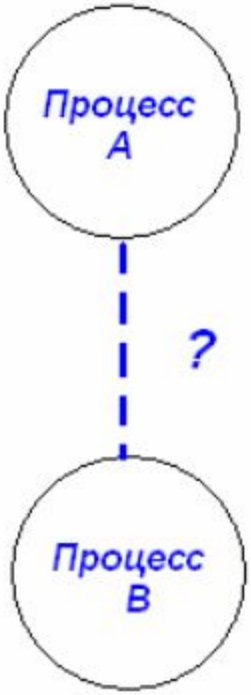
Корреляционный и регрессионный анализ

1. Есть два самостоятельных процессов



2. Устанавливается, есть ли между ними значимая связь

3. Устанавливается, что является воздействующим фактором (аргументом), а что результативным (функцией). Форма представления - уравнение

Корреляции

До сих пор нас в выборках интересовала только одна зависимая переменная.

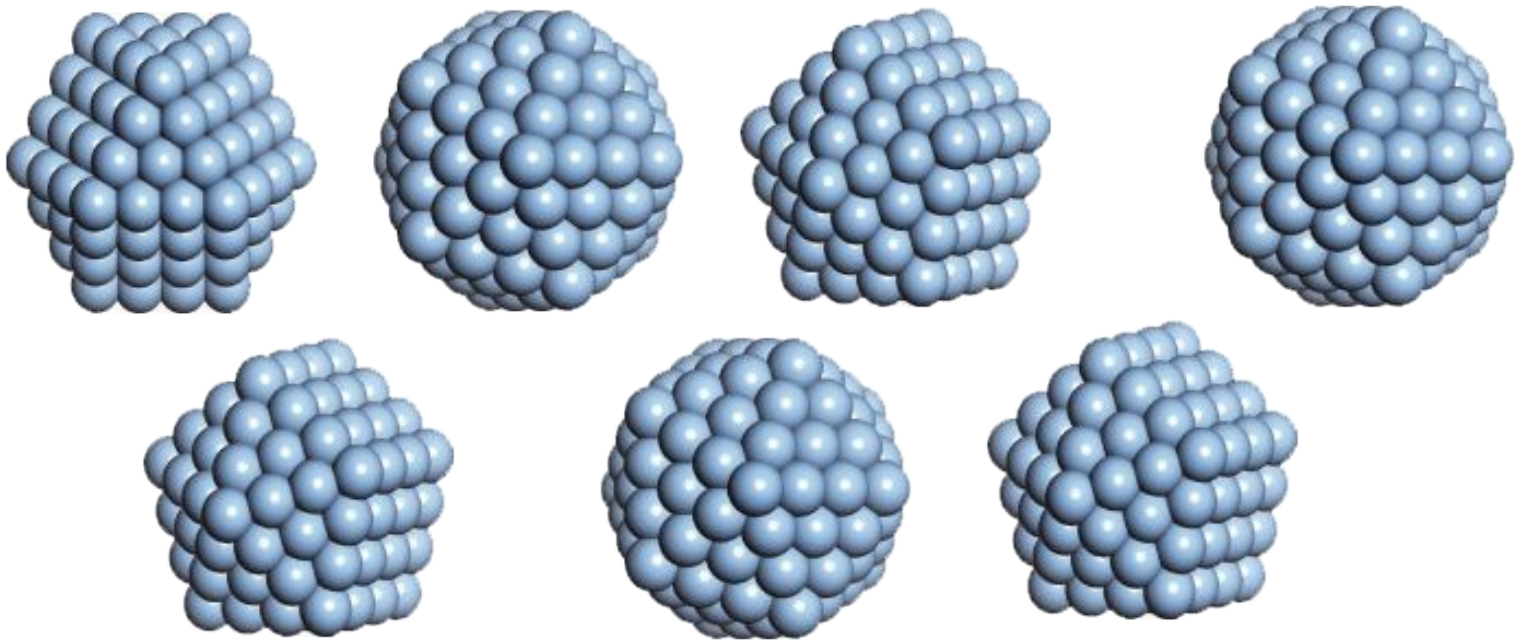
Мы изучали, отличается ли распределение этой переменной в одних условиях от распределения той же переменной в других условиях.

Обратимся к ситуации, когда зависимых переменных будет ДВЕ и более.

Вопрос: в какой степени эти переменные связаны между собой?

Это могут быть измерения одной особи или связанных пар.

Мы исследуем нано частицы. И хотим узнать, связаны ли между собой размер частиц и форма частиц?



Вопрос: в какой степени две переменные СОВМЕСТНО ИЗМЕНЯЮТСЯ?

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ характеризует силу связи между переменными.

Это просто параметр описательной статистики

Большой коэффициент корреляции между размером частицы и ее формой позволяет нам

предсказывать, что большая частица, скорее всего, будет иметь шарообразную

Основные термины

Изучение связей между переменными, интересует исследователя с точки зрения отражения соответствующих причинно-следственных отношений.

Корреляционная зависимость – это согласованные изменения двух (парная корреляционная связь) или большего количества признаков (множественная корреляционная связь). Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) другой(-их) переменной(-ых).

Корреляционный анализ – статистический метод, позволяющий с использованием коэффициентов корреляции определить, существует ли зависимость между переменными и насколько она сильна.

Коэффициент корреляции – двумерная описательная статистика, количественная мера взаимосвязи (совместной изменчивости) двух переменных.

Коэффициент корреляции

1. Может принимать значения от -1 до +1
2. Знак коэффициента показывает направление связи (прямая или обратная)
3. Абсолютная величина показывает силу связи
4. Всегда основан на парах чисел (измерений 2-х переменных от одного объекта или 2-х переменных от разных, но связанных объектов)

r – коэффициент корреляции

Рост братьев: коэффициент корреляции - r?

