2.12. Корреляция.

Понятие корреляционной связи

В статистике различают функциональную и стохастическую связи.

Функциональной называют такую связь, при которой имеется однозначное соответствие между факторными и результативными признаками.

При *стохастической связи* причинная зависимость между факторными и результативными признаками проявляется не в каждом отдельном случае, а лишь при большом числе наблюдений. В каждом конкретном случае при изменении одной переменной вторая может принимать в определенных пределах любые значения с некоторой вероятностью.

Корреляционной связью называют такой частный случай стохастической связи, при которой различным значениям факторного признака соответствуют различные средние значения результативного признака.

По направлению выделяют связь прямую и обратную.

При прямой связи увеличение или уменьшение факторного признака приводит к увеличению или уменьшению результативного признака (или его среднего значения).

При обратной связи увеличение факторного признака приводит к уменьшению результативного.

По аналитическому выражению связи могут быть линейными и нелинейными.

Если статистическая связь между явлениями может быть приближенно выражена прямой линией, то связь называется *линейной*, если же она выражается уравнением какой-либо другой линии (параболы, гиперболы и т. д.), то связь называют *нелинейной.*

Принято различать:

а) парную корреляцию - связь между результативным и факторным признаками;

б) частную корреляцию - связь между результативным признаком и одним факторным признаком при фиксированном значении всех других факторных признаков;

в) множественную корреляцию - связь между результативным признаком и двумя и более факторными признаками.

Задачей эконометрического анализа является определение аналитического выражения уравнения связи, которое может зависеть от одного факторного признака (однофакторная регрессия) или от двух и более факторных признаков (множественная регрессия).

В некоторых случаях можно ограничиться лишь качественными результатами о наличии корреляции между признаками и ее направлении.

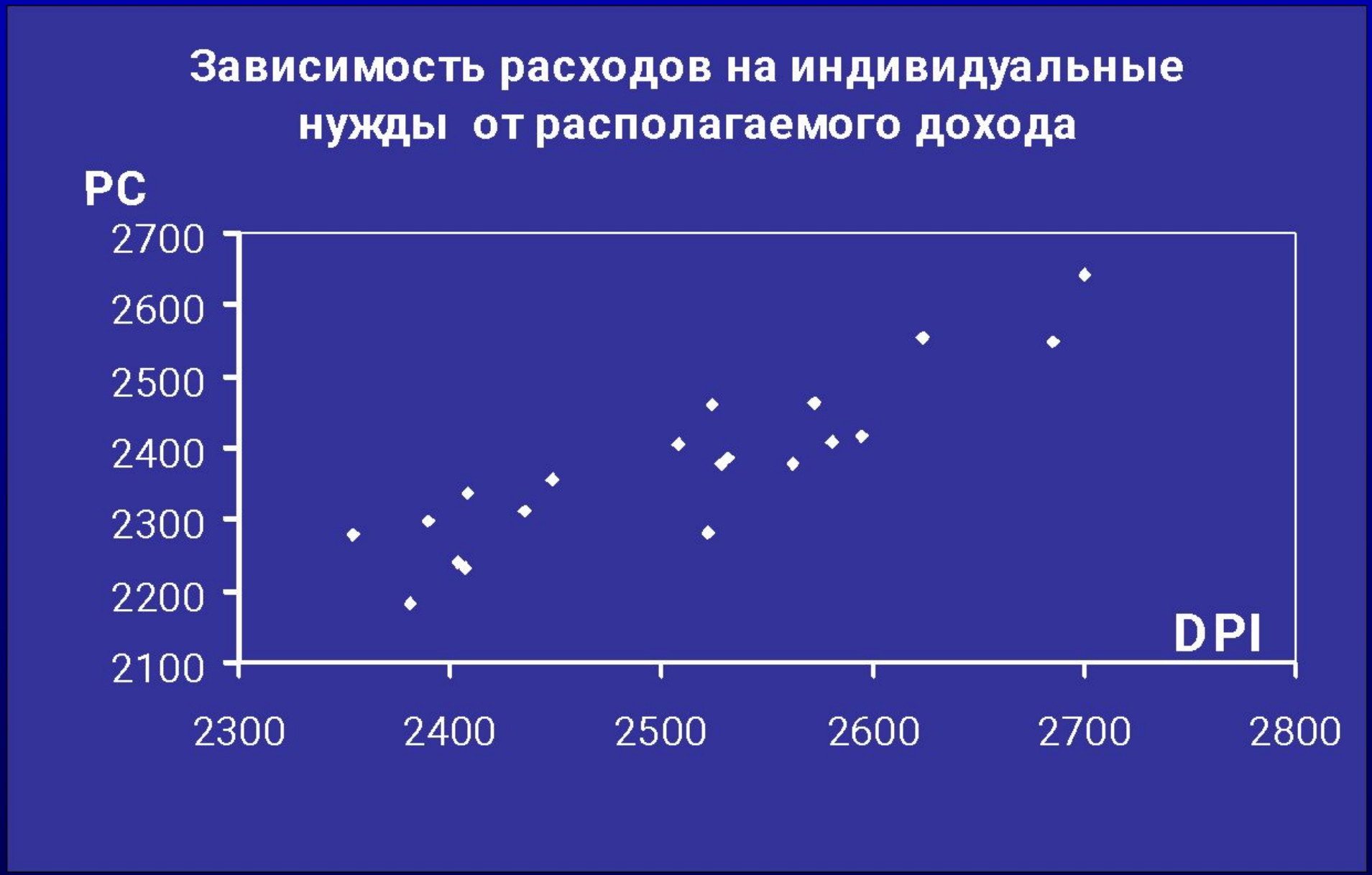
Для получения такой информации используются метод построения поля корреляции т.е. точечной диаграммы. Причем по оси X откладывается значение факторного признака а по оси Y результативного.

Вернемся к примеру рассмотренному во введении. На основании данных о годовом располагаемом доходе и годовых расходах на личное потребление в 1999 г. для 20 семей (в условных единицах), требуется выяснить существует ли взаимосвязь между располагаемым доходом и расходами на личное потребление.

Обозначения: DPI (disposable personal income) - доходы PC (personal consumption) - расходы; усл. ед.

