 = 1$ и $R^2 = 0$. В первом случае $Y - \bar{Y}$ и $\hat{Y} - \bar{\hat{Y}}$ совпадают и лежат в $\mathcal{L}(\tilde{X})$, а значит, регрессионная функция на рассматриваемой выборке в точности воспроизводит все значения зависимой переменной. Во втором случае, когда $R^2 = 0$, $Y - \bar{Y}1_n$ и $\hat{Y} - \bar{\hat{Y}}1_n$ ортогональны, а значит, используемые регрессоры непригодны для объяснения вариации в Y в используемой выборке.

Основная проблема применения R^2 заключается в том, что его значение почти всегда увеличивается (и не может уменьшаться) от добавления в модель новых переменных, даже если эти переменные никакого отношения к объясняемой переменной не имеют. Поэтому для сравнения моделей с разным количеством регрессоров лучше использовать скорректированный коэффициент детерминации R^2_{adj}, штрафующий за дополнительно включённые переменные. Пусть n – количество наблюдений, а k – количество регрессоров (включая константу). Тогда

$$R^2_{adj} := 1 - \left(1 - R^2\right)\frac{(n-1)}{(n-k)} \leq R^2.$$

Если рассматривать аналогичные (37) геометрические соотношения между исходными (нецентрированными) векторами Y, $\hat{Y}$ и $\hat{\varepsilon}$, то можно получить так называемый нецентрированный коэффициент детеминации $R^2_u = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2}$. Он редко используется в вычислениях, поскольку не является инвариантным к добавлению константы к зависимой переменной.[14]

5.2 Инференция в линейной модели с нормальными ошибками

Переходя к вопросам инференции в линейной модели, следует напомнить свойства наиболее простого случая, возникающего при нормальности ошибок модели.

Теорема 5.3. *Пусть задана линейная модель:*

$$Y = X\beta + \varepsilon, \tag{38}$$

причём ошибки имеют центрированное нормальное распределение при фиксированном X:

$$\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2 I_n). \tag{39}$$

[14] Davidson R., MacKinnon J.G. (2021). *Estimation and Inference in Econometrics*. Oxford University Press, p.14.

В таком случае условное распределение вектора оценок МНК имеет вид:

$$\hat{\beta} \mid X \sim \mathcal{N}\left(\beta, \sigma^2 (X^\top X)^{-1}\right)$$

Для отдельного j-го коэффициента это означает, что $\hat{\beta}_j \mid X \sim \mathcal{N}\big(\beta_j, \sigma^2 [(X^\top X)^{-1}]_{jj}\big)$.

Доказательство. Заметим, что в такой постановке β однозначно идентифицируется условиями ортогональности ε и $\mathbf{X}$ в предположении невырожденности $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$.

Оценка МНК является линейной функцией от вектора ошибок:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = \beta + (X^\top X)^{-1} X^\top \varepsilon.$$

Поскольку условное распределение ε – многомерное нормальное, любая его линейная трансформация (при фиксированной матрице X) также имеет нормальное распределение. Вычисляя условное математическое ожидание и ковариационную матрицу, мы получаем параметры распределения МНК-оценок:

$$\mathbb{E}[\hat{\beta} \mid X] = \beta + (X^\top X)^{-1} X^\top \mathbb{E}[\varepsilon \mid X] = \beta,$$
$$\mathbb{V}(\hat{\beta} \mid X) = (X^\top X)^{-1} X^\top \mathbb{V}(\varepsilon \mid X) X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1}.$$

□

Теорема 5.4. *(Лемма Фишера для линейных моделей).*

В линейной модели с нормальными ошибками (38)-(39), где общее количество регрессоров (включая константу) равно $k+1$, а размер случайной выборки равен n, стандартная оценка дисперсии $S^2 = \frac{SSR}{n-k-1}$ обладает следующими свойствами:

1. $\frac{(n-k-1)S^2}{\sigma^2} \mid X \sim \chi^2(n-k-1)$ *(хи-квадрат с $n-k-1$ степенями свободы).*
2. *МНК-оценка $\hat{\beta}$ и S^2 условно независимы при фиксированном наборе регрессоров X.*

Доказательство этого результата аналогично доказательству теоремы о распределении F-статистик в моделях с нормальными ошибками, приведённому в приложении.

Лемма Фишера позволяет строить точные доверительные интервалы и тестировать гипотезы о значениях компонент вектора β. Рассмотрим проверку нулевой гипотезы $H_0 : \beta_j = \beta_{j0}$ (часто $\beta_{j0} = 0$). Стандартизированная оценка имеет стандартное нормальное распределение:

$$Z_j = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 [(X^\top X)^{-1}]_{jj}}} \mid X \sim \mathcal{N}(0,1).$$

Поскольку σ^2 неизвестно, мы заменяем его на оценку S^2. Отношение стандартной нормальной случайной величины к корню из независимой величины χ^2, деленной на её степени свободы, по определению дает распределение Стьюдента (t-распределение):

$$t_j = \frac{Z_j}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{S^2 [(X^\top X)^{-1}]_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{\widehat{s.e.}(\hat{\beta}_j)}$$

При истинности H_0, статистика t_j имеет **точное** t-распределение с $n-k-1$ степенями свободы:

$$t_j \mid X \sim t(n-k-1).$$

Это позволяет проводить тесты, сравнивая $|t_j|$ с нужными квантилями распределения $t(n-k-1)$ и вычислять p-значения для выборок любого объема $n > k+1$, не прибегая к асимптотическим аппроксимациям.

Инвертируя процедуру проверки гипотез, можно построить точный $(1-\alpha)$-доверительный интервал для параметра β_j. Исходя из вероятностного утверждения $\mathrm{P}(t_{\alpha/2} \le t_j \le -t_{\alpha/2} \mid X) = 1-\alpha$, где $t_{\alpha/2}$ – квантиль уровня $\alpha/2$ распределения $t(n-k-1)$, получаем:

$$CI_{1-\alpha}(\beta_j) = \left[\hat{\beta}_j + t_{\alpha/2} \cdot \widehat{s.e.}(\hat{\beta}_j), \quad \hat{\beta}_j - t_{\alpha/2} \cdot \widehat{s.e.}(\hat{\beta}_j)\right].$$

Теорема 5.5. *(Стандартный F-тест линейных гипотез при нормальных ошибках регрессии; доказательство в приложении).*

Рассмотрим регрессионную модель с нормальными ошибками,

$$Y = X\beta + Z\gamma + \varepsilon, \quad \varepsilon \mid X, Z \sim \mathcal{N}(0, \sigma^2 I_n). \tag{40}$$

Предположим, что X и Z – матрицы регрессоров размером $n \times k$ и $n \times j$, и мы хотим тестировать $H_0 : \beta = 0_k$ против $H_a : \beta \neq 0_k$. Для этого методом наименьших квадратов оценим исходную модель (40), а также короткую модель

$$Y = Z\tilde{\gamma} + u, \tag{41}$$

и обозначим получающиеся суммы квадратов остатков за SSR и SSR^* соответственно.

Утверждение: F-статистика $F = \frac{SSR^* - SSR/k}{SSR/n-k-j}$ при H_0 имеет $F(k, n-k-j)$ распределение.

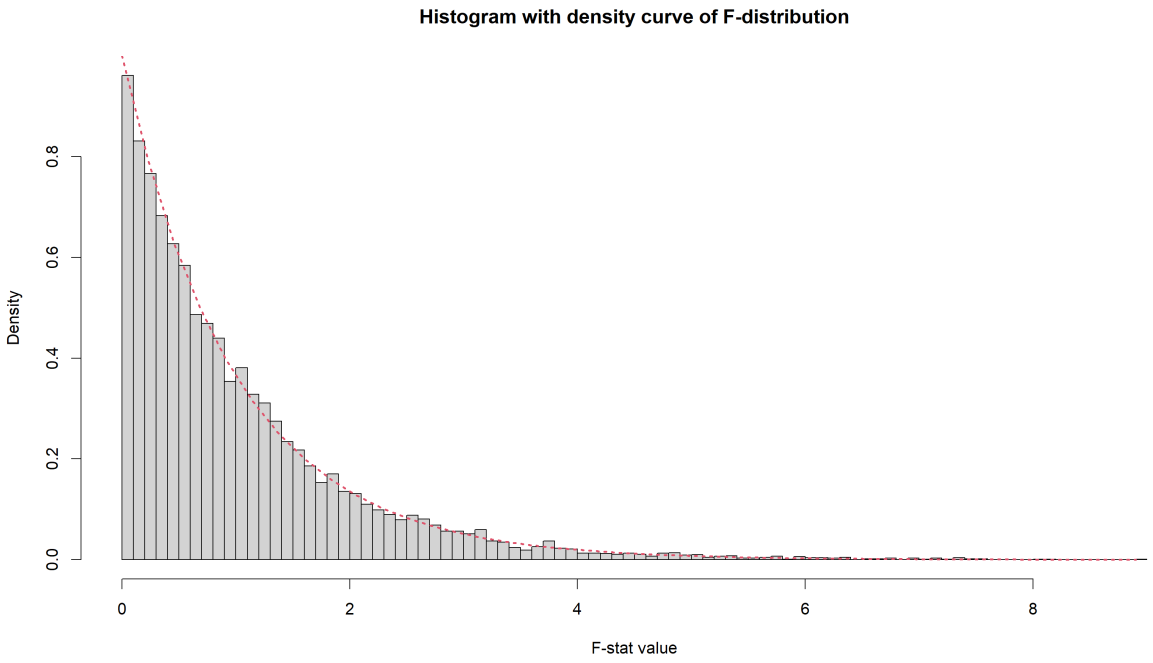


Рис. 7: Гистограмма значений F-статистики, полученных в 10000 Монте-Карло экспериментах с размером выборки $n = 100$. Пунктирная красная линия – теоретическая плотность случайной величины с соответствующим F-распределением.

```
# A numeric example on F-test
# Draw data according to a DGP Y_i = α + γ_1 z_i1 + γ_2 z_i2 + ε_i but estimate a different model
```

$Y_i = \tilde{\alpha} + \beta_1 x_{i1} + \beta_2 x_{i2} + \tilde{\gamma}_1 z_{i1} + \tilde{\gamma}_2 z_{i2} + \tilde{\varepsilon}_i,$

```r
# with x, z, ε ~ i.i.d. N(0,1).
N <- 10000
n <- 100
#N is the number of Monte-Carlo samples, n is the number of observations in one sample

X <- matrix(rnorm(2*N*n, 0, 1), ncol = 2)
Z <- cbind(rep(1, N*n), matrix(rnorm(2*N*n, 0, 1), ncol = 2))
epsilon <- rnorm(N*n, 0, 1)

alpha <- 1
gamma_1 <- 4
gamma_2 <- 5
Y <- Z %*% c(alpha, gamma_1, gamma_2) + epsilon
db <- cbind(Y,X,Z)
one_sample_est <- function(start, step = n, ddbb = db){
  yy <- ddbb[start:(start+step-1), 1]
  xx <- ddbb[start:(start+step-1), 2:3]
  zz <- ddbb[start:(start+step-1), 4:6]
  xz <- cbind(xx,zz)
  res_short <- (diag(x = 1, ncol = n, nrow = n) - zz %*% solve(t(zz) %*% zz) %*% t(zz)) %*% yy
  res_long <- (diag(x = 1, ncol = n, nrow = n) - xz %*% solve(t(xz) %*% xz) %*% t(xz)) %*% yy
  return((norm(res_short, type = "F")^2-norm(res_long, type = "F")^2)*(n-5)/
           (2*norm(res_long, type = "F")^2))
}
est_array <- sapply(seq(1,(n*N),n), one_sample_est)
hist(est_array, breaks = max(N/100,10), prob = TRUE,
  main = "Histogram with density curve of F-distribution",
  xlab = "F-stat value")

x2 <- seq(min(est_array), max(est_array), length = 40)
fun <- df(x2, 2, n-5)
# density function of F(2, n − k)

lines(x2, fun, col=2, lty="dotted", lwd = 2)
```

Часто F-тесты линейных моделей применяются для проверки совместной незначимости всех регрессоров. Для этого в качестве Z берется константа, а в качестве X – все остальные регрессоры, и теструется $H_0 : \beta = 0_k$. Если тест не отвергает H_0, это значит, что регрессоры в модели являются совместно незначимыми, и объясняющая способность "длинной" модели – такая же, как у модели, в которой Y регрессируется только на константу.

Подводя итоги, можно сказать, что предположение о нормальности ошибок – мощный, но уязвимый инструмент.

- Если ошибки нормальны, мы получаем **точные** распределения t и F в конечных выборках, независимо от того, насколько мало n.
- Если ошибки не нормальны, получаемые t и F статистики имеют распределения, отличающиеся от теоретических, и фактические уровни значимости и другие показатели, используемые для инференции, могут быть далеки от расчётных. В то же время, если n достаточно велико, то можно использовать

асимптотические версии соответствующих тестов, не требующие точного знания распределения ошибок.

5.3 Статистические свойства МНК-оценок

При оценке регрессии объектом оценивания является неизвестная функция, однако ввиду сделанного предположения о её линейности, нам достаточно рассмотреть вероятностные характеристики оценок её параметров. Существует теоретический результат, гарантирующий, что если для оценивания выбрана правильная модель, то МНК-оценки будут наиболее эффективными в некотором, достаточно широком классе оценок – а значит, и оценённая модель будет наилучшей возможной (в некотором смысле) моделью для зависимой переменной. Нюанс, делающий этот результат несущественным в практическом отношении, заключается в том, что исследователю предлагается "угадать" корректную форму функции регрессии $\mathbf{Y}$ для выполнения условий теоремы.

Предположим, что регрессионная модель удовлетворяет следующим условиям:

ГМ1 Модель является линейной относительно оцениваемых параметров;

ГМ2 $\{(Y_i, X_i)\}_{i=1}^n$ – *н.о.р.* выборка, где X_i – k-мерный вектор регрессоров $(X_{i1}, \ldots, X_{ik})$;

ГМ3 Нет полной мультиколлинеарности, то есть выборочные реализации компонент X_i (которые являются векторами-столбцами матрицы регрессоров) образуют линейно независимый набор векторов в $\mathbb{R}^n$ с вероятностью 1;

ГМ4 Нулевое условное среднее ошибки: $\mathbb{E}(\varepsilon_i \mid X_{i1}, \ldots, X_{ik}) = 0$;

ГМ5 Гомоскедастичность ошибки: $\mathbb{V}(\varepsilon_i \mid X_{i1}, \ldots, X_{ik}) = const$.

Теорема 5.6. *(Гаусс, Марков; доказательство в приложении)*
Если предположения ГМ1-ГМ5 выполнены, то МНК-оценки являются оптимальными – то есть, имеют самую низкую дисперсию в классе линейных (по Y) несмещенных оценок.

Популярный вид оценки дисперсии регрессионной ошибки также основан на использовании предположений ГМ4-ГМ5.

Лемма 5.7. *Пусть в линейной модели $Y = X\beta + \varepsilon$ с k регрессорами $X := (X_1, \ldots, X_k)$ ошибка удовлетворяет предположениям ГМ4 и ГМ5:*

$$\mathbb{E}(\varepsilon \mid X) = 0, \quad \mathbb{V}(\varepsilon \mid X) = \sigma^2 I_n.$$

Тогда несмещённой оценкой дисперсии ошибки σ^2 по случайной выборке размером n является

$$\hat{\sigma}^2 := \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-k} = \frac{SSR}{n-k}.$$

Доказательство. МНК-оценка имеет вид $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$,

$$\hat{\varepsilon} = Y - X\hat{\beta} = \left(I_n - X(X^\top X)^{-1}X^\top\right)Y = (I_n - P_X)(X\beta + \varepsilon) = (I_n - P_X)\varepsilon,$$

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}^\top \varepsilon = \left((I_n - P_X)\varepsilon\right)^\top (I_n - P_X)\varepsilon = \varepsilon^\top (I_n - P_X)\varepsilon.$$

где $P_X = X(X^\top X)^{-1}X^\top$ – проектор на $\mathcal{L}(X)$.

Чтобы найти математическое ожидание $\hat{\varepsilon}^\top\hat{\varepsilon}$, удобно использовать свойство следа матрицы $\mathrm{tr}(\cdot)$:

- След от числа (скаляра) – само это число;
- Для любых $n \times k$ и $k \times n$ матриц A и B выполнено $\mathrm{tr}(AB) = \mathrm{tr}(BA)$;
- Если матрица Q имеет конечный размер, то операторы математического ожидания и следа всегда коммутируют: $\mathbb{E}\big(\mathrm{tr}(Q)\big) = \mathrm{tr}\big(\mathbb{E}(Q)\big)$.

Поскольку $\hat{\varepsilon}^\top\hat{\varepsilon}$ является скаляром, $\hat{\varepsilon}^\top\hat{\varepsilon} = \mathrm{tr}(\hat{\varepsilon}^\top\hat{\varepsilon}) = \mathrm{tr}\big(\varepsilon^\top(I_n - P_X)\varepsilon\big) = \mathrm{tr}\big((I_n - P_X)\varepsilon\varepsilon^\top\big)$;

$$\mathbb{E}[\hat{\varepsilon}^\top\hat{\varepsilon} \mid X] = \mathbb{E}\left[\mathrm{tr}\big((I_n - P_X)\varepsilon\varepsilon^\top\big) \mid X\right] = \mathrm{tr}\left[\mathbb{E}\big((I_n - P_X)\varepsilon\varepsilon^\top \mid X\big)\right] = \mathrm{tr}\left[(I_n - P_X)\mathbb{E}(\varepsilon\varepsilon^\top \mid X)\right] =$$

$$= \mathrm{tr}\left[(I_n - P_X)\sigma^2 I_n\right] = \sigma^2\mathrm{tr}(I_n - P_X).$$

Очевидно, $\mathrm{tr}(I_n - P_X) = \mathrm{tr}(I_n) - \mathrm{tr}(P_X) = n - \mathrm{tr}\big(X(X^\top X)^{-1}X^\top\big) = n - \mathrm{tr}\big((X^\top X)^{-1}X^\top X\big) = n - k$, поскольку $X^\top X$ – это $k \times k$ матрица, и $(X^\top X)^{-1}X^\top X = I_k$ – $k \times k$ единичная матрица. Поэтому $\mathbb{E}[\hat{\varepsilon}^\top\hat{\varepsilon} \mid X] = \sigma^2(n-k)$, и её безусловное математическое ожидание – такое же, а следовательно, $\hat{\sigma}^2$ – несмещённая оценка σ^2.

```
# Check that σ̂² = (Σ_{i=1}^n ε̂_i²)/(n−k) is indeed the same estimator of error variance, that we get with lm() command
n <- 100
k <- 5
# number of observations and regressors in a linear model
X <- matrix(rnorm(k*n,0,1), ncol = k)
epsilon <- rnorm(n,0,1)
Y <- X %*% seq(1,k,1) + epsilon
# DGP is Y = Xθ + ε with i.i.d. normal X and ε, and θ = (1, 2, ..., k)^⊤.

ols_full <- lm(Y~X-1)
summary(ols_full)$sigma^2
# Value of error variance estimate obtained via lm() function
Qx <- diag(x=1,nrow = n, ncol = n) - X %*% solve(t(X) %*% X) %*% t(X)
e_hat <- Qx %*% Y
(t(e_hat) %*% e_hat)/(n-k)
# The same value, calculated manually
```

Значительно большую практическую ценность имеют результаты, не требующие выполнения предположения ГМ4, главными из которых являются состоятельность и асимптотическая нормальность МНК-оценок.

Утверждение 5.8. При выполнении условий ГМ1-ГМ3, МНК-оценки в множественной регрессии являются состоятельными и асимптотически нормальными. Если дополнительно выполнено условие ГМ4, то оценки являются несмещёнными.

Доказательство. Рассмотрим стандартную модель

$$Y = X\beta + \varepsilon,$$

в которой X – случайная матрица регрессоров $n \times k$.

МНК-оценкой вектора параметра β является вектор

$$\hat{\beta} = (X^\top X)^{-1}X^\top Y.$$

Условное математическое ожидание такой оценки при выполнении ГМ4 имеет вид

$$\mathbb{E}(\hat{\beta} \mid X) = \mathbb{E}\big((X^\top X)^{-1} X^\top Y \mid X\big) = (X^\top X)^{-1} X^\top \mathbb{E}(Y \mid X) = (X^\top X)^{-1} X^\top \mathbb{E}(X\beta + \varepsilon \mid X) =$$

$$= (X^\top X)^{-1} X^\top X\beta + (X^\top X)^{-1} X^\top \mathbb{E}(\varepsilon \mid X) = \beta.$$

Применяя закон повторных математических ожиданий, получаем $\mathbb{E}\hat{\beta} = \mathbb{E}\big[\mathbb{E}(\hat{\beta} \mid X)\big] = \mathbb{E}\beta = \beta$. Отметим, что этот результат также выполняется при гетероскедастичных ошибках, то есть без условия ГМ5.

Для состоятельности МНК-оценок условие ГМ4 $\mathbb{E}(\varepsilon \mid \mathbf{X}) = 0_n$ можно заменить на более слабое условие некоррелированности $\mathbb{E}(\mathbf{X}^\top \varepsilon) = 0_k$.[15]

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = \beta + (X^\top X)^{-1} X^\top \varepsilon = \beta + \left(\frac{X^\top X}{n}\right)^{-1} \left(\frac{X^\top \varepsilon}{n}\right).$$

При этом

$$\frac{X^\top \varepsilon}{n} = \left(\frac{1}{n}\sum_{i=1}^n X_{i1}\varepsilon_i, \ldots, \frac{1}{n}\sum_{i=1}^n X_{ik}\varepsilon_i\right)^\top \xrightarrow{p} \big(\mathbb{E}(X_{i1}\varepsilon_i), \ldots, \mathbb{E}(X_{ik}\varepsilon_i)\big)^\top = 0_k, \tag{42}$$

в силу закона больших чисел. Кроме того,

$$\frac{X^\top X}{n} = \frac{1}{n}\begin{pmatrix} X_{\cdot 1}^\top X_{\cdot 1} & X_{\cdot 1}^\top X_{\cdot 2} & \ldots & X_{\cdot 1}^\top X_{\cdot k} \\ X_{\cdot 2}^\top X_{\cdot 1} & X_{\cdot 2}^\top X_{\cdot 2} & \ldots & X_{\cdot 2}^\top X_{\cdot k} \\ \ldots & \ldots & \ldots & \ldots \\ X_{\cdot k}^\top X_{\cdot 1} & X_{\cdot k}^\top X_{\cdot 2} & \ldots & X_{\cdot k}^\top X_{\cdot k} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n X_{i1}^2 & \sum_{i=1}^n X_{i1}X_{i2} & \ldots & \sum_{i=1}^n X_{i1}X_{ik} \\ \sum_{i=1}^n X_{i2}X_{i1} & \sum_{i=1}^n X_{i2}^2 & \ldots & \sum_{i=1}^n X_{i2}X_{ik} \\ \ldots & \ldots & \ldots & \ldots \\ \sum_{i=1}^n X_{ik}X_{i1} & \sum_{i=1}^n X_{ik}X_{i2} & \ldots & \sum_{i=1}^n X_{ik}^2 \end{pmatrix} \xrightarrow{p}$$

$$\xrightarrow{p} \begin{pmatrix} \mathbb{E}(X_{i1}^2) & \mathbb{E}(X_{i1}X_{i2}) & \ldots & \mathbb{E}(X_{i1}X_{ik}) \\ \mathbb{E}(X_{i2}X_{i1}) & \mathbb{E}(X_{i2}^2) & \ldots & \mathbb{E}(X_{i2}X_{ik}) \\ \ldots & \ldots & \ldots & \ldots \\ \mathbb{E}(X_{ik}X_{i1}) & \mathbb{E}(X_{ik}X_{i2}) & \ldots & \mathbb{E}(X_{ik}^2) \end{pmatrix},$$

то есть, в общем случае эта компонента сходится по вероятности к невырожденной матрице $k \times k$, которую, используя вектор регрессоров для i-го наблюдения X_i, можно записать как $\mathbb{E}(X_i X_i^\top)$. Произведение $\left(\frac{X^\top X}{n}\right)^{-1}\left(\frac{X^\top \varepsilon}{n}\right)$ в силу теоремы о непрерывном отображении сходится к нулю по вероятности:

$$\left(\frac{X^\top X}{n}\right)^{-1}\left(\frac{X^\top \varepsilon}{n}\right) \xrightarrow{p} \mathbb{E}(X_i X_i^\top) \cdot 0_k = 0_k,$$

откуда следует $\hat{\beta} \xrightarrow{p} \beta$.

Асимптотическая нормальность МНК-оценок доказывается схожим способом. Здесь снова условие ГМ4 можно заменить на некоррелированность ошибок и X, а ГМ5 является избыточными, но оба этих условия облегчают вычисление параметров предельного распределения. Рассмотрим разность между оценкой и истинным значением, умноженную на стабилизирующий коэффициент $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{X^\top X}{n}\right)^{-1}\left(\frac{X^\top \varepsilon}{\sqrt{n}}\right) \tag{43}$$

В полученном выражении первая компонента произведения сходится по вероятности к $\mathbb{E}(X_i X_i^\top)$, а вторая,

[15] Некоррелированность следует из ГМ4, так как при $\mathbb{E}(\varepsilon \mid X) = 0_n$ выполняется $\mathbb{E}(X^\top \varepsilon) = \mathbb{E}[\mathbb{E}(X^\top \varepsilon \mid X)] = \mathbb{E}[X^\top \mathbb{E}(\varepsilon \mid X)] = 0$.