Вася

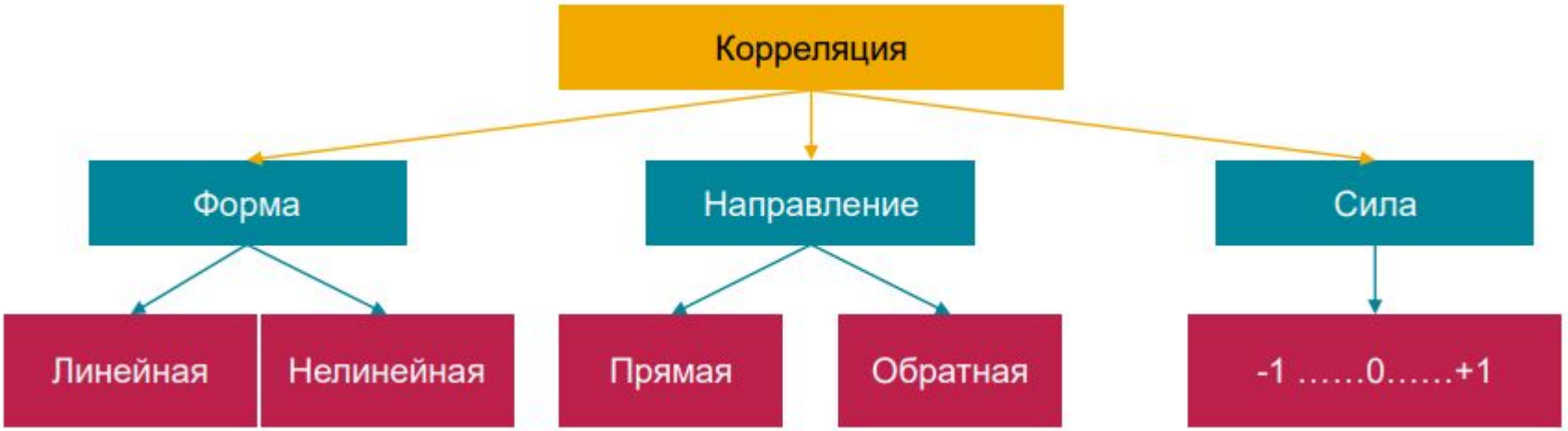
Юра

1. r=1.0: если Вася высокого роста, значит, Юра тоже высокий, это не предположение, а факт.

2. r=0.7: если Вася высокий, то, скорее всего, Юра тоже высокий.

3. r=0.0: если Вася высокий, то мы… не

Характер связи между переменными



При положительной линейной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого.

При отрицательной линейной корреляции более высоким значениям одного признака соответствуют более низкие значения другого, а более низким значениям одного признака – высокие значения другого.

Виды связи между переменными

- ✔ Прямая причинно-следственная связь - переменная X определяет значение переменной У.

Пример: Наличие воды ускоряет рост растений. Яд вызывает смерть. Температура воздуха прямо влияет на скорость таяния льда.

- ✔ Обратная причинно-следственная связь - переменная У определяет значение переменной X.

Пример: Исследователь может думать, что чрезмерное потребление кофе вызывает нервозность. Но, может быть, очень нервный человек выпивает кофе, чтобы успокоить свои нервы?

Виды связи между переменными

✔ Связь, вызванная третьей (скрытой) переменной.

Пример: существует зависимость между числом утонувших людей и числом выпитых безалкогольных напитков в летнее время. Однако, обе переменные связаны с жарой и потребностью людей во влаге?

✔ Связь, вызванная несколькими скрытыми переменными.

Пример: Исследователь может обнаружить значимую связь между оценками студентов в университете и оценками в школе. Но действуют и другие переменные: IQ, количество часов занятий, влияние родителей, мотивация, возраст, авторитет преподавателей.

✔ Связи нет, наблюдаемая зависимость случайна.

Пример: Исследователь может найти связь между увеличением количества людей, которые занимаются спортом и увеличением количества людей, которые совершают преступления. Но здравый смысл говорит, что любая связь между этими двумя переменными является случайной.

