Описательная статистика

Описательная статистика — это краткие описательные коэффициенты, обобщающие заданный набор данных, который может быть либо представлением всей совокупности, либо выборкой совокупности.

Описательная статистика подразделяется на меры центральной тенденции и меры рассеяния. Меры центральной тенденции включают среднее значение, медиану и моду, в то время как меры рассеяния включают стандартное отклонение, дисперсию, минимальные и максимальные переменные, эксцесс и асимметрию.

Меры центральной тенденции описывают положение центра распределения для набора данных. Человек анализирует частоту каждой точки данных в распределении и описывает ее, используя среднее значение, медиану или моду, которые измеряют наиболее распространенные закономерности анализируемого набора данных.

Меры рассеяния (или меры разброса) помогают анализировать, насколько рассредоточено распределение набора данных. Например, хотя показатели центральной тенденции могут дать человеку среднее значение набора данных, они не описывают, как данные распределяются в наборе.

Таким образом, хотя среднее значение данных может быть 65 из 100, все же могут быть точки данных как 1, так и 100. Меры изменчивости помогают сообщить об этом, описывая форму и разброс набора данных. Диапазон, квартили , абсолютное отклонение и дисперсия — все это примеры мер изменчивости.

Рассмотрим следующий набор данных: 5, 19, 24, 62, 91, 100. Диапазон этого набора данных равен 95, который рассчитывается путем вычитания наименьшего числа (5) в наборе данных из наибольшего (100).

Меры среднего уровня

Среднее значение

Вероятно, большинство из вас использовало такую важную описательную статистику, как среднее.

Среднее - очень информативная мера "центрального положения" наблюдаемой переменной, особенно если сообщается ее доверительный интервал. Исследователю нужны такие статистики, которые позволяют сделать вывод относительно популяции в целом. Одной из таких статистик является среднее.

Доверительный интервал для среднего представляет интервал значений вокруг оценки, где с данным уровнем доверия, находится "истинное" (неизвестное) среднее популяции.

Например, если среднее выборки равно 23, а нижняя и верхняя границы доверительного интервала с уровнем p=.95 равны 19 и 27 соответственно, то

можно заключить, что с вероятностью 95% интервал с границами 19 и 27 накрывает среднее популяции.

Если вы установите больший уровень доверия, то интервал станет шире, поэтому возрастает вероятность, с которой он "накрывает" неизвестное среднее популяции, и наоборот.

Хорошо известно, например, что чем "неопределенней" прогноз погоды (т.е. шире доверительный интервал), тем вероятнее он будет верным. Заметим, что ширина доверительного интервала зависит от объема или размера выборки, а также от разброса (изменчивости) данных. Увеличение размера выборки делает оценку среднего более надежной. Увеличение разброса наблюдаемых значений уменьшает надежность оценки.

Вычисление доверительных интервалов основывается на предположении нормальности наблюдаемых величин. Если это предположение не выполнено, то оценка может оказаться плохой, особенно для малых выборок.

При увеличении объема выборки, скажем, до 100 или более, качество оценки улучшается и без предположения нормальности выборки.

Довольно трудно «ощутить» числовые измерения, пока данные не будут содержательно обобщены. Диаграмма часто полезна в качестве отправной точки. Мы можем также сжать информацию, используя важные характеристики данных. В частности, если бы мы знали, из чего состоит представленная величина, или если бы мы знали, насколько широко рассеяны наблюдения, то мы бы смогли сформировать образ этих данных.

Среднее арифметическое, которое очень часто называют просто «среднее», получают путем сложения всех значений и деления этой суммы на число значений в наборе.

Это можно показать с помощью алгебраической формулы. Набор n наблюдений переменной X можно изобразить как X1, X2, X3, ..., Xn. Например, за X можно обозначить рост индивидуума (см), X1 обозначит рост 1-го индивидуума, а Xi — рост i-го индивидуума. Формула для определения среднего арифметического наблюдений $\bar{X}$ (произносится «икс с чертой»):

$\bar{X}$ = (X1 + X2 + ... + Xn) / n

Можно сократить это выражение:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

где Σ (греческая буква «сигма») означает «суммирование», а индексы внизу и вверху этой буквы означают, что суммирование производится от i = 1 до i = n. Это выражение часто сокращают еще больше:

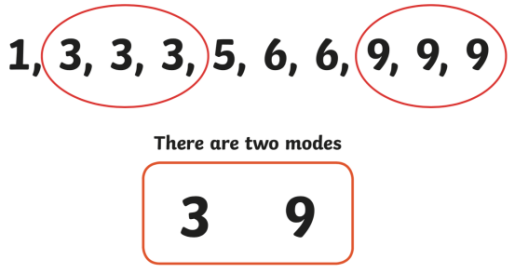
$$\bar{X} = \frac{\sum x_i}{n} \quad \text{или} \quad \bar{X} = \frac{\sum x}{n}$$

Математическое ожидание

Математическое ожидание - одно из важнейших понятий в теории вероятностей, означающее среднее (взвешенное по вероятностям возможных значений) значение случайной величины. На практике математическое ожидание обычно оценивается как среднее арифметическое наблюдаемых значений случайной величины

Мода

Мода - это значение, которое чаще всего встречается в наборе данных. Набор данных может иметь одну, более одной моды или вообще не иметь моды.



В статистике данные могут распределяться по-разному. Наиболее часто упоминаемым распределением является классическое нормальное распределение (гауссова кривая). В этом и некоторых других распределениях среднее (среднее) значение приходится на среднюю точку, которая также является пиковой частотой наблюдаемых значений.

Мода наиболее полезна в качестве меры центральной тенденции при изучении категориальных данных, таких как модели автомобилей или вкусы газированных напитков, для которых невозможно рассчитать среднее математическое значение, основанное на упорядочивании.