№	DPI	PC	№	DPI	PC
1	2508	2406	11	2435	2311
2	2572	2464	12	2354	2278
3	2408	2336	13	2404	2240
4	2522	2228	14	2381	2183
5	2700	2641	15	2581	2408
6	2531	2385	16	2529	2379
7	2390	2297	17	2562	2378
8	2595	2416	18	2624	2554
9	2524	2460	19	2407	2232
10	2685	2549	20	2448	2356

Графическое изображение корреляционного поля



Расположение точек на графике отражает общую тенденцию вариации факторного и результативного признаков.

Теперь хорошо видно, что корреляция (взаимосвязь) признаков существует, но хотелось бы получить количественную оценку тесноты этой связи.

Для количественной оценки тесноты корреляции в случае, когда связь линейна вычисляют коэффициент корреляции r.

Определим линейный коэффициент корреляции как среднее значение произведения нормированных отклонений результативного и факторного признаков от их средних значений:

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)}{n}.$$

Линейный коэффициент корреляции может принимать значения в пределах от -1 до +1 .

При наличии функциональной связи коэффициент корреляции равен по модулю единице, а при отсутствии связи - нулю.

Эмпирическая схема определения тесноты связи

Величина коэффициента корреляции	Характер связи
До $\|\pm 0{,}3\|$	Практически отсутствует
$\|\pm 0{,}3\| - \|\pm 0{,}5\|$	Слабая
$\|\pm 0{,}5\| - \|\pm 0{,}7\|$	Умеренная
$\|\pm 0{,}7\| - \|\pm 1\|$	Сильная

Задача На основе приведенной ниже таблицы найти линейный коэффициент корреляции расходов на питание и годовых доходов.

№ Семьи	Доход семьи, руб. (X)	Расходы на питание, руб. (Y)
1	30 000	8 500
2	25 000	7 000
3	40 000	9 500
4	60 000	11 500
5	33 000	8 000
6	45 000	9 500

Найдем среднее значение и дисперсию признаков X и Y, используя стандартные функции Excel Срзнач () и Диспр (). В результате получаем следующие значения

$$\bar{x} = 38833.3,\ \bar{y} = 9000,$$

$$\sigma_x^2 = 131\,805\,556,\ \sigma_y^2 = 2\ 000\ 000;$$

$$r = 0{,}970.$$

Коэффициент корреляции можно найти и с помощью стандартной функции Коррел(). Как и следовало ожидать, корреляция между доходами и расходами на питание является сильной.

2.13. Статистическая проверка гипотез

Под статистической гипотезой понимают различного рода предположения о характере или параметрах распределения случайной величины , которые можно проверить, опираясь на результаты выборочного наблюдения.

Статистическая проверка гипотез носит вероятностный характер и поэтому всегда существует риск совершить ошибку. Однако с помощью статистической теории можно оценить вероятность принятия ложного решения. Если эта вероятность мала, то решение можно считать статистически обоснованным.

При проверке гипотез ошибки могут быть двоякого рода:

а) ошибка первого рода – проверяемая гипотеза (ее обычно называют нулевой гипотезой) является в действительности верной, но в результате статистической проверки принимается решение об отказе от нее (нулевая гипотеза отвергается).

б) Ошибка второго рода — нулевая гипотеза в действительности является ошибочной, но в результате статистической проверки она принимается.

Статистическая проверка гипотез осуществляется на основании некоторых критериев.

Для построения такого критерия необходимо:

а) сформулировать нулевую гипотезу (ее обычно обозначают символом H_0);

б) сформулировать альтернативную гипотезу (ее обычно обозначают символом H_1);

в) выбрать уровень значимости α, контролирующей допустимую ошибку первого рода;

г) определить область допустимых значений и критическую область для изучаемого показателя;

д) принять то или иное решение на основании сравнения наблюдаемого и критического значения показателя.

Уровнем значимости α будем называть такое малое значение вероятности попадания критерия в критическую область при условии справедливости гипотезы, что появление этого события можно расценивать как существенное расхождение выдвинутой гипотезы с результатом выборочного наблюдения. Обычно уровень значимости принимают равным 0,05 или 0,01.

К критической области относят те значения изучаемого показателя, которые при условии верности гипотезы являются весьма мало вероятными

Вероятность α совершить ошибку первого рода т. е. отвергнуть гипотезу H_0 когда она верна, называется уровнем значимости критерия.

Мощностью критерия называется вероятность 1 – β не допустить ошибку 2-го рода т.е. отвергнуть гипотезу H_0, когда она неверна.

Если принять юридическую терминологию, то α - это вероятность осудить невиновного, а β - вероятность оправдать виновного.

Величина ошибки первого и второго рода однозначно определяется выбором критической области. Совершенно естественно их хочется сделать одновременно по возможности малыми. Однако это требование является противоречивым. Уменьшение одной величины приводит к росту другой. Лишь увеличение объема выборки позволяет уменьшать обе величины одновременно.

Важно отметить, что ***проверка статистической гипотезы не дает логического доказательства ее верности или неверности.***

К понятию критической области

2.14. Статистическая оценка значимости линейного коэффициента корреляции

Для ответа на вопрос о значимости коэффициента корреляции необходимо при заданном уровне значимости проверить нулевую гипотезу H_0 (о равенстве нулю генерального коэффициента корреляции) при конкурирующей гипотезе H_1 (об отличии от нуля генерального коэффициента корреляции).

Если нулевая гипотеза будет отвергнута, то это означает, что выборочный коэффициент корреляции значимо отличается от нуля.

Для проверки нулевой гипотезы рассмотрим величину

$$t = r_B \cdot \frac{\sqrt{n-2}}{\sqrt{1-r_B^2}}$$

При справедливости нулевой гипотезы случайная величина t подчиняется распределению Стьюдента с $k = n\text{-}2$ степенями свободы, где n – объем выборки; (предполагается, что в генеральной совокупности распределение является нормальным).

Отсюда следует простое правило: для того, чтобы при заданном уровне значимости проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $r_в \neq 0$, следует вычислить эмпирическое значение критерия

$$t_{эмп} = r_в \cdot \frac{\sqrt{n-2}}{\sqrt{1-r_в^2}}.$$

Затем по таблице критических точек распределения Стьюдента при данном числе степеней свободы и уровне значимости найти значение критической точки $t_{кр}$. Если

$$|t_{эмп}| > t_{кр},$$

то нулевую гипотезу следует отвергнуть и это значит, что выборочный коэффициент корреляции значим. В противном случае отличие от нуля выборочного коэффициента корреляции можно объяснить действием случайных причин.