Примеры корреляций

а) строгая положительная корреляция

б) положительная корреляция

в) слабая положительная корреляция

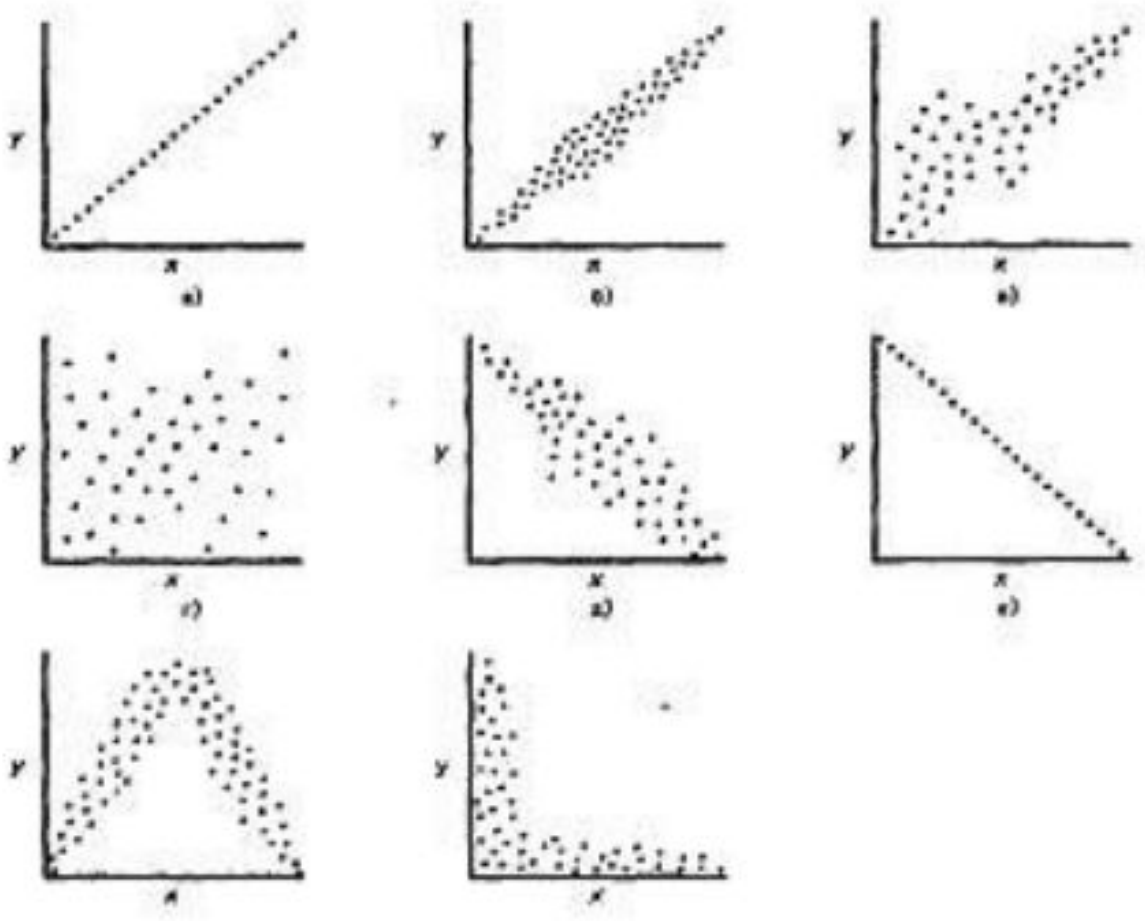
г) нулевая корреляция

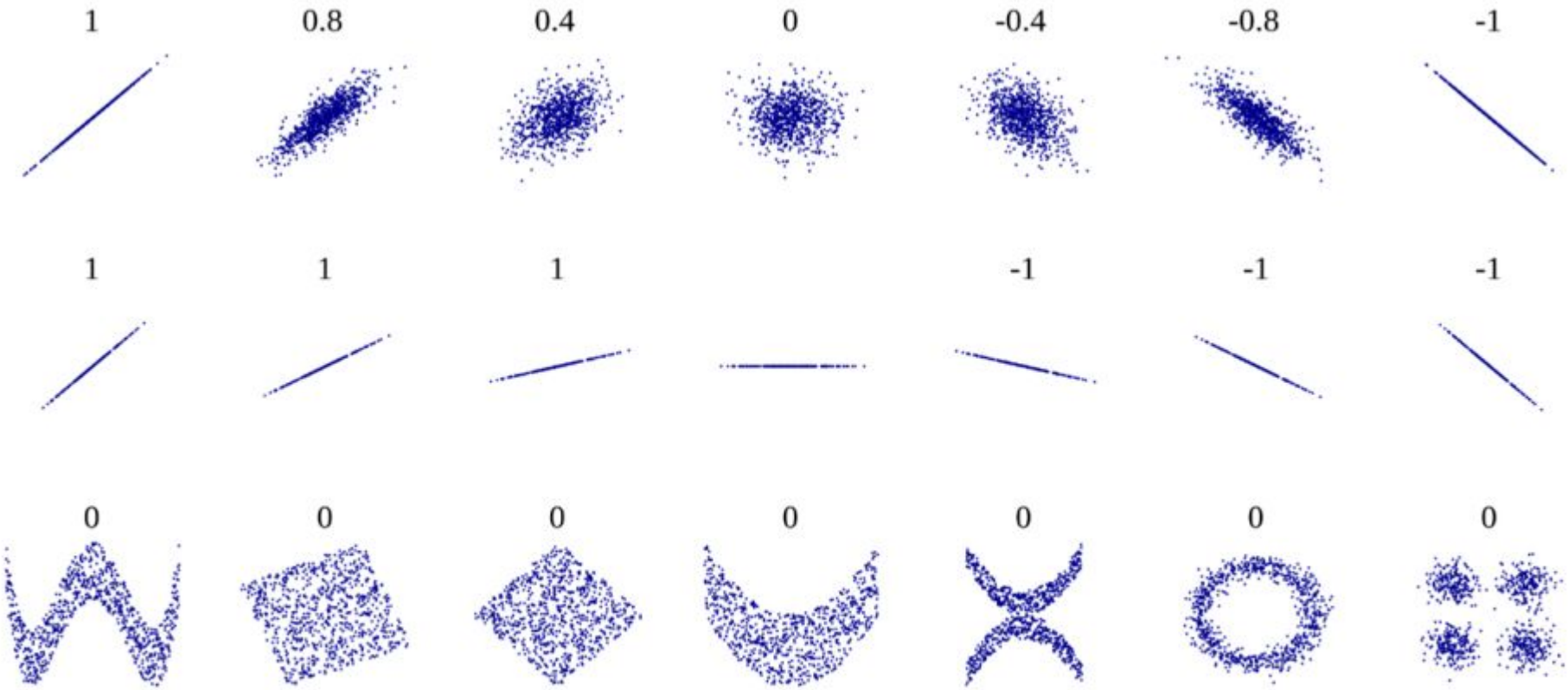
д) отрицательная корреляция

е) строгая отрицательная корреляция

ж) нелинейная корреляция

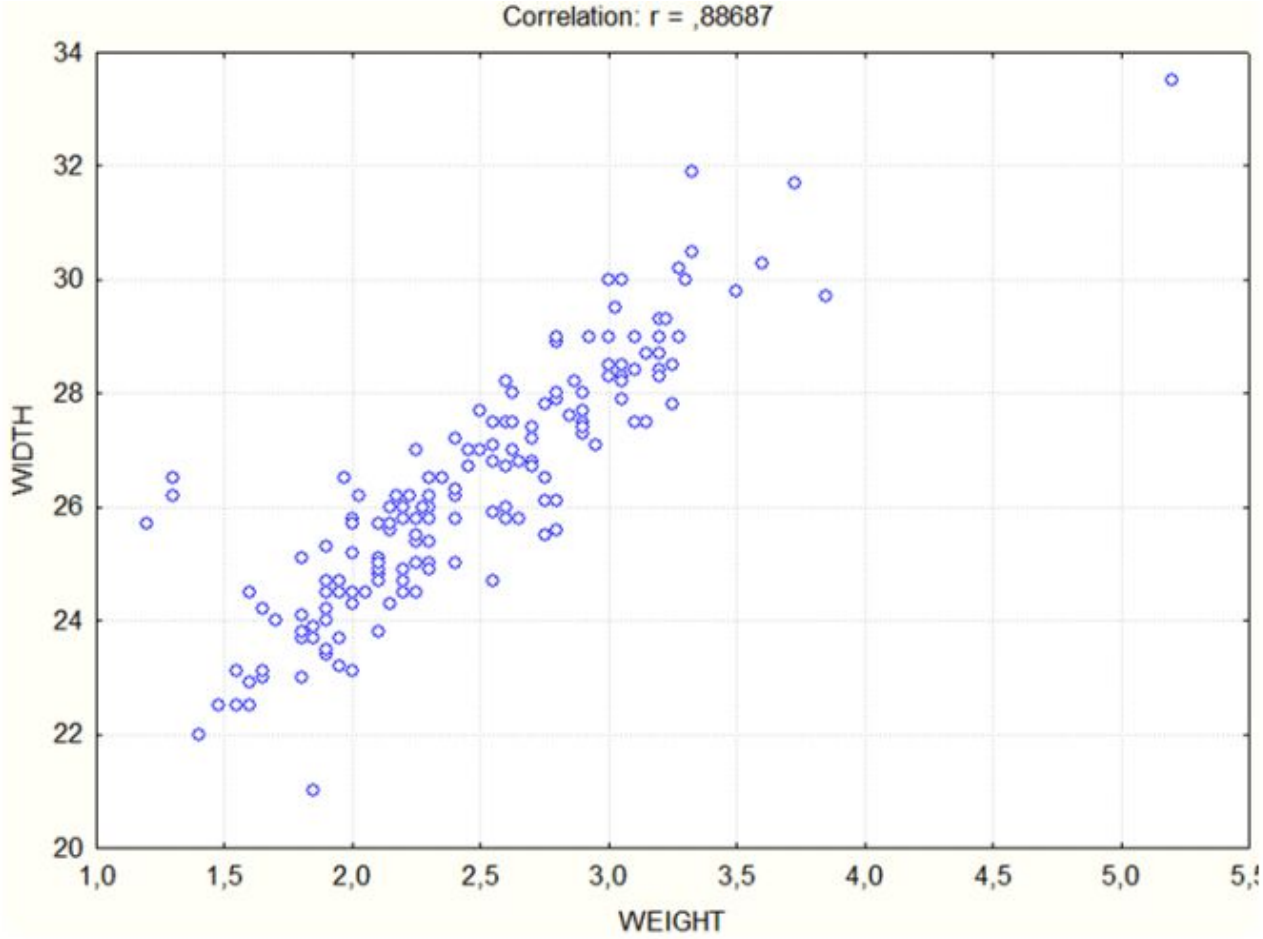
з) нелинейная корреляция




1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

Скаттерплот

(= диаграмма рассеяния)



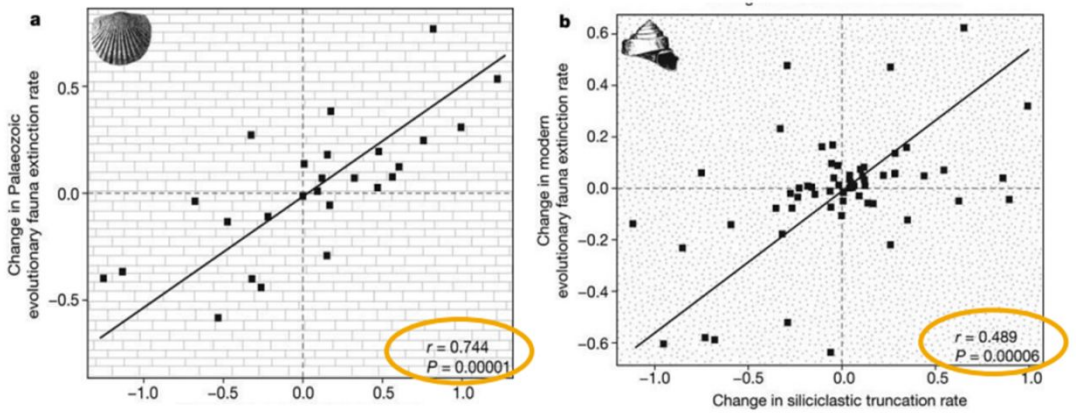
Две характеристики: – наклон (направление связи) и ширина

(сила связи) воображаемого эллипса

Сила корреляции

- ✔ Сила связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции.
- ✔ Коэффициент корреляции (r) – это показатель, величина которого варьируется в пределах от –1 до +1.
- ✔ Если коэффициент корреляции равен 0, обе переменные линейно независим

ЗНАЧЕНИЕ (по модулю)	ИНТЕРПРЕТАЦИЯ
до 0,2	очень слабая корреляция
до 0,5	слабая корреляция
до 0,7	средняя корреляция
до 0,9	высокая корреляция
свыше 0,9	очень высокая корреляция



Коэффициенты корреляции

1. Для порядковых данных используются следующие коэффициенты корреляции:

✔ ρ - коэффициент ранговой корреляции Спирмена
