АНАЛИЗ МАССИВА ДАННЫХ

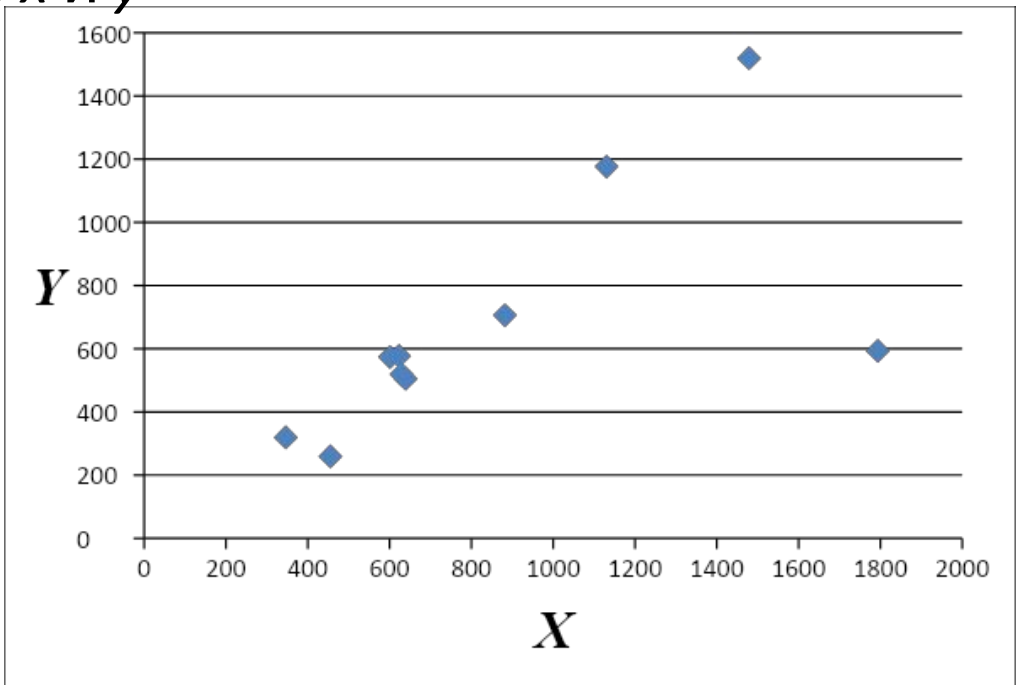
Анализ массива данных, описывающих процесс предметной области, заключается **в выявлении грубых ошибок** (промахов, выбросов, аномальных наблюдений).

Грубая ошибка (промах, выброс, аномальное наблюдение) – это ошибка результата отдельного наблюдения, входящего в массив, которая для данных условий резко отличается от остальных наблюдений этого массива.

Источники грубой ошибки, промаха, выброса:

1. ошибки оператора (неправильная запись результата наблюдения),
2. ошибки измерений (резкие изменения условий снятия показаний),
3. умышленное искажение показаний наблюдений,
4. резкие отличия показаний объектов исследования.

Грубая ошибка в ряде случаев может быть сразу видна, если построить точечную диаграмму поля рассеяния факторов *x* и *y*



Наличие такой ошибки может сильно исказить результат математического моделирования.

Поэтому рекомендуется любую совокупность наблюдений проверять на наличие грубых ошибок с помощью статистических критериев.

Статистические критерии на наличие грубой погрешности

Выдвигаемые гипотезы:

H_0 - грубой ошибки (промаха, выброса) **нет**;

H_1 - грубая ошибка (промах, выброс) **есть**.

1. Критерий **Диксона.** Используется при **n≤10**

Условие отклонения гипотезы H_0: $\frac{x_n - x_{n-1}}{x_n - x_1} > Z_q$

Критические значения критерия Диксона (Z_{nq})

	q – уровень значимости гипотезы			
	0,10	***0,05***	***0,02***	***0,01***
4	0,68	0,76	0,85	0,89
6	0,48	0,56	0,64	0,70
8	0,40	0,47	0,54	0,59
10	0,35	0,41	0,48	0,53

ПРИМЕР.

При анализе расхода газа были получены результаты (л): 22; 24; 26; 28; 48. Последний результат вызывает определенные сомнения и подлежит проверке на грубую погрешность. Использовать критерий Диксона.