как следует из (42), имеет вид

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i1}\varepsilon_i, \ldots, \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{ik}\varepsilon_i\right)^\top,$$

где каждая из компонент вектора удовлетворяет условиям центральной предельной теоремы, и поэтому такой вектор сходится по распределению. Тогда из теоремы Слуцкого следует, что произведение сходится по распределению к нормальному вектору Дисперсию предельного распределения можно найти, пользуясь разложением по условию X и несмещённостью оценки:

$$\mathbb{V}\big(\sqrt{n}(\hat\beta-\beta)\big) = \mathbb{E}\big[\mathbb{V}\big(\sqrt{n}(\hat\beta-\beta)\mid X\big)\big] + \mathbb{V}\big[\mathbb{E}\big(\sqrt{n}(\hat\beta-\beta)\mid X\big)\big] = \mathbb{E}\big[\mathbb{V}\big(\sqrt{n}(\hat\beta-\beta)\mid X\big)\big],$$

$$\mathbb{V}\big(\sqrt{n}(\hat\beta-\beta)\mid X\big) = n(X^\top X)^{-1}X^\top\mathbb{V}(\varepsilon\mid X)\big((X^\top X)^{-1}X^\top\big)^\top =$$

$$n\sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1} = n\sigma^2(X^\top X)^{-1} \Rightarrow \mathbb{E}\big[\mathbb{V}\big(\sqrt{n}(\hat\beta-\beta)\mid X\big)\big] = \mathbb{E}\big[n\sigma^2(X^\top X)^{-1}\big].$$

Далее, предполагая возможность поменять местами предел и математическое ожидание,[16] получаем

$$\lim_{n\to\infty}\mathbb{E}\big[n\sigma^2(X^\top X)^{-1}\big] = \mathbb{E}\left[\lim_{n\to\infty}\sigma^2\left(\frac{X^\top X}{n}\right)^{-1}\right] = \mathbb{E}\big[\sigma^2\big(\mathbb{E}(X_iX_i^\top)\big)^{-1}\big] = \sigma^2\big(\mathbb{E}(X_iX_i^\top)\big)^{-1}.$$

Очень часто на практике вместо инференции на полную популяцию, которую можно получить, применяя асимптотическое распределение, используется инференция на субпопуляцию, имеющую значения регрессоров X, использованные для оценки модели – то есть, фактически, допускается изменение только значений ошибок регрессии. В таком случае нам нужно рассматривать условное распределение оценки

$$\hat\beta \mid X = (X^\top X)^{-1}X^\top Y \mid X,$$

которая имеет структуру линейной функции от случайного вектора $(\sum_{i=1}^n X_{i1}Y_i, \ldots, \sum_{i=1}^n X_{ik}Y_i)^\top$, для компонент которого выполнены условия ЦПТ. Тогда при больших n можно считать распределение $\hat\beta \mid X$ приближенно нормальным с параметрами

$$\mathbb{E}(\hat\beta\mid X) = \beta \quad \text{при нулевом условном математическом ожидании ошибок,}$$

$$\mathbb{V}(\hat\beta\mid X) = (X^\top X)^{-1}X^\top\mathbb{V}(\varepsilon\mid X)X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1} \quad \text{при гомоскедастичности ошибок.}$$

Если ошибки предполагаются гетероскедастичными, то в качестве оценки ковариационной матрицы $\mathbb{V}(\hat\beta \mid X)$ можно взять выборочный аналог $(X^\top X)^{-1}X^\top\mathbb{V}(\varepsilon\mid X)X(X^\top X)^{-1}$ – оценки Уайта.

5.4 Тестирование гипотез в линейных моделях

Асимптотическое распределение МНК-оценок используется для построения t-статистик, которые можно применять для тестирования гипотез вида $H_0 : \beta_j = \beta_{j0}$, где альтернатива может быть односторонней или двухсторонней. Тогда при H_0 в модели с гомоскедастичными ошибками выполнено

$$t = \frac{\hat\beta_j - \beta_{j0}}{\sqrt{\big[\hat\sigma^2(X^\top X)^{-1}\big]_{jj}}} \xrightarrow{d} \mathcal{N}(0,1), \tag{44}$$

[16] Техническим условием для этого является то, что минимальное из собственных чисел матрицы $\mathbb{E}(X_iX_i^\top)$ отделено от нуля.

где в знаменателе дроби находится стандартное отклонение оценки $\hat{\beta}_j$ – то есть, j-й элемент на главной диагонали матрицы условной ковариации $\sigma^2(X^\top X)^{-1}$. При этом, поскольку дисперсия ошибки модели как правило незивестна, то вместо неё используется оценка $\hat{\sigma}^2 = \frac{SSR}{n-k}$.

Конструкции наподобие (44) получили такое название в связи с тем, что ранние исследования линейных моделей часто предполагали нормальность ошибок. В этом случае, используя аналог леммы Фишера, можно показать, что на малых выборках t-статистика имеет распределение Стьюдента $t(n-k)$, также называемое t-распределением.[17] В результате, некоторое время в литературе использовались разные обозначения такой тестовой статистики для случая нормальных ошибок и для остальных случаев – когда используется асимптотическая инференция (в этом случае использовалось название z-статистика, поскольку Z – типичное обозначение стандартной нормальной случайной величины). Наиболее часто на практике t-статистики используются для проверки гипотезы о значимости определённого регрессора в модели, то есть для тестирования $H_0 : \beta_j = 0$.

Далее рассмотрим три классических теста, которые можно использовать для проверки гипотез с линейными нелинейными ограничениями. Все три из них – тест Вальда, тест отношения правдоподобия (англ.: likelihood ratio, LR) и тест множителей Лагранжа (англ.: Lagrange multiplier, LM) имеют одинаковое асимптотическое распределение тестовой статистики при H_0, но могут иметь разную мощность в конечных выборках.

Лемма 5.9. *МНК-оценка в модели с ограничениями.*

Рассмотрим линейную модель с k регрессорами в предположениях ГМ1-ГМ3 и некоррелированности ошибок с регрессорами:

$$Y = X\beta_0 + u.$$

Предположим, что априорно известно, что популяционный параметр β удовлетворяет ограничению вида $R\beta = r$, где R – $q \times k$ матрица ранга q. Тогда МНК-оценка ограниченной модели имеет следующий вид:

$$\hat{\beta}_R = \hat{\beta} - (X^\top X)^{-1} R^\top \big[R(X^\top X)^{-1} R^\top\big]^{-1} \big(R\hat{\beta} - r\big).$$

Доказательство

Функция Лагранжа:

$$\mathcal{L}(\beta, \lambda) = (Y - X\beta)^\top (Y - X\beta) + 2\lambda^\top (R\beta - r).$$

Условия первого порядка:

$$\frac{\partial \mathcal{L}}{\partial \beta} = -2X^\top (Y - X\hat{\beta}_R) + 2R^\top \lambda_R = 0 \Rightarrow X^\top X \hat{\beta}_R = X^\top Y - R^\top \lambda_R.$$

Умножая на $(X^\top X)^{-1}$, получаем выражение для оценки:

$$\hat{\beta}_R = \hat{\beta} - (X^\top X)^{-1} R^\top \lambda_R. \tag{45}$$

Для найденной оценки также должно выполняться ограничение $R\hat{\beta}_R = r$:

$$R\hat{\beta} - R(X^\top X)^{-1} R^\top \lambda_R = r \Rightarrow \hat{\lambda}_R = \big[R(X^\top X)^{-1} R^\top\big]^{-1} (R\hat{\beta} - r). \tag{46}$$

[17] Это не противоречит (44), поскольку распределение Стьюдента сходится к стандартному нормальному с ростом параметра степени свободы.

Подставляем (46) обратно в (45):

$$\hat{\beta}_R = \hat{\beta} - (X^\top X)^{-1} R^\top \left[R(X^\top X)^{-1} R^\top \right]^{-1} (R\hat{\beta} - r). \tag{47}$$

Аналогичную оценку можно получить, когда ограничения задаются нелинейной гладкой функцией $g(\beta) = r$, где $g : \mathbb{R}^k \to \mathbb{R}^q$; в этом случае $\hat{\beta}_R = \hat{\beta} - (X^\top X)^{-1} G^\top \left[G(X^\top X)^{-1} G^\top \right]^{-1} \big(g(\hat{\beta}) - r \big) + o_p(n^{-1/2})$, где G – якобиан функции g, взятый в точке $\hat{\beta}$, который для обратимости $G(X^\top X)^{-1} G^\top$ должен иметь полный ранг. □

Теорема 5.10. *(Три асимптотических теста для линейных моделей). Рассматривается стандартная линейная модель с k-мерным вектором параметров β:*

$$Y = X\beta + \varepsilon. \tag{48}$$

Для модели выполняются условия ГМ1-ГМ3, ГМ5 и ошибки не коррелируют с регрессорами. Также задана непрерывно дифференцируемая в точке β функция $g : \mathbb{R}^k \to \mathbb{R}^q$. Для теста нулевой гипотезы $H_0 : g(\beta) = 0$ против альтернативы $H_1 : g(\beta) \neq 0$, можно использовать следующие статистики, имеющие при H_0 асимптотическое распределение $\chi^2(q)$:

- *Статистика Вальда:*

$$W = g(\hat{\beta})^\top \left[G(\hat{\beta}) \widehat{\mathbb{V}}(\hat{\beta}) G(\hat{\beta})^\top \right]^{-1} g(\hat{\beta}),$$

  где $G(\beta) = \frac{\partial g(\beta)}{\partial \beta^\top}$ – матрица Якоби функции ограничений, которая для корректной работы теста должна быть невырожденной. Также, $\hat{\beta}$ – это МНК-оценка β, $\hat{\mathbb{V}}(\hat{\beta})$ – стандартная оценка её дисперсии. При гетероскедастичных ошибках тест также можно использовать, применяя в качестве $\hat{\mathbb{V}}(\hat{\beta})$ оценку Уайта.

- *Статистика отношения правдоподобия:*

$$LR = 2\big(\ell(\hat{\beta}) - \ell(\hat{\beta}_R) \big) = n \ln \frac{SSR(\hat{\beta}_R)}{SSR(\hat{\beta})},$$

  где ℓ – (квази)-логарифмическая функция правдоподобия выборки, $\hat{\beta}$ – оценка МНК, а $\hat{\beta}_R$ – МНК-оценка β в предположении H_0. Также, $SSR(\cdot)$ – сумма квадратов остатков в модели, как функция от используемой оценки.

- *Статистика LM-теста в случае, когда набор регрессоров X содержит константу, и ограничение $g(\beta) = 0$ не мешает условию ортогональности остатков к константе:*

$$LM = S(\hat{\beta}_R)^\top I_n(\hat{\beta}_R)^{-1} S(\hat{\beta}_R) = n R_a^2,$$

  где S – (квази)-скор-функция; I_n – информация Фишера для параметра β условно при фиксированном X, R_a^2 – коэффициент детерминации во вспомогательной регрессии МНК-остатков модели (48), оценённой в предположении H_0, на регрессоры X.

Критерий, выбирающий H_1 в случае, когда тестовая статистика превышает квантиль уровня $1 - \alpha$ для распределения $\chi^2(q)$, имеет асимптотический уровень значимости, равный α.

Доказательство.

- Большие значения $g(\hat{\beta})$ свидетельствуют против H_0, но требуется учёт неоднородности расстояния, связанного с ковариационной матрицей этой статистики. Тест Вальда представляет собой стандартизованную квадратичную форму от $g(\hat{\beta})$. Используем асимптотическую нормальность МНК-оценок:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0_k, \Sigma),$$

где Σ – асимптотическая ковариационная матрица, $\widehat{\mathbb{V}}(\hat{\beta})$ – её состоятельная оценка.

Применяя дельта-метод с функцией $g(\beta)$, получаем

$$\sqrt{n}\big(g(\hat{\beta}) - g(\beta)\big) \xrightarrow{d} \mathcal{N}\big(0_q, G(\beta)\Sigma G(\beta)^\top\big).$$

Далее, используем невырожденность $G(\beta)$ для построения $\big(G(\beta)\Sigma G(\beta)^\top\big)^{-1/2}$ и стандартизации:

$$\big(G(\beta)\Sigma G(\beta)^\top\big)^{-1/2}\sqrt{n}\big(g(\hat{\beta} - g(\beta)\big) \xrightarrow{d} \mathcal{N}(0_q, I_q).$$

Квадрат длины этого случайного вектора по определению имеет асимптотическое распределение $\chi^2(q)$. Подставляя $g(\beta) = 0$ при H_0 и заменяя $G(\beta)$ и Σ состоятельными оценками, получаем статистику Вальда.

- LR-тест измеряет потерю качества подгонки, вызванную удержанием параметра на гиперповерхности $g(\beta) = 0$, соответствующей H_0. Используем функцию квази-правдоподобия для модели с нормальными ошибками:

$$L_n(\beta, \sigma^2) = \prod_{i=1}^{n} f_{\varepsilon|X}(Y_i - X_i^T\beta) f_X(x_i) = \frac{f_X(x)}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2} \Rightarrow$$

$$\Rightarrow \ell_n(\beta, \sigma^2) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}SSR(\beta) + \text{const}.$$

ОМП-оценкой σ^2 в такой модели является $\hat{\sigma}^2 = \frac{SSR}{n}$; подставляя это значение в выражение для $\ell_n(\beta, \sigma^2)$, получаем

$$\ell_n(\beta) = -\frac{n}{2}\ln SSR(\beta) + \text{const}.$$

LR-статистика:

$$LR = 2\left[\ell_n(\hat{\beta}) - \ell_n(\hat{\beta}_R)\right] = n\ln\frac{SSR(\hat{\beta}_R)}{SSR(\hat{\beta})}. \tag{49}$$

Используя $\ln(1+z) = z + o(z)$ для $z = \frac{SSR(\hat{\beta}_R) - SSR(\hat{\beta})}{SSR(\hat{\beta})}$, получаем при H_0

$$LR = n\frac{SSR(\hat{\beta}_R) - SSR(\hat{\beta})}{SSR(\hat{\beta})} + o_p(1).^{18} \tag{50}$$

Поскольку $SSR(\beta)$ квадратична по β, разложение Тейлора вокруг $\widehat{\beta}$ точно:

$$SSR(\beta) = SSR(\hat{\beta}) + (\beta - \hat{\beta})^\top X^\top X(\beta - \hat{\beta}),$$

[18] Далее будет показано, что при H_0 выполняется $SSR(\hat{\beta}_R) - SSR(\hat{\beta}) = O_p(1)$, кроме того, поскольку $\frac{SSR(\hat{\beta})}{n} \xrightarrow{p} \sigma^2$, то $\frac{SSR(\hat{\beta}_R) - SSR(\hat{\beta})}{SSR(\hat{\beta})} = O_p(1/n)$, откуда следует заявленный порядок малости.

(здесь используется равенство нулю градиента SSR в $\hat{\beta}$). Подставим $\beta = \hat{\beta}_R$ и используем связь между $\hat{\beta}$ и $\hat{\beta}_R$ из (45):

$$SSR(\hat{\beta}_R) - SSR(\hat{\beta}) = g(\hat{\beta})^\top \left[G(\hat{\beta})(X^\top X)^{-1} G(\hat{\beta})^\top\right]^{-1} g(\hat{\beta}) + o_p(n^{-1/2}).$$

Первое слагаемое в правой части является (при H_0) квадратичной формой от нормального вектора с нулевым средним, а значит – ограничено по вероятности.[19]

Подставляя полученное выражение в (50), получаем

$$LR = g(\hat{\beta})^\top \left[G(\hat{\beta})\hat{\sigma}^2(X^\top X)^{-1} G(\hat{\beta})^\top\right]^{-1} g(\hat{\beta}) + o_p(1).$$

Первое слагаемое в полученном выражении совпадает со статистикой Вальда (при оценке ковариационной матрицы, предполагающей гомоскедастичность ошибок), для которой асимптотическое распределение $\chi^2(q)$ было доказано ранее.

- LM-тест основан на следующей идее: если мы находимся в квазиоптимальной точке $\hat{\beta}_R$, то насколько крут склон функции правдоподобия, уводящий нас в сторону от ограничения? В линейной модели градиент ℓ (скор-функция) пропорционален $X^\top \hat{\varepsilon}_R$, где $\hat{\varepsilon}_R$ – остатки ограниченной модели. Тестовая статистика показывает, значимо ли данный градиент отличается от нуля.

  Из условий Каруша-Куна-Таккера, $S(\hat{\beta}_R) = G(\hat{\beta}_R)^\top \lambda_R$, поэтому $LM = \lambda_R^\top G(\hat{\beta}_R) I(\hat{\beta}_R)^{-1} G(\hat{\beta}_R)^\top \lambda_R$.

  Раскладываем скор-функцию и функцию ограничения в $\hat{\beta}$ около $\hat{\beta}_R$:

$$S(\hat{\beta}) = S(\hat{\beta}_R) + H(\hat{\beta}_R)(\hat{\beta} - \hat{\beta}_R) + o_p(\|\hat{\beta} - \hat{\beta}_R\|) \Rightarrow \hat{\beta} - \hat{\beta}_R = -H(\hat{\beta}_R)^{-1} S(\hat{\beta}_R) + o_p(\|\hat{\beta} - \hat{\beta}_R\|),$$

  откуда, в частности, $o_p(\|\hat{\beta} - \hat{\beta}_R\|) = o_p(n^{-1/2})$.

$$g(\hat{\beta}) = g(\hat{\beta}_R) + G(\hat{\beta}_R)(\hat{\beta} - \hat{\beta}_R) + o_p(\|\hat{\beta} - \hat{\beta}_R\|) = -G(\hat{\beta}_R) H(\hat{\beta}_R)^{-1} G(\hat{\beta}_R)^\top \lambda_R + o_p(n^{-1/2}),$$

  откуда $\lambda_R = -\left[G(\hat{\beta}_R) H(\hat{\beta}_R)^{-1} G(\hat{\beta}_R)^\top\right]^{-1} g(\hat{\beta}) + o_p(n^{1/2})$. Порядок остатка изменился из-за умножения на $\left[G(\hat{\beta}_R) H(\hat{\beta}_R)^{-1} G(\hat{\beta}_R)^\top\right]^{-1}$, где $G(\hat{\beta}_R) = O(1)$, $H(\hat{\beta}_R) = \frac{\partial^2 \ell_n}{\partial \beta \partial \beta^\top} = -\frac{X^\top X}{\hat{\sigma}_R^2} = O(n)$, где $\hat{\sigma}_R^2 = \frac{\hat{\varepsilon}_R^T \hat{\varepsilon}_R}{n}$ – ОМП-оценка дисперсии ошибки в модели с ограничением. Заменяя $-H$ на I в выражении для λ_R, и подставляя его в LM, получим

$$LM = g(\hat{\beta})^\top \left[G(\hat{\beta}_R)\hat{\sigma}_R^2(X^\top X)^{-1} G(\hat{\beta}_R)^\top\right]^{-1} g(\hat{\beta}) + o_p(1).$$

  Полученная форма аналогична статистике Вальда – а значит при H_0 имеет асимпттотическое распределение $\chi^2(q)$.

  Теперь докажем равенство $LM = nR_a^2$.

  Для линейной модели выполняется: $\hat{\varepsilon}_R = Y - X\hat{\beta}_R$, $\hat{\sigma}_R^2 = \frac{\hat{\varepsilon}_R^\top \hat{\varepsilon}_R}{n}$, $S(\hat{\beta}_R) = \frac{1}{\hat{\sigma}_R^2} X^\top \hat{\varepsilon}_R$, $I(\hat{\beta}_R)^{-1} = \hat{\sigma}_R^2 (X^\top X)^{-1}$.

  Подставляем в выражение для тестовой статистики:

$$LM = \frac{1}{\hat{\sigma}_R^2} \hat{\varepsilon}_R^\top X (X^\top X)^{-1} X^\top \hat{\varepsilon}_R = \frac{1}{\hat{\sigma}_R^2} \hat{\varepsilon}_R^\top P_X \hat{\varepsilon}_R,$$

[19] Матрица $G(\hat{\beta}) \left(\frac{X^\top X}{n}\right)^{-1} G(\hat{\beta})^\top$ является невырожденной в силу условия ГМ3 и полноты ранга $G(\hat{\beta})$.

где $P_X = X(X^\top X)^{-1}X^\top$ – матрица ортогональной проекции на столбцы X.

Поскольку $\hat{\varepsilon}_R$ имеет нулевое среднее, то $\hat{\varepsilon}_R^\top \hat{\varepsilon}_R$ – это SST вспомогательной регрессии. Так как X содержит константу, то $P_X \hat{\varepsilon}_R$ также имеет нулевое среднее, и тогда $\hat{\varepsilon}_R^\top P_X \hat{\varepsilon}_R$ соответствует SSE. Коэффициент детерминации:

$$R_a^2 = \frac{\hat{\varepsilon}_R^\top P_X \hat{\varepsilon}_R}{\hat{\varepsilon}_R^\top \hat{\varepsilon}_R} = \frac{\hat{\varepsilon}_R^\top P_X \hat{\varepsilon}_R}{n\hat{\sigma}_R^2} \Rightarrow LM = \frac{\hat{\varepsilon}_R^\top P_X \hat{\varepsilon}_R}{\hat{\sigma}_R^2} = nR_a^2.$$

6 Оценка эффектов воздействия с использованием неэкспериментальных данных

6.1 Введение

Эмпирический опыт последних десятилетий показывает, что в задачах построения прогнозных моделей классические линейные модели почти всегда уступают алгоритмам машинного обучения (случайные леса, градиентный бустинг, нейронные сети), поскольку последние готовы жертвовать интерпретируемостью и несмещенностью ради минимизации общей ошибки прогноза. Теоретические результаты также пессимистичны: теорема Гаусса-Маркова свидетельствует, что для того, чтобы линейная регрессия оказалась наилучшей несмещённой моделью для зависимой переменной, исследователю нужно как-то угадать правильную форму регрессионной функции в фундаментально нелинейном мире, полном пороговых эффектов и сложных взаимодействий – и даже в случае успеха, методы машинного обучения могут оказаться лучше по качеству прогноза за счёт более удачного компромисса между смещением и дисперсией.

Однако в анализе данных, используемом в общественных науках (экономике, социологии и др.) линейная модель остаётся незаменимым и доминирующим инструментом для другого типа задач – оценки причинно-следственных связей с использованием неэкспериментальных данных.

Ранее мы рассматривали постановку задачи оценки среднего эффекта воздействия (ATE) и её решение при помощи рандомизированных контролируемых испытаний (A/B тестов). Линейные модели можно применять и при обработке результатов A/B тестирования, однако основная проблема заключается в том, что A/B тесты дороги, а для некоторых задач они могут быть полностью недоступными. В таких случаях исследователь вынужден работать с неэкспериментальными (англ.: observational) данными, где переменная воздействия D почти всегда эндогенна: люди сами выбирают лечение, компании сами выбирают стратегии – поэтому определяя искомый параметр θ, как средний эффект воздействия D на Y в гипотетическом рандомизированном эксперименте, мы получаем ситуацию, когда в парной регрессии D почти всегда коррелирована с ненаблюдаемыми факторами в ошибке ε.

Пример 6.1. Покажите, что в примере 2.1 МНК-оценка параметра θ в модели (51) равна разности средних показателей здоровья в подвыборках подвергшихся и не подвергшихся госпитализации респондентов – а следовательно, является смещённой оценкой ATE.

$$Y_i = \beta_0 + \theta D_i + \varepsilon_i. \tag{51}$$

Тем не менее, в таких ситуациях линейные модели предоставляют исследователю множество подходов для оценки эффектов воздействия, главными из которых является использование контрольных и инструментальных переменных.

В первом подходе в линейную модель типа (51) добавляются (контрольные) переменные X, отражающие систематическое различие между объектами, подвергнутыми и не подвергнутыми воздействию. В идеале, наличие X в множественной регрессии позволяет добиться условной некоррелированности D и $\tilde{\varepsilon}$, делая МНК-оценку θ состоятельной.

$$Y_i = \beta_0 + \theta D_i + X_i^\top \beta + \tilde{\varepsilon}_i. \tag{52}$$

При таком подходе происходит важное концептуальное разделение компонент регрессионной модели. Коэффициент θ в (52) – это **целевой параметр**, ради которого собраны данные и делается исследование, а коэффициенты β при контрольных переменных – это так называемые **мешающие параметры** (англ.: nuisance parameters), не имеющие самостоятельной ценности. В парадигме машинного обучения все переменные равноправны в своей борьбе за снижение ошибки прогноза, но для причинно-следственных выводов

можно пожертвовать качеством оценки β, отвечающей за внутренний механизм связи между X и Y. Благодаря теореме Фриша-Ву-Ловелла, оценка $\hat{\theta}$ в модели (52) точно равна оценке в парной регрессии Y на D, где из обеих переменных предварительно удалили (англ.: partialled out) их линейные проекции на X; таким образом, роль X заключается в том, чтобы полностью объяснять часть D, коррелирующую с ошибкой модели $\tilde{\varepsilon}$.

В рассмотренном примере с госпитализацией, причиной смещённости "наивной" оценки ATE являлось систематическое различие между исходным показателем здоровья людей в двух группах – поэтому в качестве контрольных переменных можно было бы использовать широкий набор показателей здоровья, свидетельствующих о необходимости госпитализации пациента, а также набор социально-демографических и логистических факторов, влияющих на вероятность того, что врач решит оставить пациента в больнице, или что пациент сам этого потребует. Таким образом, составной частью задачи оценки эффекта воздействия становится построение и оценка прогнозной модели уже для переменной D. Теоретические основы подобных методов были разработаны в 2010х годах (двойное LASSO, двойное машинное обучение) и представлены в следующих разделах.

Другой подход к оценке эффектов воздействия заключается в поиске (инструментальных) переменных Z, заведомо некоррелированных с ошибками модели, эффект которых на переменную Y происходит опосредованно с участием переменной D. В рассмотренном примере возможный вариант такой переменной – загрузка приёмного покоя в час поступления пациента. Можно аргументировать, что этот показатель слабо связан с уровнем здоровья конкретного пациента, и, таким образом, дефицит ресурсов больницы используется как рандомизатор переменной воздействия D.

6.2 Проблема выбора модели и ошибочной спецификации

Проблема выбора модели при оценке эффектов воздействия с использованием неэкспериментальных данных заключается в выборе набора контрольных переменных $X = (1, X_1, \ldots, X_k)$ из общего числа доступных данных, а также в выборе функциональной формы зависимости (в уровнях – или в логарифмах, включает ли модель нелинейные преобразования от переменных, и т. д.). Именно при выборе модели на первый план выходит необходимость научного понимания анализируемой задачи и совершается подавляющее большинство исследовательских ошибок.

До недавнего времени в прикладных исследованиях доминировал принцип, что "данные должны говорить сами за себя". Это привело к появлению нескольких методологий, основанных на последовательном статистическом тестировании, которые до сих пор встречаются в устаревших учебниках и научных работах. Можно выделить три ключевых проблемы, связанные с таким подходом:

- Применение методов выбора модели без чёткого понимания различий между прогнозной и причинно-следственной методологией анализа;
- Накопление ошибки при последовательном тестировании и волюнтаристическое использование результатов тестов. Так, сложившуюся в конце 20го века практику Э. Лимер описывает следующим образом:[20] «Искусство прикладного анализа практикуется за компьютерным терминалом и предполагает подгонку множества, иногда тысяч статистических моделей. Одна или несколько моделей, которые исследователь находит удовлетворительными, отбираются для представления результатов».
- Проблема **неравномерной инференции**, обнаруженная в 2003-2005 годах в работах Х. Либа и Б. Потчера, заключающаяся в том, что асимптотическое распределение оценок целевых параметров, по-

[20] Leamer, E. E. (1983). "Let's Take the Con Out of Econometrics". The American Economic Review, 73(1), p. 31–43.

лучаемое в модели, выбранной с помощью тестов или информационных критериев, зависит от истинных параметров модели и не может быть однозначно определено. В такой ситуации исследователь может получить точечную оценку интересующей его величины, но любой доверительный интервал или статистический тест становятся ненадёжными, и могут иметь характеристики, сильно отличающиеся от расчётных.

По этой причине, подходы к выбору модели в этом разделе приводятся лишь с целью общего ознакомления с историей проб и ошибок, сопровождавших развитие эмпирической методологии.