Пример. Например, в следующем списке чисел модой является 16, поскольку оно встречается в наборе больше раз, чем любое другое число:

- 3, 3, 6, 9, **16, 16, 16** , 27, 27, 37, 48

Набор чисел может иметь более одного режима (это называется *бимодальным* , если есть два режима), если есть несколько чисел, которые встречаются с одинаковой частотой и чаще, чем другие в наборе.

- **3, 3, 3** , 9, **16, 16, 16** , 27, 37, 48

В приведенном выше примере и число 3, и число 16 являются модами, поскольку каждое из них встречается три раза, и никакое другое число не встречается чаще.

Если ни одно число в наборе чисел не встречается более одного раза, у этого набора нет моды:

- 3, 6, 9, 16, 27, 37, 48

Набор чисел с двумя модами является **бимодальным** , набор чисел с тремя режимами — **тримодальным** , а любой набор чисел с более чем одним режимом — **мультимодальным**.

Расчет моды довольно прост. Расположите все числа в заданном наборе по порядку; это может быть от низшего к высшему или от высшего к низшему, а затем подсчитать, сколько раз каждое число появляется в наборе. Тот, который появляется больше всего, является модой.

Медиана

Медиана — это среднее число в отсортированном, восходящем или нисходящем списке чисел, и оно может быть более информативным для этого набора данных, чем среднее значение. Это точка, выше и ниже которой падает половина (50%) наблюдаемых данных, и, таким образом, она представляет собой среднюю точку данных.

Медиану часто сравнивают с другими описательными статистическими данными, такими как среднее (среднее), мода и стандартное отклонение.

Пример. Чтобы найти медианное значение в списке с **нечетным** количеством чисел, нужно найти число, которое находится в середине с одинаковым количеством чисел по обе стороны от медианы. Чтобы найти медиану, сначала расположите числа по порядку, обычно от меньшего к большему.

Например, в наборе данных

{3, 13, 2, 34, 11, 26, 47}

порядок сортировки становится

{2, 3, 11, 13, 26, 34, 47}.

Медиана — это число в середине

{2, 3, 11, **13** , 26, 34, 47},

которое в данном случае равно 13, поскольку с каждой стороны есть три числа.

Чтобы найти медианное значение в списке с **четным** количеством чисел, нужно определить среднюю пару, сложить их и разделить на два. Опять же, расположите числа в порядке от меньшего к большему.

Например, в наборе данных

{3, 13, 2, 34, 11, 17, 27, 47}

порядок сортировки становится

{2, 3, 11, 13, 17, 27, 34, 47}.

Медиана — это среднее двух чисел в середине

{2, 3, 11, **13** , **17** , 26 34, 47},

что в данном случае равно пятнадцати {(13 + 17) ÷ 2 = 15}.

Медиана тесно связана с квартилями или разделением наблюдаемых данных на четыре равные части. Медиана будет центральной точкой, при этом первые два квартиля окажутся ниже нее, а вторые два выше нее.

В нормальном распределении («колоколообразная кривая») медиана, среднее значение и мода имеют одно и то же значение и приходятся на самую высокую точку в центре кривой.

Когда среднее и медиана отличаются?

В наборе искаженных данных среднее значение и медиана обычно будут разными. Среднее значение рассчитывается путем сложения всех значений в данных и деления на количество наблюдений. Если есть значительные выбросы или если данные слипаются вокруг определенных значений, среднее (среднее) не будет средней точкой данных.

Например, в наборе данных {0, 0, 0, 1, 1, 2, 10, 10} среднее значение будет 24/8 = 3. Однако медиана будет равна 1 (среднее значение).

Вот почему многие экономисты отдают предпочтение медианному показателю дохода или богатства страны, поскольку он лучше отражает фактическое распределение доходов.

Меры изменчивости

Квартиль

Квартиль — это статистический термин, который описывает разделение наблюдений на четыре определенных интервала на основе значений данных и того, как они соотносятся со всем набором наблюдений.

Медиана является надежной оценкой местоположения, но ничего не говорит о том, как данные по обе стороны от ее значения разбросаны или рассредоточены. Вот где вступает квартиль. Квартиль измеряет разброс значений выше и ниже среднего путем разделения распределения на четыре группы.

Точно так же, как медиана делит данные пополам, так что 50% измерений лежат ниже медианы, а 50 % лежат выше нее, квартиль разбивает данные на четверти, так что 25 % измерений меньше нижнего квартиля, 50% меньше медианы, а 75% меньше верхнего квартиля.

Теперь мы можем наметить четыре группы, образованные из квартилей. Первая группа значений содержит наименьшее число до Q1; ко второй группе относятся Q1 до медианы; третий набор — медиана Q3; четвертая категория включает в себя Q3 до самой высокой точки данных всего набора.

Каждый квартиль содержит 25% от общего числа наблюдений. Как правило, данные располагаются от меньшего к большему:

1. **Первый квартиль** : самые низкие 25% чисел.
2. **Второй квартиль** : от 0% до 50% (до медианы)
3. **Третий квартиль** : от 0% до 75%
4. **Четвертый квартиль** : самые высокие 25% чисел.

Дисперсия

Термин дисперсия относится к статистическому измерению разброса между числами в наборе данных. В частности, дисперсия измеряет, насколько далеко каждое число в наборе от среднего (среднего) и, следовательно, от любого другого числа в наборе. Дисперсия часто обозначается этим символом: σ^2 . Он используется как аналитиками, так и трейдерами для определения волатильности и безопасности рынка.

Квадратный корень из дисперсии представляет собой стандартное отклонение (SD или σ), которое помогает определить постоянство доходности инвестиций в течение определенного периода времени.

В статистике дисперсия измеряет отклонение от среднего или среднего значения. Он рассчитывается путем взятия разностей между каждым числом в наборе данных и средним значением, затем возведения в квадрат разностей, чтобы сделать их положительными, и, наконец, деления суммы квадратов на количество значений в наборе данных.