Применим изложенный выше подход к рассматриваемой задаче . Подставляя численные значения $n = 6,\ r_в = 0{,}970$, получаем $t_{эмп}$ = 7,988. Зададимся уровнем значимости 0,01. По таблице критических точек распределения Стьюдента находим, что при числе степеней свободы K=4, уровне значимости равном 0,01 значение $t_{кр}$ =4,404 . Поэтому нулевая гипотеза должна быть отвергнута, и можно говорить, что в генеральной совокупности существует прямая связь между доходами семьи и затратами на питание.

3. Парный Регрессионный анализ

Рассмотрим теперь задачу об определении уравнения линии регрессии. Теоретической линией регрессии называется такая линия, вокруг которой группируются точки корреляционного поля и которая указывает основное направление связи. Чаще всего уравнение регрессионной линии определяется по методу наименьших квадратов.

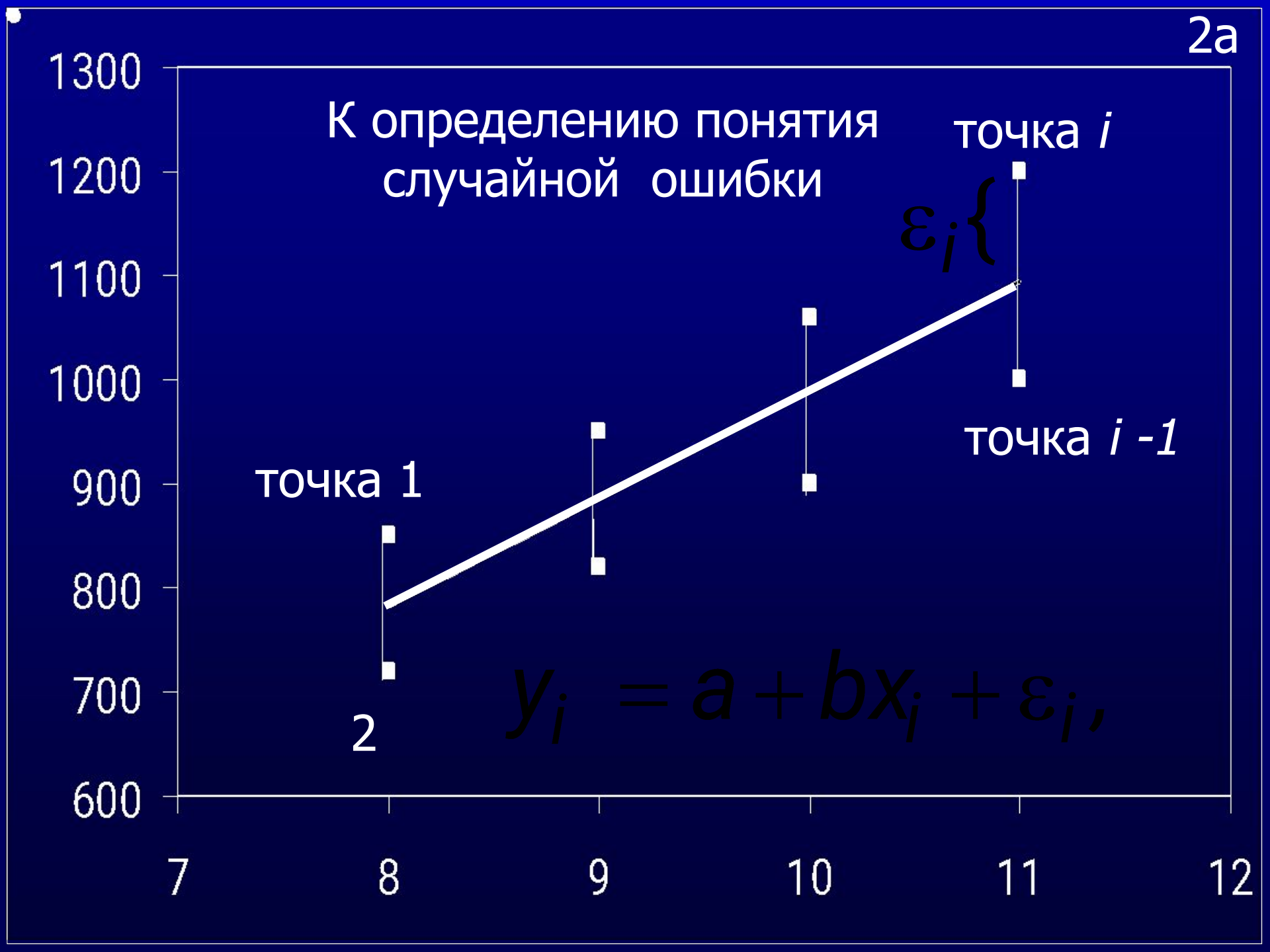
Обсудим применения этого метода для случая, когда предполагается линейная связь между факторным и результативным признаками. Пусть имеется два набора данных x_i и y_j, $i=1,2...n$ Требуется найти уравнение прямой

$$y_i^T = a + bx_i,$$

для которой сумма квадратов отклонений (ошибок)

$$S = \sum_{i=1}^{n} (y_i^T - y_i)^2$$

была бы минимальной.


К определению понятия
случайной ошибки
точка i
точка i -1
точка 1
2
ε_i
$y_i = a + bx_i + \varepsilon_i,$
1300
1200
1100
1000
900
800
700
600
7
8
9
10
11
12

Очевидно, что S является функцией двух переменных, и поэтому условие минимума дает два уравнения:

$$\frac{dS}{da} = 0; \quad \frac{dS}{db} = 0.$$

После несложных преобразований получаем систему нормальных уравнений способа наименьших квадратов для определения двух неизвестных параметров прямой a и b:

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i; \quad \sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2.$$

Действительно. Подставим $y_i^T = a + bx_i$

в выражение для S и продифференцируем это выражение по a:

$$\frac{d}{da}S = \sum_{i=1}^{n} \frac{d}{da}(a + bx_i - y_i)^2 =$$

$$= 2\sum_{i=1}^{n}(a + bx_i - y_i) = 0.$$

Отсюда получаем первое уравнение:

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i.$$

Аналогично выводится и второе уравнение.

