

Лекция 2. Модель парной линейной регрессии. Свойства оценок МНК

Введение

На прошлой лекции мы узнали, что эконометрика моделирует зависимости между переменными. Сегодня мы изучим самую простую и мощную из таких моделей - *парную линейную регрессию*. Мы не только поймем, как она работает, но и выясним, почему именно метод наименьших квадратов (МНК) является золотым стандартом для ее оценки.

1. Модель парной линейной регрессии: Суть и компоненты

Вернемся к нашему примеру с зарплатой и образованием. Мы предполагаем существование линейной зависимости:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Давайте "вскроем" эту модель и посмотрим, из чего она состоит.

Аналогия: Лук и стрелы.

Y_i (*Зависимая переменная*) - это ваша мишень. Реальное значение, которое вы наблюдаете (например, фактическая зарплата человека №5).

$\beta_0 + \beta_1 X_i$ (*Систематическая составляющая*) - это ваш *прицел*. Это часть зарплаты, которая *объясняется* образованием. Вы предсказываете ее на основе вашей модели.

β_0 - Базовая зарплата (при $X=0$). Куда смотрит прицел по вертикали, когда образование равно нулю.

β_1 - "цена" года образования. Насколько поднимается прицел при каждом дополнительном годе учебы.

ε_i (случайная ошибка) - это *случайный ветер*, который сносит стрелу. Все, что влияет на зарплату, но не учтено в модели (талант, удача, связи). Мы его не наблюдаем напрямую.

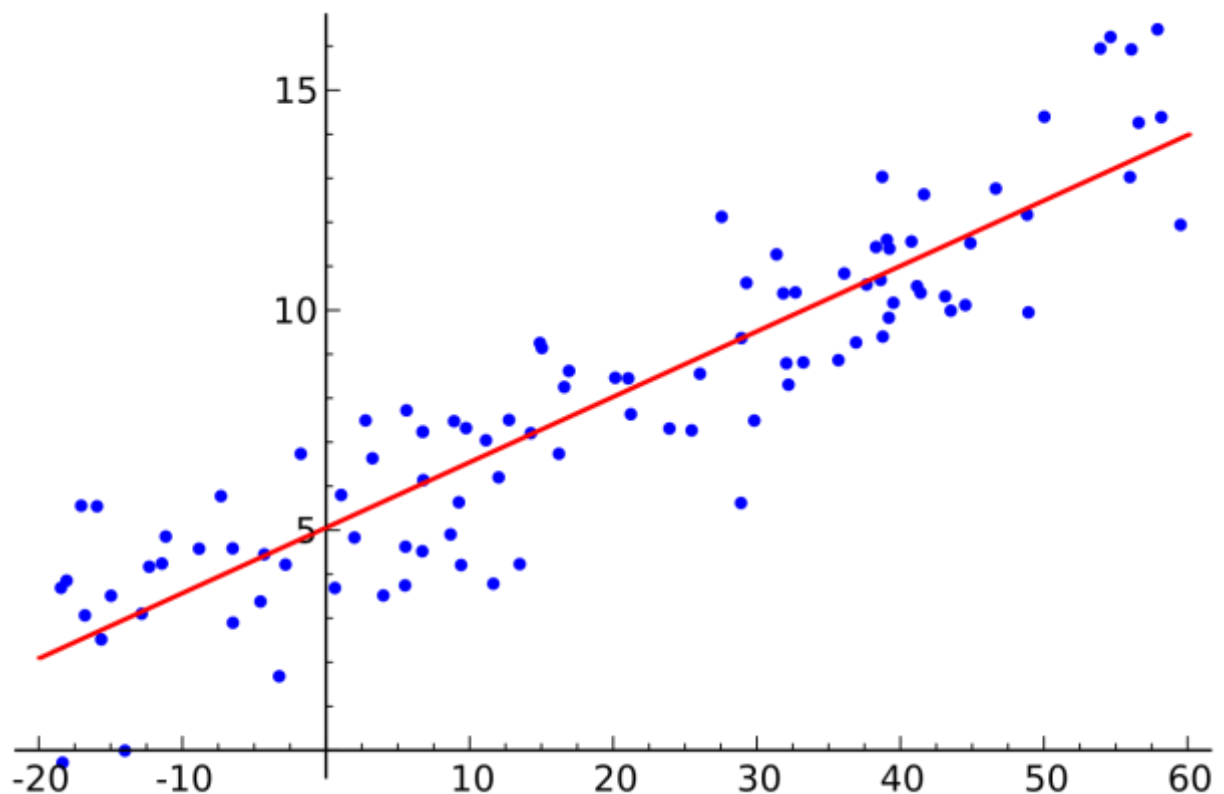


Рисунок 2.1. Диаграмма рассеяния с точками данных и линией регрессии

На этой стандартной иллюстрации (Рисунок 2.1) идеально видны все компоненты модели парной регрессии, о которых мы говорим:

- **Синие точки:** это наши исходные данные (наблюдения X_i, Y_i , та самая "мишень".
- **Наклонная красная линия:** это наша оцененная модель регрессии $\hat{Y} = b_0 + b_1X$, то есть **систематическая составляющая**. Она показывает предсказанное значение Y для каждого X .
- **Остатки (e_i)** - визуальное представление случайной ошибки e_i для каждого наблюдения. Они показывают расстояние по вертикали между реальным значением Y_i (синяя точка) и предсказанным значением

\hat{Y}_i (красная линия). Задача МНК как раз и состоит в том, чтобы сделать сумму квадратов этих длин минимальной.

2. Задача исследователя и метод наименьших квадратов (МНК)

Мы не знаем истинных значений параметров β_0 и β_1 . Это параметры *генеральной совокупности*. Наша задача - по имеющейся у нас *выборке* данных $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ найти их наилучшие *оценки*. Обозначим их как b_0 и b_1 .

Метод наименьших квадратов (МНК) - это процедура нахождения таких оценок b_0 и b_1 , при которых *сумма квадратов остатков (SSR)* минимальна.

Что такое остаток (e_i)?

Остаток — это выборочная оценка ненаблюдаемой ошибки ε_i (Рисунок 2.2).

$$e_i = Y_i - \hat{Y}_i$$

где $\hat{Y}_i = b_0 + b_1 X_i$ — это *предсказанное значение* Y для i -го наблюдения.

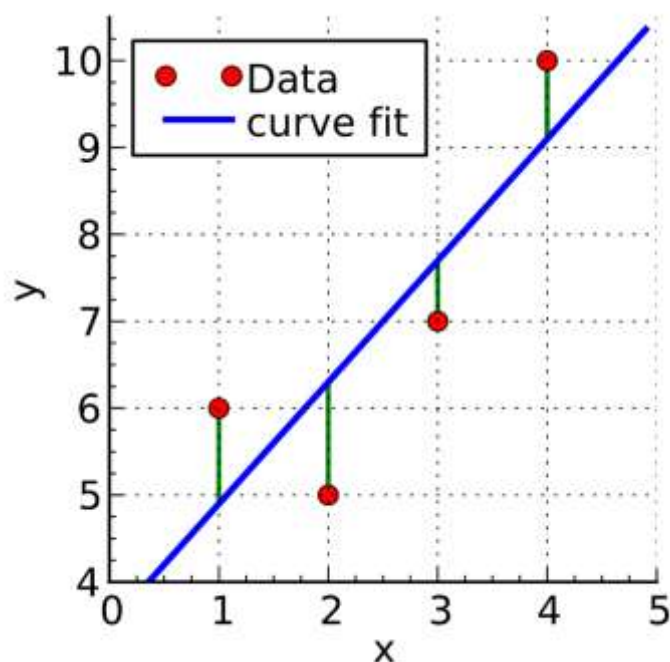


Рисунок 2.2. Остатки представляют собой расстояния от выборочных данных до регрессионной прямой

Математически мы минимизируем функцию:

$$Q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 \rightarrow \min$$

Это классическая задача на поиск экстремума. Берем частные производные по b_0 и b_1 , приравниваем их к нулю и получаем систему нормальных уравнений.

3. Вывод формул для оценок МНК

Решая систему нормальных уравнений, мы получаем знаменитые формулы:

1. Формула для углового коэффициента (b_1):

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Альтернативно, через ковариацию и дисперсию:

$$b_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

Интуитивное объяснение: коэффициент b_1 показывает, как сильно ковариация между X и Y выражена в вариации самого X . Если X и Y совместно колеблются сильно (высокая ковариация), а разброс X невелик, то каждое движение X будет сильно "толкать" Y .