✔ τ - коэффициент ранговой корреляции Кендалла
✔ γ - коэффициент ранговой корреляции Гудмена – Краскела

2. Для переменных с интервальной и номинальной шкалой используется коэффициент корреляции Пирсона (корреляция моментов произведений).

3. Если, по меньшей мере, одна из двух переменных имеет порядковую шкалу, либо не является нормально распределённой, используется ранговая корреляция Спирмана или τ-Кендалла. Применение коэффициента Кендалла предпочтительно, если в исходных данных имеются выбросы.

Типы шкал		Мера связи
Переменная X	Переменная Y	
Интервальная или отношений	Интервальная или отношений	Коэффициент Пирсона
Ранговая, интервальная или отношений	Ранговая, интервальная или отношений	Коэффициент Спирмена
Ранговая	Ранговая	Коэффициент Кендалла
Дихотомическая	Дихотомическая	Коэффициент φ,
Дихотомическая	Ранговая	Рангово-бисериальный коэффициент
Дихотомическая	Интервальная или отношений	Бисериальный коэффициент
Интервальная	Ранговая	Не разработан

Коэффициент корреляции Пирсона

Коэффициент корреляции r-Пирсона является мерой прямолинейной связи между переменными: его значения достигают максимума, когда точки на графике двумерного рассеяния лежат на одной прямой линии.