Определение 6.2. Алгоритм forward selection: выбор модели начинается с пустого набора контрольных переменных, к которому по-очереди добавляются переменные с наименьшим p-значением (наибольшими по модулю t-статистиками). Остановка алгоритма происходит, когда очередная добавляемая переменная оказывается статистически незначимой.

Определение 6.3. Алгоритм backward elimination: из модели с максимальным набором контрольных переменных по одной удаляются имеющие наибольшее p-значение. Иногда встречается разновидность, в которой удаление группы переменных сопровождается проверкой их совместной значимости с помощью F-тестов. Критерий остановки алгоритма аналогичен предыдущему случаю.

Эти процедуры по своей сути являются алгоритмами выбора прогнозной модели для зависимой переменной Y, и никаким образом не направлены на решение задачи ортогонализации переменной воздействия D и ошибок модели. Помимо прочих проблем, такие методы могут приводить к исключению из модели переменных, критически важных для устранения смещения целевого коэффициента, если они слабо предсказывают Y.

Определение 6.4. Информационные критерии (Акаике, Шварца, Ханнана-Куинна и др.) – это методы оценки переобучения моделей, основанные на значениях функции правдоподобия и штрафа за количество параметров модели.

Пример 6.5. Критерий выбора линейной модели, асимптотически эквивалентный критерию Акаике (AIC) для гауссовской линейной модели.

Пусть задана стандартная модель множественной регрессии

$$Y = X\beta + \varepsilon, \tag{53}$$

где Y – вектор ответов размера $n \times 1$, X – полная матрица регрессоров размера $n \times p$, β – вектор коэффициентов полной модели, ε – ошибки со свойствами $\mathbb{E}(\varepsilon \mid X) = 0$, $\mathbb{V}(\varepsilon \mid X) = \sigma^2 I_n$.

Величина $X\beta$ – это лучший прогноз, то есть истинное условное среднее. При этом предполагается, что набор регрессоров избыточен, и β является **разреженным** – то есть, часть его координат равна нулю. Рассмотрим некоторую подмодель, то есть подвыборку регрессоров. Обозначим её матрицу за X_m; она состоит из $k \leq p$ столбцов полной матрицы X. Также обозначим за $b_m = (X_m^\top X_m)^{-1} X_m^\top Y$ МНК-оценку по этой подмодели.

Нас интересует ожидаемая ошибка предсказания подмодели относительно полного лучшего прогноза:

$$R_m = \mathbb{E}\|X\beta - X_m b_m\|^2.$$

Цель – выбрать подмодель m, минимизирующую функцию риска R_m.

Решение. Для упрощения обозначений далее все операции проводим условно на фиксированные значения X. Также для удобства считаем что b_m и X_m дополнены нулями до размерностей β и X соответственно, и тогда $Xb_m = X_m b_m$.

Введём проекционную матрицу $H_m = X_m(X_m^\top X_m)^{-1}X_m^\top$. Проекция Y имеет вид $X_m b_m = H_m Y = H_m X\beta + H_m\varepsilon$.

Разложение риска на смещение и дисперсию:

$$R_m = \mathbb{E}\|X\beta - Xb_m\|^2 = \mathbb{E}\,\|(I-H_m)X\beta - H_m\varepsilon\|^2 =$$

$$= \|(I-H_m)X\beta\|^2 - 2\mathbb{E}\left[\left((I-H_m)X\beta\right)^\top H_m\varepsilon\right] + \mathbb{E}\left[\varepsilon^\top H_m^\top H_m\varepsilon\right]. \tag{54}$$

Второе слагаемое равно нулю, потому что $\mathbb{E}(\varepsilon \mid X) = 0$; для третьего слагаемого применяем рассуждения аналогично 10.7:

$$\mathbb{E}\left[\varepsilon^\top H_m^\top H_m\varepsilon\right] = \operatorname{tr}(\mathbb{E}\left[\varepsilon^\top H_m\varepsilon\right]) = \operatorname{tr}\left(H_m\mathbb{E}[\varepsilon\varepsilon^\top]\right) = \operatorname{tr}\left(H_m\sigma^2 I_n\right) = \sigma^2\operatorname{tr}(H_m) = \sigma^2 k.$$

Получили классическое разложение риска на квадрат смещения и дисперсионную составляющую

$$R_m = \|(I-H_m)X\beta\|^2 + \sigma^2 k.$$

Проблема в том, что β неизвестен, поэтому напрямую вычислить показатель смещения нельзя – но можно оценить через RSS подмодели.

$$RSS_m = \|Y - X_m b_m\|^2 = \|(I-H_m)X\beta + (I-H_m)\varepsilon)\|^2.$$

Так как $I - H_m$ тоже симметричная идемпотентная матрица, то аналогично (54) получаем

$$\mathbb{E}(RSS_m) = \|(I-H_m)X\beta\|^2 + \sigma^2(n-k).$$

Отсюда получаем выражение для R_m:

$$R_m = \mathbb{E}(RSS_m) - \sigma^2(n-k) + \sigma^2 k.$$

Упрощая, и учитывая, что $\sigma^2 n$ не зависит от выбора подмодели, получаем, что минимизация R_m эквивалентна минимизации $\mathbb{E}(RSS_m) + 2k\sigma^2$.

Заменяя $\mathbb{E}(RSS_m)$ на наблюдаемое RSS_m, получаем критерий выбора

$$\min_m\left\{RSS_m + 2k\sigma^2\right\} \Leftrightarrow \min_m\left\{\frac{RSS_m}{n} + \frac{2k\sigma^2}{n}\right\}.$$

Это и есть идея AIC: штраф за каждый дополнительный параметр равен $2\frac{\sigma^2}{n}$ в шкале средней квадратичной ошибки, – иначе говоря, ищется компромисс между точностью и сложностью выбираемой модели.

Упражнение. Вычислите $\mathbb{E}\|\hat\beta - \beta\|^2$, где $\hat\beta$ – МНК-оценка по полному набору X.

Другой аспект выбора модели заключается в проверке соответствия оценённой модели используемым данным и адекватности сделанных предположений.

Определение 6.6. Пост-оценочная диагностика в линейных моделях – это совокупность процедур, выполняемых после оценивания параметров модели, направленных на проверку выполнения модельных предполо-

жений, оценку качества подгонки, выявление аномальных и влиятельных наблюдений и диагностику ошибок спецификации модели. Иначе говоря, это этап анализа, на котором проверяется, насколько оценённая линейная модель адекватна данным и можно ли использовать её для статистических выводов и прогноза.

Пример 6.7. Тест RESET (англ.: regression equation specification error test).

Логика теста RESET основывается на следующей идее: если модель специфицирована верно, то никакие нелинейные функции от уже включенных регрессоров не должны обладать дополнительной объясняющей силой. В таком случае можно использовать прогнозные значения $\hat{Y} = X\hat{\beta}$, как универсальный заменитель для всей систематической части модели, и проверять, будут ли нелинейные функции от $\hat{Y}$ значимо объяснять оставшуюся часть Y.

Алгоритм теста выглядит так:

- Оцениваем исходную модель $Y = X\beta + \varepsilon$ и сохраняем предсказания $\hat{Y}$;
- Оцениваем вспомогательную регрессию, добавляя в нее степени $\hat{Y}$ (обычно квадраты и кубы):

$$Y = X\beta + \gamma_1\hat{Y}^2 + \gamma_2\hat{Y}^3 + u;$$

- Проверяем нулевую гипотезу с помощью стандартного F-теста (или асимптотического теста Вальда):

$$H_0 : \gamma_1 = 0 \text{ и } \gamma_2 = 0.$$

Если H_0 отвергается, тест RESET сигнализирует о наличии ошибки спецификации (нелинейные преобразования X значимо улучшают прогноз Y).

Пример 6.8. Тесты гетероскедастичности и автокорреляции.

Существует большое количество тестов для проверки свойств ошибок модели. Достаточно просто устроены тесты для проверки гетероскедастичности: в одном варианте (тест Бройша-Пагана) квадраты стандартизированных МНК-остатков (т.е. поделенных на $\hat{\sigma}^2 = {}^{SSR}/_{n}$) регрессируются на исходные переменные X. В другом варианте (тест Уайта), квадраты остатков регрессируются на все исходные регрессоры, их квадраты и перекрёстные произведения. При нулевой гипотезе (гомоскедастичность) в обоих случаях вспомогательные регрессии должны быть незначимыми; в тесте Уайта для проверки H_0 используется стандартный LM-тест. Если H_0 отвергается, то зависимость квадратов остатков в выборке от используемых регрессоров свидетельствует об аналогичной зависимости для среднего значения квадратов ошибок – что соответствует гетероскедастичности в модели.

При оценке линейных моделей на панельных данных и временных рядах тесты на автокорреляцию ошибок помогают уточнить структуру оптимальной модели. Тестируется гипотеза $H_0 : \mathbb{E}\varepsilon_t\varepsilon_{t-k} = 0$; здесь ошибки заиндексированы параметром t, а k – номер лага. Для проверки H_0 можно использовать тест Бройша-Годфри: в нём оценивается вспомогательная регрессия

$$\hat{\varepsilon}_t = X_t^\top\beta + \sum_{i=j}^{k} a_j\hat{\varepsilon}_{t-j} + u_t,$$

где $\hat{\varepsilon}_t$ и X_t – остатки и регрессоры в основной модели соответственно; u_t – ошибки вспомогательной регрессии. Далее проверяется равенство нулю всех a_j с помощью LM-теста. Наличие значимых a_j свидетельствует о наличии автокорреляции у ошибок основной модели.

Пример 6.9. Тесты на нормальность.

В разделе 5.2, посвященном точному выводу, упоминались t- и F-статистики, которые имеют соответствующие распределения в конечных выборках только при условии нормальности ошибок: $\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2 I)$. По этой причине в прошлом была популярной процедура тестирования на нормальность распределения остатков модели, имеющая цель "оправдать" применение результатов, специфических для моделей с нормальными ошибками.

Один из возможных способов проверки нормальности – тест Харке-Бера, основанный на моментах распределения. Нормальное распределение имеет коэффициент скошенности (асимметрии) $S = 0$ и куртозис $K = 3$ (или эксцесс куртозиса $K - 3 = 0$). Тест проверяет, насколько выборочные асимметрия $\hat{S}$ и эксцесс $\hat{K}$ остатков отклоняются от нормальных эталонов:

$$JB = \frac{n}{6}\left(\hat{S}^2 + \frac{(\hat{K}-3)^2}{4}\right).$$

При истинности H_0 (ошибки нормальны) статистика асимптотически имеет распределение хи-квадрат с двумя степенями свободы: $JB \xrightarrow{d} \chi^2(2)$.

Для малых выборок тест Харке-Бера может быть неточным. В таких случаях используют тесты, сравнивающие эмпирическую функцию распределения остатков с теоретической нормальной. Так, тест Шапиро-Уилка основан на корреляции между упорядоченными остатками и квантилями нормального распределения и считается одним из самых мощных тестов для проверки нормальности распределения при малом размере выборки ($n < 50$). Тест Колмогорова-Смирнова измеряет расстояние в равномерной норме между эмпирической и теоретической функциями распределения; с нормировкой на $\sqrt{n}$ такое расстояние при H_0 имеет известное асимптотическое распределение.

Определение 6.10. Ошибочная спецификация модели – это ситуация, в которой оцениваемая модель не включает в себя истинный механизм генерации данных. Иными словами, функция g в уравнении оцениваемой модели (55) при любом θ не совпадает с функцией f в (56), соответствующей модели, породившей наблюдаемые данные.

$$Y = g(X, D, \theta) + u, \tag{55}$$

$$Y = f(X, D, \theta_0) + \varepsilon. \tag{56}$$

Таблица 3: Типы ошибочной спецификации модели и их последствия

Тип	**Суть проблемы**	**Последствия для прогнозной модели**	**Последствия для каузальной модели**
Неверная функциональная форма	Неверно выбрана функциональная форма связи: линейная спецификация вместо нелинейной, пропуск взаимодействий, неверная трансформация переменных	Смещение прогнозов, особенно при экстраполяции за пределы области данных; систематическая ошибка прогноза и ухудшение прогнозной точности вне выборки	В самом общем случае – понижение эффективности оценки; более серьезные последствия зависят от специфики модели и проблемы
Пропуск релевантных переменных	В модель не включены переменные, влияющие на зависимую переменную и коррелированные с переменной воздействия	Не применимо	Смещённые и несостоятельные оценки эффектов воздействия

Тип ошибочной спецификации	**Суть проблемы**	**Последствия для прогнозной модели**	**Последствия для каузальной модели**
Включение нерелевантных переменных	В модель включены переменные, не влияющие на зависимую переменную	Рост дисперсии прогнозов, переобучение – ухудшение прогнозной способности на новых данных	Снижение эффективности оценок
Неверная спецификация ошибки модели	Неверные предпосылки об ошибке: гетероскедастичность, автокорреляция, неверное распределение ошибок, неучтённая кластеризация	Неточные прогнозные интервалы (недооценка или переоценка неопределённости)	Неточные оценки стандартных ошибок (могут быть и заниженными, и завышенными), неверные доверительные интервалы и тесты; сами оценки обычно остаются состоятельными
Ошибки измерения	Переменные измерены с ошибкой (ошибки в зависимой или независимых переменных).	Ухудшение прогнозной точности	Ошибка в зависимой переменной увеличивает дисперсию; ошибка в регрессорах ведёт к смещению оценок к нулю и несостоятельности.
Эндогенность регрессоров	Регрессоры коррелированы с ошибкой модели (одновременность, обратная причинность, пропущенные переменные, ошибки измерения).	Не применимо	Смещённые и несостоятельные оценки эффектов воздействия; каузальная интерпретация невозможна без идентификационных стратегий
Ошибка отбора выборки	Выборка нерепрезентативна для генеральной совокупности; отбор в выборку коррелирует с ошибкой модели	Смещение прогнозов при переносе на генеральную совокупность; плохая генерализация на целевую популяцию	Смещённые и несостоятельные оценки каузальных эффектов, необходима коррекция
Неверная спецификация динамики временных рядов	Неверно специфицирована временная динамика: неверный порядок лагов, неучтённая нестационарность, единичные корни, структурные сдвиги	Ухудшение прогнозной точности	Ложные каузальные выводы: мнимая регрессия (например, переменные, имеющие тренд, могут казатся зависимыми друг от друга); смещённые оценки динамических эффектов

Далее рассмотрены несколько ситуаций, в которой ошибочная спецификация причинно-следственной модели может привести к некорректным оценкам.

Пример 6.11. Некоторые возможные ошибки при оценке эффекта образования D на зарплату Y. Базовая модель, рассматриваемая далее – это парная регрессия Y на D.

- Отсутствие в модели контроля на уровень дохода родителей – пример пропуска релевантной переменной. Социальный капитал и связи существенно определяют доступ к качественному школьному образованию, репетиторам и престижным вузам. В то же время дети из обеспеченных семей часто имеют лучший старт на рынке труда благодаря связям родителей или наследству. Пропуск этого фактора приведет к тому, что эффект образования смешается с эффектом семейного происхождения.

- Использование в той же модели контроля текущей должности – пример ошибки функциональной формы модели. Образование влияет на зарплату, в том числе через должность. Контролируя её, мы изме-

ряем лишь прямой эффект образования, игнорируя косвенный – таким образом исключая из оценки часть истинного эффекта.

- Использование в той же модели контроля текущего статуса занятости S, вероятно, приведёт к занижению оцениваемого эффекта. Предположим, что на зарплату влияют образование D и мотивация X работника, а наличие работы S – бинарная переменная, зависящая (положительно) от образования и удовлетворённости предлагаемой зарплатой Y. Упрощённо, пусть каждая из переменных бинарна (0 соответствует "низкому" значению, 1 – "высокому"), а также $Y = D \vee X$ и $S = D \vee Y$. Тогда можно заметить, что $S = D \vee (D \vee X) = Y$, и в регрессии Y на D и S мы получим оценки 0 и 1 соответственно (S идеально объясняет Y) – при том, что зависимость между Y и D в популяции положительна.

  Такой вид ошибочной спецификации называется смещением из-за включения коллайдера, см. также 7.1.

- Оценка модели только по подвыборке работающих приведёт к занижению эффекта из-за смещённой выборки. Этот вид ошибки спецификации иногда также связывают со смещением из-за включения коллайдера.

  Продолжая предыдущий пример с бинарными показателями, можно заметить, что из 4 возможных сочетаний D и Y, в подвыборке работающих отсутствуют наблюдения с $D = 0, Y = 0$. Зависимость между Y и D на таком наборе данных не может быть положительной.

6.3 Проблема неравномерной инференции

Рассмотрим процедуру выбора модели на основе алгоритма backward elimination.

Пример 6.12. Исследователь оценивает эффект воздействия D на Y и сомневается в необходимости использования контрольной переменной X – то есть, выбор происходит между «короткой» моделью простой регрессии Y на D и «длинной» моделью регрессии Y на D и X.

$$Y_i = \theta_L D_i + \beta_g X_i + u_i, \quad \text{«длинная» модель, L,} \tag{57}$$

$$Y_i = \theta_S D_i + v_i \quad \text{«короткая» модель, S.} \tag{58}$$

Предполагается, что в длинной модели условия экзогенности D обеспечены. Тем не менее, если в популяции $\beta_g = 0$, то короткая модель более эффективна. Для оценки используется случайная выборка размером n.

Модели (57) и (58) являются вложенными; выбор между ними может быть основан на значении t-статистики коэффициента при X в длинной модели в сравнении с критическим значением c_n:

$$|t| = \frac{|\hat\beta_g|}{\widehat{\text{s.e.}}(\hat\beta_g)}, \quad \text{выбираем S при } |t| \le c_n. \tag{59}$$

Здесь $\hat\beta_g$ – МНК-оценка коэффициента при X в (57), $\widehat{\text{s.e.}}(\hat\beta_g)$ – состоятельная оценка её стандартного отклонения. Значения c_n можно взять зависящими от размера выборки для того, чтобы процедура выбора была состоятельной (вероятность выбора "правильной" модели стремится к 1 с ростом n). Интересно заметить, что выбор между (57) и (58), основанный на минимальном значении информационного критерия Шварца, эквивалентен использованию $c_n = \sqrt{\log n}$.

В случае, когда $|t| > c_n$, мы считаем, что β_g значимо отличается от нуля, и выбираем в качестве оценки эффекта воздействия D на Y МНК-оценку θ_L в (57), а при $|t| \leqslant c_n$ в качестве оценки используем МНК-оценку θ_S в (58), то есть итоговая оценка целевого параметра θ (эффект воздействия D на Y) имеет следующий вид:

$$\hat{\theta} = \hat{\theta}_S \mathbb{1}\{|t| \leq c_n\} + \hat{\theta}_L \mathbb{1}\{|t| \geq c_n\}.$$

Потребуем $c_n \to \infty$, $c_n = o(\sqrt{n})$ и покажем состоятельность выбора, используя асимптотическую нормальность МНК-оценок:

$$\sqrt{n}(\hat{\beta}_g - \beta_g) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma_\beta^2), \quad \mathrm{P}(|t| \leq c_n) = \mathrm{P}\left(-c_n \leq \frac{\hat{\beta}_g}{\widehat{s.e.}(\hat{\beta}_g)} \leq c_n\right) =$$

$$= \mathrm{P}\left(-c_n - \sqrt{n}\frac{\beta_g}{\sigma_\beta} \leq \frac{\hat{\beta}_g}{\frac{\sigma_\beta}{\sqrt{n}} + o_p(1/\sqrt{n})} - \sqrt{n}\frac{\beta_g}{\sigma_\beta} \leq c_n - \sqrt{n}\frac{\beta_g}{\sigma_\beta}\right) = \Phi\left(c_n - \sqrt{n}\frac{\beta_g}{\sigma_\beta}\right) - \Phi\left(-c_n - \sqrt{n}\frac{\beta_g}{\sigma_\beta}\right) + o(1), \quad (60)$$

где σ_β^2 – асимптотическая дисперсия $\hat{\beta}_g$, а $\Phi(\cdot)$ – функция распределения стандартной нормальной случайной величины.

При $\beta_g = 0$ правая часть (60) сходится к $\Phi(+\infty) - \Phi(-\infty) = 1$; при $\beta_g \neq 0$ это же выражение сходится к 0.

В работе Х. Либа и Б. Потчера[21] было показано, что по сравнению с безальтернативным использованием длинной модели, предложенная процедура выбора между (57) и (58) приводит, с одной стороны, к росту точности прогноза Y и уменьшению среднеквадратичной ошибки оценки θ – в этом смысле выбор модели работает как инструмент регуляризации, сжимая оценку до нуля там, где сигнал слабый. С другой стороны, выбор приводит к тому, что распределение $\hat{\theta}$ становится крайне чувствительным к тому, насколько значение параметра β_g близко к границе исключения (к нулю). Этот результат видно при анализе двойной асимптотики с $n \to \infty$ и $\beta_g \to 0$. Рассмотрим последовательность случайных выборок $\{(Y_i, D_i, X_i)\}_{i=1}^n$ разного размера n из совместных распределений P_n, соответствующих длинной модели (57), но с уменьшающимся к нулю коэффициентом β_g. В такой ситуации предел в (60) зависит от скорости сходимости $\beta_g(n)$ к нулю. В частности, легко подобрать такую асимптотику, когда этот предел будет числом, отличным от 0 и 1: для этого нужно, чтобы $\beta_g(n)$ убывала к нулю со скоростью $\frac{c_n}{\sqrt{n}}$. В такой ситуации безусловное предельное распределение $\hat{\theta}$ будет смесью двух нормальных с разными параметрами.

Проблема неравномерности в контексте представленного примера проявляется следующим образом: на практике, исследователь имеет лишь одну выборку из P_n с фиксированным значением β_g, которое можно считать элементом числовой последовательности с произвольным асимптотическим поведением, – что естественно никак не влияет на свойства получаемой оценки $\hat{\theta}$ в конечной выборке. Однако, при больших n и при β_g достаточно близких к нулю, распределение получающейся оценки должно быть близким к одному из возможных асимптотических пределов – нормальному с центром θ_S, соответствующему быстро убывающей к нулю $\beta_g(n)$, либо к бимодальному распределению, являющемуся смесью двух нормальных – соответствующему медленно убывающей к нулю $\beta_g(n)$, рис. 8.

Таким образом, оценка исследуемого параметра обладает паталогическими свойствами, когда коэффициент при переменной X оказывается умеренно близким к нулю – но для практического анализа данных это единственная ситуация, когда формальный подход к выбору контрольной переменной представляют интерес. Это связано с тем, что если бы зависимость Y от X была бы неотличима от нуля, то у исследователя изначально не должно было быть сомнений в том, что соответствующим контролем можно пренебречь, и

[21] Leeb, H., Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. Econometric Theory, 21(1), 21–59.

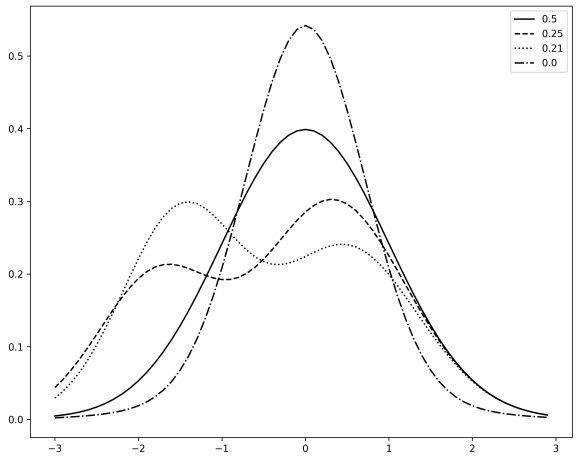


Рис. 8: Плотность распределения $\sqrt{n}(\hat{\theta} - \theta_L)$ при $n = 100$; $c_n = \sqrt{\log n}$, $\text{corr}(X_i, D_i) = 0.7$. Значения $\frac{\beta_g}{\sigma_\beta}$ равны 0, 0.21, 0.25 и 0.5. Источник – Leeb, Pötscher (2005)

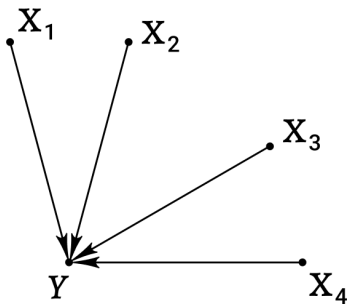
точно так же, в ситуации, когда коэффициент зависимости Y от X далек от нуля, то тестирование, вероятно, было бы излишним, – так как соответствующая контрольная переменная должна была бы быть включена в модель из общих соображений.

7 Использование графических моделей

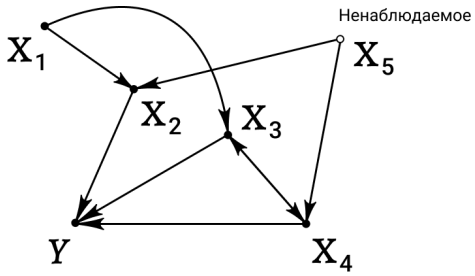
7.1 Причинно-следственные диаграммы и структурные статистические модели

Ещё один подход к идентификации структурных параметров использует модели на орграфах (диаграммы причинно-следственных связей), в которых каждая из вершин соответствует одной из переменных, представляющих интерес в конкретной задаче; при этом никакая переменная не ставится в соответствие с более чем одной вершиной, а рёбра графа могут быть отождествлены с теми или иными зависимостями между ними. Такие модели являются частью более общего класса графических моделей, в которых направленные связи используются для задания марковских свойств совместного распределения. В нашем случае, связанная пара вершин показывает о наличии статистической связи между соответствующими переменными, а направленный путь отождествляется с причинно-следственной связью от вершин-родителей к детям. Иногда в таких моделях используются двунаправленные рёбра – для демонстрации того, что существуют другие переменные, не представленные в графе, влияющие на обе переменные, связанные данным ребром, рис. 9. Ацикличность орграфов необходима, поскольку в обратном случае в переменных нельзя разделить причину и следствие.

Применение графических моделей для описания структуры зависимостей в данных впервые встречается в работах Сьюэла и Филипса Райта, и было популяризовано в 2000-х годах работами Дж. Перла, некоторые идеи из которых представлены далее. Модели причинно-следственных связей, основанные на таких графах, называются DAG-моделями (англ. directed acyclic graph).