Дисперсия рассчитывается по следующей формуле:

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{N}$$

where:

x_i = Each value in the data set

$\overline{x}$ = Mean of all values in the data set

N = Number of values in the data set

Вы также можете использовать приведенную выше формулу для расчета дисперсии в областях, отличных от инвестиций и торговли, с некоторыми небольшими изменениями. Например, при расчете выборочной дисперсии для оценки дисперсии совокупности знаменатель уравнения дисперсии становится равным N-1, так что оценка является несмещенной и не занижает дисперсию совокупности.

Преимущества и недостатки дисперсии. Статистики используют дисперсию, чтобы увидеть, как отдельные числа соотносятся друг с другом в

наборе данных, а не используют более широкие математические методы, такие как распределение чисел по квартилям. Преимущество дисперсии в том, что она рассматривает все отклонения от среднего значения как одинаковые, независимо от их направления. Квадраты отклонений не могут в сумме равняться нулю.

Однако одним из недостатков дисперсии является то, что она придает дополнительный вес выбросам. Это цифры, далекие от среднего. Возведение этих чисел в квадрат может исказить данные. Еще одна ловушка использования дисперсии заключается в том, что ее сложно интерпретировать. Пользователи часто используют его в первую очередь для извлечения квадратного корня из его значения, которое указывает на стандартное отклонение данных. Как отмечалось выше, инвесторы могут использовать стандартное отклонение, чтобы оценить, насколько постоянна доходность с течением времени.

Визуально, чем больше дисперсия, тем «жирнее» будет распределение вероятностей. В финансах, если что-то вроде инвестиций имеет большую дисперсию, это может быть интерпретировано как более рискованное или волатильное.

Стандартное отклонение

Стандартное отклонение — это квадратный корень из дисперсии. Иногда это более полезно, так как при извлечении квадратного корня единицы измерения удаляются из анализа. Это позволяет проводить прямое сравнение между разными вещами, которые могут иметь разные единицы измерения или разные величины. Например, если сказать, что увеличение X на одну единицу увеличивает Y на два стандартных отклонения, вы сможете понять взаимосвязь между X и Y независимо от того, в каких единицах они выражены.

Стандартное отклонение — это статистическое измерение в финансах, которое применительно к годовой доходности инвестиций проливает свет на историческую волатильность этих инвестиций .

Чем больше стандартное отклонение ценных бумаг, тем больше разница между каждой ценой и средним значением, которое показывает больший ценовой диапазон. Например, волатильная акция имеет высокое стандартное отклонение, тогда как отклонение стабильной акции обычно довольно низкое.

Стандартное отклонение рассчитывается путем извлечения квадратного корня из значения, полученного в результате сравнения точек данных, с коллективным средним значением генеральной совокупности. Формула:

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

where:

x_i = Value of the i^{th} point in the data set

$\overline{x}$ = The mean value of the data set

Использование стандартного отклонения. Стандартное отклонение — особенно полезный инструмент в стратегиях инвестирования и торговли, поскольку он помогает измерять волатильность рынка и ценных бумаг, а также прогнозировать тенденции производительности.

Более низкое стандартное отклонение не обязательно предпочтительно. Все зависит от инвестиций и готовности инвестора взять на себя риск. Имея дело с величиной отклонения в своих портфелях, инвесторы должны учитывать свою терпимость к волатильности и свои общие инвестиционные цели. Более агрессивным инвесторам может быть удобной инвестиционная стратегия, которая выбирает инструменты с волатильностью выше среднего, в то время как более консервативным инвесторам может не понравиться.

Стандартное отклонение — одна из ключевых фундаментальных мер риска, которую используют аналитики, портфельные менеджеры, консультанты. Инвестиционные фирмы сообщают о стандартном отклонении своих взаимных фондов и других продуктов. Большая дисперсия показывает, насколько доходность фонда отклоняется от ожидаемой нормальной доходности. Поскольку это легко понять, эта статистика регулярно сообщается конечным клиентам и инвесторам.

Большое стандартное отклонение указывает на то, что наблюдаемые данные сильно отличаются от среднего значения. Это указывает на то, что наблюдаемые данные весьма разбросаны. Небольшое или низкое стандартное отклонение вместо этого указывает на то, что большая часть наблюдаемых данных плотно сгруппирована вокруг среднего значения.

Стандартное отклонение важно, потому что оно может помочь пользователям оценить риск. Рассмотрим вариант инвестирования со средней годовой доходностью 10% в год. Однако это среднее значение было получено на основе доходности за последние три года в размере 50%, -15% и -5%. Рассчитав стандартное отклонение и поняв низкую вероятность фактического среднего значения в 10% в любой отдельно взятый год, вы будете лучше подготовлены к принятию обоснованных решений и выявлению скрытого риска.

Регрессионный анализ

Регрессия — это статистический метод, используемый в финансах, инвестициях и других дисциплинах, который определияет силу и характер связи между одной зависимой переменной (обычно обозначаемой Y) и рядом других переменных (известных как независимые переменные).

В частном случае, когда фактор единственный (без учёта константы), говорят о парной или простейшей линейной регрессии:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Когда количество факторов (без учёта константы) больше 1-го, то говорят о множественной регрессии.

Парная регрессия представляет собой регрессию между двумя переменными – y и x, т. е. модель вида:

$$y = f(x),$$

где y – зависимая переменная (результативный признак);

x – независимая, или объясняющая, переменная (признак-фактор). Знак «^» означает, что между переменными x и y нет строгой функциональной зависимости, поэтому практически в каждом отдельном случае величина y складывается из двух слагаемых:

$$y = y_x + \varepsilon,$$

где y – фактическое значение результативного признака;

y_x – теоретическое значение результативного признака, найденное исходя из уравнения регрессии;

ε – случайная величина, характеризующая отклонения реального значения результативного признака от теоретического, найденного по уравнению регрессии.