Таким образом, получаем следующую систему нормальных уравнений для определения коэффициентов регрессии

$$a \cdot n + b \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i;$$

$$a \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Решая систему двух уравнений относительно неизвестных коэффициентов a и b, получаем расчетные формулы

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}; \quad a = \bar{y} - b\bar{x},$$

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

Параметр b называют коэффициентом регрессии. Коэффициент регрессии используют для определения параметра эластичности

$$K_Э = b\frac{\bar{x}}{\bar{y}}.$$

Между коэффициентом регрессии и линейным параметром корреляции существует простое соотношение:

$$b = r\frac{\sigma_y}{\sigma_x},$$ где σ_y, σ_x –

коэффициенты среднего квадратического отклонения факторного и результативного признаков.

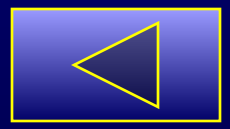
Воспользуемся данными табл. на слайде 16 и 7 найдем параметры линейной регрессионной модели для этой задачи. Коэффициент корреляции и другие необходимые параметры мы вычисляли ранее см. слайд 17 : Напомним результат

$$\bar{x} = 38833.3,\ \bar{y} = 9000,$$

$$\sigma_x^2 = 131\,805\,556,\ \sigma_y^2 = 2\ 000\,000;$$

$$r = 0{,}970.$$

В результате получаем параметры уравнения регрессии



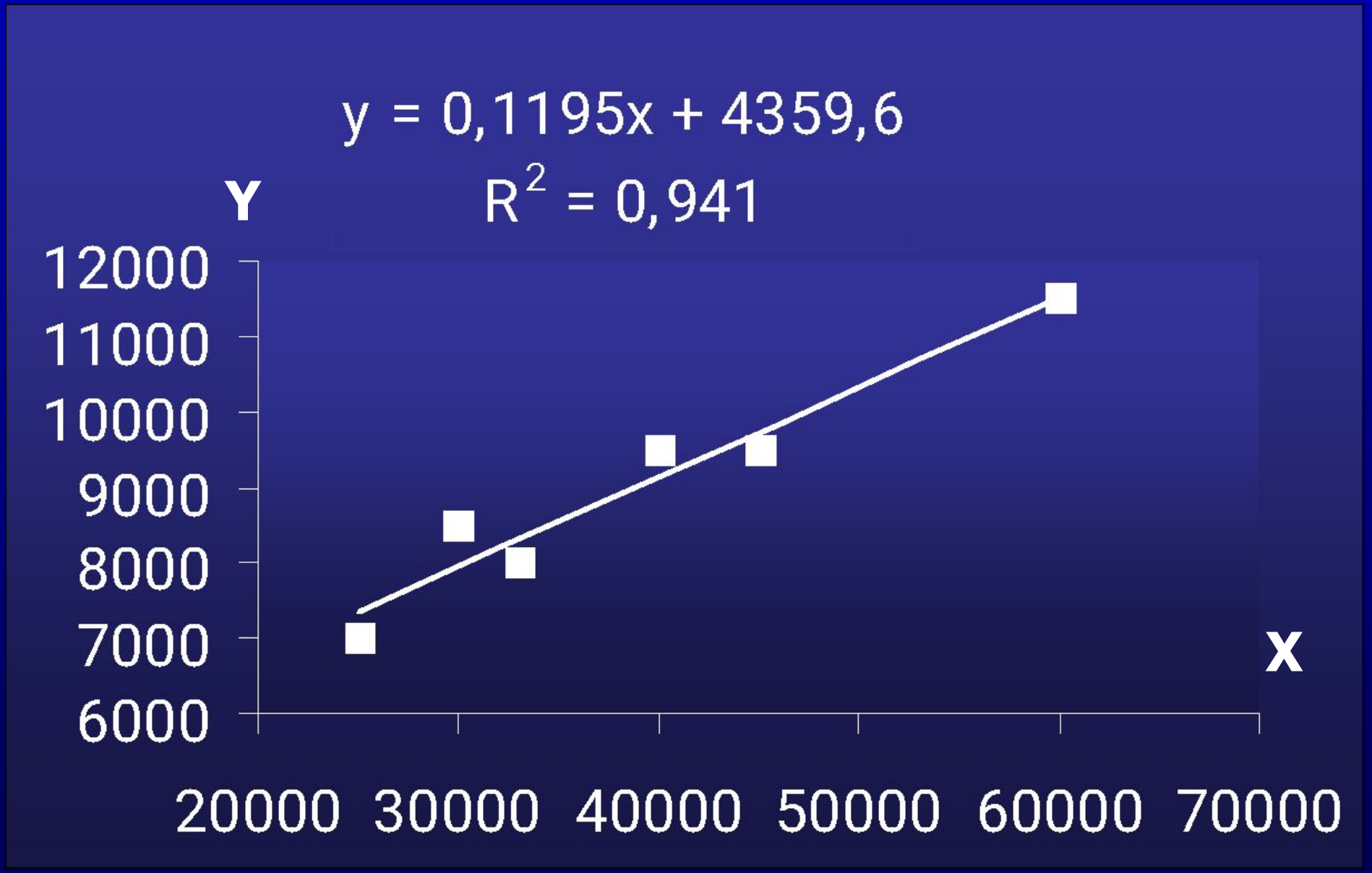
$$b = \sqrt{\frac{2000000}{131805556}} \cdot 0{,}970 = 0{,}119;$$

$$a = 9000 - 0{,}119 \cdot 38833{,}333 = 4359{,}642$$

Следовательно уравнение регрессии будет иметь вид

$$y^T = 0{,}119 \cdot x + 4359{,}642$$

Регрессионное уравнение, полученное с помощью Excel



Хотя выше был рассмотрен лишь с случай линейной функции, во многих случаях можно использовать эти же формулы для коэффициентов регрессии, выполнив простую замену переменных. Пусть, например, изучаемая модель описывается степенной функцией

$$y = C \cdot x^k,$$

где C – некоторая константа. Чтобы привести задачу построения кривой регрессии к линейному случаю для этой модели, достаточно по осям координат откладывать не значения результативного и факторного признаков, а их логарифмы (процедура линеаризации).