2. Формула для константы (b_0):

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Интуитивное объяснение: линия регрессии всегда проходит через точку средних значений (\bar{X}, \bar{Y}) . Мы знаем средний уровень Y (\bar{Y}). Константа b_0 — это просто корректировка, которая подтверждает, что при среднем X модель дает средний Y .

Пример расчета (упрощенный):

Пусть у нас есть 3 наблюдения:

Образование (X)	Зарплата (Y)
12	40
14	50
16	60

$$\bar{X} = (12+14+16)/3 = 14$$

$$\bar{Y} = (40+50+60)/3 = 50$$

$$\text{Cov}(X,Y) = [(12-14)(40-50) + (14-14)(50-50) + (16-14)(60-50)] / (3-1) = [20 + 0 + 20] / 2 = 20$$

$$\text{Var}(X) = [(12-14)^2 + (14-14)^2 + (16-14)^2] / (3-1) = [4+0+4]/2 = 4$$

$b_1 = 20 / 4 = 5$ (Каждый дополнительный год образования ассоциирован с увеличением зарплаты на 5 тыс. у.е.)

$$b_0 = 50 - 5 * 14 = -20$$

Наша оцененная регрессия: $\hat{Y} = -20 + 5X$

4. Свойства оценок МНК: Теорема Гаусса-Маркова

Теперь самый важный вопрос: почему МНК так хорош? Ответ дает фундаментальная *Теорема Гаусса-Маркова*.

При выполнении *предположений классической линейной модели регрессии (КЛМР)*, оценки МНК b_1 и b_0 являются *наилучшими линейными несмещенными оценками* (BLUE - Best Linear Unbiased Estimators).

Давайте разберем эту сложную фразу по косточкам. Сначала - предположения.

Предположения КЛМР:

1. Линейность по параметрам: модель имеет вид $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.
2. Случайное выборочное наблюдение: данные (X_i, Y_i) являются случайной выборкой из генеральной совокупности.
3. Нулевое условное математическое ожидание ошибок: $E(\varepsilon_i | X_i) = 0$. Это ключевое предположение! Оно означает, что объясняющая переменная X не содержит информации о значении ошибки ε . Все факторы, систематически влияющие на Y и коррелированные с X , должны быть включены в модель. Нарушение этого предположения ведет к смещенности оценок.
4. Гомоскедастичность: дисперсия ошибки постоянна для всех X : $\text{Var}(\varepsilon_i | X_i) = \sigma^2$. (Картинка с "коридором постоянной ширины" вокруг линии регрессии).
5. Отсутствие автокорреляции: Ошибки для разных наблюдений не коррелированы между собой: $\text{Cov}(\varepsilon_i, \varepsilon_j | X_i, X_j) = 0$, где $i \neq j$.
6. Отсутствие совершенной мультиколлинеарности (в парной регрессии тривиально): в модели с несколькими регрессорами ни один из них не должен быть точной линейной комбинацией других.

Что означают свойства BLUE?

Линейность (L): Оценки МНК b_0 и b_1 являются линейными функциями от значений Y_i . Это удобно для анализа.

Несмещенность (U): $E(b_0) = \beta_0$ и $E(b_1) = \beta_1$.

Аналогия: стрельба по мишени. Несмещенный стрелок — это тот, чьи попадания в среднем группируются вокруг центра мишени. Смещенный стрелок будет постоянно промахиваться в одну и ту же сторону.

Эффективность (B — Наилучшие): среди всех возможных линейных и несмещенных оценок, оценки МНК имеют наименьшую дисперсию. Это

значит, что их распределение наиболее "плотное" вокруг истинного значения параметра.

Аналогия: несмещенный стрелок с маленьким разбросом (высокой эффективностью) лучше, чем несмещенный стрелок с большим разбросом.

Вывод из теоремы Гаусса-Маркова: если выполняются предположения 1-5 (и особенно ключевое предположение №3), то нет других линейных и несмещенных оценок, которые были бы точнее, чем МНК. Это делает его оптимальным методом.

5. Что еще нам нужно оценить? Дисперсия ошибок

Мы нашли b_0 и b_1 . Но в нашей модели есть еще один важный параметр - дисперсия случайной ошибки σ^2 (показывает "разброс" точек вокруг линии). Ее несмещенной оценкой является:

$$\widehat{\sigma^2} = s^2 = \frac{\sum_{i=1}^n e_i^2}{n - 2}$$

где $(n - 2)$ - это число *степеней свободы*. Мы теряем 2 степени свободы, потому что для расчета остатков нам пришлось оценить два параметра (b_0 и b_1). Корень из этой величины, s , называется *стандартной ошибкой регрессии* и измеряет "среднее" расстояние от точек до линии регрессии.