Коэффициент корреляции

В эконометрическом исследовании вопрос о наличии или отсутствии зависимости между анализируемыми переменными решается с помощью методов корреляционного анализа. Только после утвердительного ответа на этот вопрос имеет смысл определять вид зависимости. Корреляционный анализ подробно изучается в курсе математической статистики, напомним некоторые его положения.

Корреляционный анализ *– метод, посвященный изучению взаимосвязей между случайными величинами, применяемый тогда, когда данные наблюдений или эксперимента можно считать случайными и выбранными из совокупности, распределенной по многомерному нормальному закону*.

В корреляционном анализе исследуют следующие варианты зависимостей.

1. *Парную корреляцию* – связь между двумя признаками: результативным и факторным или двумя факторными.

2. *Частную корреляцию* – зависимость между результативным и одним факторным признаком при фиксированных значениях других факторных признаков.

3. *Множественную корреляцию* – зависимость между результативным и множеством факторных признаков.

Основная задача корреляционного анализа *состоит в выявлении тесноты связи между случайными переменными путем точечной и интервальной оценки различных (парных, частных, множественных) коэффициентов корреляции*.

Наиболее разработанной в эконометрике является методология парной линейной корреляции, рассматривающая влияние переменной x на переменную y. В теории вероятностей показателем тесноты *линейной* зависимости между двумя случайными величинами являлся коэффициент корреляции, в математической статистике таким показателем является выборочный коэффициент корреляции.

Выборочным коэффициентом (линейной) корреляции называется величина, рассчитываемая по формуле

$$r = r_{xy} = r_B = r(x, y) = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y},$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ , \ \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \ ,$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \ , \ \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$$

– выборочные средние,

$$\sigma_x = \sqrt{\overline{x^2} - \bar{x}^2} \ , \ \sigma_y = \sqrt{\overline{y^2} - \bar{y}^2}$$

– выборочные средние квадратические отклонения, полученные по наблюдаемым значениям x и y соответственно.

Величина

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

называется ***выборочной ковариацией***, а формула

$$\operatorname{cov}(x, y) = \overline{xy} - \bar{x} \cdot \bar{y}$$

– упрощенная формула для расчета ковариации.

Ковариация характеризует зависимость признаков и зависит от единиц измерения x и y, чтобы получить безразмерную характеристику зависимости ковариацию делят на произведение средних квадратических отклонений признаков, в результате получают коэффициент линейной корреляции:

$$r = r(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}.$$

Выборочный коэффициент линейной корреляции r *является показателем тесноты связи признаков в линейной форме* (на фоне влияния остальных признаков, входящих в модель).

На рисунке 1 приведены две диаграммы рассеяния, отражающие корреляционную зависимость переменных y и x. Очевидно, что в случае а) зависимость между переменными менее тесная и коэффициент корреляции должен быть меньше, чем в случае б), так как точки корреляционного поля а) дальше отстоят от линии регрессии, чем точки поля б).

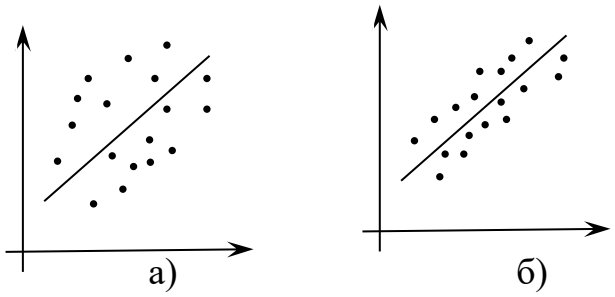


Рисунок 1

Отметим основные свойства выборочного коэффициента корреляции, которые при достаточно большом объеме выборки n, аналогичны свойствам коэффициента корреляции для двух случайных величин.

Свойства выборочного коэффициента корреляции

- ✓ Коэффициент корреляции принимает значения на отрезке $[-1,1]$, то есть $-1 \le r \le 1$.
- ✓ Чем ближе значение $|r|$ к единице, тем более тесная *линейная* зависимость между изучаемыми признаками. В зависимости от того, насколько $|r|$ приближается к единице, говорят, что линейная связь *практически отсутствует* $(0 \le |r| < 0{,}3)$, *слабая* $(0{,}3 \le |r| < 0{,}5)$, *умеренная* $(0{,}5 \le |r| < 0{,}7)$, *тесная* $(0{,}7 \le |r| < 0{,}9)$ и *весьма тесная* $(0{,}9 \le |r| < 0{,}99)$.

✓ Если $r > 0$, то корреляционная связь между переменными называется *прямой*, а если $r < 0$ – *обратной*. При прямой связи увеличение одной из переменных ведет к увеличению условной средней другой, при обратной – наоборот.

✓ Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится. Коэффициент корреляции есть безразмерная характеристика тесноты линейной связи.

✓ При $r_{xy} = \pm 1$ корреляционная связь представляет линейную функциональную зависимость, при этом все точки поля корреляции лежат на одной прямой. И наоборот, если x и y связаны линейной функциональной зависимостью, то $|r| = 1$.

✓ Парный коэффициент корреляции является симметричной характеристикой, то есть $r_{xy} = r_{yx}$.

✓ При совпадающих признаках коэффициент корреляции равен единице – $r_{xx} = 1$.

✓ При $r_{xy} = 0$ *линейная* корреляционная связь отсутствует, а признаки x и y называют ***некоррелированными***. Но это не означает отсутствие вообще корреляционной, а тем более статистической зависимости. Например, нелинейная корреляционная связь может быть очень тесной.

✓ Если случайные величины x и y статистически независимы, то $r_{xy} = 0$. Обратное верно не всегда.

✓ *В случае нормального распределения из некоррелированности x и y следует их независимость*.