Действительно, прологарифмировав уравнение степенной зависимости, имеем линейную зависимость для логарифмов

$$\ln y = k \ln x + \ln C.$$

Аналогично можно подобрать подходящую замену переменных и во многих других случаях.

Некоторые примеры линеаризации будут рассмотрены в качестве примера на лекциях и практических занятиях.

3. 1. Оценка значимости регрессионной модели. Коэффициент детерминации

В рассматриваемой линейной модели регрессии вариация зависимой переменной **y** не может быть объяснена только действием фактора **x,** поскольку действуют и другие неучтенные моделью причины вариации величины **y.**

Поэтому в общем случае уравнение регрессии будет иметь вид

$$y_i = a + b \cdot x_i + \varepsilon_i, \quad \text{где} \quad \varepsilon_i,$$

случайный член, (необъясненный остаток) характеризующий отклонение эмпирических точек от функции регрессии.

Отметим основные постулаты, которые должны выполняться для того, чтобы можно было считать применение регрессионного анализа обоснованным.

1. В рассматриваемой регрессионной модели случайными величинами являются y_i и ε_i, а x_i случайной величиной не является.

2. Математическое ожидание $M(\varepsilon) = 0.$

3. Дисперсия возмущения ε или зависимой переменной y_i постоянна и не зависит от номера точки i (условие гомоскедастичности или равноизменчивости возмущения)

$$D(\varepsilon) = \sigma^2.$$

4. Возмущения ε_i и ε_j являются независимыми. Отсюда следует, что

$$M(\varepsilon_i \cdot \varepsilon_j) = 0.$$

5. Возмущение ε_i или зависимая переменная y_i распределены по нормальному закону. Последнее условие позволяет произвести оценку статистической значимости модели и коэффициентов регрессии.

Регрессионная модель удовлетворяющая этим пяти требованиям называется **классической нормальной линейной регрессионной (КНЛР) моделью.**

Для **КНЛР -** модели доказано несколько важных математических теорем, которые мы примем без доказательства.

Теорема Гаусса-Маркова

Если регрессионная модель удовлетворяет условиям 1 - 4, то полученные оценки для коэффициентов a и b имеют наименьшую дисперсию среди всех линейных несмещенных оценок. Иначе говоря, эти оценки являются эффективными (наилучшими среди других возможных).

Одной из задач регрессионного анализа является оценка адекватности модели. Для проверки того, насколько хорошо кривая регрессии представляет набор эмпирических данных, определяется коэффициент детерминации (пользователи электронных таблиц EXCEL знают ее как фактор детерминации R^2).

Оценка адекватности линейной модели регрессии на основе вычисления фактора детерминации и оценка значимости уравнения регрессии с помощью критерия Фишера основаны на использовании идей дисперсионного анализа. В своей сущности эти идеи достаточно просты и мы их изложим в применении к линейной модели регрессии

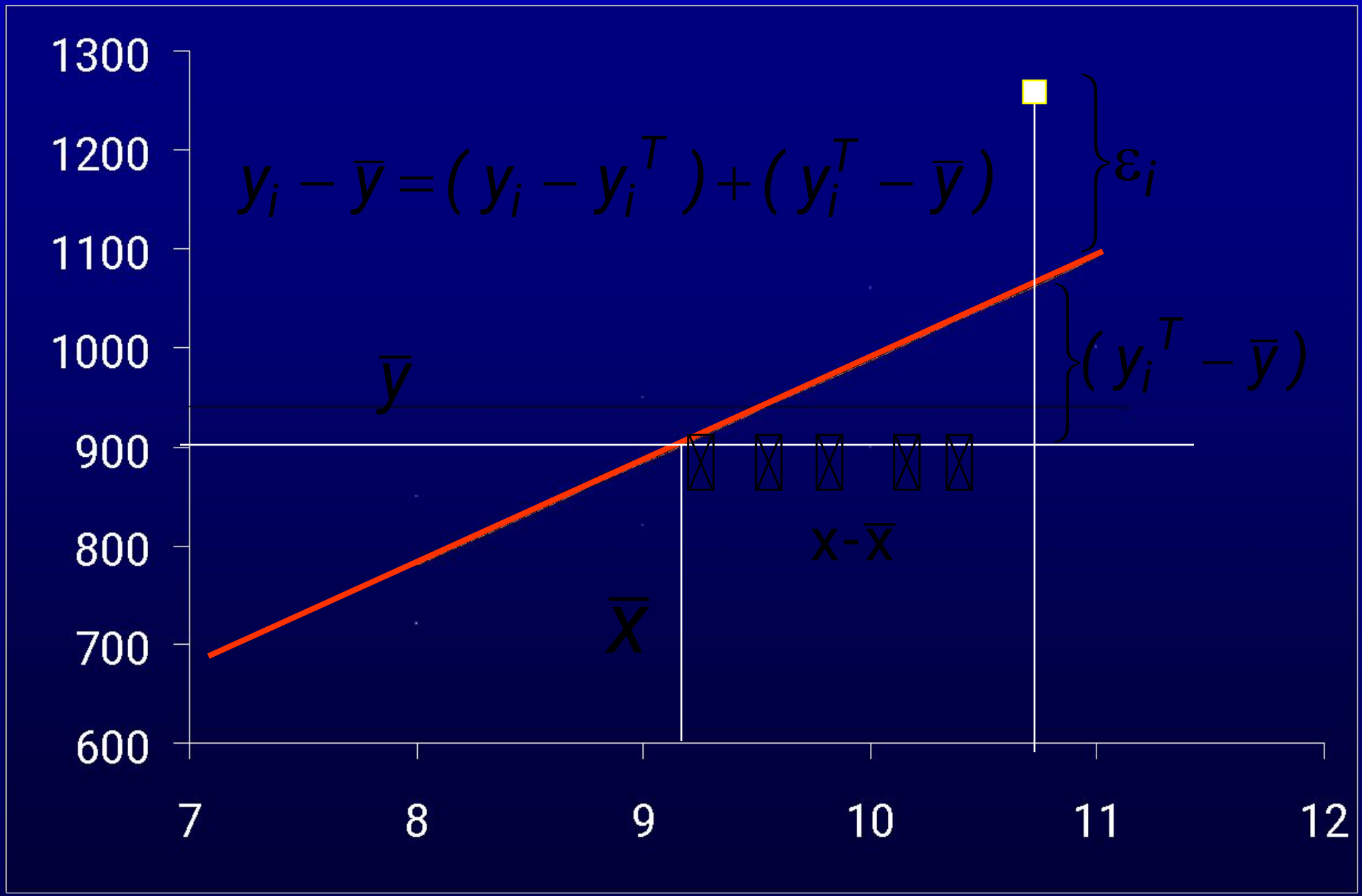
Основная идея метода состоит в том, чтобы разделить общую вариацию факторного признака на часть, которая объясняется регрессионной моделью (действием изучаемого фактора), и часть не находящую объяснения в данной модели (объясняется действием неучтенных факторов):