$$r = \frac{\sum z_{X_i} z_{Y_i}}{n-1}$$

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$$

$$z_{X_i} = \frac{X_i - \bar{X}}{s_X}$$

$$z_{Y_i} = \frac{Y_i - \bar{Y}}{s_Y}$$

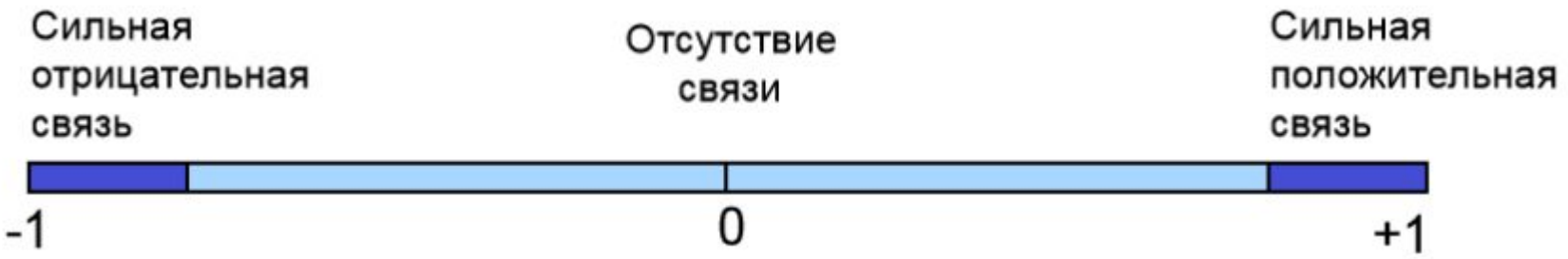
стандартное отклонение для веса

стандартное отклонение для роста

для каждого X и Y (для каждого респондента)

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right]}}$$

Интерпретация результатов



Значение r – Пирсона характеризует уровень связи между переменными:

- ✔ 0,75 – 1.00 очень высокая положительная
- ✔ 0,50 – 0.74 высокая положительная
- ✔ 0,25 – 0.49 средняя положительная
- ✔ 0,00 – 0.24 слабая положительная
- ✔ 0,00 – -0.24 слабая отрицательная
- ✔ -0,25 – -0.49 средняя отрицательная
- ✔ -0,50 – -0.74 высокая отрицательная
- ✔ -0,75 – -1.00 очень высокая отрицательная

Оценка статистической значимости коэффициента корреляции

Критическое значение t-критерия определяется из таблицы значений t-распределения для выбранного уровня значимости α и числа степеней свободы f=n–2

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

ВАЖНО ЗАПОМНИТЬ!

- ✔ Коэффициент корреляции r - Пирсона оценивает только линейную связь переменных. Нелинейную связь данный коэффициент выявить не может.
- ✔ Коэффициент корреляции Пирсона очень чувствителен к аутлаерам (выбросам).
- ✔ Корреляция не подразумевает наличия причинно-следственной связи между переменными.
- ✔ Нельзя путать коэффициент корреляции Пирсона с критерием Пирсона ХИ-квадрат

Регрессионный анализ

Регрессионный анализ – инструмент для количественного предсказания значения одной переменной на основании другой.

Для этого в линейной регрессии строится прямая – **линия регрессии**.

Простая линейная регрессия:

Даёт нам правила, определяющие линию регрессии, которая ЛУЧШЕ ДРУГИХ предсказывает одну переменную на основании другой (переменных всего две).

По оси Y располагают переменную, которую мы хотим предсказать (зависимую), а по оси X – переменную, на основе которой будем предсказывать (независимую).

Основные термины

Уравнение регрессии - это математическая формула, применяемая к независимым переменным, чтобы лучше спрогнозировать зависимую переменную, которую необходимо смоделировать.

Зависимая переменная (Y) — это переменная, описывающая процесс, который мы пытаемся предсказать или понять.

Независимые переменные (X) - это переменные, используемые для моделирования или прогнозирования значений зависимых переменных. В уравнении регрессии они располагаются справа от знака равенства и часто называются объяснительными переменными. Зависимая переменная - это функция независимых переменных.

Коэффициенты регрессии (β) — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.

Невязки - существует необъяснимое количество зависимых величин, представленных в уравнении регрессии как случайные ошибки ε.

Регрессия и корреляция

То есть,
РЕГРЕССИЯ – предсказание одной переменной на основании другой. Одна переменная – независимая, а другая – зависимая.

Пример: чем дольше проводить измельчение объекта, тем меньший размер частиц получишь

КОРРЕЛЯЦИЯ – показывает, в какой степени две переменные СОВМЕСТНО ИЗМЕНЯЮТСЯ. Нет зависимой и независимой переменных, они эквивалентны.

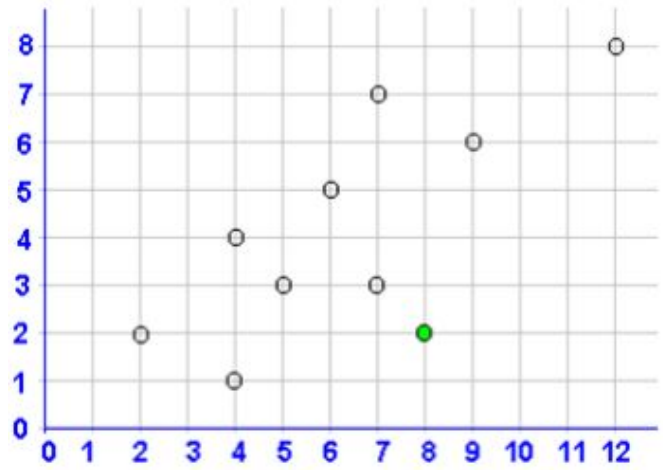
Пример: размер частиц коррелирует положительно с его массой

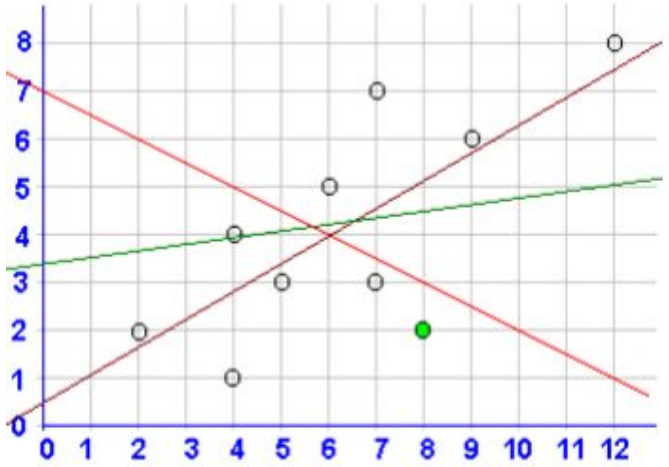
Пример:

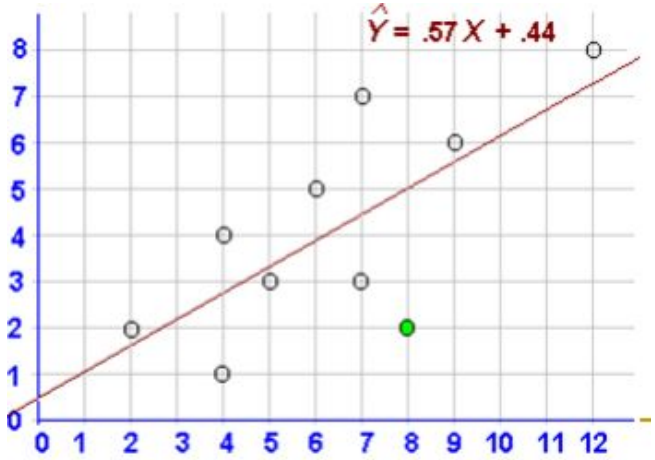
Мы изучаем размер частиц объекта от времени его измельчения. Мы хотим узнать, как влияет время измельчения на размер частиц?

У нас две переменные –

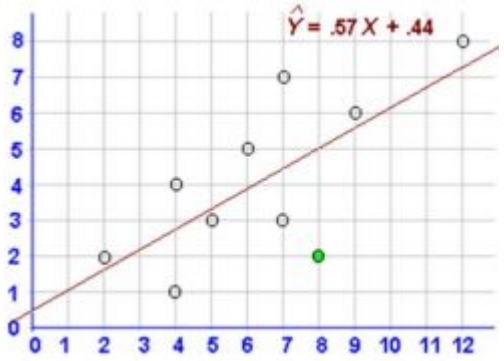
1. Время измельчения, ч (independent);
2. Размер частиц, нм (dependent)







Простая линейная регрессия



Y – зависимая переменная

X – независимая переменная

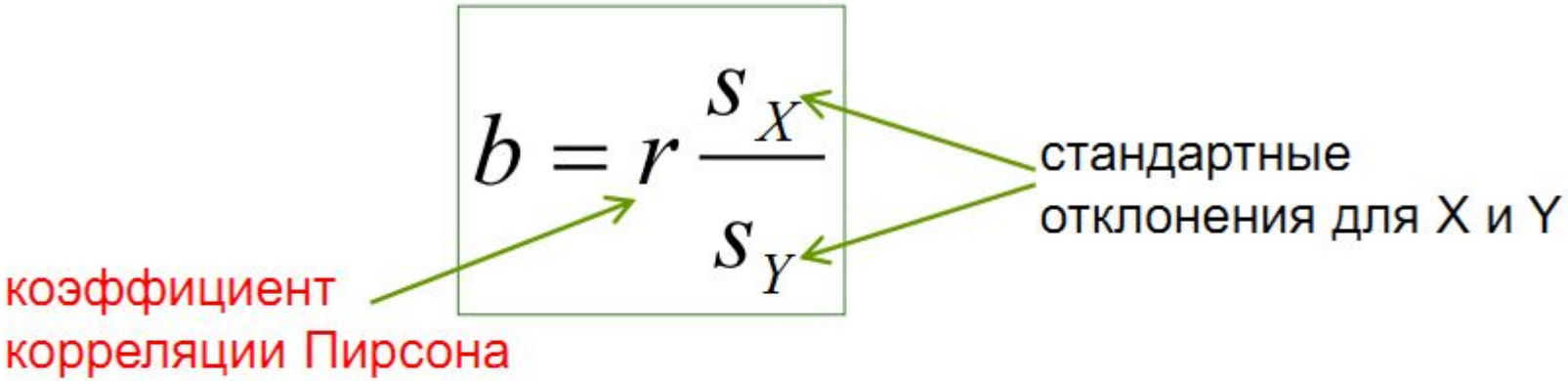
a и b - коэффициенты регрессии

$$Y_i = a + bX_i$$

b – характеризует НАКЛОН прямой;

a – определяет точку пересечения прямой с осью OY;

Задача сводится к поиску коэффициентов a и b.



$$\bar{Y} = a + b\bar{X} \longrightarrow a = \bar{Y} - b\bar{X}$$

Линия регрессии всегда проходит через точку, то есть через середину графика.

b – определяет, насколько изменится Y на единицу X;

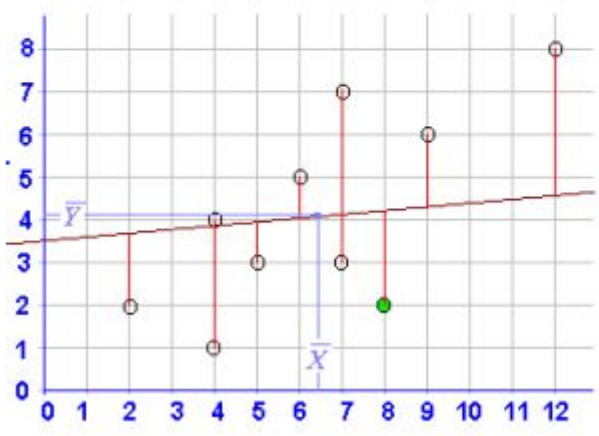
имеет тот же знак, что и r.

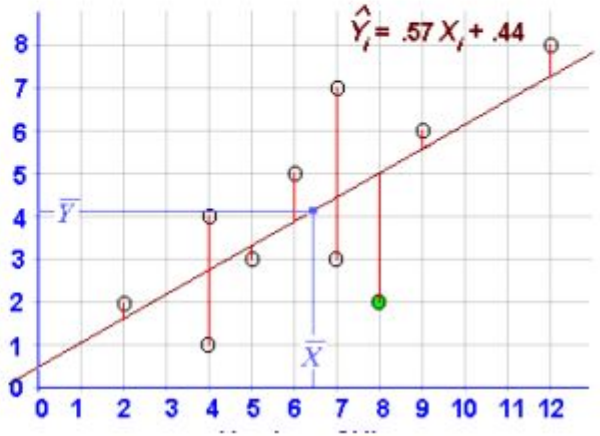
Как определить наилучшую линию регрессии?