Многомерный случай

В случае анализа зависимости компонент m-мерного случайного вектора $x = (x_1, x_2, ..., x_m)$ используется ***матрица парных коэффициентов корреляции***

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} \\ r_{21} & 1 & r_{23} & \dots & r_{2m} \\ r_{31} & r_{32} & 1 & \dots & r_{3m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{m1} & r_{m2} & r_{m3} & \dots & 1 \end{pmatrix},$$

где $r_{ij} = r(x_i, x_j)$ – парный коэффициент корреляции между признаками x_i, x_j.

Матрица парных коэффициентов корреляции обладает свойствами, вытекающими из свойств парного коэффициента корреляции.

Свойства матрицы парных коэффициентов корреляции

1. Матрица R симметрична относительно главной диагонали, так как $r_{ij} = r_{ji}$.
2. На главной диагонали R стоят единицы, так как $r_{ii} = 1$.

Если компоненты вектора x попарно независимы (тогда $r_{ij} = 0,\ i \neq j$), то R – единичная матрица.

Парная линейная регрессия. Оценка параметров

Модель парной линейной регрессии является наиболее распространенным видом зависимости. Пусть имеются n наблюдений над переменными x и y, то есть пары $(x_i, y_i),\ i = 1,2,\ldots,n$. Рассмотрим на примере.

*Пример***.** Исследуется зависимость расходов домашнего хозяйства на покупку продовольственных товаров y (% к общему объему расходов) от размера среднемесячной заработной платы одного работающего x (у.е.). Опытные данные, за рассматриваемый год по десяти районам области N представлены в таблице.

Таблица 1

№	x (у.е.)	y (%)
1	4,5	68,8
2	5,9	58,3
3	5,7	62,6
4	7,2	52,1
5	6,2	54,5
6	6,0	57,1
7	7,8	51,0
8	7,5	50,7
9	8,1	48,6
10	7,9	49,1

Решение. Отобразим пары наблюдений точками на графике, получим диаграмму рассеяния, представленную на рисунке 16.

На основании анализа диаграммы рассеяния можем сделать предположение, что в среднем y описывается линейной функцией от x, то есть имеет место уравнение парной линейной регрессии

$$M(y/x) = \beta_0 + \beta_1 x,$$

где $M(y/x)$ – условное математическое ожидание случайной величины y при заданном значении x.

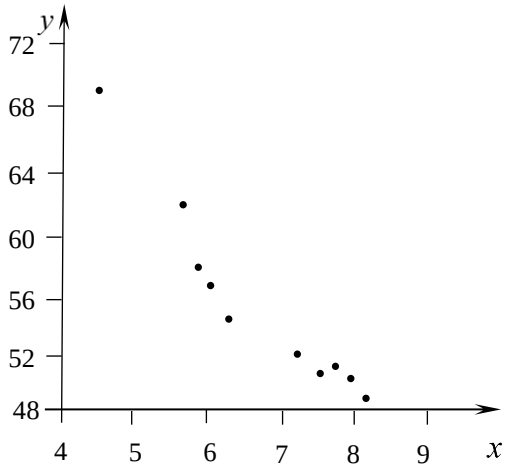


Рисунок 1 – Диаграмма рассеяния

Для отражения факта, что каждое индивидуальное значение y_i отклоняется от соответствующего математического ожидания, необходимо ввести случайное слагаемое ε_i, и тогда для наблюдений (x_i, y_i) уравнение регрессии имеет вид

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i .$$

Соотношение

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

– ***теоретическая линейная регрессионная модель***,

β_0, β_1 – ***теоретические параметры регрессии***,

ε_i – ***случайный член*** (*случайная составляющая, случайная переменная, случайная ошибка*) для i-го наблюдения. В общем виде теоретическую линейную модель будем представлять в виде:

$$y = \beta_0 + \beta_1 x + \varepsilon .$$

Для определения значений теоретических коэффициентов регрессии необходимо знать и использовать все значения переменных y и x генеральной совокупности, что практически невозможно. В этом случае речь может идти об оценке (приближенном выражении) функции регрессии на основе имеющейся выборки данных. Таким образом, **задача линейного регрессионного анализа** состоит в том, чтобы по имеющимся статистическим данным (x_i, y_i) объема n построить ***эмпирическое уравнение регрессии***

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

где $\hat{y}_i$ – оценка условного математического ожидания $M(y/x=x_i)$ или ***прогноз*** значения y_i в точке x_i, $\hat{\beta}_0, \hat{\beta}_1$ – ***оценки неизвестных параметров*** β_0, β_1 или ***эмпирические параметры*** (***коэффициенты) линейной регрессии***. Следовательно,

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i,$$

где $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ – оценки ε_i, которые называют ***остатками*** регрессии (***отклонениями***), $i = \overline{1,n}$. Не следует путать остатки e_i со случайными составляющими ε_i. Остатки e_i, так же как и переменные ε_i, являются случайными величинами, однако разница состоит в том, что остатки, в отличие от составляющих ε_i наблюдаемы.

Оценить параметры β_0, β_1 в данном случае означает выбрать «наилучшие» значения параметров – $\hat{\beta}_0, \hat{\beta}_1$, при которых линия регрессии $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ будет ближайшей к точкам наблюдений y_i по их совокупности, где $e_i = y_i - \hat{y}_i$ – отклонение наблюдаемой точки y_i от значения $\hat{y}_i$, найденного по эмпирическому уравнению регрессии .

Например, оценки $\hat{\beta}_0, \hat{\beta}_1$ могут быть найдены из условий минимизации одной из следующих сумм:

1) $E = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} |y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i|$,

2) $E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$,

3) $E = \sum_{i=1}^{n} g(e_i) = \sum_{i=1}^{n} g(y_i - \hat{y}_i) = \sum_{i=1}^{n} g(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$, где g – «мера», в соответствии с которой отклонение e_i формирует тип функционала E.