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

или $Q_T = Q_R + Q_E$,

Деление вариации Y на объясняемую и необъясняемую регрессией части

7а



При возведении в квадрат и последующем суммировании получаем

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

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

Преобразуем последнее слагаемое. Первое произведение представим в виде

$$(y_i^T-\bar{y})=b(x_i-\bar{x}).$$

Этот результат прямо следует из рисунка на предыдущем слайде.

Для преобразования второго сомножителя преобразуем сначала последнее выражение

$$y_i^T = \bar{y} + b(x_i - \bar{x}),$$

И подставим этот результат в рассматриваемый член. В результате получаем

$$y_i - y_i^T = (y_i - \bar{y}) + b(x_i - \bar{x}).$$

Теперь подставим оба преобразованных сомножителя в изучаемую сумму. В итоге получаем

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

$$2b\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - 2b^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0,$$

Поскольку, как было показано ранее, коэффициент *b* может быть представлен в виде

$$b = r\frac{\sigma_y}{\sigma_x}, \qquad b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Величина Q_R дает сумму квадратов отклонений, объясненной моделью (Regression sum of squares). Будем использовать для ее обозначения аббревиатуру RSS.

Q_E – характеризует влияние неучтенных факторов. Ее называется чаще всего суммой квадратов ошибок (Error sum of squares). Для ее обозначения будем использовать абривеатуру ESS

Величину Q_T в левой части формулы будем называть полной суммой квадратов (Total sum of squares) и использовать для ее обозначения аббревиатуру TSS.

Очевидно, что если $Q_R >> Q_E$, то уравнение регрессии статистически значимо и фактор x оказывает существенное влияние на результат y.

Для получения количественной оценки, выдвинем нулевую гипотезу H_0 утверждающую, что влияние фактора x является несущественным.

В условиях справедливости выдвинутой гипотезы оценка дисперсии в генеральной совокупности не должна зависеть от способа получения этой оценки.

Напомним, что для получения несмещенной оценки дисперсии, сумму квадратов отклонений от средней следует делить не на число наблюдений, а на число степеней свободы, т. е. число наблюдений за вычетом числа наложенных на эти наблюдения связей.

Составим схему дисперсионного анализа, позволяющие получить несмещенные оценки дисперсии зависимой переменной.

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Q_R	$\sum_{i=1}^{n}(y_i^T - \bar{y})^2$	m-1	$\frac{Q_R}{m-1}$
Q_E	$\sum_{i=1}^{n}(y_i - y_i^T)^2$	n-m	$\frac{Q_E}{n-m}$
Q_T	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	n-1	$\frac{Q_T}{n-1}$

Рассмотрим две оценки дисперсии

$$s_R^2 = \frac{Q_R}{m-1}; \quad s_E^2 = \frac{Q_E}{n-m},$$

где m число параметров в уравнении регрессии, n – число наблюдений. Обе эти величины являются случайными и распределены по закону хи-квадрат с m-1 и n-m числом степеней свободы. Отношение этих величин подчиняется статистике Фишера-Снедекора и обычно используется для оценки значимости регрессионной модели. Критерий Фишера)

$$F = \frac{Q_R \cdot (n-m)}{Q_E \cdot (m-1)}.$$

Задача. Используя приведенные данные оценить значимость линейной модели связи расходов на питание и доходов семьи

№ Семьи	Доход семьи, руб. (X)	Расходы на питание, руб. (Y)
1	30 000	8 500
2	25 000	7 000
3	40 000	9 500
4	60 000	11 500
5	33 000	8 000
6	45 000	9 500