РЕШЕНИЕ:

1. Имеем: $x_n=48$, $x_{n-1}=28$, $x_1=22$. $\frac{x_n-x_{n-1}}{x_n-x_1}=\frac{48-28}{48-22}=0,77$

2. Задаемся уровнем значимости q=0,05.

Критическое значение критерия Диксона дан для n=4 (0,76) и n=6 (0,56). Для получения критического значения Диксона для n=5 берется среднее:

$Z_q=(0,76+0,56)/2=0,66$

3. Поскольку расчетное значение критерия Диксона больше критического: 0,77>0,66, то гипотезу H_0 о том, что грубой ошибки нет отклоняем.

Следовательно, результат 48 л является в данном случае грубой ошибкой и не должен учитываться при последующих расчетах.

2. Критерий ***Шовине.***

Используется при **n≤10**

Условие отклонения гипотезы H_0:

n=3 $\left|\bar{x} - x_i\right| > 1{,}6S$

n=6 $\left|\bar{x} - x_i\right| > 1{,}7S$

n=8 $\left|\bar{x} - x_i\right| > 1{,}9S$

n=10 $\left|\bar{x} - x_i\right| > 2{,}0S$

$$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}};$$

- исправленное среднее квадратическое отклонение

Замечание:

при расчете $\bar{x}, S$ сомнительное значение учитывается

ПРИМЕР.

При измерении количества пассажиропотока (тыс.чел.) получен: 10; 11; 12; 12; 15. Определить является ли результат 15 тыс.чел. промахом? Использовать критерий Шовине.

РЕШЕНИЕ:

1. Рассчитать $\bar{x}, S$

Получим: $\bar{x} = 12 тыс.чел.; S = 1{,}87 тыс.чел.$

2. Рассчитать показатель $|\bar{x} - x_i|$

Получим: $|\bar{x} - x_i| = 3{,}0$

3. Рассчитать показатель: $1{,}7S$

Получим: $1{,}7S = 3{,}18$

4. Поскольку расчетное $|\bar{x} - x_i|$ меньше 1,7S: 3,0<3,18, то гипотезу H_0 о том, что грубой ошибки нет не отклоняем (принимаем).

5. **Вывод:** результат x=15 тыс.чел. не является грубой ошибкой и должен быть учтен при последующих расчетах.

3. Критерий ***Романовского.*** Используется при **n≤20**

Условие отклонения гипотезы H_0: $\frac{|\bar{x} - x_i|}{S} \geq \beta_q$

$S_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$; - исправленное среднее квадратическое отклонение

Замечание:

при расчете $\bar{x}, S$ сомнительное значение НЕ учитывается

Критические значения критерия Романовского

q (β_q)	***n***						
	4	***6***	***8***	***10***	***12***	***15***	***20***
0,01	1,73	2,16	2,43	2,62	2,75	2,90	3,08
0,02	1,72	2,13	2,37	2,54	2,66	2,80	2,96
0,05	1,71	2,10	2,27	2,41	2,52	2,64	2,78
0,10	1,69	1,00	2,17	2,29	2,39	2,49	2,62

ПРИМЕР.

При продажах стиральных машин были получены следующие результаты (тыс.шт): 10,07; 10,08; 10,10; 10,12; 10,13; 10,15; 10,16; 10,17; 10,20; 10,40. Не является ли промахом максимальное значение 10,40 тыс.шт.? Использовать критерий Романовского.

РЕШЕНИЕ:

1. Рассчитать $\bar{x}, S$

   Получим: $\bar{x} = 10{,}13 тыс.шт.; S = 0{,}17 тыс.шт.$

2. Задаемся уровнем значимости q=0,05.

3. Рассчитать показатель и сравнить с критическим β_q=2,41

$$\frac{|\bar{x} - x_i|}{S} = 1{,}59 < \beta_q = 2{,}41$$

4. Поскольку расчетное значение критерия Романовского меньше критического: 1,59<2,41, то гипотезу H_0 о том, что грубой ошибки нет принимаем.

5. **Вывод**: результат 10,40 тыс.шт. не является грубой ошибкой и должен быть учтен при последующем перерасчете числовых $\bar{x}, S$ характеристик:

4. Критерий ***Трех сигм.***

Используется при **n>20…50**

Условие отклонения гипотезы H_0: $\left|\overline{x} - x_i\right| > 3\sigma$

$$\sigma_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n}}$$

- среднее квадратическое отклонение

Замечание:

при расчете $\overline{x}, \sigma$ сомнительное значение НЕ учитывается.

ПРИМЕР.