6. Коэффициент детерминации R^2 : насколько хорошо модель описывает данные?

После того как мы построили линию регрессии, нам нужна мера, которая покажет, насколько хорошо эта линия подходит к данным.

Общая сумма квадратов (TSS) измеряет общую вариацию в зависимой переменной Y :

$$TSS = \sum (Y_i - \bar{Y})^2$$

Объясненная сумма квадратов (ESS) измеряет вариацию Y , которая объясняется нашей моделью (регрессией):

$$ESS = \sum (\hat{Y}_i - \bar{Y})^2$$

Остаточная сумма квадратов (RSS) измеряет необъясненную вариацию (ту, что остается в ошибках):

$$RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum (e_i)^2$$

Важно: $TSS = ESS + RSS$

Коэффициент детерминации R^2 — это доля общей вариации зависимой переменной Y , которая объясняется нашей регрессионной моделью.

$$R^2 = ESS/TSS = 1 - (RSS/TSS)$$

Интерпретация:

- **$R^2 = 1$:** Идеальная модель. Все точки данных лежат на линии регрессии (все изменения Y объясняются X).
- **$R^2 = 0$:** Модель ничего не объясняет. Вариация Y полностью обусловлена случайными ошибками.
- **$0 < R^2 < 1$:** Чем ближе R^2 к 1, тем лучше модель описывает данные.

Пример: Если $R^2 = 0.75$, это означает, что 75% колебаний переменной Y объясняется изменениями переменной X , а остальные 25% обусловлены неучтенными в модели факторами (случайной ошибкой).

7. Статистические гипотезы: является ли зависимость значимой?

Мы получили оценку наклона b_1 . Но это оценка по выборке. Как мы можем быть уверены, что истинный параметр генеральной совокупности β_1 не равен нулю? Если $\beta_1 = 0$, то линейной зависимости между X и Y не существует.

Для проверки этого мы используем **t-тест**.

Формулируем гипотезы:

- **Нулевая гипотеза (H_0):** $\beta_1 = 0$ (нет линейной зависимости).
- **Альтернативная гипотеза (H_1):** $\beta_1 \neq 0$ (есть линейная зависимость).

Стандартная ошибка коэффициента (se):

Это оценка стандартного отклонения распределения нашей оценки b_1 . Она показывает, насколько точно мы измерили коэффициент. Чем меньше se, тем точнее оценка.

$$se(b_1) = \sqrt{s^2 / \sum (X_i - \bar{X})^2},$$

где $s^2 = RSS / (n - 2)$ - оценка дисперсии ошибок.

t-статистика для коэффициента β_1 вычисляется как:

$$t = (b_1 - \beta_1) / se(b_1)$$

Для проверки гипотезы $H_0: \beta_1 = 0$ мы используем:

$$t_{\text{тест}} = b_1 / se(b_1)$$

Интерпретация:

- t-статистика показывает, на сколько стандартных ошибок наша оценка b_1 отклонена от гипотетического значения $\beta_1 = 0$.

- Чем больше абсолютное значение t-статистики ($|t|$), тем менее вероятно, что мы получили такое большое b_1 чисто случайно, при условии, что H_0 верна.

Принятие решения:

Мы сравниваем рассчитанное значение t-статистики с **критическим значением t-распределения** с $n-2$ степенями свободы и выбранным уровнем значимости (обычно $\alpha=0.05$).

- Если $|t| > t_{\text{крит}}$, то мы **отвергаем H_0** . Коэффициент является **статистически значимым**.
- Если $|t| \leq t_{\text{крит}}$, то у нас **нет оснований отвергать H_0** .

p-value (достоверность значимости):

На практике чаще используют **p-value** — вероятность получить значение коэффициента как минимум такое же экстремальное, как наше b_1 , при условии, что H_0 верна.

- **p-value $< \alpha$ (например, 0.05)** → отвергаем H_0 , коэффициент значим.
- **p-value $\geq \alpha$** → не отвергаем H_0 .

Пример: если для b_1 мы получили p-value = 0.01, это означает, что если бы истинный β_1 был равен 0, то шанс получить такую сильную зависимость по выборке составил бы всего 1%. Это веское доказательство в пользу существования зависимости.