Линейное регрессионное уравнение было получено ранее и имеет вид

$$y^T = 0{,}119 \cdot x + 4359{,}642$$

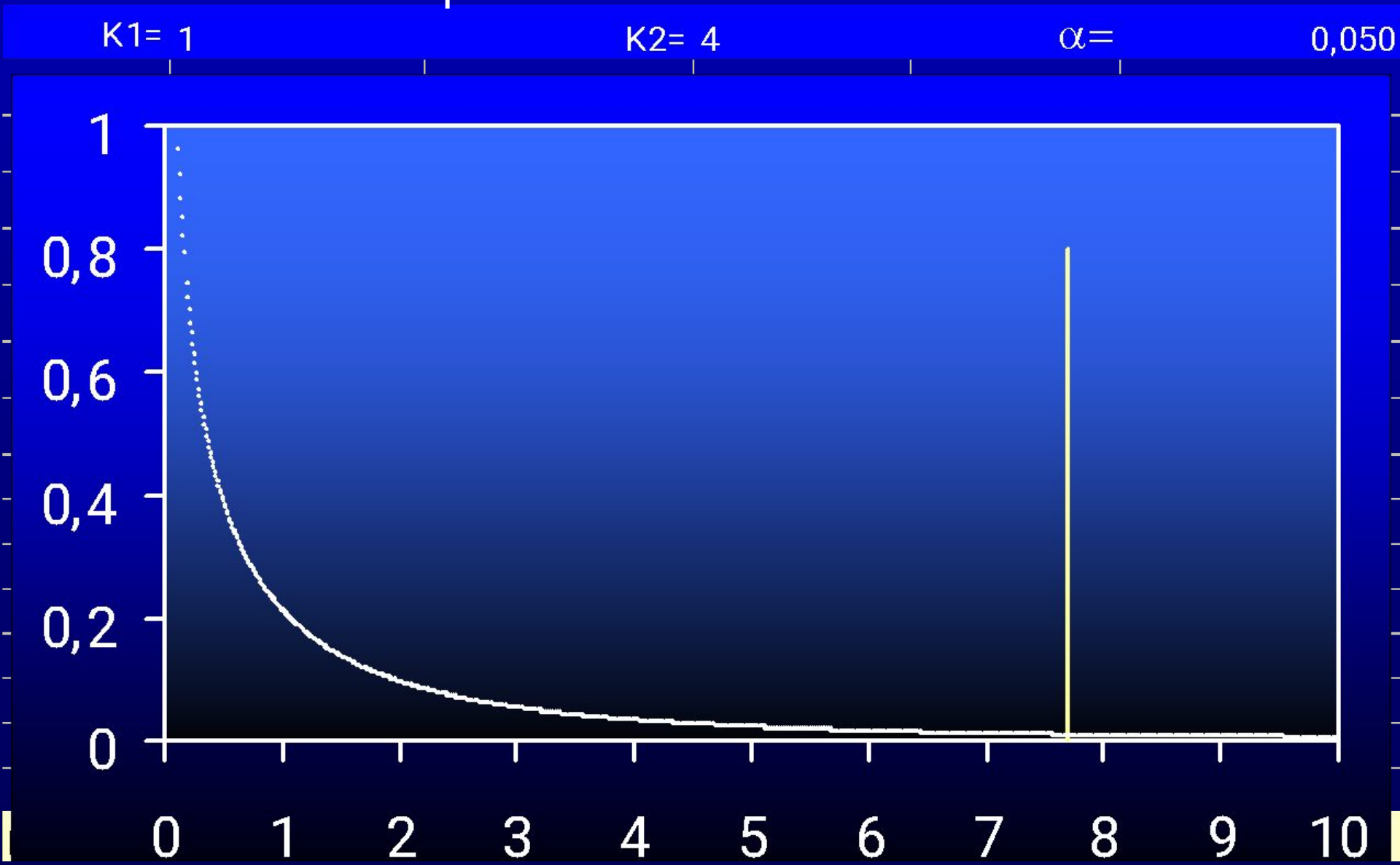
Используя электронные таблицы Excel, находим суммы квадратов отклонений . Найдем расчетное значение критерия Фишера *F*, учитывая, что в нашем случае $m = 2$, $n = 6$

$$F_{расч} = \frac{Q_R \cdot (n-m)}{Q_E \cdot (m-1)} = \frac{11292202}{707797{,}68} \cdot 4 = 63{,}82$$

Величина F подчиняется распределению Фишера –Снедекора для $K_1=1$, $K_2=4$.

Используя функцию Excel FРАСПОБР(0,05;1;4) Получаем критическое значение статистики Фишера - Снедекора для уровня значимости 0,05 $F_{крит} = 7,72$. Поскольку эмпирическое значение значительно превышает критическое, то гипотезу об отсутствии связи между признаками Y и X следует отбросить и признать, что регрессионное уравнение является значимым.

График плотности распределения Фишера - Снедекора для $k_1=1$, $k_2=4$. Критическая область справа от желтой линии.



Для проверки значимости линейного уравнения регрессии можно использовать и функцию ЛИНЕЙН ( ) электронных таблиц Excel.

Кроме значения критерия Фишера, эта функция возвращает и ряд других параметров регрессионной модели, важных для ее правильной статистической оценки. Применение функции ЛИНЕЙН ( ) для оценки значимости линейной модели рассмотрим на примере.

Задача

Имеются следующие данные об общем объеме розничного товарооборота региона по месяцам в 1997 г., млрд. руб.:

1	2	3	4	5	6
22,8	24,9	31,0	29,5	30,5	35,6
7	8	9	10	11	12
36,4	42,6	45,1	47,3	51,0	53,4

Оцените значимость линейной регрессионной модели и значимость коэффициентов модели при уровне значимости 0,05.

Для нахождения параметров линейной модели применим функцию **Линейн** электронных таблиц Excel.

Ниже приведены параметры возвращаемые функцией ЛИНЕЙН и их смысл.

2,80	19,31
0,14	1,05
0,97	1,70
387,18	10,00
1120,84	28,95

b	**a**
m_b	m_a
R^2	s_y
F	n-2
Q_R	Q_E

Уравнение регрессии имеет вид

$$y^T = 2{,}7996 \cdot x + 19{,}310\epsilon$$

Для оценки значимости регрессионной модели найдем критическую точку распределения Фишера при уровне значимости 0,05 и числе степеней свободы $k_1=1$ и $k_2=10$, используя функцию Excel FРАСПОБР(0,05;1;10), которая возвращает значение 4,96. Поскольку эмпирическое значение коэффициента Фишера в рассматриваемой задаче равно 387,18, и превышает во много раз критическое значение, то необходимо признать, что рассматриваемая связь значима.

Как уже указывалось, одной из наиболее эффективных оценок адекватности регрессионных моделей, мерой качества уравнения регрессии является фактор детерминации R^2 . Для расчета этого коэффициента используются величины Q_R Q_E и Q_T :

$$R^2 = \frac{Q_R}{Q_T} = 1 - \frac{Q_E}{Q_T}.$$

Коэффициент детерминации изменяется в пределах от 0 до 1. Чем ближе коэффициент к единице, тем выше качество регрессионной модели.

В случае парной регрессии легко показать, что коэффициент детерминации равен квадрату коэффициента корреляции.

Действительно, вспоминая уравнение для определения коэффициента a и регрессионное уравнение

$$a = \bar{y} - b\bar{x}, \quad y^T = a + bx, \text{ получаем}$$

$$y^T - \bar{y} = b(x - \bar{x}).$$

Подставляя последний результат в определение коэффициента детерминации, получаем:

$$R^2 = \frac{Q_R}{Q_T} = \frac{\sum_{i=1}^{n}(y_i^T - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} =$$

$$= \frac{b^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \left(\frac{b^2\sigma_x^2}{\sigma_y^2}\right) = r^2.$$

Следует заметить, что оценка качества регрессионного уравнения с помощью критерия Фишера или коэффициента детерминации возможно только в том случае, когда коэффициент a уравнения регрессии не равен нулю, поскольку только в этом случае возможно представление

$$y^T - \bar{y} = b(x - \bar{x}),$$

Которое использовалось для доказательства возможности разбиения

$$Q_T = Q_R + Q_E.$$

3. 2. Проверка значимости коэффициентов регрессии

Интервальная оценка для коэффициентов регрессии и индивидуальных значений зависимой переменной.