Проверить по критерию Трех сигм показатели душевого дохода (x) и индекс человеческого развития (y), представленные в таблице.

Страна	Душевой доход долл., *x*	Индекс человеческого развития (ИЧР), *y*
ОАЭ	1600	0,866
Таиланд	7100	0,833
Уругвай	6750	0,833
Ливия	6130	0,801
Колумбия	6110	0,848
Иордания	4190	0,73
Египет	3850	0,514
Марокко	3680	0,566
Перу	3650	0,717
Шри-Ланка	3280	0,711
Филиппины	2680	0,672
Боливия	2600	0,589
Китай	2600	0,626
Зимбабве	2200	0,513
Пакистан	2150	0,445
Уганда	1370	0,328
Нигерия	1350	0,393
Индия	1350	0,446
Бангладеш	1050	0,335

РЕШЕНИЕ:

1. Построить точечную диаграмму (x;y) и сделать предположение о наличии промаха для x и y.

2. Рассчитать показатели для промахов $\left|\overline{x} - x_i\right|$ $\left|\overline{y} - y_i\right|$

3. Рассчитать показатели σ_x, $3\sigma_x$ σ_y, $3\sigma_y$,

4. **Сделать выводы**.

5. Критерий ***Ирвина.***

Используется при **n>20…50**

Условие отклонения гипотезы H_0: $\frac{x_{n+1} - x_n}{\sigma} > \theta_p$

$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$ - среднее квадратическое отклонение

Замечание:

при расчете $\bar{x}, \sigma$ сомнительное значение учитывается

Критические значения критерия Ирвина (λ_p)

n	***Доверительная вероятность, p***	
	0,95	***0,99***
2	2,8	3,7
3	2,2	2,9
10	1,5	2,0
20	1,3	1,8
30	1,2	1,7
50	1,1	1,6
100	1,0	1,5
400	0,9	1.3
1000	0,8	1.2

Порядок расчета

1. Исходные данные ранжируются в порядке убывания или возрастания.

2. Из полученного ряда выбирают два наибольших или два наименьших значения.

3. Рассчитывается показатель критерия Ирвина.

4. Грубой ошибкой считается показатель x_i , если значение критерия превышает значение τ_p

ПРИМЕР.

Использовать критерий Ирвина для выявления промахов для исходных данных предыдущего примера.

РЕШЕНИЕ:

1. Рассчитать $\bar{x}, S$

   Получим: $\bar{x} = 10{,}13 тыс.шт.; S = 0{,}17 тыс.шт.$

2. Задаемся уровнем значимости q=0,05.

3. Рассчитать показатель и сравнить с критическим $\beta_q = 2{,}41$

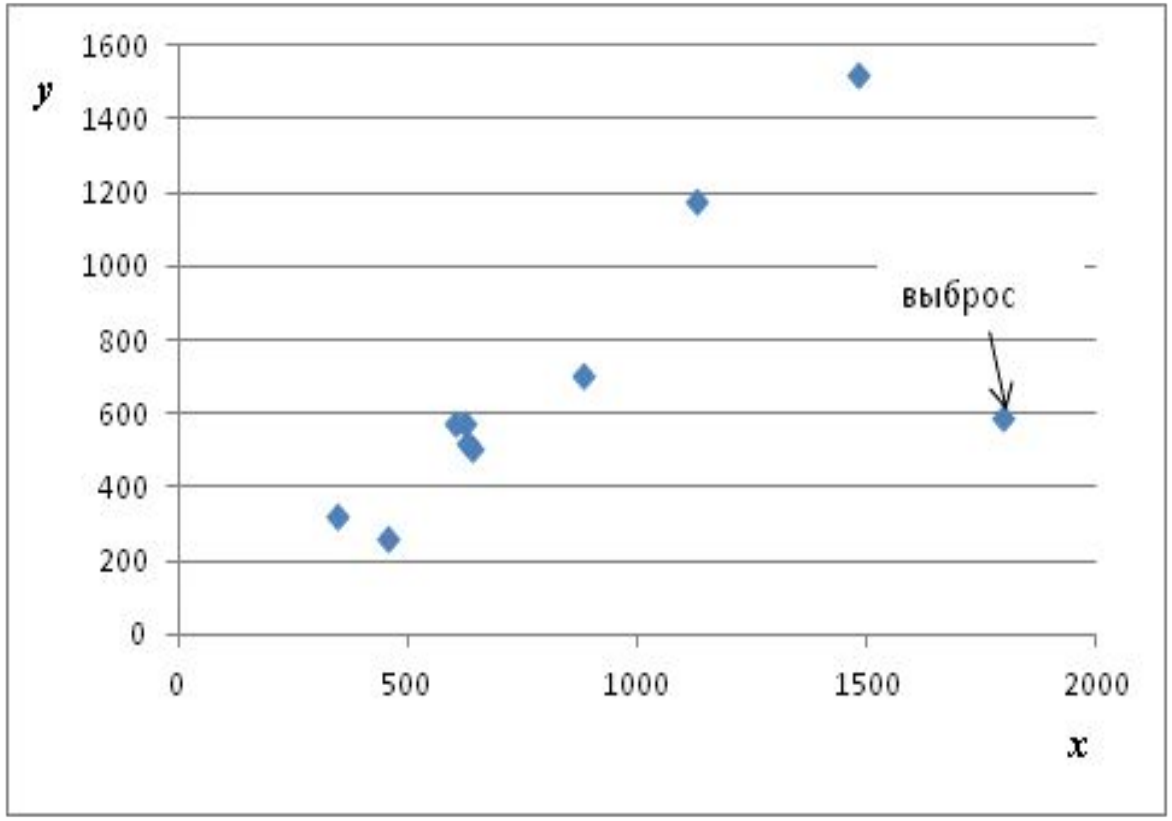
$$\frac{|\bar{x} - x_i|}{S} = 1{,}59 < \beta_q = 2{,}41$$