(a) Эффект каждой из переменных является причинно-следственным по отношению к Y



(b) Влияние X на Y осложнено совместным взаимодействием и присутствием ненаблюдаемой компоненты

Рис. 9: Различные возможные структуры зависимости между наблюдаемыми переменными $X = (X_1, X_2, X_3, X_4)$ и Y

Для того чтобы проиллюстрировать применение графических моделей, рассмотрим k переменных и их совместное распределение $\mathrm{P}_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k)$, где нумерация переменных выбрана произвольным образом. Пользуясь стандартным определением условной вероятности, мы можем представить совместное распределение в виде произведения условных распределений:

$$\mathrm{P}_{X_1,X_2,\ldots,X_k}(x_1, x_2, \ldots, x_k) = \prod_{i=1}^{k} \mathrm{P}_{X_i|X_1,\ldots,X_{i-1}}(x_i \mid x_1, \ldots, x_{i-1}) \tag{61}$$

Разумно предположить, что условное распределение X_i зависит не от всех переменных, предшествующих ему в этой нумерации, но лишь от некоторых, набор которых можно обозначить как Pa_i. Такие переменные принято называть родителями данной переменной в марковском смысле, или просто родителями. Тогда условные распределения соответствуют уравнению (62):

$$\mathrm{P}_{X_i|X_1,\ldots,X_{i-1}}(x_i \mid x_1, \ldots, x_{i-1}) = \mathrm{P}_{X_i|Pa_i}(x_i \mid pa_i) \tag{62}$$

Понятие родителей в марковском смысле может быть формализовано как минимальный (в смысле включения) набор переменных, удовлетворяющих свойству (62). Пользуясь заданным определением, можно переписать разложение (61):

$$\mathrm{P}_{X_1,X_2,\ldots,X_k}(x_1,x_2,\ldots,x_k)=\prod_{i=1}^{n}\mathrm{P}_{X_i|Pa_i}(x_i\mid pa_i) \tag{63}$$

Для заданного ориентированного ациклического графа можно говорить о совместных распределениях, которые являются совместимыми с ним. Такие распределения можно определить, как все распределения, для которых существует такая нумерация переменных, что если вершина j является родителем вершины i в графе, то переменная j входит в число родителей переменной i в марковском смысле.

Содержательно понятие совместимости ориентированного ациклического графа с заданным совместным распределением играет ключевую роль, поскольку оно оказывается необходимым и достаточным условием для того, чтобы распределение могло быть порождено процессом, который последовательно генерирует переменные из соответствующих условных распределений, принимая во внимание значения переменных-родителей, сгенерированных на предыдущих шагах.

Одной из наиболее важных характеристик совместного распределения, которая представляет особый интерес в прикладных статистических исследованиях, является набор отношений условной независимости. Оказывается, что такую информацию можно получить, используя направленный ациклический граф, с которым заданное распределение совместимо. Для этого требуется ввести понятие блокировки пути: ненаправленный путь S на графе DAG-модели блокирован набором вершин Z, если выполнено хотя бы одно из следующих условий, рис. 10:

- в S найдётся цепочка $X_1 \to X_2 \to X_3$ или вилка $X_1 \leftarrow X_2 \to X_3$, причем $X_2 \in Z$;
- в S найдётся обратная вилка (коллайдер) $X_1 \to X_2 \leftarrow X_3$, такая что $X_2 \notin Z$, и также никакой из потомков X_2 не содержится в Z.

Такие обозначения связаны со статистической зависимостью между переменными в модели. Действительно, если переменные X_1 и X_3 имеют общую причину X_2, то они статистически зависимы – в полном соответствии с принципом Райхенбаха. В обратной ситуации, две безусловно независимые переменные X_1 и X_3, имеющие общий эффект X_2, становятся статистически зависимыми условно на значение X_2.

(a) Вилка (конфаундер) (b) Обратная вилка (коллайдер)

Рис. 10: Виды разветвлений в DAG-модели

Будем также говорить, что наборы переменных X и Y разделены Z, если все пути от каждого из элементов X к каждому из элементов Y блокированы элементами Z. Можно показать, что две группы случайных величин X и Y независимы условно на третью группу случайных величин Z при любом распределении, согласованном с заданным ориентированным ациклическим графом, если Z разделяет X и Y в этом графе в приведенном выше смысле. В обратную сторону, если Z не разделяет X и Y в этом графе, то существует согласованное с ним распределение, в которым X и Y не являются независимыми условно на Z.

Одна из причин удобства DAG-моделей заключается в предположении о том, что каждому ориентированному ребру в графе соответствует некий устойчивый и автономный механизм. Иными словами, можно представить ситуацию, где лишь одна из связей в графе меняется, оставляя все прочие неизменными. Организация знания в подобную модульную структуру позволяет использовать причинно-следственную диаграмму для предсказания эффекта внешних вмешательств с минимальной дополнительной информацией, поэтому подобная модель (предполагая, что она корректна) зачастую оказывается намного более информативной, чем набор уравнений.

Формальное условие соответствия диаграммы причинно-следственных связей некоторому распределению, предложенное в работах Перла, основано на марковских свойствах и совместимости распределения со структурой графа.

Предположение 7.1. (Условия соответствия направленного ациклического графа причинно-следственной модели).

Направленный ациклический граф G задаёт причинно-следственную модель, если выполнены следующие условия:

1. Все переменные, включенные в G, являются наблюдаемыми;
2. Все переменные, включенные в G, функционально зависят от своих родителей и идиосинкратических шоков – случайных величин, которые независимы от других переменных в модели и друг от друга;
3. Каждая переменная X в графе гипотетически может быть подвергнута интервенции $do(X = x)$, которая
   - заменяет вероятностное распределение в X на фиксированное значение x,
   - удаляет из графа направленные рёбра, заканчивающиеся в X,
   - не производит никаких других эффектов по отношению к другим переменным в G (кроме тех эффектов, которые связаны с влиянием X на своих потомков в оставшемся графе).

Ещё одним удобным инструментом при работе с DAG-моделями является так называемый критерий обходных путей (the back-door criterion).

7.2 Критерий обходных путей

Определение 7.2. (Критерий обходных путей).

Для упорядоченной пары переменных-вершин (D, Y) в ориентированном ациклическом графе G набор вершин X удовлетворяет критерию обходных путей, если ни одна вершина в X не является потомком D, и заблокирован каждый путь от D к Y, кроме тех, в которых Y является потомком D.

Если критерий обходных путей выполнен, то можно доказать, что причинно-следственное влияние D на Y определяется следующим образом

$$\forall B \in \mathcal{B}(\mathbb{R}), \quad \mathrm{P}_Y\big(B \mid \mathrm{do}(D)\big) = \sum_x \mathrm{P}_Y(B \mid D, X = x)\mathrm{P}_X(x), \tag{64}$$

где в левой части – условное распределение Y после воздействия, устанавливающего значение D, а в правой – суммирование по всем возможным наборам значений переменных X, основанное на формуле полной вероятности. При этом становится очевидной неоднозначность понятия причинно-следственного эффекта D на Y, поскольку несколько различных наборов вершин могут удовлетворять критерию обходных путей, рис.

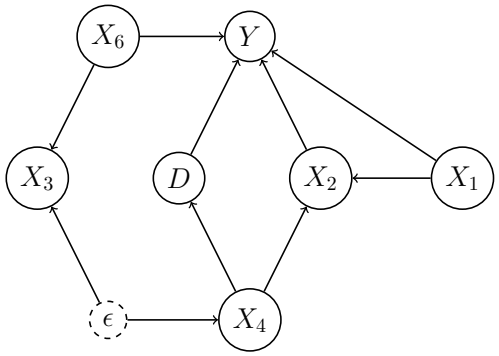


Рис. 11: Переменная X_4 и набор $\{X_1, X_2\}$ удовлетворяют критерию обходных путей для эффекта D на Y, поэтому причинно-следственное влияние D на Y можно определить в соответствии с (64) неоднозначно

11. Тем не менее, этот факт находится в полном соответствии с идеей контрфактуальности – то есть, соображению о том, что причинно-следственный эффект D на Y можно наблюдать в разных гипотетических обстоятельствах и при фиксированных значениях у различных наборов остальных характеристик.

Интегрируя 64 можно получить выражение для регрессионной функции $\mathbb{E}\big(Y \mid X, do(D)\big)$. В случае, когда эта функция является линейной, регрессионная модель может иметь вид

$$Y = \theta_0 D + X^\top \beta_0 + \varepsilon, \tag{65}$$

где $\theta_0 \in \mathbb{R}$ – коэффициент, отражающий влияние D на Y, $\beta_0 \in \mathbb{R}^k$ – параметр модели, не имеющий каузальной интерпретации (поскольку по отношению к X критерий обходных путей может быть не выполнен), а ε – ненаблюдаемая ошибка, обладающая свойством $\mathbb{E}\big(\varepsilon \mid do(D = d), X\big) = 0$. Переменные X, не представляющие самостоятельный интерес для исследователя, но добавляемые в статистическую модель с целью идентификации эффекта воздействия, называются контрольными переменными. Отсюда легко видно, что в типичном случае объясняющие переменные в множественной регрессии оказываются неравноценными: оценки одних коэффициентов могут иметь причинно-следственную интерпретацию, а других – нет.

DAG-модели совместимы с другими причинно-следственными моделями, в том числе с моделью Рубина, и позволяют систематизировать задачу отбора переменных при статистической оценке эффектов воздействия. В некоторых случаях возможно решение обратной задачи – построения возможных вариантов DAG-модели, основывась на наборе свойств условной зависимости/независимости между переменными в заданном наборе данных. Такой автоматизированный поиск структуры причинно-следственных связей в данных реализован, среди прочих, на платформах DoWhy, CausalML, causal-learn, Econ-ML, CausalImpact. Важно отметить, что задача построения DAG модели по данным не может иметь уникального решения, поскольку распределение вектора, совместимое с заданным DAG, совместимо и с более насыщенными DAG. Кроме того, интересный с философской точки зрения результат, заключается в том, что если некоторый DAG оказывается совместимым с имеющимися данными, то это так же верно для DAG, получаемого инверсией направлений рёбер исходной модели.[22] Таким образом, окончательный выбор DAG-структуры всегда должен быть связан с экспертной оценкой.

[22] Kalisch M., Buhlmann P. (2008). *Robustification of the PC-algorithm for directed acyclic graphs.* Journal of Computational and Graphical Statistics. 17: 773-789.

7.3 Структурные причинно-следственные модели

DAG-модели и критерий обходных путей являются удобными инструментами для построения статистических моделей, называемых структурными, которые пригодны для оценки эффектов воздействия. Такие модели состоят из одного или нескольких уравнений, описывающих направленные связи между переменными. Они отличаются от обычных регрессионных моделей лишь набором предположений о свойствах ошибок (или, что эквивалентно, коэффициентов), которые для линейного случая можно записать в следующем виде:

$$Y := X^\top \beta + \varepsilon. \tag{66}$$

Здесь Y – зависимая переменная, $X = (X_1, \dots, X_\ell)^\top$ – переменные воздействия/регрессоры, $\beta = (\beta_1, \dots, \beta_\ell)^\top$ – регрессионные коэффициенты. Знак присвоения в (66) подчеркивает направленную зависимость между переменными, и идентифицирует β, как такой параметр, при котором установка для X произвольного значения $x \in \mathbb{R}^\ell$, приводит к тому, что среднее значение Y становится равным $x^\top \beta$.

Ненаблюдаемая переменная ε, называемая ошибкой модели, объединяет эффект всех переменных, являющихся причинами Y, но не включенных в модель. Параметр β из (66) также можно идентифицировать, задавая свойства ε с использованием do-нотации для интервенций:

$$\mathbb{E}\big(\varepsilon \mid do(X = x)\big) = 0,$$

что соответствует нулевому условному математическому ожиданию ошибки в модели, подвергнутой интервенции $do(X = x)$ и идентифицирует β через набор моментных соотношений, соответствующим ортогональности между ошибкой и регрессорами в таком сценарии.

Стандартная рекомендация при оценке эффектов воздействия с использованием неэкспериментальных данных заключается в представлении гипотетического рандомизированного эксперимента, соответствующего рассматриваемой задаче.[23] Такой подход достаточно полезен и для интерпретации результатов, и для проверки адекватности исследуемой модели – хотя соответствующий рандомизированный эксперимент не всегда оказывается очевидным. Так, в исследовании, где изучается, насколько футбольные тренеры более склонны ставить автора победного гола в стартовый состав следующей игры, по сравнению с авторами голов, оказавшимися не последними,[24] важно отделить влияние именно того факта, что гол оказался победным – то есть, был забит при равном счете, и после него не было других голов, – то соответствующим рандомизированным испытанием можно считать эксперимент, в котором в матчах, выигранных командой с преимуществом в один мяч, и в которых было забито более одного гола, авторство забитых мячей перераспределяется случайным образом среди игроков команды.

Определение 7.3. (Ациклическая) структурная модель, соответствующая графу $G = (V, E)$, – это набор случайных величин $\{X_j\}_{j \in V}$, соответствующих уравнениям

$$X_j := f_j(Pa_j, \varepsilon_j), \quad j \in V,$$

где Pa_j – родители по отношению к вершине j направленного графа, f_j – неизвестные (измеримые) функции, а ошибки ε_j являются независимыми в совокупности.

Определение 7.4. Контрфактуальная структурная модель, получаемая вследствие интервенции $do(X_j =$

[23] Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

[24] Smirnov A. (2024). *Striking Success: How Goals Shape Coach Bias in Football.* Working paper.

x_j) – это новая структурная модель, основанная на модифицированном графе $G(x_j) = (V, E^*)$ и наборе (контрфактуальных) переменных $(X_k^*)_{k \in V^*}$, где

- все рёбра, входящие в вершину j удалены, то есть $\forall i \in V,\ e_{ij}^* \in E^*,\ e_{ij}^* = 0$;
- все остальные рёбра сохранены, $\forall i, k \in V:\ k \neq j,\ e_{ik}^* = e_{ik}$;
- контрфактуальные случайные величины определяются следующим образом:

$$X_k^* := f_k(Pa_k^*, \varepsilon_k), \quad \text{при } k \neq j,$$
$$X_j^* := x_j,$$

где Pa_k^* – родители X_k^* в E^*.

Для иллюстрации представленных понятий рассмотрим пример, основанный на так называемой “треугольной структурной модели” из [**?**].

Пример 7.5. Пусть структурная модель задана с помощью уравнений (67)-(69):

$$X := \varepsilon_x, \tag{67}$$
$$D := \alpha X + \varepsilon_d, \tag{68}$$
$$Y := \beta D + \gamma X + \varepsilon_y. \tag{69}$$

Такая структурная модель соответствует графу на рис. 12 слева:

Рис. 12: Слева: DAG-модель зависимости между рассматриваемыми переменными. Справа: результат интервенции $do(D = d)$

Для удобства предположим, что все переменные в модели совместно нормальны и имеют нулевое математическое ожидание, ошибки $\varepsilon_x, \varepsilon_y, \varepsilon_d$ независимы, а числовые коэффициенты α, β, γ таковы, что дисперсия у X, Y и D равна единице. Тогда легко видеть, что $\mathbb{E}(Y \mid D = d) = (\alpha\gamma + \beta)d$, но если установить для D значение d, то уравнение (68) теряет смысл, и заменяя в (69) значение D на константу, получаем, что функция регрессии примет вид $\mathbb{E}\big(Y \mid do(D = d)\big) = \beta d$, а средний эффект воздействия D на Y равен β. Полученной контрфактуальной модели соответствует граф на рис. 12 справа.

Сопоставляя свойства структурной линейной модели с моделью потенциальных исходов Рубина, легко видеть, что структурные модели являются моделями не для наблюдаемых, а для потенциальных исходов $Y(d)$, соответствующих различным возможным значениям переменной воздействия D. Тогда идентифицируемость β можно связать с условиями устойчивости индивидуальной величины воздействия. В рассмотренной модели (67)-(69) они выполнены благодаря предположению о независимости ошибок, и при оценке регрессии Y на D и X, оценка β методом наименьших квадратов (МНК) будет состоятельной в силу свойств метода.В то же время, в парной регрессии Y на D МНК-оценка будет иметь предел по вероятности, равный $\alpha\gamma + \beta$, что связано с нарушением предположения об экзогенности D: ошибка в такой модели содержит X и из-за

этого имеет ненулевую корреляцию с D, что приводит к систематической зависимости между значением воздействия D и потенциальными исходами $Y(D)$.

В рассмотренном примере имела место типичная для прикладных исследований ситуация, когда для получения состоятельной оценки искомого эффекта D на Y требуется включить в модель дополнительную (контрольную) переменную X, не представляющую самостоятельного интереса. В более типичном случае совместная нормальность не предполагается, и функции условного математического ожидания могут быть произвольными. Пусть, как обычно, D – переменная воздействия, Y – зависимая переменная, а $Y(d)$ – её потенциальные значения при $D = d$. Модель потенциальных исходов имеет вид

$$Y(d) := \beta_0 d + \varepsilon, \quad \mathbb{E}\varepsilon = 0,$$

где β_0 – значение параметра, интересующее исследователя, которое соответствует среднему изменению Y при увеличении d на единицу. Линейность этого эффекта может быть обоснована бинарностью переменной воздействия, или же в ситуации, когда нам важны лишь небольшие отклонения от некоторого значения d, что позволяет использовать линейное приближение. Предполагаем также, что переменная воздействия не является (безусловно) экзогенной – иначе β_0 можно было бы состоятельно оценить в простой регрессии Y на D, – но у исследователя есть данные, частично объясняющие Y, то есть $\varepsilon = g(X) + u$, где X – вектор наблюдаемых случайных величин, $g : \mathbb{R}^k \to \mathbb{R}$ – неизвестная (измеримая) функция, вид которой не представляет самостоятельного интереса, а u – ошибка модели, обладающая свойством $\mathbb{E}(u \mid X) = 0$.

Предположение 7.6. Предположение об условной экзогенности переменной воздействия в структурной причинно-следственной модели.

1. Корректная спецификация модели (англ.: model consistency) – наблюдаемые данные Y_i сгенерированы в соответствии со структурной моделью $Y(d)$,

$$Y_i = Y(D_i);$$

2. Условная некоррелированность:

$$\operatorname{cov}\big(D, Y(d) \mid X\big) = 0,$$

(наблюдаемая переменная воздействия D не коррелирует с ошибками модели условно на наблюдаемый набор контролей). В литературе на английском языке используются близкие по смыслу понятия conditional exogeneity и conditional ignorability, подразумевающие условную независимость D и $Y(d)$ при фиксированных X.

При таких предположениях в модели

$$Y = Y(D) := \beta_0 D + g(X) + u,$$

ошибка u ортогональна (в L^2-смысле) обоим регрессорам X и D, а значит параметр причинно-следственного эффекта β_0 является идентифицированным.

Оценка причинно-следственной связи на неэкспериментальных данных включает этап экспертного решения о выполнении условия экзогенности переменной воздействия в модели. Если такое условие может быть выполнено, то в оцениваемую модель следует включить контрольные переменные, обеспечивающие экзогенность переменной, то есть переменные, статистически зависимые с D и Y одновременно, причем не относящиеся к механизму воздействия D на Y.

В то же время, представленные условия дают возможность проверки полученных результатов следующим образом: в ходе дальнейшего исследования структуры зависимостей между переменными в рассматриваемой задаче могут быть обнаружены новые потенциально релевантные контрольные переменные. При этом, в случае, когда изначально предложенная структура зависимостей является верной, включение таких переменных в модель не должно приводить к существенному изменению оценок причинно-следственного эффекта. Подобная проверка может лишь увеличить убедительность полученного результата, но не окончательно доказать его (поскольку перебрать все потенциально возможные наборы контрольных переменных на практике невозможно, кроме гипотетической ситуации, когда удаётся подобрать модель, полностью объясняющую зависимую переменную), однако даёт достаточно эффективный механизм выявления ошибочных результатов.

В определении и рассмотренных примерах очевидна аналогия процедуры интервенции с её описанием для причинно-следственных моделей на графах. В этом отношении можно сказать, что все три рассмотренных вида каузальных моделей дополняют друг друга и позволяют выбрать наиболее удобный вариант для конкретной задачи. В каждом из этих случаев набор переменных дополняется структурой, задающей причинно-следственные отношения между ними, – что соответствует концепции Перла о том, что каузальная модель не может быть задана лишь в терминах совместного распределения.

7.4 DAG для моделей с инструментальными переменными

Метод инструментальных переменных активно используется в прикладных исследованиях и является одним из наиболее важных и надёжных подходов к оценке параметров, описывающих направленные зависимости. Общая идея метода заключается в поиске экзогенного источника вариации, который влияет на зависимую переменную Y опосредованно через переменную воздействия D. Этот метод применим, в том числе, в ситуациях, когда оценка влияния D на Y напрямую невозможна из-за наличия ненаблюдаемых переменных, одновременно влияющих на Y и D.

В простейшем случае, подходящем для применения метода инструментальных переменных, структурная модель имеет линейные зависимости, и основана на DAG-модели 13.

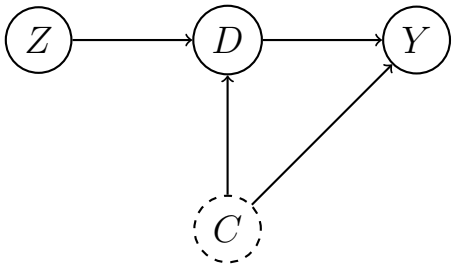


Рис. 13: В модели статистическая связь между D к Y складывается из причинно-следственного эффекта D на Y и корреляции за счёт вляния на них общей причины C, причём влияние C устранить невозможно из-за того, что соответствующая переменная является ненаблюдаемой. Переменную Z можно использовать в качестве инструмента для D

Ф. Райт обнаружил, что в такой модели, которую можно описать уравнениями

$$Y_i := \alpha_0 + \alpha_1 D_i + \varepsilon_i, \quad \operatorname{cov}(\varepsilon_i, D_i) \neq 0, \tag{70}$$

$$D_i := \beta_0 + \beta_1 Z_i + u_i, \tag{71}$$

эффект Z на Y равен $\alpha_1\beta_1$, а эффект Z на D равен β_1, причем оба этих параметра идентифицированы

в силу экзогенности Z, и их можно состоятельно оценить. Таким образом, состоятельную оценку α_1 можно получить в виде частного выборочных ковариаций, или двухшаговым методом наименьших квадратов (англ.: two stage least squares, TSLS), в котором сначала оценённая МНК модель (71) используется для построения прогнозных значений $\hat{D}_i$, а потом оценивается регрессия Y на $\hat{D}$. В простейшем случае эти два метода дают одинаковую оценку,

$$\hat{\alpha}_1^{IV} = \frac{\widehat{\mathrm{cov}}(Y_i, Z_i)}{\widehat{\mathrm{cov}}(D_i, Z_i)} \xrightarrow{p} \alpha_1. \tag{72}$$

Оценки, построенные по типу (72), называются IV-оценками от англ. instrumental variable. Однако, чаще для оценки эффекта воздействия в таких случаях используется метод 2SLS (two stage least squares, двухстадийный метод наименьших квадратов), который предполагает вначале оценить с помощью МНК уравнение (71) и построить прогнозные значения $\hat{D}_i$, а затем использовать их в модели (70) вместо исходных переменных D_i. В простом случае IV и 2SLS оценки совпадают.

В более сложном случае имеется k регрессоров X, ℓ инструментальных переменных Z для переменной воздействия, и модель имеет вид

$$Y = \theta D + X\beta + \varepsilon, \quad \mathbb{E}(\varepsilon \mid X, Z) = 0.$$

Для получения 2SLS-оценки нужно предсказать D с помощью Z, то есть $\hat{D} = P_Z D$, где $P_Z = Z(Z^\top Z)^{-1} Z^\top$, и далеее оценить МНК регрессию Y на $\hat{D}$ и X. Для выражения оценки при $\hat{D}$ можно воспользоваться теоремой Фриша-Ву-Ловелла: искомая оценка равна оценке в регрессии $Q_X Y$ на $Q_X \hat{D}$, где $Q_X = I - X(X^\top X)^{-1} X^\top$ – проектор на ортогональное дополнение к столбцам X. Тогда

$$\hat{\theta} = (\hat{D}^\top Q_X \hat{D})^{-1} \hat{D}^\top Q_X Y.$$

На графе подобной модели возможно одновременное наличие наблюдаемых и ненаблюдаемых мешающих переменных, но этот случай часто возможно свести к предыдущему при помощи перехода к остаткам от проекции.

Пример 7.7. В качестве примера можно рассмотреть модель с инструментальной переменной и мешающим фактором: пусть структура зависимостей имеет вид

$$\begin{aligned} Y &:= \alpha_0 D + \delta C + g_Y(X) + \varepsilon_Y, \\ D &:= \beta_0 Z + \gamma C + g_D(X) + \varepsilon_D \\ C &:= f_C(X) + \varepsilon_C \\ X &:= f_X(Z) + \varepsilon_X \\ Z &:= \varepsilon_Z. \end{aligned}$$

Условно на значение X, ошибки модели $\varepsilon_Y, \varepsilon_D, \varepsilon_Z, \varepsilon_C$ попарно не коррелированы и имеют нулевое среднее. Требуется оценить структурный параметр $\alpha_0 = \frac{\partial \mathbb{E}\left(Y | do(D=d)\right)}{\partial d}$. Соответствующая DAG-модель приведена на рис. 14.

Удаление влияния X из переменных осуществляется с помощью оценки функций регрессии $\mathbb{E}(D \mid X)$, $\mathbb{E}(Y \mid X)$ и вычисления соответствующих остатков. Несложно видеть, что структура зависимостей между остатками переменных D, Y, а также переменными Z и C соответствует DAG-модели 13, и поэтому искомый параметр может быть оценён аналогично (72).

В более общем случае, когда предположение о линейной зависимости между ключевыми переменными в

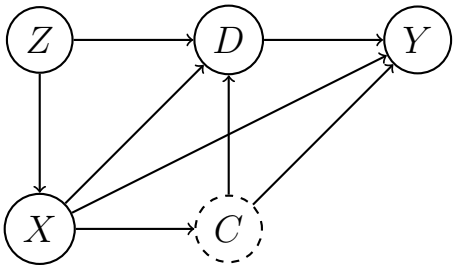


Рис. 14: Модель с наблюдаемыми и ненаблюдаемыми конфаундерами

модели оказывается слишком ограничивающим, существуют подходы для оценки эффекта воздействия, основанные на монотонности зависимости D от инструмента Z. Примером такого подхода является метод оценки локального среднего эффекта воздействия (англ.: local average treatment effect, LATE), предложенный Г. Имбенсом и Дж. Ангристом[25], который позволяет оценить эффект воздействия среди «соглашающихся» (англ.: compliers). Это часть популяции, отличная от той, которая использовалась для расчета ATT ранее: терминология основана на специфическом примере, рассмотренном в оригинальной работе, где исследовался эффект от участия в обучающей программе на некоторый показатель Y, а в качестве инструмента для бинарной переменной воздействия D использовалась бинарная переменная Z, являющаяся индикатором получения данным человеком приглашения участвовать в программе. Как показали авторы, в предположении, что вероятность $D = 1$ не уменьшается по Z, в модели можно идентифицировать эффект $\mathbb{E}\big(Y(1) - Y(0) \mid D(1) > D(0)\big)$, где $\{D(1) > D(0)\}$ – событие, при котором человек участвует в программе при наличии приглашения, и не участвует в обратном случае. Схожий подход используется в инструментальной квантильной модели (англ.: instrumental variable quantile regression, IVQR), в которой предполагаемая строгая монотонность влияния пропущенных переменных на зависимую переменную позволяет идентифицировать (обычно параметризованную) зависимость Y от переменной воздействия.