На минимизации первой суммы основан *метод наименьших модулей*. Плюсом данного метода является робастность, то есть нечувствительность к «выбросам». К минусам – сложность вычислительной процедуры, неоднозначность выводов.

На минимизации второй суммы основан *метод наименьших квадратов*. Этот метод является наиболее простым с вычислительной точки зрения. Кроме того, его оценки обладают рядом оптимальных статистических свойств. Простота математических выводов делает возможным построить развитую теорию, позволяющую провести тщательную проверку различных статистических гипотез. Минусом данного метода является чувствительность к «выбросам».

В третьей сумме в качестве функции g можно использовать функцию Хубера, которая при малых отклонениях квадратична, а при больших линейна:

$$g(x)=\begin{cases} x^2,\ |x|<c, \\ 2cx-c^2,\ x\geq c, \\ -2cx-c^2,\ x\leq -c. \end{cases}$$

Функция Хубера является попыткой совместить достоинства первых двух функционалов.

Среди других методов следует отметить метод моментов (ММ) и метод максимального правдоподобия (ММП), рассматриваемые в курсе математической статистики.

Метод наименьших квадратов

Метод наименьших квадратов — это форма математического регрессионного анализа, используемая для определения линии наилучшего соответствия набору данных, обеспечивающая визуальную демонстрацию взаимосвязи между точками данных. Каждая точка данных представляет отношение между известной независимой переменной и неизвестной зависимой переменной.

Метод наименьших квадратов — это статистическая процедура поиска наилучшего соответствия набора точек данных путем минимизации суммы смещений или невязок точек на построенной кривой.

Согласно методу наименьших квадратов (МНК) *в качестве оценок неизвестных параметров* β_0, β_1 *следует брать такие значения* $\hat{\beta}_0, \hat{\beta}_1$, *которые минимизируют сумму квадратов отклонений фактических значений результативного признака* y_i *от значений* $\hat{y}_i$, *рассчитанных по уравнению регрессии* (рис.2), то есть доставляют минимум функционала

$$Q = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \longrightarrow \min_{\hat{\beta}_0,\, \hat{\beta}_1}.$$

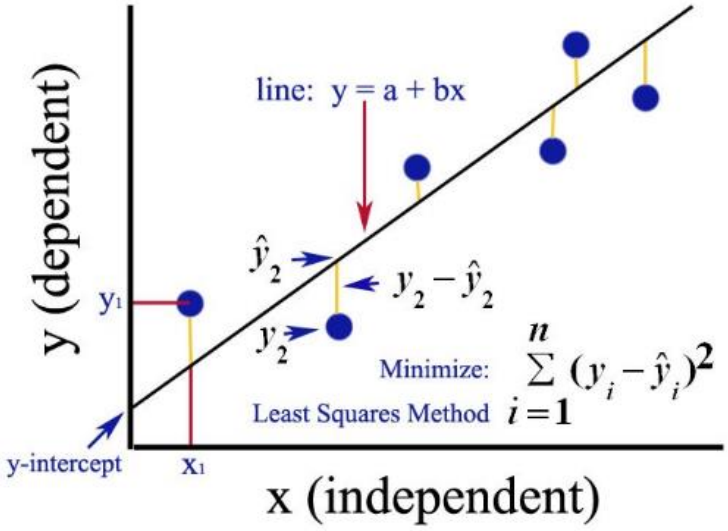


Рисунок 2 – Визуализация МНК

Функция $Q(\hat{\beta}_0, \hat{\beta}_1)$ является квадратичной функцией параметров $\hat{\beta}_0$, $\hat{\beta}_1$ (x_i, y_i – известные данные наблюдений). Так как $Q(\hat{\beta}_0, \hat{\beta}_1)$ непрерывна, выпукла и ограничена снизу, $Q(\hat{\beta}_0, \hat{\beta}_1) \geq 0$, то она имеет минимум. Необходимым условием существования минимума функции двух переменных является равенство нулю ее частных производных по неизвестным параметрам $\hat{\beta}_0$ и $\hat{\beta}_1$.

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{\beta}_0} = -2\sum\limits_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \\ \dfrac{\partial Q}{\partial \hat{\beta}_1} = -2\sum\limits_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0. \end{cases} \Rightarrow \begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ \hat{\beta}_0 \sum\limits_{i=1}^{n} x_i + \hat{\beta}_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases}$$

Полученная система уравнений называется ***системой нормальных уравнений*** МНК. Разделив оба уравнения последней системы на n, имеем:

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}, \\ \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \overline{x^2} = \overline{xy}. \end{cases}$$

Решая систему относительно $\hat{\beta}_0$ и $\hat{\beta}_1$, получим:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \dfrac{\operatorname{cov}(x, y)}{\sigma_x^2}, \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \end{cases}$$

Таким образом, МНК-оценки параметров $\hat{\beta}_0$ и $\hat{\beta}_1$ определяются по формулам (3.10). Отметим, что, как известно из курса математической статистики, *в случае нормального закона распределения случайных величин ε_i, оценки МНК и ММП совпадают*.

Зная выборочный коэффициент линейной корреляции, определяемый по (3.7), формулу для $\hat{\beta}_1$ можно записать в виде:

$$\hat{\beta}_1 = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_x} = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = r \frac{\sigma_y}{\sigma_x}.$$

Эмпирический (***выборочный***) ***коэффициент регрессии*** $\hat{\beta}_1$, стоящий в уравнении (3.9) при x, показывает среднее изменение результата с изменением фактора на одну единицу измерения. Знак при коэффициенте регрессии $\hat{\beta}_1$ показывает направление связи: $\hat{\beta}_1 > 0$ – связь прямая, $\hat{\beta}_1 < 0$ – связь обратная.