4. Поскольку расчетное значение критерия Романовского меньше критического: 1,59<2,41, то гипотезу H_0 о том, что грубой ошибки нет принимаем.

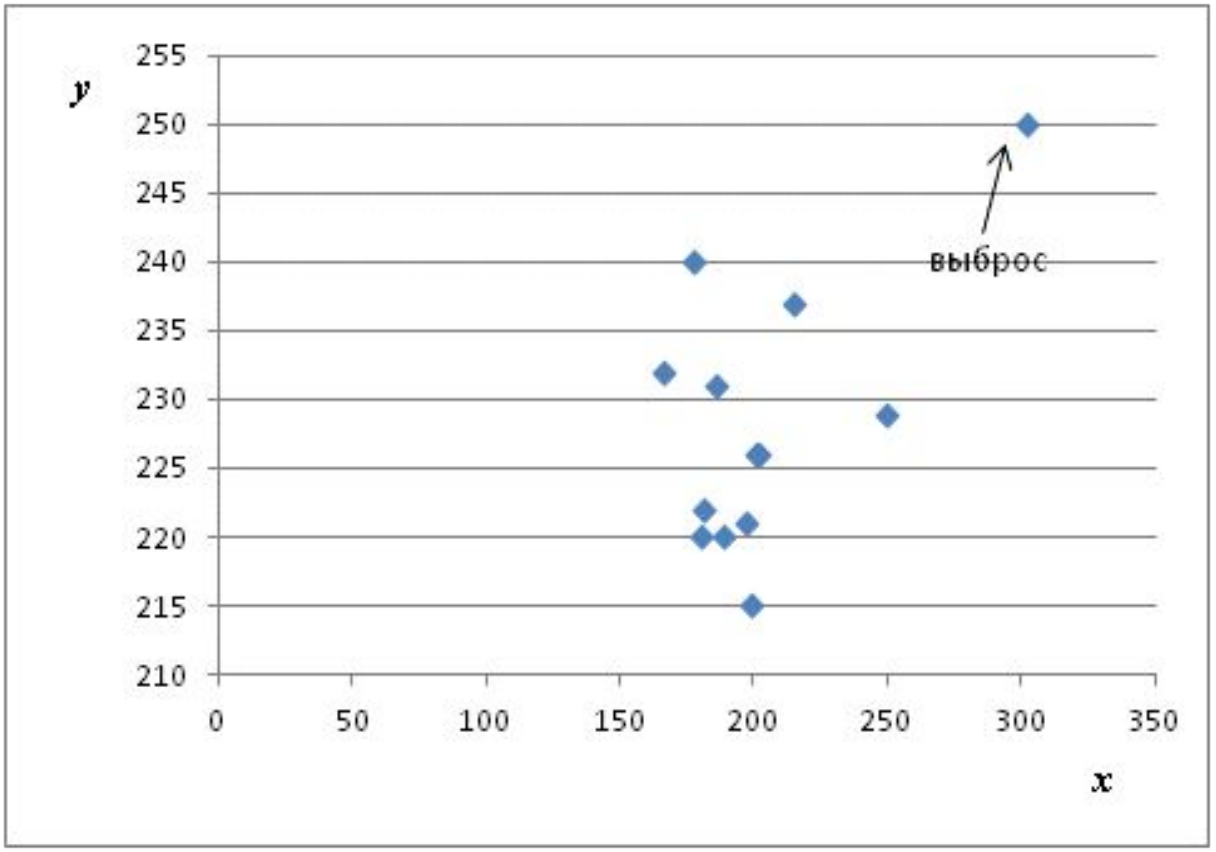
5. **Вывод**: результат 10,40 тыс.шт. не является грубой ошибкой и должен быть учтен при дальнейшем исследовании.