Полный алгоритм расчета p-value

После того как мы вычислили t-статистику для проверки гипотезы о коэффициенте, следующим шагом является расчет p-value. Этот процесс можно разбить на четкие шаги.

Шаг 1: Формулировка гипотез

- Нулевая гипотеза (H_0): $\beta_1 = 0$ (нет линейного влияния X на Y).
- Альтернативная гипотеза (H_1): $\beta_1 \neq 0$ (есть линейное влияние X на Y). Это двусторонний тест.

•

Шаг 2: Расчет t-статистики

$$t_{\beta_1} = b_1 / se(b_1),$$

где:

- b_1 — оценка МНК для наклона.
- $se(b_1)$ — стандартная ошибка коэффициента b_1 , вычисляемая как: $se(b_1) = \sqrt{s^2 / \sum (X_i - \bar{X})^2}$
- $s^2 = RSS / (n - 2)$ — оценка дисперсии ошибок.
- $n - 2$ — количество степеней свободы.

Шаг 3: Определение распределения t-статистики

При справедливости нулевой гипотезы ($H_0: \beta_1 = 0$) и выполнении предположений классической линейной модели, рассчитанная t-статистика следует **t-распределению Стьюдента** (Рисунок 2.3) с $v = n - 2$ степенями свободы.

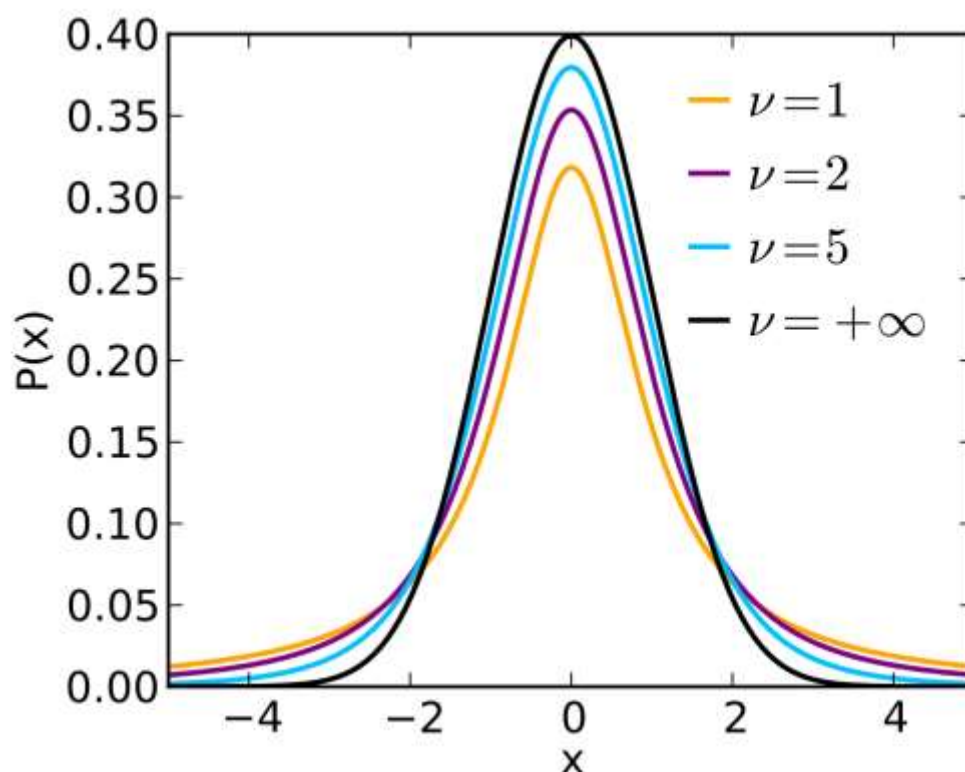


Рисунок 2.3. t-распределение Стьюдента

t-распределение похоже на нормальное, но имеет более тяжелые "хвосты", что учитывает дополнительную неопределенность из-за оценки стандартной ошибки по выборке. Чем больше степеней свободы ($n-2$), тем больше оно приближается к нормальному.

Шаг 4: Непосредственный расчет p-value

p-value - это вероятность получить значение t-статистики, по абсолютной величине равное или превосходящее фактически наблюдаемое значение $t_{\text{набл}}$, при условии, что нулевая гипотеза верна.