Коэффициент $\hat{\beta}_0$ – ***свободный член уравнения регрессии*** (3.9), который указывает на значение результирующего признака при нулевом значении фактора. Это важный показатель для выбора вида уравнения регрессии. Например, если в результате вычислений коэффициент $\hat{\beta}_0$ оказался отрицательным, а

экономический смысл задачи диктует положительность или равенство нулю показателя $\hat{\beta}_0$, значит, выбор вида уравнения был неудачен. Например, в регрессионной модели производительности труда следует ожидать свободный член равным нулю, если равны нулю производственные площади или численность рабочих.

Анализ вариации зависимой переменной. Коэффициент детерминации

Определим полную сумму квадратов (total sum of squares) отклонений y_t от выборочного среднего значения $\overline{y}$:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

Полную сумму квадратов можно представить как

$$SST = SSR + SSE,$$

где $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ – сумма квадратов, обусловленная включенными в модель объясняющими переменными (regression sum of squares),

$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ – сумма квадратов остатков (error sum of squares), $\overline{y}$ – выборочное среднее наблюдений переменной y_t.

Это представление справедливо только в случае, если уравнение множественной линейной регрессии содержит свободный член β_0.

Коэффициент детерминации, – это величина, определяемая по формуле

$$R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS} .$$

Он характеризует качество подгонки регрессионной модели к наблюдаемым значениям y_i. Отметим, что коэффициент R^2 корректно определен только в том случае, если свободный член (константа) включен в уравнение регрессии.

Коэффициент детерминации указывает на долю вариации зависимой переменной, объясняемую включенными в модель факторами. Он принимает значения между 0 и 1. Если $R^2 = 0$, то это означает, что модель не улучшает качество предсказания y_t по сравнению с тривиальным $\hat{y}_t = \overline{y}$. Если $R^2 = 1$, то это означает точную подгонку уравнения, т.е. все $e_t = 0$. Однако на R^2 нельзя ориентироваться как на главный критерий при сравнении двух различных структур модели. Коэффициент детерминации целесообразно использовать только совместно с дополнительным анализом регрессионного уравнения.

При сравнении качества двух регрессий на основе статистики R^2 следует помнить:

Проверка значимости уравнения регрессии

Для проверки значимости уравнения регрессии используют F-критерий. Нулевая гипотеза состоит в том, что коэффициенты при всех регрессорах равны нулю, то есть проверяется гипотеза

$$H_0:\ \beta_1 = 0,\ \beta_2 = 0, ...,\ \beta_m = 0.$$

Следовательно, F-статистика, определяемая по формуле, имеет распределение Фишера:

$$F = \frac{R^2}{1-R^2}\frac{n-m-1}{m} =$$
$$= \frac{RSS}{ESS}\frac{n-m-1}{m} = \frac{RSS/m}{ESS/(n-m-1)} =$$
$$= \frac{\dfrac{(\hat{Y}-\bar{y})^{\mathrm{T}}(\hat{Y}-\bar{y})}{\sigma^2}\cdot\dfrac{1}{m}}{\dfrac{e^{\mathrm{T}}e}{\sigma^2}\cdot\dfrac{1}{n-m-1}} \in F(m;\ n-m-1),$$

и ее можно использовать для проверки гипотезы. Гипотеза H_0 отвергается, например, при уровне значимости α, если $F > F(\alpha;\ m;\ n-m-1)$, где $F(\alpha;\ m;\ n-m-1)$ – $\alpha\%$-я точка распределения Фишера $F(m;\ n-m-1)$

Замечание. Число степеней свободы остаточной суммы квадратов ESS равно разности между числом наблюдений и числом линейных связей между ними, участвующими в определении $ESS = \sum(y_i - \hat{y}_i)^2$, то есть $n-(m+1) = n-m-1$ (для определения $\hat{y}_i$ требуется решить систему $m+1$ линейных уравнений для определения оценок $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$). Для множественной регрессии сумма $TSS = \sum(y_i - \bar{y})^2$ имеет $n-1$ степеней свободы, так как в этой сумме все наблюдения связаны одной связью (при определении значения $\bar{y}$). Для суммы $RSS = \sum(\hat{y}_i - \bar{y})^2$ число степеней свободы равно $(m+1)-1 = m$, так как в выражение $\hat{y}_i$ входят $m+1$ оценок неизвестных параметров и одна линейная связь, определяемая $\bar{y}$.

Статистику (4) используют для проверки гипотезы о статистической значимости коэффициента детерминации:

$$H_0:\ R^2 = 0.$$

Это равнозначно проверке значимости уравнения регрессии и полностью совпадает с описанной выше проверкой.

Если уравнение регрессии незначимо, то есть все коэффициенты при регрессорах для генеральной совокупности равны нулю, то на этом анализ уравнения регрессии заканчивается. Если же нулевая гипотеза

$$H_0: \ \beta_1 = 0, \ \beta_2 = 0, ..., \ \beta_m = 0$$

отвергается, то представляют интерес проверка значимости отдельных коэффициентов регрессии и построение интервальных оценок для значимых коэффициентов.

Проверка адекватности линейной регрессионной модели на основе статистических тестов.

Проверку модели на адекватность по F-критерию Фишера целесообразно осуществлять по следующему алгоритму.

Алгоритм тестирования по F-критерию Фишера

Шаг 1. Формулировка нулевой и альтернативной гипотез

H_0: $\hat{a}_0 = \hat{a}_1 = \ldots = \hat{a}_m = 0$, то есть ни один фактор модели не влияет на показатель, или все параметры модели незначимы.

H_A: хотя бы одно значение $\hat{a}_j$ отличное от нуля, то есть $\hat{a}_j \neq 0, \ j = \overline{0;m}$.

Шаг 2. Выбор подходящего уровня значимости

Уровнем значимости α называется вероятность сделать ошибку 1-го рода, то есть отклонить правильную гипотезу. Величина $P = 1 - \alpha$ называется уровнем доверия или доверительной вероятностью.