Существенным недостатком метода инструментальных переменных является более высокая дисперсия оценок по сравнению с обычными оценками методом наименьших квадратов. Один из путей снижения дисперсии у таких оценок заключается в построении оптимального инструмента на основе набора потенциальных инструментальных переменных $Z = (Z_1, \dots, Z_k)$; простейшим примером такого данного подхода является двухшаговый МНК. Традиционный анализ в этой области предполагает непараметрическое задание оптимального инструмента (который в большинстве случаев соответствует регрессионной функции переменной воздействия на используемые инструменты), что эквивалентно созданию словаря из большого количества технических инструментов, то есть трансформаций исходных инструментальных переменных

$$T = (T_1, \dots, T_p)^\top = \big(T_1(Z), \dots, T_p(Z)\big)^\top. \tag{73}$$

Число используемых технических инструментов p может значительно превосходить число наблюдений выборки n. Такая ситуация возможна в случае, когда общее число потенциальных инструментальных переменных Z_ℓ, $\ell \in \{1, \dots, k\}$, оказывается велико само по себе, либо большое количество технических переменных T_ℓ получается в результате разложения функций от Z в ряды. В таком случае технические инструменты могут иметь вид вейвлетов, сплайнов, бинарных переменных, степеней и произведений компонент Z. Такой анализ близок к стандартным подходам непараметрической статистики и эконометрики, опирающимся на приближения функций возрастающим с размером выборки числом элементов некоторого ряда, обеспечивая тем самым асимптотически точное представление функции. При этом, как правило, требуется, чтобы чис-

[25] Imbens G. W., Angrist J. D. (1994). *Identification and Estimation of Local Average Treatment Effects.* Econometrica. 62: 467-475.

ло элементов ряда, используемое для приближения, росло с размером выборки достаточно медленно, что ограничивает применимость подобных методов в выборках умеренных размеров.

Однако, несмотря на концептуальную привлекательность подобных непараметрических методов с точки зрения стандартной асимптотической теории, нежелательные свойства оценок в малых выборках задокументированы в значительном числе работ. Осложнения могут возникнуть в случае, когда некоторые из используемых инструментов оказываются слабыми (например, имеют близкую к нулю корреляцию с инструментируемой переменной в линейной модели), что требует использования подхода, связанного с асимптотикой слабых инструментов.

8 Оценка моделей с мешающим параметром

8.1 Полупараметрические модели

С помощью DAG (или иных структурных подходов) можно определить, какие переменные необходимо включить в модель: это могут быть как контрольные переменные, блокирующие обходные пути, так и инструментальные переменные, когда часть конфаундеров остаётся ненаблюдаемой. Однако граф ничего не говорит о том, в какой функциональной форме эти переменные должны входить в модель. Ошибочное задание формы связи (например, необоснованное предположение о линейности там, где она отсутствует) способно привести к смещённым оценкам эффекта, неверным выводам о его значимости и даже к ложному обнаружению или пропуску причинно-следственной связи. Это связано с тем, что неверная параметризация не позволяет полностью исключить компоненту переменной воздействия, коррелирующую с ошибкой модели, – тем самым нарушаются условия идентификации и возникает скрытое смещение. Здесь на первый план выходит полупараметрический подход, ключевая идея которого заключается в разделении модели на две части:

- Параметрическая часть модели содержит конечномерный (часто скалярный) оцениваемый параметр и имеет строгие ограничения на функциональную форму. Эта часть используется для описания зависимости Y от переменной воздействия D;
- Непараметрическая часть модели включает зависимость от контрольных переменных X, а также вид распределения ошибки. Обычно предполагается, что "вклад" X описывается с помощью произвольной (измеримой) неизвестной функции – бесконечномерного параметра.

Пример 8.1. (Оценка частично-линейной модели).

Уравнение для зависимой переменной Y имеет следующий вид:

$$Y_i = \theta D_i + g(X_i) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid X_i, D_i) = 0, \quad i \in \{1, \dots, n\}. \tag{74}$$

Здесь D и $X = (X_1, \dots, X_p)$ – объясняющие переменные, ε – ошибка модели, $g : \mathbb{R}^p \to \mathbb{R}$ – неизвестная измеримая функция, а θ – интересующий скалярный параметр.

Первый способ оценки θ в такой модели, предложенный в работе Г. Вахбы,[26] имеет следующий вид: выборка разбивается на основную (индексы $i \in I$) и вспомогательную ($i \in \overline{I}$), и используется итеративная процедура:

- предполагается начальное значение $\hat{\theta}_0$ для θ и с помощью сплайновых полиномов (вместо них можно использовать метод ядерной оценки регрессионной функции, или любой другой подходящий метод машинного обучения) на вспомогательной выборке оценивается $\hat{g}_0(X_i) = \mathbb{E}(Y_i - \hat{\theta}_0 D_i \mid X_i)$;
- используя вторую часть выборки вычисляются значения остатков $Y_i - \hat{g}_0(X_i)$, после чего с помощью МНК можно получить уточнённую оценку θ:

$$\hat{\theta}_1 := \left(\frac{1}{n_I} \sum_{i \in I} D_i^2 \right)^{-1} \frac{1}{n_I} \sum_{i \in I} D_i \big(Y_i - \hat{g}_0(X_i) \big);$$

- процедура повторяется с начала, используя полученную на предыдущей итерации оценку θ до достижения сходимости алгоритма.

[26] Wahba G. (1984). *Cross validated spline methods for direct and indirect sensing experiments.* In "Statistical Signal Processing", E.J. Wegman and J. Smith, eds., Marcel Dekker. 53: 179-190.

Скорость сходимости по распределению у такой оценки θ медленнее стандартной $n^{-1/2}$ из-за смещения при оценке g:

$$\sqrt{n}\left(\hat{\theta}-\theta\right)=\underbrace{\left(\frac{1}{n}\sum_{i\in I}D_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I}D_i\varepsilon_i}_{a}+\underbrace{\left(\frac{1}{n}\sum_{i\in I}D_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I}D_i\big(g(X_i)-\hat{g}(X_i)\big)}_{b} \tag{75}$$

Первая компонента в (75) сходится по ЦПТ к нормальной случайной величине, но вторая (которую можно рассматривать как смещение из-за регуляризации при оценке параметра высокой размерности) не центрирована и в общем случае расходится. Обозначая скорость сходимости $\hat{g}$ к g за $n^{-\psi_g}$, где $\psi_g<\frac{1}{2}$, получаем, что b имеет асимптотику $\sqrt{n}n^{-\psi_g}$ и расходится с ростом n.

П. Робинсон[27] предложил другой метод, обеспечивающий асимптотическую нормальность оценки θ в частично-линейной модели. Из уравнения (74) следует, что функция регрессии $\mathbb{E}(Y_i\mid X_i)$ имеет вид $\mathbb{E}(Y_i\mid X_i)=\theta\mathbb{E}(D_i\mid X_i)+g(X_i)$, и вычитая её из исходного уравнения, можно получить

$$Y_i-\mathbb{E}(Y_i\mid X_i)=\theta\big(D_i-\mathbb{E}(D_i\mid X_i)\big)+\varepsilon_i. \tag{76}$$

В алгоритме Робинсона две функции регрссии $m_Y(x)=\mathbb{E}(Y\mid X=x):=\theta m_D(x)+g(x)$ и $m_D(x):=\mathbb{E}(D\mid X=x)$ оцениваются на полной выборке методом Надараи-Ватсона[28] и используются для вычисления остатков $\tilde{Y}_i:=Y_i-\hat{m}_Y(X_i)$ и $\tilde{D}_i:=D_i-\hat{m}_D(X_i)$. После этого вычисляется МНК-оценка $\hat{\theta}$ в регрессии $\tilde{Y}_i$ на $\tilde{D}_i$.

Рассмотрим доказательство асимптотической нормальности оценки Робинсона с дополнительным приёмом, упрощающим анализ: будем считать, что выборка случайным образом разделена на две части с индексами из I и $\bar{I}$ и размерами n_I и $n_{\bar{I}}\approx n/2$. Регрессионные функции оцениваются на второй подвыборке, итоговая МНК-оценка строится по первой подвыборке.[29]

Ошибки оценок регрессионных функций $r_{D_i}=\hat{m}_D(X_i)-m(X_i)$ и $r_{Y_i}=\hat{m}_Y(X_i)-m(X_i)$ должны удовлетворять следующим условиям:

$$\mathbb{E}r_{D_i}^2=o(n^{-\frac{1}{2}}),\quad \mathbb{E}r_{Y_i}^2=o(n^{-\frac{1}{2}}). \tag{77}$$

Обозначим $\eta_i=D_i-\mathbb{E}(D_i\mid X_i)$ компоненту D, ортогональную X; $\mathbb{V}(\eta_1\mid X)=\sigma_\eta^2$. Тогда

$$\tilde{D}_i=\eta_i-r_{D_i},\quad \tilde{Y}_i-\theta\tilde{D}_i=\varepsilon_i-r_{Y_i}+\theta r_{D_i}. \tag{78}$$

МНК-оценка $\hat{\theta}$:

$$\hat{\theta}=\Big(\sum_{i\in I}\tilde{D}_i^2\Big)^{-1}\sum_{i\in I}\tilde{D}_i\tilde{Y}_i\Rightarrow\sqrt{n}(\hat{\theta}-\theta)=\frac{\frac{1}{\sqrt{n_I}}\sum_{i\in I}\tilde{D}_i(\tilde{Y}_i-\theta\tilde{D}_i)}{\frac{1}{n_I}\sum_{i\in I}\tilde{D}_i^2}.$$

Подставим в числитель полученной дроби выражения из (78):

$$\frac{1}{\sqrt{n_I}}\sum_{i\in I}\tilde{D}_i(\tilde{Y}_i-\theta\tilde{D}_i)=\frac{1}{\sqrt{n_I}}\sum_{i\in I}(\eta_i-r_{D_i})(\varepsilon_i-r_{Y_i}+\theta r_{D_i})=$$

$$=\frac{1}{\sqrt{n_I}}\sum_{i\in I}\eta_i\varepsilon_i-\frac{1}{\sqrt{n_I}}\sum_{i\in I}\eta_i r_{Y_i}+\frac{1}{\sqrt{n_I}}\sum_{i\in I}\theta\eta_i r_{D_i}-\frac{1}{\sqrt{n_I}}\sum_{i\in I}r_{D_i}\varepsilon_i+\frac{1}{\sqrt{n_I}}\sum_{i\in I}r_{D_i}r_{Y_i}-\frac{1}{\sqrt{n_I}}\sum_{i\in I}\theta r_{D_i}^2.$$

[27] Robinson P.M. (1988). *Root-N-Consistent Semiparametric Regression.* Econometrica. 56: 931-954.

[28] Надарая Э. А. *Непараметрические оценки плотности вероятности и кривой регрессии.* Издательство Тбилисского университета, 1965. — Вып. «Теория вероятностей и математическая статистика».

[29] Такой приём широко используется в современных алгоритмах двойного обучения, и обычно сочетается с кросс-фитингом – когда роли двух подвыборок меняются, а итоговая оценка получается усреднением. Это позволяет более эффективно использовать имеющуюся в выборке информацию и снизить дисперсию оценки.

Благодаря заданным скоростям для r_{Di} и r_{Yi} все слагаемые, кроме первого, сходятся к нулю. Например,

$$\left|\frac{1}{\sqrt{n_I}}\sum_{i\in I} r_{D_i} r_{Y_i}\right| \le \frac{\sqrt{n_I}}{2}\left(\frac{1}{n_I}\sum_{i\in I} r_{D_i}^2 + \frac{1}{n_I}\sum_{i\in I} r_{Y_i}^2\right) = \sqrt{n} o_p(n^{-\frac{1}{2}}) = o_p(1),$$

где для связи асимптотики $\mathbb{E} r_{Y_i}^2$ и выборочного среднего можно использовать неравенство Маркова.

Для другой компоненты удобно аргументировать сходимость к нулю в L^2 :

$$\mathbb{E}\left[\frac{1}{\sqrt{n_I}}\sum_{i\in I}\eta_i r_{Y_i} \mid X_{I_1}\right] = 0, \quad \mathbb{V}\left[\frac{1}{\sqrt{n_I}}\sum_{i\in I}\eta_i r_{Y_i} \mid X_{I_1}\right] = \frac{1}{n_I}\sum_{i\in I} r_{Y_i}^2 \sigma_\eta^2 = o_p(1).$$

Из сходимости к нулю в L^2 следует аналогичная сходимость по вероятности. Итак,

$$\frac{1}{\sqrt{n_I}}\sum_{i\in I}\tilde{D}_i(\tilde{Y}_i - \theta\tilde{D}_i) = \frac{1}{\sqrt{n_I}}\sum_{i\in I}\eta_i\varepsilon_i + o_p(1).$$

Асимптотика знаменателя:

$$\frac{1}{n_I}\sum_{i\in I}\tilde{D}_i^2 = \frac{1}{n_I}\sum_{i\in I}(\eta_i - r_{D_i})^2 = \frac{1}{n_I}\sum_{i\in I}\eta_i^2 - \frac{2}{n_I}\sum_{i\in I}\eta_i r_{D_i} + \frac{1}{n_I}\sum_{i\in I} r_{D_i}^2.$$

Перекрёстное произведение и квадрат r_{Di} сходятся к нулю по вероятности, поэтому

$$\frac{1}{n_I}\sum_{i\in I}\tilde{D}_i^2 \xrightarrow{p} \mathbb{E}\eta_1^2 = \sigma_\eta^2.$$

Для итогового результата осталось заметить, что числитель асимптотически эквивалентен нормированной сумме *н.о.р.* величин с нулевым средним и конечной дисперсией, и применить лемму Слуцкого. □

В основе метода Робинсона лежит идея удаления мешающего воздействия контрольных переменных X из **обоих** переменных Y и D. Учитывая, что условное математическое ожидание является ортогональной проекцией в L^2-пространстве, можно понять, что предложенная процедура аналогична теореме Фриша-Ву-Ловелла в пространстве случайных величин, интегрируемых с квадратом.

8.2 Ортогонализация по Нейману

Ключевое различие между оценками Вахбы и Робинсона, приводящее к тому, что только вторая имеет асимптотически нормальное распределение, заключается в том, что конструкция Робинсона оказывается малочувствительной к неточности оценки неизвестных функций. Такой подход позже получил название ортогонализации по Нейману, и связан с исследованиями Е. Неймана по построению тестовых статистик в задачах, содержащих мешающий параметр.

Рассмотрим их подробнее: пусть исследователю доступна *н.о.р.* выборка $\{X_i\}_{i=1}^n$, где X – случайные вектора с плотностью $p(x; a, b)$, причём $b \in \mathbb{R}^s$ – мешающий параметр распределения, а a – скалярный параметр, представляющий интерес. Для теста гипотезы $H_0 : a = a_0$ можно использовать стандартизованную сумму

$$Z_n(b) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{f(X_i) - \mathbb{E}_b\big(f(X_i)\big)}{\sigma_b}, \tag{79}$$

где $f(\cdot)$ – произвольная измеримая функция, такая, что $f(X)$ имеет конечную дисперсию σ_b^2 при H_0; также $\mathbb{E}_b\big(f(X_i)\big)$ – математическое ожидание $f(X_i)$ при H_0 и настоящем значении b. Истинные параметры распре-

деления обозначим за (a_0, b_0) и предполагаем, что $p(x; a, b)$ дважды дифференцируема по b в окрестности b_0 и соответствующие градиенты ограничены.

В силу центральной предельной теоремы для независимых величин, $Z_n(b_0)$ при H_0 сходится по распределению к стандартной нормальной случайной величине, – однако содержит неизвестный параметр b, который требуется заменить оценкой. Нейман отмечал, что при работе с данными, связанными с задачами общественных наук, часто приходится иметь дело с нестандартными распределениями, далеко не всегда позволяющими надеяться на существование оценки b, не приводящей к смещению в (79) в широком диапазоне возможных значений a. Компромиссная оценка b, предложенная им, должна обладать свойством $\sqrt{n}(\hat{b} - b - A(a - a_0)) = O_p(1)$, где A – некоторая константа, отвечающая за смещение. Статистика $Z_n(\hat{b})$ имеет такое же асимптотическое распределение, как и $Z_n(b_0)$, тогда и только тогда, когда $\nabla_b \mathbb{E}_b f(X) = 0$ – условие ортогональности по Нейману. Ортогональности можно добиться, заменяя в (79) $f(\cdot)$ на функцию $g(\cdot)$, имеющую следующий вид:

$$g(x, b) := f(x, b) - \mathbb{E}_b f(X) - \sum_{j=1}^{s} c_j \left.\frac{\partial \log p}{\partial b_j}\right|_{\xi=0}, \quad j \in \{1, 2, \ldots, s\}, \tag{80}$$

где c_j – коэффициенты ортогональной проекции в L^2 функции $f(X, b)$ на $\frac{\partial \log p}{\partial b_j}$.

Критерии, в которых значения $Z_n(\hat{b})$ сравнивается с $q_{\alpha/2}$ квантилями $\mathcal{N}(0, 1)$, Нейман предложил называть $C(\alpha)$-тестами.[30]

Теорема 8.2. *Статистика $Z_n(\hat{b})$, рассчитанная с использованием асимптотически нормальной при H_0 оценки $\hat{b}$ и ортогонализованной функции $g(\cdot)$ сходится по распределению к $\mathcal{N}(0, 1)$ равномерно по b.*

Доказательство. Доказательство основано на разложении статистики в ряд Тейлора. При H_0 и $\sigma_b = 1$ выполняется

$$Z_n(\hat{b}) = Z_n(b_0) - \left\langle \frac{1}{n} \sum_{i=1}^{n} \nabla_b \mathbb{E}_b\big(f(X_i)\big), \sqrt{n}(\hat{b} - b_0) \right\rangle + o_p(1).$$

Условия на $p(x; a, b)$ позволяют дифференцировать под знаком интеграла, поэтому

$$\nabla_b \mathbb{E}_b\big(f(X_i)\big) = \int f(x) \nabla_b p(x; a_0, b)\, dx = \int f(x) \nabla_b \log p(x; a_0, b) p(x; a_0, b)\, dx.$$

Полученное выражение соответствует $\mathbb{E}_b\big(f(X_i) \nabla_b \log p(X; a_0, b)\big)$. Если $f(X)$ ортогональна в L^2 ко всем компонентам $\nabla_b \log p(X; a_0, b)$, то математическое ожидание равно нулю, и $Z_n(\hat{b})$ и $Z_n(b_0)$ имеют одинаковое асимптотическое распределение. □

Упражнение. Как модифицировать доказательство для случая, когда $\sigma_b \neq 1$?

Представленное доказательство показывает центральную роль ортогонализации скор-функции, приводящее к занулению линейной компоненты разложения Тейлора. Схожая идея позже была предложена Черножуковым и др.[31] для оценки полупараметричских моделей: моментное условие, идентифицирующее целевой параметр, должно быть ортогональным в неймановском смысле к мешающим параметрам модели. Такой контекст позволяет заметить принципиальное различие между оценками Вахбы и Робинсона в частично-

[30] Neyman, J. (1979). *$C(\alpha)$ tests and their use.* Sankhyā: The Indian Journal of Statistics. 41: 1–21.

[31] Chernozhukov V., Chetverikov D., Demirer M., Duflo E., Hansen C., Newey W., Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters.* The Econometrics Journal. 21: C1–C68.

линейной модели: в первом случае оценка $\hat{\theta}$ получается минимизацией регуляризованного функционала

$$\sum_{i=1}^{n} \big(Y_i - \theta D_i - g(X_i)\big)^2 + \lambda \|g\|,$$

которому соответствует популяционный момент, основанный на условии первого порядка

$$\mathbb{E}\Big[D\big(Y - \theta D - g(X)\big)\Big] = 0, \tag{81}$$

В уравнении (81) имеется единственный мешающий параметр g. Проверим ортогональность: пусть $h(X)$ – произвольное направление. Вычисляем производную Гато в истинной точке (θ_0, g_0):

$$\frac{\partial}{\partial t}\mathbb{E}\Big[D\Big(Y - \theta_0 D - \big(g_0(X) + t h(X)\big)\Big)\Big]\Big|_{t=0} = -\mathbb{E}\big[Dh(X)\big] \neq 0.$$

Во втором случае (оценка Робинсона) оценка θ получается в регрессии остатков, поэтому соответствующее моментное условие имеет следующий вид:

$$\mathbb{E}\Big[\big(D - m_D(X)\big)\Big(Y - m_Y(X) - \theta\big(D - m_D(X)\big)\Big)\Big] = 0. \tag{82}$$

Здесь мешающий параметр – это (m_Y, m_D). Проверим ортогональность, подставляя возмущённые функции $m_Y + t h_Y$, $m_D + t h_D$ и вычисляя производную в точке, соответствующей истинным значениям параметров.

Пусть $\varepsilon_0 = Y - m_Y(X) - \theta_0(D - m_D(X))$. Тогда $\mathbb{E}[\varepsilon_0 \mid D, X] = 0$ и $\mathbb{E}[\varepsilon_0(D - m_D(X))] = 0$.

Возмущённая моментная функция: $\psi = (D - m_D - t h_D)\big(Y - (m_Y + t h_Y) - \theta_0(D - m_D - t h_D)\big)$. Берём производную по t в $t = 0$ и математическое ожидание:

$$\frac{d}{dt}\mathbb{E}\psi\Big|_{t=0} = \mathbb{E}\Big[- h_D \varepsilon_0 - (D - m_D)(h_Y + \theta_0 h_D)\Big].$$

Раскладываем:

- $\mathbb{E}[h_D \varepsilon_0] = \mathbb{E}[h_D \mathbb{E}[\varepsilon_0 \mid X]] = 0$;
- $\mathbb{E}[(D - m_D) h_Y] = \mathbb{E}[h_y \mathbb{E}[D - m_D \mid X]] = 0$;
- $\mathbb{E}[(D - m_D)(\theta_0 h_D)] = \mathbb{E}[\theta_0 h_X \mathbb{E}[D - m_D \mid X]] = 0$.

Таким образом, производная по любому направлению (h_Y, h_D) равна нулю – условие ортогональности выполнено. Далее будет показано, что это условие, как и в случае $C(\alpha)$-тестов, гарантирует, что ошибки оценивания мешающих функций (даже при непараметрическом оценивании с медленной сходимостью) не влияют на асимптотическое распределение $\hat{\theta}$, которое остаётся $\sqrt{n}$-нормальным. В следующей части раздела мы рассмотрим метод LASSO оценки неизвестной функции, позволяющий во многих случаях добиться достаточно высокой скорости сходимости прогноза к истинному значению. После этого перейдём к формулировке и доказательству свойств метода двойного машинного обучения в более общем виде.

9 Статистические модели высокой размерности

9.1 Неравенства концентрации меры

Для оценки неизвестной функциональной зависимости между переменными в данных обычно выбирают один из двух подходов: параметрический или непараметрический.

Первый вариант предполагает жесткую параметризацию модели (например, использование линейных функций с конечным числом параметров), но может приводить к проблемам, если используемый функциональный класс недостаточно хорошо может приблизить искомую функцию. Второй вариант – разложение неизвестной функции по базису достаточно богатого бесконечномерного функционального класса, такого как полиномы, ряды Фурье, или простые функции. Поскольку число параметров в таких моделях бесконечно, их теоретический анализ традиционно опирается на асимптотику, предполагающую рост числа оцениваемых коэффициентов с размером выборки.

Классический асимптотический подход, основанный на центральной предельной теореме, предполагает, что размер выборки n стремится к бесконечности при фиксированной размерности пространства признаков p. Однако в современных задачах размерность p часто сопоставима с n или значительно его превышает ($p \geq n$). В таких условиях асимптотические результаты перестают быть применимы: предельные распределения либо не существуют, либо сходятся слишком медленно, чтобы давать полезные гарантии для конечных выборок.

Возникает необходимость в неасимптотических вероятностных гарантиях – неравенствах, которые выполняются для любого конечного n и p, а не только в пределе $n \to \infty$. Ключевым инструментом здесь выступают концентрационные неравенства – верхние оценки для вероятности отклонения от своего среднего значения у случайных величин. Для того чтобы такие оценки были нетривиальными и достаточно быстрыми (экспоненциально убывающими), необходимо, чтобы распределения имели “лёгкие хвосты”.

Большинство распределений, встречающихся на практике, попадают в один из двух классов:

- **Субгауссовские величины**: хвосты убывают как $\exp(-ct^2)$. Примеры: нормальное распределение, ограниченные случайные величины, суммы независимых ограниченных величин.
- **Субэкспоненциальные величины**: хвосты убывают как $\exp(-ct)$. Примеры: экспоненциальное распределение, хи-квадрат, произведения субгауссовских величин.

Для этих классов существуют строгие концентрационные неравенства (Хёфдинга, Бернштейна, Чернова), позволяющие контролировать максимумы по многим координатам одновременно. Это критически важно в непараметрической статистике для ситуаций, когда нужно гарантировать, что ни одна из p координат не отклонится слишком сильно, даже если p растёт экспоненциально по n.

9.2 Субгауссовские и субэкспоненциальные распределения

Определение 9.1. Центрированная случайная величина ξ называется субгауссовской, если существует положительная константа σ (параметр масштаба), такая что производящая функция моментов ξ удовлетворяет условию:

$$\mathbb{E}e^{tX} \leq \exp\left(\frac{\sigma^2 t^2}{2}\right) \qquad \forall t \in \mathbb{R}.$$

Наименьшая такая константа σ называется субгауссовой нормой X и обозначается, как $\|X\|_{\psi_2}$. Существуют другие эквивалентные определения субгауссовости; одно из них напрямую связано с характеризацией

хвостов: существуют константы $c, C > 0$ такие, что

$$\mathrm{P}(|X| > u) \leq C \exp\left(-\frac{u^2}{c^2}\right) \qquad \forall u > 0.$$

Упражнение: докажите эквивалентность этих определения (с точностью до констант).

Субгауссовские распределения занимают важное место в современной статистике и машинном обучении, поскольку они позволяют строго контролировать хвосты сумм и максимумов таких случайных величин, обеспечивая оценки, которые лежат в основе выбора регуляризационных параметров, доказательства состоятельности LASSO-оценок и построения доверительных интервалов в условиях, когда размерность данных сравнима или превосходит объём выборки.

Примеры субгауссовских случайных величин и распрделений:

- Нормальное распределение: $X \sim \mathcal{N}(0, \sigma^2)$, $\|X\|_{\psi_2} = \sigma$;
- Ограниченные случайные величины: если $|X| \leq B$ п.н., то X субгауссова с $\|X\|_{\psi_2} \leq B$;
- Гладкие функции от субгауссовских величин;
- Конечные смеси субгауссовых распределений;
- Кончные суммы независимых субгауссовских величин.