Для двустороннего теста формула расчета выглядит так:

$$p - \text{value} = 2 * P(T > |t_{\text{набл}}|)$$

где T — случайная величина, имеющая t-распределение с $n-2$ степенями свободы.

Алгоритм расчета:

1. Вычислите модуль (абсолютное значение) вашей t-статистики: $|t_{\text{наблюд}}|$.
2. Используя статистическое программное обеспечение (R, Python, Stata) или даже Excel, найдите площадь под кривой t-распределения (с $n-2$ степенями свободы) **справа** от полученного значения $|t_{\text{наблюд}}|$. Эта площадь равна $P(T > |t_{\text{наблюд}}|)$.
3. Умножьте эту площадь на 2, чтобы учесть оба "хвоста" распределения (возможность того, что истинное β_0 могло бы быть как положительным, так и отрицательным).

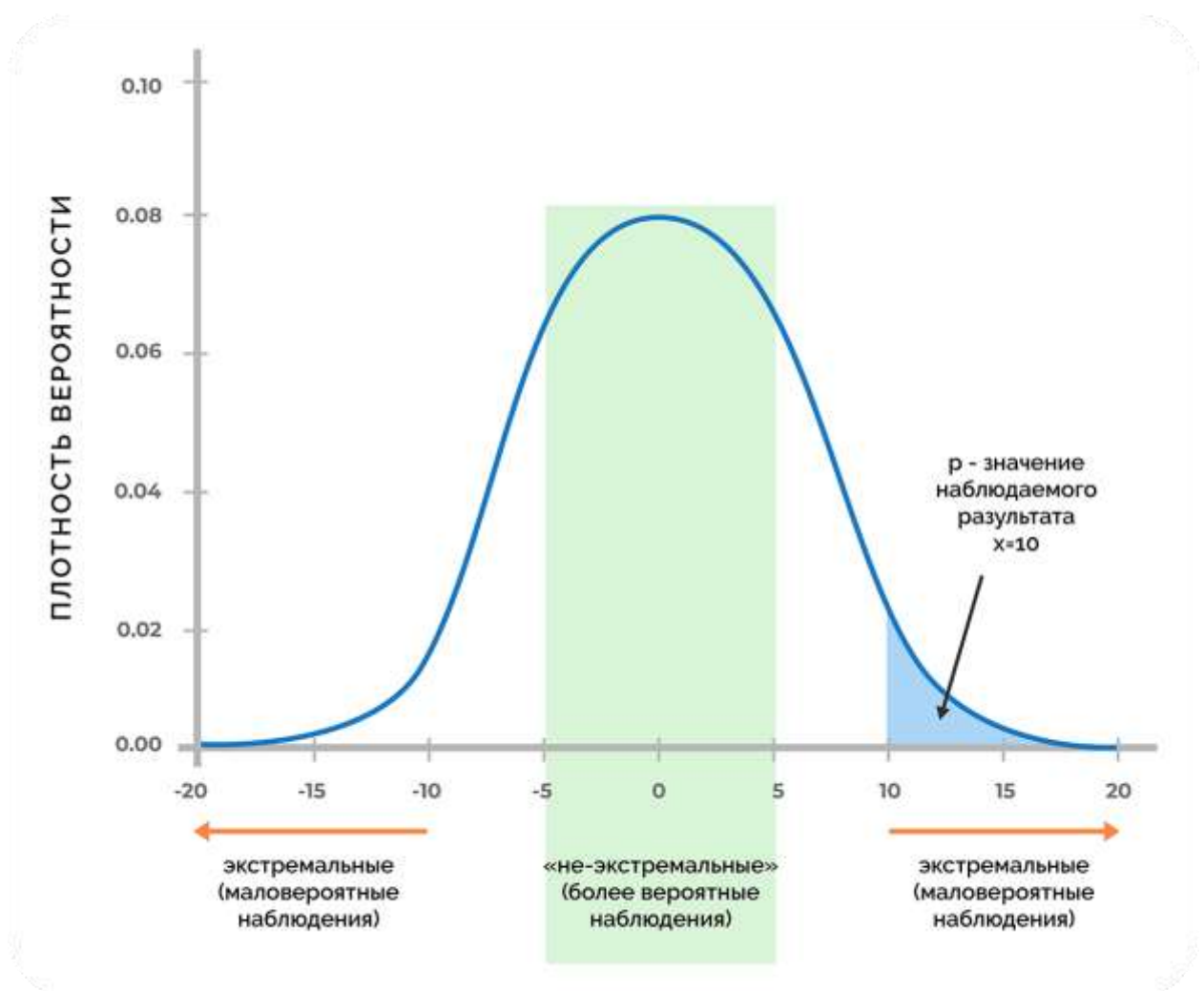


Рисунок 2.4. p-value для двустороннего теста

На этом графике (Рисунок 2.4.) синяя область в хвосте - это и есть p-value. Фактическое значение t-статистики отмечено вертикальной линией. P-value - это суммарная площадь этих двух синих областей.