Шаг 3. Расчет расчетного значения F-критерия

Расчетное значение F-критерия, так называемое F-отношение, определяется по формуле:

$$F_{расч} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m},$$

где R^2 – коэффициент множественной детерминации.

Примечание. При использовании электронных таблиц Excel расчетное значение F-критерия можно найти в таблице Дисперсионный анализ вывода итогов пакета Анализ данных - регрессии.

Шаг 4. Определение по статистическим таблицам F-распределения Фишера критического значения F-критерия

Критическое значение F-критерия находят по статистическим таблицам F-распределения Фишера за соответствующими значениями:

– доверительной вероятности P;

– степеней свободы $k_1 = m$ и $k_2 = n - m - 1$.

Шаг 5. Сравнение расчетного значения F-критерия с критическим и интерпретация результатов тестирования

Если $F_{расч} < F_{кр}$, то нет оснований отклонить нулевую гипотезу о том, что ни один фактор модели не является значимым, то есть с принятой надежностью можно утверждать, что модель неадекватна;

Если $F_{расч} > F_{кр}$, то нулевая гипотеза о незначительности факторов отклоняется есть с принятой надежностью можно утверждать, что модель адекватна.

Проверка значимости оценок параметров модели по t-критерия Стьюдента

Т-распределение Стьюдента позволяет протестировать гипотезы о значимости каждого параметра модели и построить их интервалы доверия.

Алгоритм тестирования по t-критерию Стьюдента

Шаг 1. Формулировка нулевой и альтернативной гипотез

H_0: $\hat{a}_0 = 0,\ j = \overline{0;\ 2}$, – оценка j-го параметра является статистически незначимой, и j-й фактор никак не влияет на показатель y;

H_1: $\hat{a}_0 \neq 0,\ j = \overline{0;\ 2}$ – оценка j-го параметра является статистически значимой, и j-й фактор никак влияет на показатель y.

Шаг 2. Выбор подходящего уровня значимости

Уровень значимости избирается аналогично F-критерия.

Шаг 3. Расчет расчетного значения t-критерия

Расчетные значения t-критерия определяются по формулам:

$$\hat{t}_{\hat{a}_j} = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{a}_j}},\quad j = \overline{0;m}$$

.

При анализе двухфакторной модели расчетные значения t-критерия определяются по формулам:

$$\hat{t}_{\hat{a}0} = \frac{\hat{a}_0}{\hat{\sigma}_{\hat{a}_0}},\qquad \hat{t}_{\hat{a}1} = \frac{\hat{a}_1}{\hat{\sigma}_{\hat{a}_1}},\qquad \hat{t}_{\hat{a}2} = \frac{\hat{a}_2}{\hat{\sigma}_{\hat{a}_2}}.$$

Шаг 4. Определение по статистическим таблицам t-распределения Стьюдента критического значения t-критерия

Критическое значение t-критерия находят по статистическим таблицам t-распределения Стьюдента по соответствующим значениями:

– доверительной вероятности P;

– числом степеней свободы $k = n - m - 1$.

Шаг 5. Сравнение расчетного значения t-критерия с критическим и интерпретация результатов тестирования

Если $|t_{\text{расч}}| < t_{\text{кр}}$, то нет оснований отклонять нулевую гипотезу, то есть с принятой надежностью можно утверждать, что оценка j-го параметра является статистически незначимой, j-й фактор не влияет на показатель *y*.

Если $|t_{\text{расч}}| > t_{\text{кр}}$, то нулевая гипотеза отклоняется, то есть с принятой надежностью можно утверждать, что оценка j-го параметра является статистически значимой, j-й фактор влияет на показатель *y*.

При отклонении H_0 коэффициент β_1 считается ***статистически значимым***, что указывает на наличие определенной линейной связи между y и x: $M(y/x) = \beta_0 + \beta_1 x$. Для уравнения парной линейной регрессии тестирование статистической значимости коэффициента β_1 *эквивалентно тестированию значимости построенного линейного уравнения регрессии* в целом, так как именно в коэффициенте β_1 скрыто влияние фактора x на результативную переменную y.

Если $|t_{\hat{\beta}_1}| < t_{\alpha;n-2}$, гипотеза $H_0 : \beta_1 = 0$ не отвергается, и уравнение регрессии считают статистически незначимым – на этом регрессионный анализ заканчивается.

Для значимого уравнения регрессии представляет интерес построение интервальной оценки коэффициента β_1 и дальнейший регрессионный анализ.

Замечание. Компьютерные статистические и эконометрические пакеты при проверке гипотезы H_0 о статистической значимости коэффициента регрессии β_1 вычисляют наблюдаемое значение критерия Стьюдента $t_{\hat{\beta}_1}$ и двухстороннее P-значение t-статистики (P-level), то есть вероятности того, что случайная величина, распределенная по закону $t(n-m-1)$, принимает значение по абсолютной величине большее, чем $|t_{\hat{\beta}_1}| = \left|\frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}\right|$, то есть $P = P(|t(n-m-1)| > |t_{\hat{\beta}_1}|)$. Если эта вероятность мала (меньше выбранного уровня значимости, например $\alpha = 0{,}05$), то коэффициент считается значимым; в противном случае – незначимым.

Вообще, еще раз напомним, при проверке статистической гипотезы H_0 следует пользоваться следующим правилом:

$$\begin{cases} H_0 \text{ "отклоняется", если } P \le \alpha; \\ H_0 \text{ "не отклоняется", если } P > \alpha. \end{cases}$$

Замечание. При проверке статистической значимости коэффициентов парной регрессии «на глаз» рассчитанные $\left|t_{\hat{\beta}_0}\right|$, $\left|t_{\hat{\beta}_1}\right|$ сравнивают с двойкой, так как $t_{0,05;\,n-2} \approx 2$ для больших n. Если, например, $\left|t_{\hat{\beta}_1}\right| > 2$, то β_1 статистически значим.