Распределения с хвостами тяжелее гауссовских (например, распределение Лапласа или t-распределение) не являются субгауссовскими.

Определение 9.2. Центрированная случайная величина Z ($\mathbb{E}Z = 0$) называется субэкспоненциальной, если существует константа $K > 0$ такая, что для всех $m \geq 3$

$$\mathbb{E}|Z|^m \leq \frac{m!}{2} K^{m-2} \sigma^2, \tag{3}$$

где $\sigma^2 = \mathbb{E}Z^2$ – дисперсия Z. Эквивалентно: хвосты убывают экспоненциально: $\mathrm{P}(|Z| > u) \leq Ce^{-cu}$.

Лемма 9.3. *Если X и ε – субгауссовские случайные величины, то их произведение $X\varepsilon$ является субэкспоненциальным.*

Доказательство. Введём обозначения для субгауссовских хвостовых оценок у X и ε:

$$\mathrm{P}(|X| \geq t) \leq 2\exp\left(-\frac{t^2}{K_1^2}\right), \quad \mathrm{P}(|\varepsilon| \geq t) \leq 2\exp\left(-\frac{t^2}{K_2^2}\right).$$

Достаточно показать экспоненциальное убывание у произведения $\mathbb{P}(|X\varepsilon| \geq t)$:

$$\begin{aligned}\{|X\varepsilon| \geq t\} \subset \{|X| \geq \sqrt{t}\} \cup \{|\varepsilon| \geq \sqrt{t}\} \Rightarrow \mathrm{P}(|X\varepsilon| \geq t) &\leq \mathrm{P}(|X| \geq \sqrt{t}) + \mathrm{P}(|\varepsilon| \geq \sqrt{t}) \\ &\leq 2\exp\left(-\frac{t}{K_1^2}\right) + 2\exp\left(-\frac{t}{K_2^2}\right).\end{aligned}$$

Обозначим $K^2 = \max(K_1^2, K_2^2)$. Тогда оба слагаемых убывают не быстрее, чем $Ce^{-\frac{t}{K^2}}$:

$$\mathrm{P}(|X\varepsilon| \geq t) \leq 4\exp\left(-\frac{t}{K^2}\right).$$

Это в точности форма хвоста субэкспоненциальной величины (убывание как e^{-ct}, а не как e^{-ct^2}). $\square$

9.3 Неравенство Бернштейна для сумм независимых субэкспоненциальных величин

Теорема 9.4. *(Неравенство Бернштейна).*

Пусть $Z_1, \dots, Z_n$ – независимые центрированные субэкспоненциальные случайные величины. Для каждой из них существуют константы $K_i > 0$ и $\sigma_i^2 = \mathbb{E}Z_i^2$ такие, что для всех $m \geq 1$

$$\mathbb{E}|Z_i|^m \leq \frac{m!}{2} K_i^{m-2} \sigma_i^2. \tag{83}$$

Обозначим $S_n = \sum_{i=1}^n Z_i$. Тогда для любого $t > 0$ справедливо

$$\mathrm{P}\left(S_n \geq t\right) \leq \exp\left(-\frac{t^2}{2\left(\sum_{i=1}^n \sigma_i^2 + K_{\max} t\right)}\right), \quad K_{\max} = \max_{1 \leq i \leq n} K_i.$$

Аналогичное неравенство верно для $S_n \leq -t$, поэтому

$$\mathrm{P}\left(|S_n| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\left(\sum \sigma_i^2 + K_{\max} t\right)}\right).$$

Доказательство. Используя (83), покажем, что для любого λ с $0 \leq \lambda \leq \frac{1}{K_{\max}}$ выполнено

$$\mathbb{E}e^{\lambda Z_i} \leq \exp\left(\frac{\lambda^2 \sigma_i^2}{2(1 - K_i\lambda)}\right) \leq \exp\left(\lambda^2 \sigma_i^2\right) \quad \text{при } \lambda \leq \frac{1}{2K_i}.$$

Действительно, разложение $e^{\lambda Z_i} = 1 + \lambda Z_i + \sum_{m \geq 2} \frac{\lambda^m Z_i^m}{m!}$ и условие (83) дают:

$$\mathbb{E}e^{\lambda Z_i} \leq 1 + \frac{\lambda^2 \sigma_i^2}{2} \sum_{m \geq 2} (K_i\lambda)^{m-2} \leq 1 + \frac{\lambda^2 \sigma_i^2}{2(1 - K_i\lambda)} \leq \exp\left(\frac{\lambda^2 \sigma_i^2}{2(1 - K_i\lambda)}\right).$$

Разложение в ряд функции $e^{\lambda Z_i}$ можно почленно интегрировать, поскольку этот ряд при $|\lambda| < \frac{1}{K_i}$ сходится абсолютно.

При $K_i\lambda \leq \frac{1}{2}$ имеем $\frac{1}{1 - K_i\lambda} \leq 2$, следовательно $\mathbb{E}e^{\lambda Z_i} \leq \exp\left(\lambda^2 \sigma_i^2\right)$.

Поскольку Z_i независимы, для $0 \leq \lambda \leq \frac{1}{2K_{\max}}$ получаем

$$\mathbb{E}e^{\lambda S_n} = \prod_{i=1}^n \mathbb{E}e^{\lambda Z_i} \leq \exp\left(\lambda^2 \sum_{i=1}^n \sigma_i^2\right). \tag{9}$$

По неравенству Маркова:

$$\mathbb{P}(S_n \geq t) \leq e^{-\lambda t}\mathbb{E}e^{\lambda S_n} \leq \exp\left(-\lambda t + \lambda^2 \sum \sigma_i^2\right).$$

Оптимизируем по λ на интервале $[0, 1/2K_{\max}]$: функция $f(\lambda) = -\lambda t + \lambda^2 \sum \sigma_i^2$ – выпуклая, её минимум достигается при $\lambda_0 = \frac{t}{2\sum \sigma_i^2}$, если $\lambda_0 \leq \frac{1}{2K_{\max}}$, иначе на границе. Это приводит к выражению:

$$\min_{0 \leq \lambda \leq \frac{1}{2K_{\max}}} f(\lambda) = \begin{cases} -\frac{t^2}{2\sum \sigma_i^2} & \text{при } t \leq \frac{\sum \sigma_i^2}{K_{\max}}, \\ -\frac{t}{2K_{\max}} & \text{при } t > \frac{\sum \sigma_i^2}{K_{\max}}. \end{cases}$$

В компактной форме:

$$\mathrm{P}(S_n \geq t) \leq \exp\left(-\frac{t^2}{2(\sum \sigma_i^2 + K_{\max} t)}\right).$$

Это и есть неравенство Бернштейна. Кроме того, поскольку $\mathrm{P}(S_n \leq -t)$ можно ограничить аналогичным образом, то из доказанного следует, что хвосты S_n убывают экспоненциально – то есть, S_n также является субэкспоненциальной величиной.. $\square$

9.4 Оценка максимума выборочного среднего

Рассмотрим следующую задачу. Пусть имеются независимые наблюдения (X_i, ε_i), $i = 1, \ldots, n$, где $X_i \in \mathbb{R}^p$ – субгауссовские векторы (с независимыми координатами, каждая с ограниченной субгауссовской нормой), а ε_i – центрированный субгаусссский шум, независимый от X_i. Нас интересует максимальная по координатам абсолютная величина вектора

$$Z = \frac{1}{n}\sum_{i=1}^{n} X_i \varepsilon_i \in \mathbb{R}^p,$$

то есть $\|Z\|_\infty = \max_{1 \leq j \leq p} |Z_j|$. Какова вероятность того, что $\|Z\|_\infty$ превысит заданный порог?

Для каждой фиксированной координаты j, $Z_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}\varepsilon_i$. Величины $X_{ij}\varepsilon_i$ независимы по i, центрированы и субэкспоненциальны (как произведение двух субгауссовских). Обозначим $W_{ij} = X_{ij}\varepsilon_i$. Тогда для них существуют константы $\sigma_{ij}^2 = \mathrm{Var}(W_{ij})$ и K_{ij} такие, что выполняется определение (83). Поскольку X_{ij} и ε_i имеют равномерно ограниченные субгауссовы нормы, существуют универсальные константы C_σ, C_K такие, что

$$\sigma_{ij}^2 \leq C_\sigma, \quad K_{ij} \leq C_K \quad \forall i, j. \tag{12}$$

Применим неравенство Бернштейна к сумме $S_n^{(j)} = \sum_{i=1}^{n} W_{ij}$. Тогда для $Z_j = \frac{S_n^{(j)}}{n}$:

$$\mathrm{P}\left(|Z_j| \geq t\right) = \mathrm{P}\left(|S_n^{(j)}| \geq nt\right) \leq 2\exp\left(-\frac{n^2 t^2}{2\left(\sum_{i=1}^{n} \sigma_{ij}^2 + C_K nt\right)}\right).$$

Так как $\sum_{i=1}^{n} \sigma_{ij}^2 \leq nC_\sigma$, получаем

$$\mathrm{P}\left(|Z_j| \geq t\right) \leq 2\exp\left(-\frac{n^2 t^2}{2\left(nC_\sigma + C_K nt\right)}\right) = 2\exp\left(-\frac{nt^2}{2(C_\sigma + C_K t)}\right).$$

При $t \leq C_\sigma / C_K$ знаменатель ограничен сверху, и можно записать

$$\mathrm{P}(|Z_j| \geq t) \leq 2\exp(-cnt^2)$$

для некоторой константы $c > 0$.

Теперь применим объединение по координатам (union bound):

$$\mathrm{P}(\|Z\|_\infty \geq t) \leq \sum_{j=1}^{p} \mathrm{P}(|Z_j| \geq t) \leq 2p\, e^{-cnt^2}.$$

Выберем $t = C_0\sqrt{\frac{\log p}{n}}$ с достаточно большой константой $C_0 > 0$. Тогда

$$\mathrm{P}\left(\|Z\|_\infty \geq C_0\sqrt{\frac{\log p}{n}}\right) \leq 2pe^{-cC_0^2 \log p} = 2p^{1-cC_0^2}. \tag{84}$$

Если $cC_0^2 > 1$, то правая часть стремится к нулю при $p \to \infty$. Следовательно,

$$\|Z\|_\infty = O_{\mathrm{P}}\left(\sqrt{\frac{\log p}{n}}\right).$$

При дополнительном условии $\log p = o(n)$, применение леммы Бореля–Кантелли позволяет усилить выражение до ограниченности почти наверное:

$$\|Z\|_\infty = O_{\text{п.н.}}\left(\sqrt{\frac{\log(p \vee n)}{n}}\right).$$

Пусть мы хотим гарантировать, что с вероятностью, близкой к 1, для некоторого $\lambda > 0$ выполняется

$$\|Z\|_\infty \leq \frac{\lambda}{2}.$$

Из (84) следует, что достаточно взять

$$\lambda = 2C_0\sqrt{\frac{\log(p \vee n)}{n}}, \tag{19}$$

где константа C_0 выбрана так, что $cC_0^2 > 1$. В таком случае

$$\mathrm{P}\left(\|Z\|_\infty \leq \frac{\lambda}{2}\right) \geq 1 - 2p^{1-cC_0^2} \to 1.$$

Такой выбор λ называют оптимальной скоростью для теоретического обоснования регуляризационных методов (таких как LASSO), где требуется доминирование градиентного шума. В контексте регрессий λ выступает как значение штрафа, которое контролирует сложность модели.

9.5 Метод LASSO и post-LASSO

LASSO (от англ.: least absolute shrinkage and selection operator; в свободном переводе – оператор сжатия и отбора с ограничением на абсолютные значения) – это метод оценки параметров линейной регрессии, при котором целевая функция суммы квадратов остатков дополняется штрафом, пропорциональным L^1-норме вектора параметров. Этот метод был предложен Ф. Сантозой и В. Саймсом в 1986 году[32] и позже Р. Тибширани,[33] предложившим его современное название. Метод LASSO формально решает задачи регуляризации и автоматического отбора переменных, позволяя оценивать непараметрическую компоненту регрессий, удовлетворяя типичным требованиям на скорость убывания среднеквадратичных ошибок прогноза (77). Кроме того, далее будет показано, как применение этого метода позволяет осуществить выбор и оценку модели для оценки эффекта воздействия, избегая проблемы неравномерной инференции по Либу-Потчеру. Описание

[32] Santosa F., Symes W. W. (1886). *Linear inversion of band-limited reflection seismograms.* SIAM Journal on Scientific and Statistical Computing. 7: 1307–1330.

[33] Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society. Series B (Methodological). 58: 267–288.

свойств метода LASSO основано на материалах книги Черножукова и др.[34]

Пусть процесс порождающий данные (DGP) имеет вид

$$Y_i = g(Z_i) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid Z_i) = 0, \quad \mathbb{V}(\varepsilon_i \mid Z_i) = \sigma^2.$$

где Z_i – k-мерный случайный вектор, $g(\cdot)$ – неизвестная (измеримая) функция, ε_i – ошибки модели.

Исследователю доступна случайная выборка $\{(Y_i, Z_i)\}_{i=1}^n$, и требуется построить предиктивную модель $\hat{Y}(Z)$. В рамках метода LASSO эта модель является линейной (по параметрам) и имеет вид

$$\hat{Y} = X^\top \hat{\beta},$$

где X – $n \times p$ матрица технических регрессоров, каждый столбец в которой является некоторой известной функцией от столбцов аналогичной $n \times k$ матрицы реализаций Z; идея такого преобразования заключается в аппроксимации неизвестной функции $g(\cdot)$ с помощью ряда $g(Z_i) \approx \sum_{j=1}^p \beta_{0j} X_{ij}$. Кроме того, считаем, что все вектора-столбцы матрицы X центрированы и нормированы. Размерность вектора β_0, необходимая для построения достаточно богатого функционального пространства, может быть велика, и в рамках LASSO обычно предполагается, что $p \gg n$. Такая постановка задачи благоприятствует переобучению модели, поэтому при оценке требуется регуляризация. Далее будет показано, что использование L^1 – регуляризации, являющейся отличительной особенностью LASSO, приводит к тому, что вектор оценок $\hat{\beta}$ может оказаться разреженным – то есть, отличаться от нулевого вектора лишь на небольшом числе координат. Естественно предположить, что такой метод оценивания хорошо сочетается с разреженными процессами порождения данных: пусть $\exists \beta_0 \in \mathbb{R}^p$, для которого выполняется

$$Y_i = X_i \beta_0 + u_i, \quad \operatorname{cov}(X_i, u_i) = 0, \quad \beta_{0j} \neq 0 \text{ только для } j \in A, \quad |A| = s \ll p. \tag{85}$$

Иногда вместо (85) используется более слабое предположение о приближенной разреженности.[35] В обоих случаях будем считать, что существенно отличаются от нуля лишь координаты β_0, соответствующие элементам множества $A \subset \{1, \ldots, p\}$, и $|A| = s$.

Функция потерь:

$$L(b) = \sum_{i=1}^n (Y_i - X_i^\top b)^2 + \frac{\lambda}{n} \sum_{j=1}^k |b_j| \tag{86}$$

Обозначим $Q(b) = \sum_{i=1}^n (Y_i - X_i^\top b)^2$, тогда оценка имеет вид $\hat{\beta} = \arg\min_b \left\{ Q(b) + \frac{\lambda}{n} \|b\|_1 \right\}$.

В отличие от МНК и некоторых других методов, у оценок LASSO нет аналитического решения, однако функция потерь является выпуклой (поскольку является суммой выпуклых функций), что позволяет эффективно находить оценки численными методами.[36] Кроме того, в силу выпуклости, оптимальное значение функции потерь всегда уникально.

Условие первого порядка для (86), записанное в векторно-матричной форме:

$$X^\top (Y - X\hat{\beta}) = \lambda b, \tag{87}$$

[34] Chernozhukov V., Hansen C. B., Kallus N., Spindler M., Syrgkanis V. *Causal Machine Learning*. Cambridge, MA : MIT Press, [в печати]. Доступен препринт arXiv:2206.15475.

[35] Часто используют $|\beta_0|_{(j)} \leq C j^{-a}$, $C \in \mathbb{R}$, $a > 0.5$.

[36] В случае ортонормированных регрессоров, можно записать $\hat{\beta} = S_\lambda(X^\top Y)$, где $S_\lambda(a) = \operatorname{sign}(a)(|a| - \lambda)_+$ – оператор мягкого порогового значения, применяемый к вектору $X^\top Y$ покоординатно.

где $b \in \partial\|\hat{\beta}\|_1$ – субградиент L^1-нормы в точке $\hat{\beta}$. Координаты вектора b имеют следующий вид:

$$b_j \in \begin{cases} \{+1\} & \hat{\beta}_j > 0 \\ \{-1\} & \hat{\beta}_j < 0 \\ [-1,1] & \hat{\beta}_j = 0 \end{cases} \quad j \in \{1,\dots,p\}. \tag{88}$$

Из (89) и (91) получаем, что оптимальный субградиент b всегда уникально определяется значениями $\hat{\beta}$; в частности – в ситуации, когда модель LASSO имеет неуникальное решение, знаки ненулевых координат в оценках должны совпадать.

Для анализа свойств метода LASSO, покажем разреженность оценок LASSO и построим ограничение на разность между прогнозом модели и оптимальным прогнозом $X\beta_0$. В силу оптимальности $\hat{\beta}$, выполняется

$$L(\hat{\beta}) \le L(\beta_0) \Rightarrow Q(\hat{\beta}) + \frac{\lambda}{n}\|\hat{\beta}\|_1 \le Q(\beta_0) + \frac{\lambda}{n}\|\beta_0\|_1 \Rightarrow Q(\hat{\beta}) - Q(\beta_0) \le \frac{\lambda}{n}(\|\beta_0\|_1 - \|\hat{\beta}\|_1). \tag{89}$$

Обозначим $v = \hat{\beta} - \beta_0$ и $\nabla Q(b) = -2\sum_{i=1}^n X_i(Y_i - X_i^\top b)$ – градиент функции $Q(\cdot)$ в точке b. Эта функция является выпуклой, что даёт возможность получить ограничение сверху на $\|v\|_1$:

$$Q(\hat{\beta}) - Q(\beta_0) \ge \langle \nabla Q(\beta_0), \hat{\beta} - \beta_0 \rangle = -\langle -\nabla Q(\beta_0), v\rangle \ge -\| - \nabla Q(\beta_0)\|_\infty \|v\|_1. \tag{90}$$

Гиперпараметр λ выбирается так, чтобы обеспечить $\mathrm{P}\big(\frac{\lambda}{n} \ge 2\| - \nabla Q(\beta_0)\|_\infty\big) \ge 1 - a$. В простых случаях теоретическую оценку можно построить, используя неравенства концентрации меры. Так, в ряде работ[37] предлагается использовать $\lambda = \sigma\sqrt{n \ln p}$, в работе Беллони и Черножукова[38] λ пропорционален $(1-a)$-квантилю бутстрап-распределения нормы градиента среднего значения квадратов остатков $\| - \frac{2}{n}\nabla Q(\hat{\beta}^*)\|_\infty$, а в прикладных исследованиях λ, как правило, выбирают, максимизируя качество прогнозов модели с использованием кросс-валидации,

В таком случае, с высокой вероятностью выполняется

$$\frac{\lambda}{n}(\|\beta_0\|_1 - \|\hat{\beta}\|_1) \ge -\| - \nabla Q(\beta_0)\|_\infty \|v\|_1 \ge -\frac{\lambda}{2n}\|v\|_1 \Rightarrow \|\hat{\beta}\|_1 \le \|\beta_0\|_1 + \frac{1}{2}\|v\|_1 \tag{91}$$

Перепишем (91), выделяя из рассматриваемых векторов подвектора с индексами из A и $\overline{A}$, и используя $\beta_{0i} = 0$ при $i \in \overline{A}$:

$$\|\hat{\beta}\|_1 = \|v + \beta_0\|_1 = \|v^A + \beta_0^A\|_1 + \|v^{\overline{A}}\|_1 \le \|\beta_0^A\|_1 + \frac{1}{2}\|v^A\|_1 + \frac{1}{2}\|v^{\overline{A}}\|_1 \Rightarrow$$

$$\Rightarrow \frac{1}{2}\|v^{\overline{A}}\|_1 \le \frac{1}{2}\|v^A\|_1 + \|\beta_0^A\|_1 - \|v^A + \beta_0^A\|_1 \le \frac{3}{2}\|v^A\|_1,$$

где в последнем переходе использовалось неравенство треугольника

$$\|\beta_0^A\|_1 - \|v^A + \beta_0^A\|_1 \le \|\beta_0^A - v^A + \beta_0^A\|_1 = \|v^A\|_1.$$

Из доказанного следует $\|v^{\overline{A}}\|_1 \le 3\|v^A\|_1$, что соответствует тому, что координаты вектора $v = \hat{\beta} - \beta_0$, отличающиеся от нуля, относятся преимущественно к множеству A.

[37] Bickel P. J., Ritov Y., Tsybakov A. B. (2009). *Simultaneous analysis of Lasso and Dantzig selector.* Annals of Statistics. 37: 1705-1732, а также Greenshtein E., Ritov Y. (2004). *Persistence in high-dimensional linear predictor selection and the virtue of overparametrization.* Bernoulli. 10: 971-988.

[38] Belloni A., Chernozhukov V. (2013). *Least squares after model selection in high-dimensional sparse models.* Bernoulli. 19: 521-547.

Для анализа ошибок прогноза (RMSE) зависимой переменной в методе LASSO требуется введение некоторых условий регулярности на ковариационную матрицу регрессоров: предположим, что равномерно по $Z \subset X : \dim(Z) \leq s \log n$ выполняется

$$\lim_{n\to\infty} \sup_{\|a\|=1} \left| a^\top \left(\mathbb{E}_n \left[ZZ^\top \right] - \mathbb{E} \left[ZZ^\top \right] \right) a \right| = 0, \tag{92}$$

$$0 < C_1 \leq \inf_{\|a\|=1} a^\top \mathbb{E} \left[ZZ^\top \right] a \leq C_2 < \infty, \tag{93}$$

где C_1 и C_2 – некоторые константы, $\mathbb{E}_n$ – среднее по случайной выборке размером n.

Далее, пусть $\tilde{X}$ – новый набор регрессоров; вычислим среднее значение RMSE в этой точке при фиксированной обучающей выборке:

$$RMSE = \sqrt{\mathbb{E}[(\tilde{X}^\top \hat{\beta} - \tilde{X}^\top \beta_0)^2 \mid X]} = \sqrt{\tilde{\mathbb{E}}[(\hat{\beta} - \beta_0)^\top \tilde{X}\tilde{X}^\top (\hat{\beta} - \beta_0) \mid X]} = \sqrt{v^\top \mathbb{E}(\tilde{X}\tilde{X}^\top) v} \leq C_2 \|v\|_2. \tag{94}$$

Функция $Q(b)$ является квадратичной, поэтому точно равна своему ряду Тейлора

$$Q(\hat{\beta}) - Q(\beta_0) = \langle \nabla Q(\beta_0), v \rangle + v^\top [\mathbb{E}_n(XX^\top)] v \geq -\| - \nabla Q(\beta_0)\|_\infty \|v\|_1 + \hat{C}_1 \|v\|_2^2, \tag{95}$$

где $\hat{C}_1$ – некоторая константа, ограничивающая минимальный собственный вектор матрицы $\mathbb{E}_n(XX^\top)$; её существование следует из (92) и (93) при всех достаточно больших n.

Из неравенства (89) можно получить (применяя неравенство треугольника) $\frac{\lambda}{n}\|v\|_1 \geq Q(\hat{\beta}) - Q(\beta_0)$, что вместе с (95) и условием $\frac{\lambda}{n} \geq \|\nabla Q(\beta_0)\|_\infty$ даёт (с вероятностью 1-a) неравенство

$$\frac{\lambda}{n}\|v\|_1 \geq Q(\hat{\beta}) - Q(\beta_0) \geq -\frac{\lambda}{2n}\|v\|_1 + \hat{C}_1 \|v\|_2^2 \Rightarrow \|v\|_2^2 \leq \frac{3\lambda}{2\hat{C}_1 n}\|v\|_1.$$

Далее, поскольку вектор v имеет преимущественно нулевые координаты вне A, то его L^2 и L^1 нормы должны отличаться примерно в $\sqrt{s}$ раз:

$$\|v\|_1 = \left\|v^A\right\|_1 + \left\|v^{\overline{A}}\right\|_1 \leq 4 \left\|v^A\right\|_1 \leq 4\sqrt{s} \left\|v^A\right\|_2 \leq 4\sqrt{s}\|v\|_2, \tag{96}$$

где использовалось неравенство Коши-Буняковского $\|v^A\|_1 = \langle |v^A|, 1_s \rangle \leq \|1_s\|_2 \|v^A\|_2 = \sqrt{s}\|v^A\|_2$.

Отсюда получаем неравенство на вторую норму вектора v:

$$\|v\|_2 \leq \frac{6\lambda}{\hat{C}_1 n}\sqrt{s} \Rightarrow RMSE \leq \frac{6\lambda C_2}{\hat{C}_1 n}\sqrt{s} \approx C\sqrt{\frac{s \ln p}{n}}, \tag{97}$$

что, при достаточно медленно растущих по n значениях s и p обеспечивает скорость сходимости к нулю среднеквадратичного отклонения прогноза от оптимума, достаточную для применения в методе Робинсона, (77).

Помимо построения прогнозов, метод LASSO можно применять и для выбора модели. В этом случае из исходного набора технических регрессоров отбираются те, для которых оценка оказалась ненулевой. В одной из разновидностей метода – post-LASSO, эти регрессоры далее используются для оценки линейной модели с помощью МНК.

Теорема 9.5. *В условиях предыдущей модели выбор состоятелен, то есть* $\mathrm{P}(\hat{A} \supseteq A) \to 1$, *где* $\hat{A}$ *– набор (индексов) регрессоров, выбранных LASSO, а* A *– истинный носитель* β_0 *(или его разреженная аппрокси-*

мация). Кроме того, среднеквадратичное отклонение прогноза post-LASSO относительно оракула на новом наблюдении имеет асимптотику $\frac{s}{n} + o\left(\frac{s}{n}\right)$.

Доказательство. Обозначим $\delta = \hat{\beta} - \beta^0$. Из уравнений (89), (91) и выражения для оптимального λ следует, что с высокой вероятностью выполняется

$$\|v\|_1 \leq \frac{24s\sigma\sqrt{\ln p}}{\hat{C}_1\sqrt{n}}. \tag{98}$$

Для любого $j \in A$

$$|\hat{\beta}_j| = |\beta_j^0 + v_j| \geq |\beta_j^0| - |v_j| \geq \beta_{\min} - |v_j|, \quad \text{где } \beta_{\min} = \min_{i \in A} \beta_0.$$

Так как $|v_j| \leq \|v\|_1$, получаем $|\hat{\beta}_j| \geq \beta_{\min} - \|v\|_1$, следовательно, если $\|v\|_1 < \beta_{\min}$, то для всех $j \in A$ $|\hat{\beta}_j| > 0$, то есть $A \subseteq \hat{A}$. Поскольку LASSO с оптимально выбранным λ гарантирует $\|v\|_1 \xrightarrow[n\to\infty]{} 0$, то при всех достаточно высоких n с произвольно высокой вероятностью $A \subset \hat{A}$.