В линейной регрессии обычно оценивается значимость не только уравнения в целом, но и отдельных его параметров Для оценки статистической значимости коэффициентов регрессии используются случайные величины

$$t_b = \frac{|b - b_{ген}|}{m_b}, \quad t_a = \frac{|a - a_{ген}|}{m_a}$$

m_b и m_a - стандартные ошибки коэффициентов регрессии. В качестве нулевой гипотезы выдвинем предположение, что

$$b_{ген} = 0, \quad a_{ген} = 0.$$

В условиях справедливости выдвинутой гипотезы случайные величины t_b и t_a подчиняются распределению Стьюдента. Поэтому для проверки гипотезы нужно вычислить эмпирические значения t_b и t_a

$$t_b = \frac{|b|}{m_b}, \quad t_a = \frac{|a|}{m_a},$$

и затем сравнить их с критическим значением статистики Стьюдента $t_{крит}$ при заданном уровне значимости и числе степеней свободы n-2.

Для нахождения m_b найдем дисперсию коэффициента b. Для этого используем запись коэффициента b в виде

$$b = r\frac{\sigma_y}{\sigma_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Поскольку переменные X не являются случайными, то

$$\sigma_b^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_y^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{\sigma_y^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Оценим дисперсию используя формулу остаточной дисперсии. В условиях справедливости выдвигаемой гипотезы (равенства нулю коэффициента b) такая оценка является справедливой.

$$s^2 = \sigma_y^2 = \frac{Q_E}{n-2} = \frac{\sum_{i=1}^{n}(y_i - y_i^T)^2}{n-2}.$$

В итоге получаем среднеквадратическое отклонение (ошибку) для коэффициента b в виде

$$m_b = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^T)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot (n-2)}}.$$

Поэтому, если

$$t_b = \frac{|b|}{m_b} > t_{крит},$$

то коэффициент b значим.

интервальная оценка коэффициента при заданном уровне значимости ($t_{крит}$) определяется стандартными формулами

$$b - t_{крит} \cdot m_b < b_{ген} < b + t_{крит} \cdot m_b.$$

Статистическая оценка значимости коэффициента *a* производится аналогично и мы приведем формулы без дополнительных комментариев.

Найдем дисперсию коэффициента a.

$$a = \bar{y} - b\bar{x} = \frac{\sum y_i}{n} - \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \cdot \bar{x} =$$

$$= \sum \left(\frac{1}{n} - \frac{(x_i - \bar{x}) \cdot \bar{x}}{\sum (x_i - \bar{x})^2} \right) \cdot y_i .$$

После такого преобразования коэффициента *a*, можно вычислить его дисперсию. Введем обозначение

$$c_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} .$$

Учитывая, что дисперсия суммы равна сумме дисперсий, а также то, что величины X_i не являются случайными. получаем

$$\sigma_a^2 = \sigma_y^2 \sum \left(\frac{1}{n} - c_i \bar{x} \right)^2 =$$

$$\sigma_y^2 \sum \left(\frac{1}{n^2} - \frac{2c_i \bar{x}}{n} + c_i^2 \bar{x}^2 \right)$$

поскольку сумма

$$\sum_{i=1}^{n} c_i = 0,$$

после элементарных преобразований получаем

$$\sigma_a^2 = \sigma_y^2 \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

Вспоминая выражение для дисперсии находим следующую оценку для средеквадратического отклонения коэффициента *a*

$$m_a = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^T)^2 \cdot \sum_{i=1}^{n} x_i^2}{(n-2)\cdot n \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Оценка значимости и расчет доверительного интервала при заданном уровне значимости, определяется точно также как и для коэффициента *b*.

Используя электронные таблицы Excel можно избежать утомительных вычислений, поскольку функция ЛИНЕЙН ( ) возвращает и стандартные ошибки отклонений m_b m_a.

Еще более полную информацию о параметрах регрессионной модели можно получить используя функцию **РЕГРЕССИЯ** из **Пакета анализа.**

Использование этого пакета будет продемонстрировано на практических занятиях.

Построим доверительный интервал для функции регрессии т. е. интервал значений переменной y^T, который при заданной доверительной вероятности $\gamma = 1-\alpha$ накроет неизвестное значение $M(y^T)$ при заданном значении аргумента x. Для этой цели точно также как и ранее, рассмотрим случайную величину

$$t = \frac{y^T - M(y^T)}{m_{y^T}},$$

которая имеет распределение Стьюдента с k=n-2 степенями свободы.

Найдем среднеквадратическое отклонение для предсказываемых моделью значений y^T

$$m_{y^T} = \sqrt{\sigma^2_{y^T}}; \quad y^T = \bar{y} + b(x - \bar{x});$$

$$\sigma^2_{y^T} = \sigma^2_{\bar{y}} + \sigma^2_b \cdot (x - \bar{x})^2.$$

$\sigma^2_{\bar{y}}$

$\sigma^2_{\bar{y}}$

Дисперсия среднего значения факторной переменной оценивается по известной формуле

$$\sigma^2_{\bar{y}} = \frac{\sigma^2_y}{n},$$

где σ^2_y генеральная дисперсия.

Дисперсия коэффициента b вычислялась ранее и равна

$$\sigma_b^2 = \frac{\sigma_y^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

учитывая два последних результата, получаем

$$\sigma_{y_T}^2 = \sigma_y^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i - \bar{x})^2}\right);$$

$$m_{y_T} = \sqrt{\sigma_{y_T}^2}.$$

В качестве оценки для дисперсии результативного признака снова возьмем величину необъясненной дисперсии

$$s^2 = \frac{Q_E}{n-2} = \frac{\sum_{i=1}^{n}(y_i - y_i^T)^2}{n-2}.$$

В результате получаем выражение для ошибки

$$m_{y^T} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_i^T)^2}{n-2} \cdot \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}.$$

Поскольку случайная величина

$$t = \frac{y^T - M(y^T)}{m_{y^T}},$$

подчиняется распределению Стьюдента с числом степеней свободы k=n-2, то доверительный интервал для математического ожидания результативной переменной может быть записан в виде

$$y^T - t_{\text{крит}} \cdot m_{y^T} < M(y^T) < y^T + t_{\text{крит}} \cdot m_{y^T}.$$

Доверительные границы для задачи, представленной на слайде 63