ВЫЯВЛЕНИЕ ГРУБЫХ ОШИБОК В ДВУМЕРНЫХ МАССИВАХ ИСХОДНЫХ ДАННЫХ

Два взаимосвязанных массива x и y, где предполагаемый выброс или грубую ошибку можно заметить на диаграмме рассеяния.



Два взаимосвязанных массива x и y, где предполагаемый выброс или грубая ошибка менее очевидна на диаграмме рассеяния.



Для оценки выбросов двух взаимосвязанных массивов X и Y необходимо использовать критерии, характеризующие связи этих массивов.

ВОПРОС:

Какие показатели характеризуют связи двух массивов или двух факторов X и Y ?

ОТВЕТ:

1. Коэффициент корреляции r_{xy} .

2. Регрессия y по x или $y_{теор}=f(x)$.

Использование коэффициента корреляции для выявления грубой ошибки

Линейный коэффициент корреляции r_{xy} характеризует тесноту и направление связи двух факторов X и Y и вычисляется по формуле:

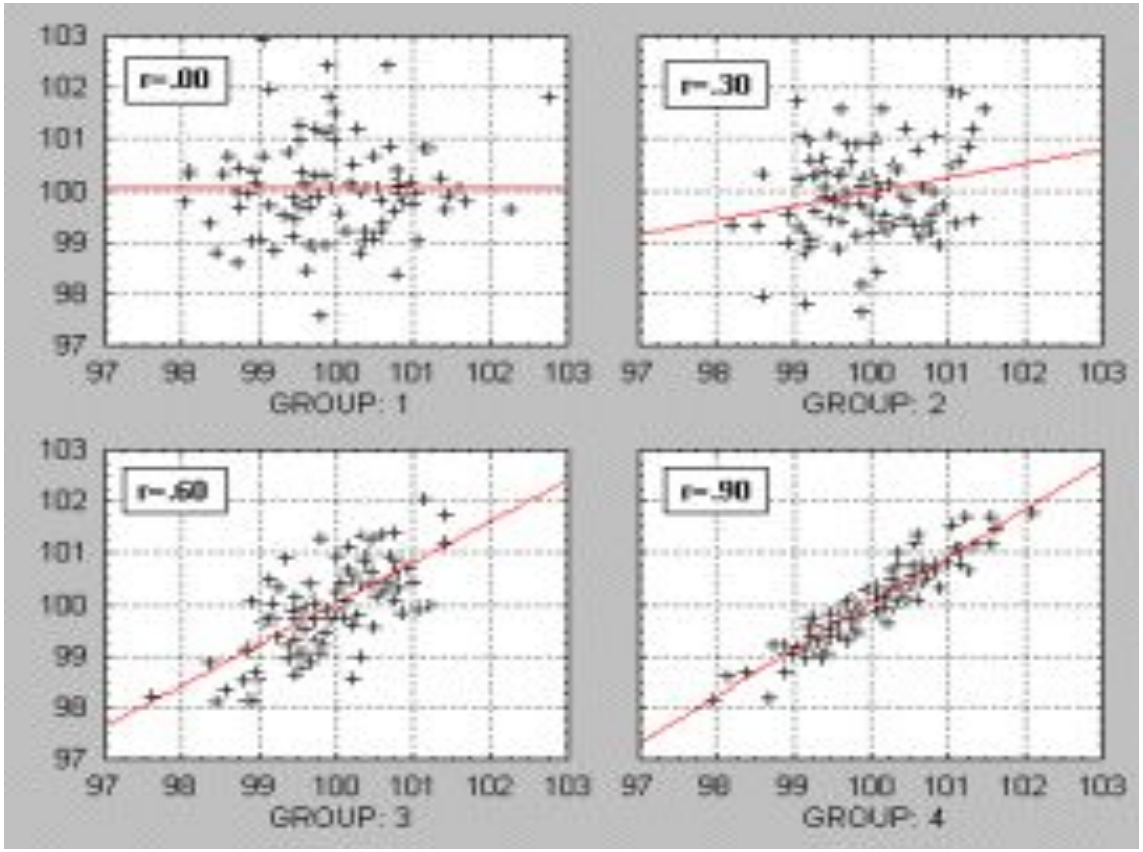
$$r_{xy} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\sigma_x \cdot \sigma_y}$$