Далее, обозначим за $\tilde{X}$ $p \times 1$ вектор новых наблюдений, независимый от обучающей выборки. Нижним индексом A будем обозначать переход к подвектору/подматрице с индексами из A. В частности, X_A – $n \times s$ подматрица регрессоров со столбцами из A,[39] $\Sigma_{AA} := \mathbb{E}[\tilde{X}_A \tilde{X}_A^\top]$ – ковариационная матрица размерностью $s \times s$, а $\hat{\beta}_A$ – подвектор post-LASSO оценки.

Условно на обучающей выборке, то есть на X и Y, вектор $\delta := \hat{\beta}_A - \beta_{0,A}$ фиксирован. Тогда среднее значение ошибки прогноза имеет вид

$$\mathbb{E}r_{\tilde{Y}}^2 := \mathbb{E}\big(\tilde{X}^\top(\hat{\beta} - \beta_0)\big)^2 = \mathbb{E}\left[\mathbb{E}\big[(\tilde{X}_A^\top \delta)^2 \mid X, Y\big]\right] = \mathbb{E}[\delta^\top \Sigma_{AA} \delta].$$

Преобразование через след:

$$\mathbb{E}\left[\delta^\top \Sigma_{AA} \delta\right] = \mathbb{E}\left[\operatorname{tr}\left(\Sigma_{AA}\delta\delta^\top\right)\right] = \operatorname{tr}\left(\Sigma_{AA}\mathbb{E}[\delta\delta^\top]\right).$$

Учитывая, что $\delta = (X_A^\top X_A)^{-1} X_A^\top \varepsilon$, при гомоскедастичных ошибках $\mathbb{E}[\delta\delta^\top \mid X_A] = \sigma^2 (X_A^\top X_A)^{-1}$, следовательно

$$\mathbb{E}r_{\tilde{Y}}^2 = \sigma^2 \mathbb{E} \operatorname{tr}\left(\Sigma_{AA}(X_A^\top X_A)^{-1}\right).$$

В силу ЗБЧ $\operatorname{tr}\big(\Sigma_{AA} n (X_A^\top X_A)^{-1}\big) \xrightarrow{p} s$, поэтому

$$\operatorname{tr}\left(\Sigma_{AA}(X_A^\top X_A)^{-1}\right) = \frac{s}{n} + o_p\left(\frac{s}{n}\right).$$

Если LASSO выбрает дополнительно к истинным переменным r – ложных, то при $r = o(s)$ скорость остаётся оракульной:

$$\mathbb{E}r_{\tilde{Y}}^2 = \frac{\sigma^2 s}{n} + o\left(\frac{s}{n}\right).$$

9.6 Метод двойного LASSO

В работе Беллони и др.[40] при исследовании задачи построения оптимального инструмента, был предложен метод двойного выбора/двойного LASSO для нахождения асимптотически нормальных оценок структурных

[39] Приём с занулением столбцов здесь неудобен, так как далее потребуется обращать матрицу Грама $X_A^\top X_A$.

[40] Belloni, A., Chen D., Chernozhukov, V., Hansen, C. (2012). *Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain.* Econometrica. 80: 2369-2429.

параметров моделей. Особенностью предложенного метода является отбор с помощью процедуры LASSO переменных как для зависимой, так и для объясняющих переменных. Авторы обнаружили, что предложенный метод близок по своей идее к процедуре ортогонализации Неймана, и в более поздних работах предложили обобщение – метод двойного машинного обучения, который позволяет использовать разнообразные методы машинного обучения для построения вспомогательных прогнозных значений внутри каузальной модели.

Модель имеет следующий вид:

$$Y_i = \theta_0 D_i + X_i^\top \beta_0 + \varepsilon_i,$$

где D_i – эндогенная переменная воздействия, для которой имеется большой набор инструментов Z_i (потенциально $p_z \gg n$), и оптимальный инструмент – $D_i^* = \mathbb{E}[D_i \mid Z_i, X_i]$. Ошибка основной модели обладает свойством $\mathbb{E}[\varepsilon_i \mid Z_i, X_i] = 0$.

Алгоритм имеет следующий вид:

- LASSO-регрессия D на (X, Z) даёт набор оценок $\hat{\pi}$ и предсказанный оптимальный инструмент

$$\hat{D}_i = X_i^\top \hat{\pi}_x + Z_i^\top \hat{\pi}_z.$$

- Двойной отбор контролей: с помощью LASSO-регрессии Y на $(\hat{D}, X)$ отбирается множество регрессоров с индексами $\hat{A}_y$, и также LASSO-регрессией D на X отбираются регрессоры с индексами $\hat{A}_d$, после чего эти наборы объединяются: $\hat{A} = \hat{A}_y \cup \hat{A}_d$. Обозначим за $X_{\hat{A}}$ подматрицу X с столбцами с индексами из $\hat{A}$.

- Строится 2SLS-оценка, где D инструментируется с помощью $\hat{D}$, а в качестве контрольных переменных используются $X_{\hat{A}}$:

$$\tilde{\theta} = (\hat{D}^\top Q_{\hat{A}} D)^{-1} \hat{D}^\top Q_{\hat{A}} Y, \tag{99}$$

  где $\hat{D} = (\hat{D}_1, \dots, \hat{D}_n)^\top$, $Q_{\hat{A}} = I - X_{\hat{A}}(X_{\hat{A}}^\top X_{\hat{A}})^{-1} X_{\hat{A}}^\top$.

Теорема 9.6. *В представленном алгоритме двойного LASSO при условиях регулярности оценка θ является асимптотически нормальной.*

Доказательство.

$$\sqrt{n}(\tilde{\theta} - \theta_0) = \Big(\frac{1}{n}\hat{D}^\top Q_{\hat{A}} D\Big)^{-1} \frac{1}{\sqrt{n}} \hat{D}^\top Q_{\hat{A}} \varepsilon, \tag{100}$$

поскольку на событии $\hat{A} \supseteq A$ выполнено $Q_{\hat{A}} X \beta_0 = 0$.

Удобно представить $\hat{D}$ как $D^* + \hat{V}$, где $D^* = X\gamma_0 + Z\pi_0$ – истинный оптимальный инструмент, $\hat{V} = Z(\hat{\pi} - \pi_0) + X(\hat{\pi}_x - \gamma_0)$ – ошибка предсказания LASSO.

Для знаменателя: $\frac{1}{n}\hat{D}^\top Q_{\hat{A}} D \xrightarrow{p} B = \mathbb{E}[\tilde{D}_i^{\perp}]^2$, где $\tilde{D}_i^{\perp}$ – остатки от регрессии истинного оптимального инструмента D_i^* на $X_{\hat{A}}$. Компоненты, содержащие $\hat{V}$, по свойствам LASSO имеют порядок $o_p(1)$, поэтому зануляются в пределе.

Для числителя: $\frac{1}{\sqrt{n}}\hat{D}^\top Q_{\hat{A}} \varepsilon = \frac{1}{\sqrt{n}}(D^*)^\top Q_{\hat{A}} \varepsilon + \frac{1}{\sqrt{n}}\hat{V}^\top Q_{\hat{A}} \varepsilon$. Второе слагаемое раскладывается как

$$(\hat{\pi} - \pi_0)^\top \frac{Z^\top Q_{\hat{A}} \varepsilon}{\sqrt{n}} + \text{аналог для } x.$$

Используя неравенство Хёфдинга/субгауссовость, $\|Z^\top Q_{\hat{A}} \varepsilon / \sqrt{n}\|_\infty = O_p(\sqrt{\log(p_z + p)})$, и $\|\hat{\pi} - \pi_0\|_1 = O_p(s_\pi \sqrt{\log(p_z + p)/n})$, получаем, что это слагаемое есть $O_p(s_\pi \log(p_z + p)/\sqrt{n}) = o_p(1)$ при условии $s_\pi \log(p_z + p) = o(\sqrt{n})$.

Таким образом,

$$\sqrt{n}(\tilde{\theta}-\theta_0) = B^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{D}_i^{\perp}\varepsilon_i + o_p(1),$$

где $\tilde{D}_i^{\perp}\varepsilon_i$ – *н.о.р.* слагаемые с нулевым средним и дисперсией $\mathbb{E}[(\tilde{D}_i^{\perp})^2\varepsilon_i^2]$. По центральной предельной теореме получаем

$$\sqrt{n}(\tilde{\theta}-\theta_0) \xrightarrow{d} \mathcal{N}\left(0,\ B^{-2}\mathbb{E}[(\tilde{D}_i^{\perp})^2\varepsilon_i^2]\right).$$

При гомоскедастичности $\mathbb{E}[\varepsilon_i^2 \mid Z_i, X_i] = \sigma^2$ дисперсия упрощается до $\sigma^2 B^{-1}$. Состоятельные оценки $\hat{B}$ и $\hat{\sigma}^2$ получаются подстановкой: $\hat{B} = \frac{1}{n}\hat{D}^{\top}Q_{\hat{A}}\hat{D}$, $\hat{\sigma}^2 = \frac{1}{n-|\hat{A}|-1}\sum\hat{\varepsilon}_i^2$, где $\hat{\varepsilon}_i$ – остатки на последнем этапе процедуры. □

Связь представленного метода с идеей ортогонализации моментного условия можно представить следующим образом: пусть $\eta^0 := \left(\beta_{YX}, \gamma_{DX}^{\top}\right)^{\top}$ – вектор популяционных значений коэффициентов при X в LASSO-моделях Y на $\hat{D}, X$ и D на X соответственно, являющийся мешающим параметром, а $\eta := \left(\hat{\beta}_{YX}, \hat{\gamma}_{DX}^{\top}\right)^{\top}$ – его значение в наблюдаемой выборке, которое можно использовать для выражения $\hat{\theta} = \hat{\theta}(\eta)$:

$$\check{Y}(\eta) = Y - X^{\top}\eta_1, \quad \check{D}(\eta) = D - X^{\top}\eta_2.$$

Оценку $\hat{\theta}(\eta)$ можно выразить, как решение выборочного моментного условия

$$\hat{M}(t,\eta) := \mathbb{E}_n\big[\big(\check{Y}(\eta) - t\check{D}(\eta)\big)\check{D}(\eta)\big] = 0,$$

соответствующего популяционному условию

$$M(t,\eta) := \mathbb{E}\big[\big(\check{Y}(\eta) - t\check{D}(\eta)\big)\check{D}(\eta)\big] = 0, \tag{101}$$

задающему неявную функцию $\theta(\eta)$. Основная идея метода двойного выбора заключается в том, что в популяционном пределе алгоритм обучения $\theta(\eta)$ нечувствителен в первом порядке к отклонениям значений мешающего параметра от своего истинного значения η^0. Это связано с тем, что $\frac{\partial}{\partial\eta}\theta(\eta^0) = 0$, поскольку

$$\frac{\partial}{\partial\eta}t\big(\eta^0\big) = -\frac{\partial}{\partial t}M\big(t,\eta^0\big)^{-1}\frac{\partial}{\partial\eta}M\big(t,\eta^0\big),$$

причём

$$\frac{\partial}{\partial\eta}M\big(\theta,\eta^0\big) = \Big(\frac{\partial}{\partial\eta_1}M\big(\theta,\eta^0\big), \frac{\partial}{\partial\eta_2}M\big(\theta,\eta^0\big)\Big),$$

где

$$\frac{\partial}{\partial\eta_1}M\big(\theta,\eta^0\big) = -\mathbb{E}\big[X\check{D}\big(\eta^0\big)\big] = -\mathbb{E}\big[X\big(D - X^{\top}\gamma_{DX}\big)\big] = 0,$$

поскольку предсказания методом LASSO сходятся в пределе к наилучшему, в среднеквадратичном смысле, прогнозу.

Аналогично,

$$\frac{\partial}{\partial\eta_2}M\left(\theta,\eta^0\right) = -\mathbb{E}\left[X\check{Y}(\eta^0)\right] + 2\mathbb{E}\left[\theta X\check{D}(\eta^0)\right] = 0,$$

поскольку в первой компоненте математическое ожидание берется от произведения X на ортогональную к X величину, а вторая компонента аналогична выражению, рассмотренному в предыдущем случае.

9.7 Метод двойного машинного обучения

Интересующий исследователя параметр $\theta_0 \in \mathbb{R}$, определяется моментным условием

$$\mathbb{E}[\psi(W; \theta_0, \eta_0)] = 0,$$

где W – наблюдаемые данные, η – (возможно, бесконечномерные) мешающие параметры, а ψ – ортогональная по Нейману скор-функция:

$$\partial_\eta \mathbb{E}[\psi(W; \theta_0, \eta_0)](\eta - \eta_0) = 0, \qquad \forall \eta. \tag{102}$$

Алгоритм оценки в двойном машинном обучении:

- Разбиваем выборку размера n на K блоков;
- Для каждого блока k на данных вне блока оцениваем $\hat{\eta}_k$ с помощью произвольного ML-алгоритма (ядерная регрессия, LASSO, случайные леса, нейросети и т. д.);
- На k-м блоке вычисляем оценку $\hat{\theta}$ как решение

$$\frac{1}{|I_k|} \sum_{i \in I_k} \psi(W_i; \hat{\theta}, \hat{\eta}_k) = 0.$$

- Финальная DML-оценка $\tilde{\theta}$ получается усреднением решений или (чаще) как решение единого выборочного уравнения с усреднённой скор-функцией (DML2):

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \psi(W_i; \tilde{\theta}, \hat{\eta}_k) = 0. \tag{103}$$

Упрощающие предположения метода:

- Данные (W_i) независимы и одинаково распределены;
- θ_0 — внутренняя точка компактного множества;
- ψ дважды непрерывно дифференцируема по θ в окрестности θ_0 с равномерно ограниченными производными. Значение $J_0 = \mathbb{E}[\partial_\theta \psi(W; \theta_0, \eta_0)]$ не равно нулю;
- Выполнено условие ортогональности по Нейману (102);
- Оценки мешающих параметров сходятся с достаточной скоростью: $\|\hat{\eta}_k - \eta_0\|_{P,2} = o_p(n^{-1/4})$, где $\|\cdot\|_{P,2}$ – $L_2(P)$-норма.

Теорема 9.7. *Оценка, получаемая представленным алгоритмом при заданных предположениях является асимптотически нормальной.*

Доказательство.
Разложим уравнение (103) в точке θ_0:

$$0 = \frac{1}{n} \sum_{k,i} \psi(W_i; \theta_0, \eta_0) + \frac{1}{n} \sum_{k,i} \big(\psi(W_i; \tilde{\theta}, \hat{\eta}_k) - \psi(W_i; \theta_0, \eta_0)\big).$$

Второе слагаемое линеаризуем:

$$\psi(W;\tilde\theta,\hat\eta)-\psi(W;\theta_0,\eta_0)=\underbrace{\partial_\theta\psi(W;\theta_0,\eta_0)(\tilde\theta-\theta_0)}_{a}+\underbrace{\big(\psi(W;\theta_0,\hat\eta)-\psi(W;\theta_0,\eta_0)\big)}_{b}+R,$$

где R включает члены второго порядка по $(\tilde\theta-\theta_0)$ и $(\hat\eta-\eta_0)$. Благодаря ортогональности (102) линейный член b имеет нулевое среднее при усреднении по W, и в кросс-фит схеме его вклад в сумму имеет порядок $O_p(\|\hat\eta-\eta_0\|_2^2)$, а не $O_p(\|\hat\eta-\eta_0\|_2)$. При нашем условии на скорость сходимости $o_p(n^{-\frac14})$ получаем

$$\frac{1}{\sqrt n}\sum[\psi(W;\theta_0,\hat\eta)-\psi(W;\theta_0,\eta_0)]=\sqrt n\,O_p(\|\hat\eta-\eta_0\|_{P,2}^2)=o_p(1).$$

То же условие скрывает квадратичные остатки. В итоге

$$\sqrt n(\tilde\theta-\theta_0)=-J_0^{-1}\frac{1}{\sqrt n}\sum_{i=1}^n\psi(W_i;\theta_0,\eta_0)+o_p(1).$$

По центральной предельной теореме Леви,

$$\frac{1}{\sqrt n}\sum_{i=1}^n\psi(W_i;\theta_0,\eta_0)\xrightarrow{d}\mathcal N\big(0,\ \Omega_0\big),\quad \Omega_0=\mathbb E[\psi(W;\theta_0,\eta_0)\psi(W;\theta_0,\eta_0)^\top].$$

Следовательно,

$$\sqrt n(\tilde\theta-\theta_0)\xrightarrow{d}\mathcal N\big(0,\ (J_0^{-1})^2\Omega_0\big).$$

Оценка дисперсии строится подстановкой: по найденным значениям $\tilde\theta$, $\hat\eta_k$, вычисляются

$$\hat J=\frac1n\sum_{k,i}\partial_\theta\psi(W_i;\tilde\theta,\hat\eta_k)\quad\text{и}\quad\hat\Omega=\frac1n\sum_{k,i}\psi(W_i;\tilde\theta,\hat\eta_k)\psi(W_i;\tilde\theta,\hat\eta_k)^\top,$$

и используется аналог сэндвич-формулы. Доверительные интервалы строятся стандартно на основе нормальной аппроксимации.

9.8 Заключение

Метод двойного LASSO и более общая парадигма двойного машинного обучения (DML) представляют собой значительный шаг вперёд в прикладном инструментарии, позволяя решать задачи причинного вывода в условиях высокой размерности данных. В отличие от классических подходов, DML сводит задачу оценки эффекта к двум задачам прогнозирования и позволяет применять гибкие ML-алгоритмы, сходящиеся со скоростью медленнее $n^{-\frac12}$, сохраняя при этом $\sqrt n$-состоятельность и асимптотическую нормальность итоговых оценок. Важно подчеркнуть, что конструкция DML опирается прежде всего на ортогональные моментные условия и кросс-фиттинг, а не на конкретный оценивающий алгоритм. Поэтому итоговый структурный параметр можно получать не только через линейную МНК-регрессию остатков, но и с помощью GMM, M-оценивания, инструментальных переменных, квантильной регрессии или других методов. Аналогично, на первом этапе допустимо использовать любые ML-методы – LASSO, ridge-регрессию, случайный лес, градиентный бустинг, нейросети или логистические модели – при условии достаточной скорости сходимости оценок мешающих параметров.

В эмпирической практике остаётся открытым вопрос выбора конкретного алгоритма машинного обучения для оценки мешающих параметров, поскольку различные модели могут давать заметный разброс в прогнозах

на конечных выборках. В духе современных прикладных исследований, оптимальный выбор первичной ML-модели следует осуществлять на основе её ошибки прогнозирования. Поскольку асимптотические свойства DML опираются на качество аппроксимации функций, сравнивать конкурирующие алгоритмы необходимо по вневыборочной среднеквадратичной ошибке, вычисленной на кросс-валидационных фолдах. Алгоритм, демонстрирующий наименьшую ошибку прогноза для мешающих параметров, обеспечивает наиболее точное удаление шума и, как следствие, наиболее надёжен при оценке структурных параметров.

10 Приложение

Теорема 10.1. *(Теорема Александрова о характеризации сходимости по распределению). Пусть $\mathcal{V}$ – топологическое векторное пространство с борелевской сигма-алгеброй $\mathcal{B}$ (порождённой всеми открытыми множествами $\mathcal{V}$). Пусть $(\Omega, \mathcal{F}, \mu)$ – вероятностное пространство с мерой μ, заданной на сигма-алгебре $\mathcal{F}$ измеримых подмножеств Ω. Случайный элемент в $\mathcal{V}$ – это измеримая функция*

$$\mathrm{Y} : \Omega \to \mathcal{V}, \quad Y^{-1}\mathcal{B} \subset \mathcal{F}.$$

Случайный элемент Y задаёт вероятностную меру P на $\mathcal{V}$:

$$\forall B \in \mathcal{B},\ P(B) = \mu\{\omega \in \Omega : Y(\omega) \in B\}.$$

Последовательность $\{Y_n; n \geq 1\}$ случайных элементов на $\mathcal{V}$, сходящуюся по распределению к Y, будем обозначать следующим образом: $Y_n \xrightarrow{d} Y$, и, как обычно, это означает, что соответствующие вероятностные меры P_n на $\mathcal{V}$ сходятся слабо к P – мере, соответствующей Y. Слабая сходимость мер P_n к P обозначает, что для любого ограниченного и непрерывного функционала f, заданного на $\mathcal{V}$, выполняется

$$\lim_{n\to\infty} \int_{\mathcal{V}} f\, dP_n = \int_{\mathcal{V}} f\, dP.$$[41]

Следующая теорема, предложенная А.Д. Александровым, характеризует слабую сходимость мер. Пусть P и $\{P_n, n \in \mathbb{N}\}$ – вероятностные меры на $\mathcal{B}(\mathcal{V})$. Следующие пять условий эквивалентны:

(i) *P_n слабо сходится к мере P;*

(ii) *$\int_{\mathcal{V}} f\, dP_n \xrightarrow[n\to\infty]{} \int_{\mathcal{V}} f\, dP$ для любого ограниченного и равномерно непрерывного функционала $f : \mathcal{V} \to \mathbb{R}$;*

(iii) *Для каждого замкнутого множества $A \in \mathcal{B}$ справедливо соотношение*

$$\limsup_{n\to\infty} P_n(A) \leq P(A);$$

(iv) *Для каждого открытого множества $G \in \mathcal{B}$ справедливо соотношение*

$$\liminf_{n\to\infty} P_n(G) \geq P(G);$$

(v) *Для любого множества $A \in \mathcal{B}$, такого, что $P(\partial A) = 0$ (здесь ∂A – граница множества A), справедливо соотношение $\lim_{n\to\infty} P_n(A) = P(A)$.*

Доказательство. Импликация (i)⇒(ii) очевидна. Докажем импликацию (ii)⇒(iii). Для этого возьмём произвольное $A \in \mathcal{B}$, $\varepsilon > 0$ и построим (равномерно) непрерывный функционал $f_\varepsilon : \mathcal{V} \to \mathbb{R}$, аппроксимирующий индикатор $\mathbb{1}_A$, рис. 15.

$$f_\varepsilon(x) = \max\left\{0, 1 - \frac{\mathrm{dist}(x, A)}{\varepsilon}\right\}, \quad \mathrm{dist}(x, A) = \inf_{y\in A} \|x - y\|.$$

[41] Если $\mathcal{V} = \mathbb{R}$, то Y_n – случайные величины, и представленная запись соответствует сходимости $\mathbb{E}f(Y_n)$ к $\mathbb{E}f(Y)$.

Обозначим $A_\varepsilon = \{x \in \mathcal{V} : \operatorname{dist}(x, A) \le \varepsilon\}$. Тогда $\forall x \in \mathcal{V}$ очевидно неравенство

$$\mathbb{1}_A(x) \le f_\varepsilon(x) \le \mathbb{1}_{A_\varepsilon}(x) \Rightarrow \int_{\mathcal{V}} \mathbb{1}_A \, dP_n \le \int_{\mathcal{V}} f_\varepsilon \, dP_n \le \int_{\mathcal{V}} \mathbb{1}_{A_\varepsilon} \, dP_n. \tag{104}$$

Из (104) и слабой сходимости P_n следует $\limsup_{n\to\infty} P_n(A) \le \limsup_{n\to\infty} \int_{\mathcal{V}} f_\varepsilon \, dP_n = \int_{\mathcal{V}} f_\varepsilon \, dP$. Также, переходя от неравенства функций в (104) к интегралам по P, получаем

$$P(A) \le \int_{\mathcal{V}} f_\varepsilon \, dP \le P(A_\varepsilon).$$

Если A – замкнутое, то из непрерывности вероятностной меры следует, что $\lim_{\varepsilon \downarrow 0} P(A_\varepsilon) = P(A)$, а значит, такой же предел имеет и $\int_{\mathcal{V}} f_\varepsilon \, dP$. Тогда, поскольку неравенство $\limsup_{n\to\infty} P_n(A) \le \int_{\mathcal{V}} f_\varepsilon \, dP$ выполняется для любого $\varepsilon > 0$, то переходя к пределу с $\varepsilon \downarrow 0$, получаем $\limsup_{n\to\infty} P_n(A) \le P(A)$ для замкнутых множеств.

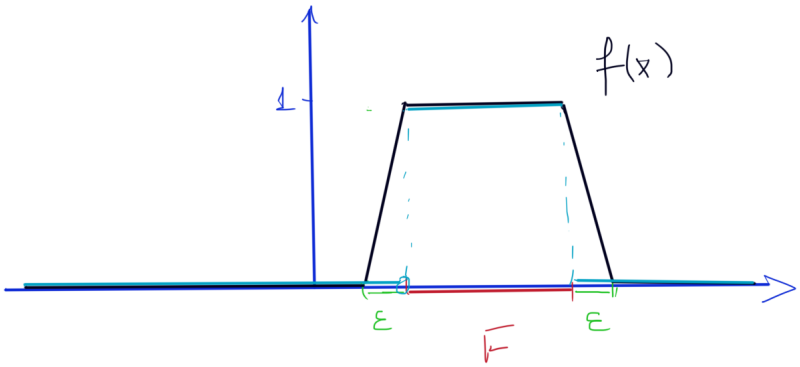


Рис. 15: Индикаторы множеств A и A_ε, а также аппроксимирующая их непрерывная функция $f_\varepsilon(x)$.

Эквивалентность (iii) и (iv) легко получить, переходя к дополнениям замкнутых множеств и учитывая, что $\limsup_{n\to\infty}\big(1 - P_n(A)\big) = 1 - \liminf_{n\to\infty} P_n(A)$ применяя закон де Моргана. Далее, докажем, что из (iii) и (iv) следует свойство (v). Если A^- и A^+ – внутренность и замыкание множества A соответственно, то из условий (iii) и (iv) получаем

$$PA^+ \ge \limsup_{n\to\infty} P_n A^+ \ge \limsup_{n\to\infty} P_n A \ge \liminf_{n\to\infty} P_n A \ge \liminf_{n\to\infty} P_n A^- \ge PA^-. \tag{105}$$

Если $P(\partial A) = 0$, то $P(A^-) = P(A^+)$, и из (105) следует $\lim_{n\to\infty} P_n(A) = P(A)$.