Используют метод наименьших квадратов – подбирают такую линию регрессии чтобы общая сумма квадратов отклонений значений зависимой переменной была наименьшей.

$$\sum e_i = 0$$

$$\sum e_i^2 \quad \text{- минимальна}$$





Суть метода наименьших квадратов

Пусть имеются n наблюдений признаков x и y. Причем известен вид уравнения регрессии - f(x), например, прямолинейная зависимость: $f(x_i)=a + b \cdot x_i$

Необходимо подобрать такие значения параметров (a и b), которые смогут минимизировать сумму квадратов отклонений фактических значений признака-результата y_i от расчетных (теоретических) значений $f(x_i)$ для всех наблюдений i=1:n

$$S = \sum_{i=1}^{n} (y_i - (a + b \cdot x_i))^2 \Rightarrow \min_{a,b}$$

Основное положение МНК

сумма квадратов отклонений εi экспериментальных точек от кривой по вертикальному направлению, т.е. сумма квадратов величин εi, должна быть наименьшей ($\Sigma \varepsilon i^2$ = минимум).

Коэффициенты a и b в МНК

$$a = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Оценка качества уравнения регрессии и коэффициент детерминации

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{\sum_i (Y_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

$$SS_{total} = \sum_i (Y_i - \bar{Y})^2$$

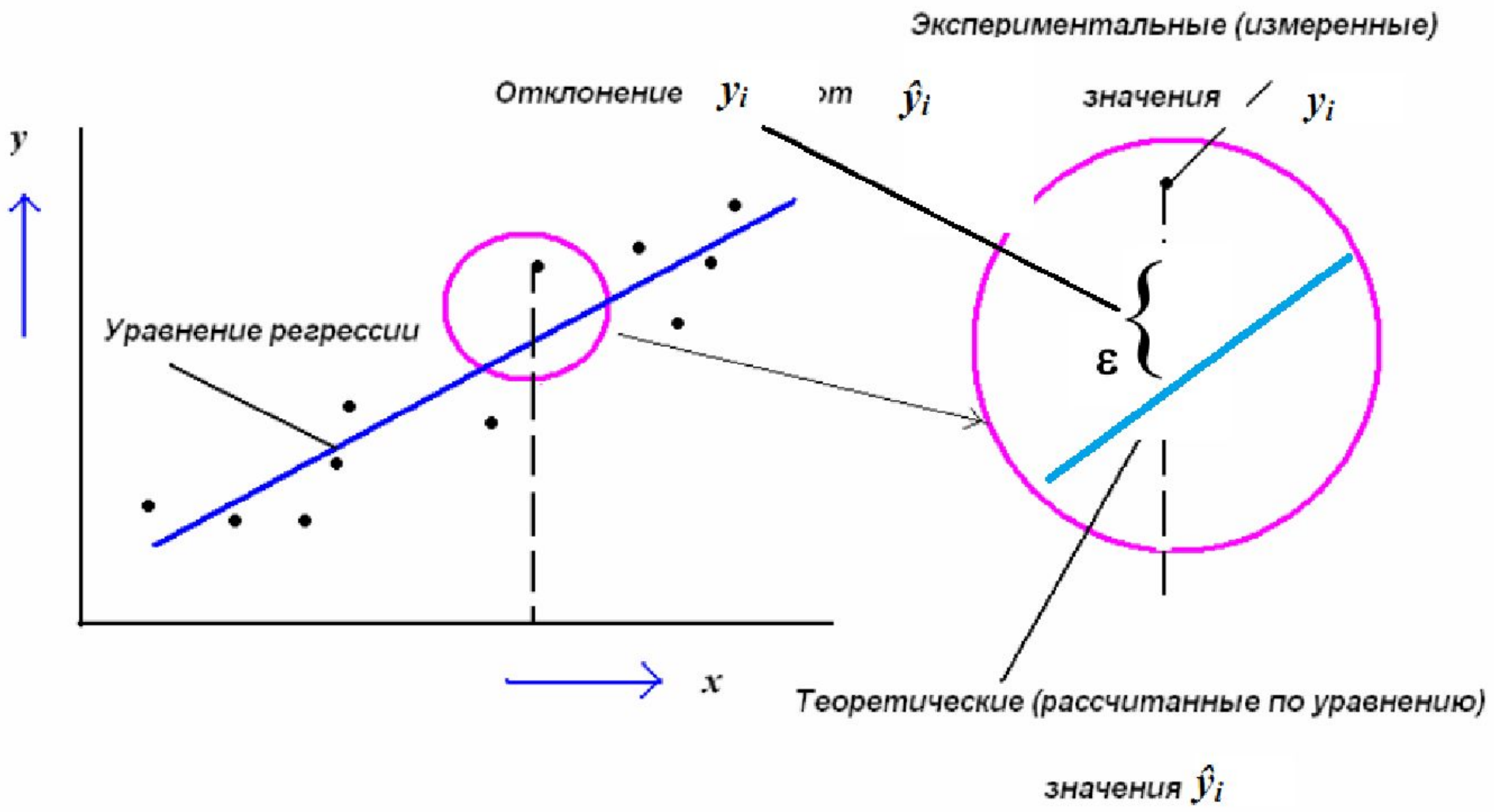
$$SS_{regression} = \sum_i (\hat{Y}_i - \bar{Y})^2$$

$$SS_{residual} = \sum_i (Y_i - \hat{Y}_i)^2$$

Коэффициент множественной детерминации R-квадрат показывает, какую долю изменчивости (можно выразить в процентах) зависимой переменной (Y) объясняет независимая переменная (регрессионная модель).