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

Качественную оценку тесноты связи величин *x* и *y* можно оценить с помощью **шкалы Чеддока**

Теснота связи	**Значение коэффициента корреляции при наличии:**	
	прямой связи	**обратной связи**
Слабая	**0,1-0,3**	**(-0,1)-(-0,3)**
Умеренная	**0,3-0,5**	**(-0,3)-(-0,5)**
Заметная	**0,5-0,7**	**(-0,5)-(-0,7)**
Высокая	**0,7-0,9**	**(-0,7)-(-0,9)**
Весьма высокая	**0,9-0,99**	**(-0,9)-(-0,99)**

Представление связи факторов на диаграммах рассеяния



Порядок выявления грубой ошибки по коэффициенту корреляции

1. Строится диаграмма рассеяния взаимосвязанных массивов X и Y.
2. По диаграмме визуально определяется предполагаемый выброс с координатами $(x_в;y_в)$.
3. Вычисляется коэффициент корреляции по исходному массиву данных r_{xy} и коэффициент корреляции r_{xy1} по данным без учета предполагаемого выброса.
4. Проверяется условие: $|r_{xy}-r_{xy1}| > 0{,}15$.

   Если условие выполняется, то проверяемую координату $(x_в;y_в)$ можно считать выбросом или грубой ошибкой и она должна быть исключена из дальнейшего рассмотрения (построения математической модели связи факторов x и y).

Повышение надежности полученного вывода:

Проверяется статистическая значимость вычисленных коэффициентов корреляции с помощью t-статистики.

1). Вычисляется *t*-критерия Стьюдента по формуле:

$$t_r = r_{xy} \cdot \frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$

2). Определяется табличное значение *t*-критерия Стьюдента $t_{табл}$ по двум аргументам: - уровень значимости α (задаются, 5%);

- степень свободы $n-2$

3). Проверяемый коэффициент корреляции статистически значим и связь между исходными массивами данных X и Y можно считать доказанной, если $t_r > t_{табл}$ (с заданной ошибкой не более α).

Использование регрессия y по x или $y_{теор}=f(x)$ для выявления грубой ошибки

Последовательность действий по выявлению грубой ошибки в исходном двумерном массиве с помощью линейной регрессии:

1. По исходному двумерному массиву строится диаграмма рассеяния с целью выявления координаты предполагаемого выброса ($x_в$;$y_в$).

2. Строится:

- линейная регрессия $y_{теор}=b_0+b_1x$ по исходным данным;
- линейная регрессия $y'_{теор}=b'_0+b'_1x$ по исходным данным, <u>но без предполагаемого выброса</u>.

3. Вычисляется остаточные компоненты по обоим уравнениям регрессии: $\varepsilon = y - y_{теор}$ и $\varepsilon' = y' - y'_{теор}$.

4. Вычисляется суммы квадратов остаточных компонентов:

$$S^2 = \sum_{i=1}^{n} \varepsilon_i^2 \qquad S'^2 = \sum_{i=1}^{n-1} \varepsilon_i'^2$$

5. Вычисляется отношение : $R = \frac{S^2}{S'^2}$

6. Оценивается статистическая значимость отношения R с помощью *F*-критерия Фишера.
Если $R > F_{табл}$, то предполагаемый выброс считается существенным и влияющим на искажение характеристики связи исходных факторов двумерного массива X и Y (с заданной ошибкой не более α).

Действия: такая координата ($x_в$;$y_в$) должна быть исключена из дальнейшего расчета.

При оценке $F_{табл}$ берутся следующие степени свободы: степень свободы числителя n_1=2 (число параметров при переменных *x*), степень свободы знаменателя $n_2 = n - n_1 - 1 = n - 3$.
Вероятность или значимость ошибки α=5%.