Шаг 5: Принятие решения

- Если $p\text{-value} < \alpha$ (уровня значимости, обычно 0.05), то мы **отвергаем** H_0 . Результат считается **статистически значимым**.
- Если $p\text{-value} \geq \alpha$, то у нас **нет оснований отвергать** H_0 .

Пример: Расчет p-value "вручную" (концептуально)

Вернемся к нашему примеру с зарплатой и образованием.

- $b_1 = 5$
 - $se(b_1) = 0.8$
 - $n = 100$ (предположим)
 - $t = 5 / 0.8 = 6.25$
 - Степени свободы: $v = 100 - 2 = 98$
1. $|t| = 6.25$
 2. Мы обращаемся к таблице t-распределения или используем программное обеспечение, чтобы найти вероятность $P(T > 6.25)$ для $v=98$. Это значение будет чрезвычайно мало, так как 6.25 — это огромное значение для t-статистики. Допустим, это 0.0000005.
 3. $p\text{-value} = 2 * 0.0000005 = 0.000001$
 - 4.

Вывод: поскольку $p\text{-value} \approx 0.000001 < 0.05$, мы уверенно отвергаем нулевую гипотезу. Влияние образования на зарплату является статистически значимым.

Важное замечание: на практике исследователи никогда не рассчитывают p-value вручную по таблицам. Статистические пакеты (например, `scipy.stats` в Python, `summary(lm())` в R) делают это автоматически

и выдают его прямо в выходной таблице регрессии рядом с коэффициентами и t-статистиками. Однако понимание того, что стоит за этой цифрой, критически важно для корректной интерпретации результатов.

8. Пример: Полный анализ модели «Зарплата vs Образование»

Допустим, мы оценили модель: $\hat{Y} = -20 + 5X$

- $b_1 = 5$: Каждый дополнительный год образования ассоциирован с увеличением зарплаты на 5 тыс. у.е.
- $R^2 = 0.65$: 65% вариации зарплаты объясняется вариацией уровня образования.
- $se(b_1) = 0.8$: Стандартная ошибка коэффициента.
- $t = 5 / 0.8 = 6.25$: t-статистика.
- $p\text{-value} < 0.0001$: Крайне мало.
-

Вывод: Мы обнаружили **статистически значимую** ($p\text{-value} < 0.05$) и **сильную** ($R^2=0.65$) положительную линейную зависимость между образованием и зарплатой. Каждый дополнительный год образования значимо увеличивает зарплату.

Резюме

Сегодня мы заложили фундамент:

1. Мы *определили* модель парной линейной регрессии и ее компоненты.
2. Мы *вывели* формулы для оценок МНК b_1 и b_0 и поняли их интуитивный смысл.
3. Мы сформулировали *предположения КЛМР*, при которых МНК работает идеально.

4. Мы доказали (концептуально), что при этих предположениях МНК-оценки являются BLUE (несмещенными и самыми точными).
5. Мы научились оценивать дисперсию ошибок.
6. Мы ввели R^2 как меру качества подгонки модели.
- 7 Мы научились использовать **t-статистику** и **p-value** для проверки статистической значимости коэффициентов.

Вопросы для самопроверки:

1. Почему мы минимизируем именно сумму квадратов остатков, а не их модулей?
2. Что означает нарушение предположения $E(\epsilon_i | X_i) = 0$? Приведите пример из экономики.
3. Рассчитайте коэффициенты МНК для данных: $(X, Y) = (1, 2), (2, 4), (3, 5)$. Проходит ли линия регрессии через точку средних?
4. Может ли R^2 быть отрицательным в модели парной линейной регрессии, оцененной МНК?
5. Если коэффициент b_1 статистически незначим ($p\text{-value} > 0.05$), означает ли это, что связи между X и Y нет?
6. Почему при проверке гипотезы о коэффициенте мы используем t -распределение, а не нормальное?