Докажем импликацию (v)⇒(i). В силу линейности интегралов по мере, можно без потери общности считать, что ограниченные функционалы $g : \mathcal{V} \to \mathbb{R}$ удовлетворяют условию $0 < g < 1$. Тогда по теореме Фубини

$$\int_{\mathcal{V}} g \, dP = \int_{\mathcal{V}} \int_0^1 \mathbb{1}_{\{g(x)>t\}}(x,t) \, dt dP(x) = \int_0^1 \int_{\mathcal{V}} \mathbb{1}_{\{g(x)>t\}}(x,t) \, dP(x) dt = \int_0^1 P\{g(x) > t\} \, dt,$$

и аналогично можно выразить интеграл g по P_n. Мы хотим доказать, что в заданных условиях происходит

$$\int_0^1 P_n\{g(x) > t\} \, dt \xrightarrow[n\to\infty]{} \int_0^1 P\{g(x) > t\} \, dt. \tag{106}$$

Разобьем $[0,1]$ на T и T', где $\forall t \in T \; P\big(\partial\{g(x) > t\}\big) = 0$ и $\forall t \in T' \; P\big(\partial\{g(x) > t\}\big) \ne 0$. Для непрерывного g выполняется $\partial\{g(x) > t\} \subset \{g(x) = t\}$, так как $\{g(x) < t\}$ и $\{g(x) > t\}$ являются открытыми множествами – а значит содержат любой свой элемент вместе с некоторой открытой окрестностью. Отсюда следует, что

P-мера границы множества $\{g(x) > t\}$ равна нулю, возможно за исключением счётного набора t – поскольку должно выполняться условие нормировки $\sum_t P\{g(x) = t\} \leq 1$, а значит, во-первых, T и T' измеримы по Лебегу, а во-вторых, интегралы ограниченных функций по T' равны нулю. Из предположения (v) следует сходимость $P_n\big(\{g(x) > t\}\big)$ к $P\big(\{g(x) > t\}\big)$ на T, откуда, используя теорему о сходимости интегралов от ограниченных функций, получаем доказательство утверждения (106).

$$\int_0^1 P_n\{g(x) > t\}\, dt \xrightarrow[n\to\infty]{} \int_0^1 P\{g(x) > t\}\, dt.$$

□

Теорема 10.2. *(О представлении симметричных матриц).*

Пусть A – произвольная симметричная матрица $n \times n$, которую можно рассматривать как матрицу линейного оператора $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$, выраженного в стандартном базисе пространства $\mathbb{R}^n$. Тогда A можно представить в виде

$$A = O^\top \Lambda O,$$

где Λ – диагональная матрица собственных чисел, а $O^\top$ – матрица собственных векторов A, являющаяся ортонормированной (то есть, $O^\top O = I_n$).

Доказательство.

Ключевым следствием симметрии матрицы является свойство

$$\forall x, y \in \mathbb{R}^n \quad \langle Ax, y\rangle = \langle x, Ay\rangle, \tag{107}$$

поскольку скалярное произведение конечномерных векторов можно всегда рассматривать как произведение вектора-строки на вектор-столбец: $\langle Ax, y\rangle = (Ax)^\top y = x^\top A^\top y = x^\top Ay$, поскольку $A^\top = A$; кроме того, $\langle x, Ay\rangle = x^\top Ay$.

Поскольку матрица A задаёт действие оператора $\mathcal{A}$ на произвольный вектор, выраженный в стандартном базисе, то свойство (107) соответствует симметричности (самосопряженности) оператора:

$$\forall x, y \in \mathbb{R}^n \quad \langle \mathcal{A}(x), y\rangle = \langle x, \mathcal{A}(y)\rangle.$$

Докажем, что все собственные числа $\mathcal{A}$ являются вещественными от противного.

Допустим, что это не так. Поскольку $\det(A - \lambda I_n)$ является полиномом степени n от λ, то у характеристического полинома всегда есть n корней, возможно комплексных. Пусть $\lambda_1 \in \mathbb{C}$ – корень уравнения $\det(A - \lambda I) = 0$, а v – соответствующий комплексный собственный вектор, $Av = \lambda_1 v$.

Всилу симметрии свойство (107) матрицы A должно выполняться и для комплексных векторов. Напомним, что для $u, v \in \mathbb{C}^n$ скалярное произведение задаётся как $\langle u, v\rangle = \sum_{i=1}^n u_i \overline{v_i}$, где $\overline{v_i}$ обозначает комплексно-сопряженное значение к v_i.

$$\langle v, Av\rangle = \langle v, \lambda_1 v\rangle = \sum_{i=1}^n v_i \overline{\lambda_1 v_i} = \overline{\lambda_1}\|v\|^2,$$

$$\langle Av, v\rangle = \langle \lambda_1 v, v\rangle = \sum_{i=1}^n \lambda_1 v_i \overline{v_i} = \lambda_1 \|v\|^2.$$

Поэтому из $\langle v, Av\rangle = \langle Av, v\rangle$ следует $\overline{\lambda_1} = \lambda_1$, а значит λ_1 – вещественное число. Таким образом, у $\mathcal{A}$ имеется n вещественных собственных чисел (но возможно, что какие-то из них имеют алгебраическую

кратность выше единицы).

Теперь докажем, что из собственных векторов $\mathcal{A}$ можно составить ортогональный базис в $\mathbb{R}^n$. Пусть w_1 – собственный вектор $\mathcal{A}$, а w – произвольный вектор из $\mathbb{R}^n$, такой, что $w_1 \perp w$. Тогда

$$\langle \mathcal{A}(w), w_1 \rangle = \langle w, \mathcal{A}(w_1) \rangle = \lambda_1 \langle w, w_1 \rangle = 0.$$

Это означает, что сужение оператора $\mathcal{A}_1 = \mathcal{A}|_{\mathcal{L}^\perp(w_1)}$ на $(n-1)$-мерное подпространство $\mathbb{R}^n$ отображает $\mathcal{L}^\perp(w_1)$ в себя. Кроме того, $\mathcal{A}_1$ является симметричным и его собственные вектора должны быть ортогональными к w_1. Выбираем произвольный собственный вектор w_2 и повторяем такое рассуждение, пока не получим ортогональный базис из собственных векторов в $\mathbb{R}^n$, который далее можно нормировать, поделив каждый вектор на численное значение своей нормы (длины). Собранная из таких векторов матрица O в таком случае будет ортонормированной.

Наличие у линейного оператора собственных векторов, образующих базис в рассматриваемом пространстве, является необходимым и достаточным условием диагонализуемости, то есть, $\Lambda = O^{-1}AO$, причём поскольку $O^{-1} = O^\top$, то из полученного результата следует доказываемое свойство $\square$

Теорема 10.3. *(О матрице оператора ортогональной проекции в $\mathbb{R}^n$).*

Оператор $\mathcal{P}$ задаёт ортогональную проекцию в $\mathbb{R}^n$ тогда и только тогда, когда его матрица P имеет следующие свойства:

- $P^\top = P$ *(симметрия);*
- $P^2 = P$ *(идемпотентность).*

Доказательство. Сначала докажем, что проектор $\mathcal{P}$ имеет симметричную идемпотентную матрицу.

Оператор ортогонального проецирования $\mathcal{P}$ линеен, поэтому в пространстве конечной размерности $\mathbb{R}^n$ его можно представить с помощью некоторой $n \times n$ матрицы P. По определению, $\mathcal{P}$ проецирует вектора на линейное подпространство, которое обозначим V, а его ортогональное дополнение – за $V^\perp$. Для V и $V^\perp$ выполняется

$$\mathbb{R}^n = V \oplus V^\perp.$$

Чтобы доказать, что P симметрична и идемпотентна, возьмём два произвольных вектора $x, y \in \mathbb{R}^n$. Их разложения на суммы векторов из V и $V^\perp$ уникальны:

$$x = x_v + x_n, \quad x_v \in V,\ x_n \in V^\perp,$$

$$y = y_v + y_n, \quad y_v \in V,\ y_n \in V^\perp,$$

Поскольку $\mathcal{P}(x) = x_v, \mathcal{P}(y) = y_v$, то такие же соотношения должны выполняться для матрицы P:

$$Px = x_v, \quad Py = y_v.$$

Рассмотрим скалярные произведения одного из векторов с проекцией второго:

$$\langle Px, y \rangle = (Px)^\top y = x_v^\top y = x_v^\top (y_v + y_n) = x_v^\top y_v,$$

$$\langle x, Py \rangle = x^\top Py = x^\top y_v = (x_v + x_n)^\top y_v = x_v^\top y_v.$$

В каждом из этих скалярных произведений один из членов (Px или Py) полностью лежит в V, поэтому компонента второго вектора из $V^\perp$ всегда зануляется, из-за чего эти скалярные произведения равны между

собой и совпадают с значением $(Px)^\top Py$. Из этого следует

$$(Px)^\top y = x^\top P^\top y = x^\top Py, \quad x_v^\top y_v = (Px)^\top Py = x^\top P^\top Py.$$

Поскольку x и y произвольные, получаем $P^\top = P$ и $P^2 = P$.

Теперь докажем, что симметричная и идемпотентная матрица P задаёт оператор ортогонального проецирования.

Матрица P симметрична, а следовательно диагонализуема: $P = O\Lambda O^{-1}$ где O – матрица собственных векторов, а Λ – диагональная матрица собственных чисел. Кроме того, матрицу O можно сделать ортогональной.[42] Для любого собственного вектора z_j выполняется $Pz_j = \lambda_j z_j \Rightarrow PPz_j = P\lambda_j z_j \Rightarrow \lambda_j = \lambda_j^2$. Из этого следует, что собственными числами P могут быть только 0 или 1. Возьмём произвольный $x \in \mathbb{R}^n$ и разложим его по ортогональному базису из собственных векторов $\{e_i\}_{i=1}^n$ (столбцы матрицы O):

$$x = \sum_{i=1}^n x_i e_i, \quad x_i \in \mathbb{R}.$$

Тогда

$$Px = \sum_{i=1}^n x_i Pe_i, \quad \text{где} \quad Pe_i = e_i \quad \text{или} \quad Pe_i = 0,$$

то есть образом P является пространство, натянутое на собственные вектора с собственным числом 1, и на все $x \in \operatorname{Im} P$ матрица P действует как тождественный оператор. Очевидно, что ядро P – это пространство, натянутое на собственные вектора с собственным числом 0, и (в силу ортогональности собственных векторов) любой вектор в $\operatorname{Ker} P$ ортогонален вектору в $\operatorname{Im} P$. Поэтому матрица P задаёт оператор ортогональной проекции на подпространство $\operatorname{Im} P$. □

Лемма 10.4. *Пусть Z – k-мерный стандартный нормальный вектор, а P – $k \times k$ матрица проектора. Тогда $Z^\top PZ \sim \chi^2\big(\operatorname{tr}(P)\big)$.*

Доказательство. Используя теоремы 10.2, 10.3, получаем, что P можно представить в виде $P = O^\top \Lambda O$, где $O^\top$ – ортонормированная матрица собственных векторов P, а Λ – диагональная матрица собственных чисел, содержащая только значения 0 и 1. Используя инвариантность следа матрицы при преобразовании базиса линейного пространства, делаем вывод, что $\operatorname{tr}(\Lambda) = \operatorname{tr}(P)$, то есть, количество единиц на главной диагонали Λ совпадает с $\operatorname{tr}P$. Далее,

$$Z^\top PZ = Z^\top O^\top \Lambda OZ = (OZ)^\top \Lambda (OZ),$$

причём легко видеть, что OZ является стандартным нормальным вектором размерностью $n \times 1$, поскольку OZ является линейным преобразованием Z, а также $\mathbb{E}(OZ) = 0_n$ и $\mathbb{V}(OZ) = O\mathbb{V}(Z)O^\top = OO^\top = I_n$. Тогда обозначим $\tilde{Z} = OZ$, и заметим, что исходное выражение свелось к квадратичной форме с диагональной матрицей коэффициентов $(\tilde{Z})^\top \Lambda \tilde{Z}$, которое имеет вид суммы $\tilde{Z}_i^2$ с весами, равными значениям на главной диагонали. Учитывая, что среди этих значений есть только нули и единицы, можно сделать вывод, что полученное выражение будет иметь распределение χ^2 с числом степеней свободы, равным числу единиц среди собственных чисел P, которое, в свою очередь, равно $\operatorname{tr}(P)$. □

[42] Доказательство: в случае собственных векторов с одинаковыми собственными числами, это следует из того факта, что в пространстве, порождённом этими векторами можно выбрать ортогональный базис. Для собственных векторов x и y с разными собственными числами λ_x и λ_y, в силу симметрии P выполняется $\langle Px, y\rangle = \langle x, Py\rangle \Rightarrow \lambda_x x^\top y = x^\top \lambda_y y$, что возможно только при $x^\top y = 0$.

Теорема 10.5. *(F-тест линейных гипотез при нормальных ошибках регрессии).*

Рассмотрим регрессионную модель с нормальными ошибками, удовлетворяющую условиям Гаусса-Маркова,

$$Y = X\beta + Z\gamma + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n). \tag{108}$$

Предположим, что X и Z – матрицы регрессоров размером $n \times k$ и $n \times j$, и мы хотим тестировать $H_0 : \beta = 0_k$ против $H_a : \exists \ell : \beta_\ell \neq 0$. Для этого методом наименьших квадратов оценим исходную модель (108), а также короткую модель

$$Y = Z\tilde{\gamma} + u, \tag{109}$$

и обозначим получающиеся суммы квадратов остатков за SSR и SSR^* соответственно.

Утверждение: F-статистика $F = \frac{(SSR^* - SSR)/k}{SSR/(n-k-j)}$ при H_0 имеет $F(k, n-k-j)$ распределение.

Доказательство. Обозначим за Q и Q_Z проекторы на $\mathcal{L}(X, Z)^\perp$ и $\mathcal{L}(Z)^\perp$. По определению, F-распределённая случайная величина является отношением двух независимых χ^2 случайных величин, поправленных на свои степени свободы. Далее мы покажем, что $(SSR^* - SSR)$ и SSR действительно являются при H_0 независимыми случайными величинами с χ^2-распределениями.

Сначала отметим, что остатки в модели (108) имеют вид $\hat{\varepsilon} = QY = Q\varepsilon$, поэтому $SSR = \varepsilon^\top Q\varepsilon$. Далее, если H_0 верна, то коэффициенты при Z и ошибки в короткой модели – такие же, как и в длинной, поэтому МНК-остатки в короткой модели $\hat{u}$ можно выразить, как $Q_Z\varepsilon$.

$$Y = Z\gamma + \varepsilon, \quad SSR^* = \varepsilon^\top Q_Z \varepsilon. \tag{110}$$

Рассмотрим $SSR^* - SSR = \varepsilon^\top Q_Z\varepsilon - \varepsilon^\top Q\varepsilon = \varepsilon^\top (Q_Z - Q)\varepsilon$. Можно заметить, что $Q_Z - Q$ также является проектором, поскольку эта матрица симметрична и идемпотентна:

$$(Q_Z - Q)(Q_Z - Q) = Q_Z^2 - Q_Z Q - QQ_Z + Q^2 = Q_Z - Q - Q + Q = Q_Z - Q.$$

Здесь $Q_Z Q = Q$ из-за того, что $\mathcal{L}(X, Z)^\perp \subset \mathcal{L}(Z)^\perp$. То же самое верно для QQ_Z, поскольку произведение двух симметричных матриц является симметричной матрицей, и поэтому $Q_Z Q = (Q_Z Q)^\top = Q^\top Q_Z^\top = QQ_Z$.

Далее, $\frac{1}{\sigma}\varepsilon$ является стандартным нормальным вектором, и квадратичная форма $\frac{1}{\sigma}\varepsilon$ с любой матрицей-проектором согласно лемме 10.4 имеет χ^2-распределение с степенью свободы, равной следу матрицы. Поэтому $\frac{1}{\sigma^2}\varepsilon^\top Q\varepsilon \sim \chi^2(n-k-j)$, $\frac{1}{\sigma^2}\varepsilon^\top (Q_Z - Q)\varepsilon \sim \chi^2(k)$. Также заметим, что $\varepsilon^\top Q\varepsilon = \|Q\varepsilon\|^2$ и $\varepsilon^\top (Q_Z - Q)\varepsilon = \|(Q_Z - Q)\varepsilon\|^2$, где $Q\varepsilon$ и $(Q_Z - Q)\varepsilon$ (условно на X и Z) являются совместно нормальными векторами с нулевой ковариацией:

$$\operatorname{cov}\big(Q\varepsilon, (Q_Z - Q)\varepsilon \mid X, Z\big) = \mathbb{E}\left[Q\varepsilon\varepsilon^\top (Q_Z - Q) \mid X, Z\right] = \sigma^2 Q(Q_Z - Q) = 0.$$

Некоррелированные совместно-нормальные вектора являются независимыми, поэтому $Q\varepsilon$ не зависит от $(Q_Z - Q)\varepsilon$ (условно на X и Z), поэтому независимыми являются также функции от них, в частности, $(SSR^* - SSR)$ и SSR. Этот результат также выполняется безусловно (для доказательства нужно применить формулу полной вероятности). Остаётся заметить, что при H_0,

$$F = \frac{(SSR^* - SSR)/k}{SSR/(n-k-j)} = \frac{\frac{\varepsilon^\top (Q_Z - Q)\varepsilon}{\sigma^2 k}}{\frac{\varepsilon^\top (Q_Z - Q)\varepsilon}{\sigma^2 (n-k-j)}} \sim \frac{\frac{\chi^2(k)}{k}}{\frac{\chi^2(n-k-j)}{n-k-j}} \stackrel{d}{=} F(k, n-k-j).$$

□

Теорема 10.6. *(Гаусс, Марков, доказательство в приложении)*

Если предположения ГМ1-ГМ5 выполнены, то МНК-оценки являются оптимальными – то есть, имеют самую низкую дисперсию в классе линейных (по Y) несмещенных оценок.

Доказательство. Рассмотрим стандартную модель множественной регрессии

$$Y = X\beta + \varepsilon, \quad \mathbb{V}(\varepsilon \mid X) = \sigma^2 I_n.$$

Здесь β – это $k \times 1$ вектор параметров. Его МНК-оценка имеет вид $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$.

Возьмём произвольную линейную комбинацию коэффициентов $c^\top\beta$, $c \in \mathbb{R}^k$. Мы хотим доказать, что $\hat{\beta}$ является наиболее эффективной линейной несмещённой оценкой β, что эквивалентно доказательству того, что $\mathbb{V}(c^\top\hat{\beta}) \leq \mathbb{V}(c^\top\tilde{\beta})$ для любой другой линейной несмещённой оценки $\tilde{\beta}$.[43] Из-за линейности, каждая координата вектора $\tilde{\beta}$ является взвешенной суммой Y_i, поэтому $\tilde{\beta} = a^\top Y$ для некоторой векторнозначной функции $a = a(X)$, задающей матрицу весов размерностью $n \times k$.

Используемые оценки являются несмещенными, то есть $\mathbb{E}(\hat{\beta}) = \mathbb{E}(\tilde{\beta}) = \beta$, и это равенство верно для всех возможных значений β и распределений X. Рассмотрим условное математическое ожидание $\tilde{\beta}$:

$$\mathbb{E}\left(a^\top Y \mid X\right) = \mathbb{E}\big(a^\top(X\beta + \varepsilon) \mid X\big) = a^\top X\beta = \beta \Rightarrow a^\top X = I_k.^{44}$$

Для $\hat{\beta}$ также верно, что её условное математическое ожидание по X равно β (это в явном виде показано в задаче #3), поэтому для сравнения $\mathbb{V}(c^\top\hat{\beta})$ и $\mathbb{V}(c^\top\tilde{\beta})$ достаточно сравнить соответствующие условные дисперсии.

$$\mathbb{V}\left(c^\top\hat{\beta} \mid X\right) = \mathbb{V}\big(c^\top(X^\top X)^{-1}X^\top Y \mid X\big) = c^\top(X^\top X)^{-1}X^\top \mathbb{V}\left(Y \mid X\right)\left(c^\top(X^\top X)^{-1}X^\top\right)^\top =$$

$$= c^\top(X^\top X)^{-1}X^\top\sigma^2 I_n X(X^\top X)^{-1}c = \sigma^2 c^\top(X^\top X)^{-1}c \tag{111}$$

$$\mathbb{V}\left(c^\top\tilde{\beta} \mid X\right) = \mathbb{V}\left(c^\top a^\top Y \mid X\right) = \sigma^2(ac)^\top ac \geq \sigma^2(ac)^\top P_X ac, \tag{112}$$

где $P_X = X(X^\top X)^{-1}X^\top$ – ортогональный проектор на пространство, порождённое столбцами X. Неравенство в (112) обеспечивается за счёт того, что для любого вектора $ac \in \mathbb{R}^n$ выполняется $\|ac\| \geq \|P_X ac\|$ в силу теоремы Пифагора. Используя свойство $a^\top X = I_k$, получаем

$$\mathbb{V}\left(c^\top\tilde{\beta} \mid X\right) \geq \sigma^2(ac)^\top P_X ac = \sigma^2 c^\top a^\top X(X^\top X)^{-1}X^\top ac = \sigma^2 c^\top(X^\top X)^{-1}c = \mathbb{V}\left(c^\top\hat{\beta} \mid X\right),$$

что и требовалось доказать. □

Лемма 10.7. *Пусть в линейной модели $Y = X\beta + \varepsilon$ с k регрессорами $X := (X_1, \ldots, X_k)$ ошибка является*

[43] Выбирая в качестве c вектора наподобие $(1, 0, \ldots, 0)^\top$, можно, например, показать эффективность оценок отдельных коэффициентов регрессии. Предположение о том, что распределение k-мерного случайного вектора $\hat{\beta}$ уникальным образом определяется совокупностью одномерных распределений $c^\top\hat{\beta}$ было впервые доказано Крамером и Вольдом в 1936 г. и имеет в англоязычной литературе название Cramer-Wold device. Подобные подходы, когда информация об одной компоненте скалярного произведения, получается на основе исследования различных вторых компонент, в математике обычно связывают с принципом двойственности.

[44] Можно привести более формальную аргументацию того, что для несмещённой оценки должно выполняться $\mathbb{E}(\tilde{\beta} \mid X) = \beta$. Обозначим $Z = a^\top X$; из несмещенности $\tilde{\beta}$ следует, что $\mathbb{E}Z = I_k$ должно выполняться при любом распределении X. Тогда $\int Z_{ij}(x)\, dF_X(x) = \delta_{ij}$ (символ Кронекера). Возьмём семейство распределений X соответствующих единичной массе в различных точках $x_0 \in \mathbb{R}^{n\times k}$, тогда $\int Z_{ij}(x)\, dF_X(x) = Z_{ij}(x_0) = \delta_{ij}$, следовательно Z должна иметь постоянные значения при различных X, откуда следует $a^\top X = I_k$.

экзогенной и гомоскедастичной, то есть

$$\mathbb{E}(\varepsilon \mid X) = 0, \quad \mathbb{V}(\varepsilon \mid X) = \sigma^2 I_n.$$

Тогда несмещённой оценкой дисперсии ошибки σ^2 по случайной выборке размером n является

$$\hat{\sigma}^2 := \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-k} = \frac{SSR}{n-k}.$$

Доказательство. МНК-оценка имеет вид $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top Y$,

$$\hat{\varepsilon} = Y - X\hat{\beta} = \left(I_n - X(X^\top X)^{-1}X^\top\right)Y = (I_n - P_X)(X\beta + \varepsilon) = (I_n - P_X)\varepsilon,$$

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}^\top\varepsilon = \left((I_n - P_X)\varepsilon\right)^\top (I_n - P_X)\varepsilon = \varepsilon^\top (I_n - P_X)\varepsilon.$$

где $P_X = X(X^\top X)^{-1}X^\top$ – проектор на $\mathcal{L}(X)$.

Чтобы найти математическое ожидание $\hat{\varepsilon}^\top\hat{\varepsilon}$, удобно использовать свойство следа матрицы $\mathrm{tr}(\cdot)$:

- След от числа (скаляра) – само это число;
- Для любых $n \times k$ и $k \times n$ матриц A и B выполнено $\mathrm{tr}(AB) = \mathrm{tr}(BA)$;
- Если матрица Q имеет конечный размер, то операторы математического ожидания и следа всегда коммутируют: $\mathbb{E}\big(\mathrm{tr}(Q)\big) = \mathrm{tr}\big(\mathbb{E}(Q)\big)$.

Поскольку $\hat{\varepsilon}^\top\hat{\varepsilon}$ является скаляром, $\hat{\varepsilon}^\top\hat{\varepsilon} = \mathrm{tr}(\hat{\varepsilon}^\top\hat{\varepsilon}) = \mathrm{tr}\big(\varepsilon^\top(I_n - P_X)\varepsilon\big) = \mathrm{tr}\big((I_n - P_X)\varepsilon\varepsilon^\top\big)$;

$$\mathbb{E}[\hat{\varepsilon}^\top\hat{\varepsilon} \mid X] = \mathbb{E}\left[\mathrm{tr}\big((I_n - P_X)\varepsilon\varepsilon^\top\big) \mid X\right] = \mathrm{tr}\left[\mathbb{E}\big((I_n - P_X)\varepsilon\varepsilon^\top \mid X\big)\right] = \mathrm{tr}\left[(I_n - P_X)\mathbb{E}(\varepsilon\varepsilon^\top \mid X)\right] =$$

$$= \mathrm{tr}\left[(I_n - P_X)\sigma^2 I_n\right] = \sigma^2 \mathrm{tr}(I_n - P_X).$$

Очевидно, $\mathrm{tr}(I_n - P_X) = \mathrm{tr}(I_n) - \mathrm{tr}(P_X) = n - \mathrm{tr}\big(X(X^\top X)^{-1}X^\top\big) = n - \mathrm{tr}\big((X^\top X)^{-1}X^\top X\big) = n - k$, поскольку $X^\top X$ – это $k \times k$ матрица, и $(X^\top X)^{-1}X^\top X = I_k$ – $k \times k$ единичная матрица. Поэтому $\mathbb{E}[\hat{\varepsilon}^\top\hat{\varepsilon} \mid X] = \sigma^2(n-k)$, и её безусловное математическое ожидание – такое же, а следовательно, $\hat{\sigma}^2$ – несмещённая оценка σ^2.

```
# Check that σ̂² = Σ_{i=1}^n ε̂_i² / (n-k) is indeed the same estimator of error variance, that we get with lm() command
n <- 100
k <- 5
# number of observations and regressors in a linear model
X <- matrix(rnorm(k*n,0,1), ncol = k)
epsilon <- rnorm(n,0,1)
Y <- X %*% seq(1,k,1) + epsilon
# DGP is Y = Xθ + ε with i.i.d. normal X and ε, and θ = (1,2,...,k)^⊤.

ols_full <- lm(Y~X-1)
summary(ols_full)$sigma^2
# Value of error variance estimate obtained via lm() function
Qx <- diag(x=1,nrow = n, ncol = n) - X %*% solve(t(X) %*% X) %*% t(X)
e_hat <- Qx %*% Y
(t(e_hat) %*% e_hat)/(n-k)
```

```
# The same value, calculated manually
```