- ✔ Под качеством уравнения регрессии понимается степень близости (соответствия) рассчитанных по данному уравнению значений признака результата f(x) фактическим (наблюдаемым) значениям y.
- ✔ Чем ближ… …рессионной модели.

r – коэффициент корреляции, $r^2 = R^2$

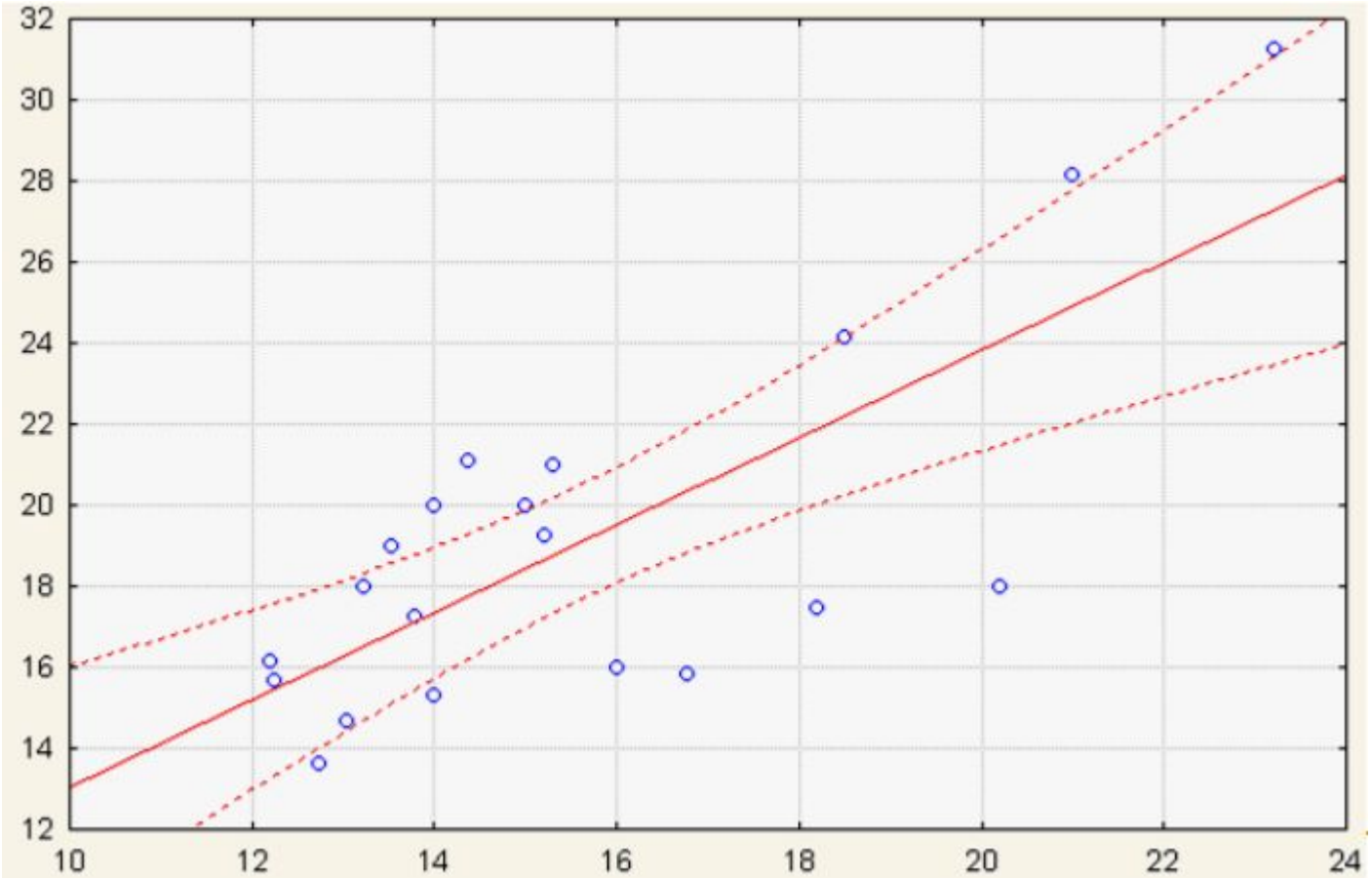

Экспериментальные (измеренные)
Отклонение y_i от $\hat{y}_i$
значения y_i
y
Уравнение регрессии
ε
x
Теоретические (рассчитанные по уравнению)
значения $\hat{y}_i$

ВАЖНЫЕ ЗАМЕЧАНИЯ

Любая регрессионная модель позволяет обнаружить только количественные зависимости, которые не обязательно отражают причинные зависимости, т.е. влияние одного фактора на другой.

Гипотезы о причинной связи признаков должны дополнительно обосновываться с помощью теоретического анализа, содержательно объясняющего изучаемое явление или процесс.

Доверительный интервал для значений зависимой переменной: строится для каждого значения X, причём наименьшая ошибка получается для среднего Y.



Сравнение двух (и более) уравнений линейной регрессии

1. Сравнение коэффициентов наклона b_1 b_2
2. Сравнение коэффициентов сдвига a_1 и a_2

На основе критерия Стьюдента

Сравнение двух линий регрессии в целом (предполагается, что если линии для 2-х выборок у нас сильно различаются, и мы объединим выборки, то общая линия по этим двум выборкам будет хуже описывать изменчивость, остаточная дисперсия будет больше)

Критерий Фишера

- Вычисляем дисперсию аналитического сигнала относительно его средних значений Sy:

$$S_y^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - y_i)^2}{m(n-1)}$$

- Градуировочная зависимость признается удовлетворительной, если

$$S_o^2 / S_y^2 \le F_{табл},$$

где табл F — табличное значение критерия Фишера для m − 2 и m(n −1) степеней свободы. Этот критерий служит для обнаружения отклонения градировочной зависимости от уравнения регрессии.

Критерий Стьюдента

1. Вычисляем коэффициент корреляции r по формуле:

$$r = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y}) / (m-1)}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 / (m-1) \cdot \sum_{i=1}^{m}(y_i - \bar{y})^2 / (m-1)}}.$$

2. Вклад градуировки в погрешность результат анализа вычисляют по формуле:

$$\Delta_{\text{гр}} = \frac{t(P,f) \cdot S_{\text{o}}}{b} \cdot \sqrt{\frac{1}{n_x} + \frac{1}{m} + \left(\frac{S_b}{b}\right)^2 \cdot \frac{(\bar{y}_x - \bar{y})^2}{S_{\text{o}}^2}},$$

где число степеней свободы; f =(m−2) y — среднее арифметическое значение аналитического сигнала для всех образцов для градуировки, полученное при градуировке, y — среднее арифметическое значение аналитического сигнала при анализе пробы, n — число измерений аналитического сигнала при анализе пробы; а Sb — значение СКО углового коэффициента b, которое вычисляется по форму.

$$S_b = S_{\text{o}} \cdot \sqrt{\frac{m}{m\sum_{i=1}^{m} x_i^2 - \left(\sum_{i=1}^{m} x_i\right)^2}}.$$

Если погрешностью градуировки, рассчитанная по формуле (3.4), сопоставима с погрешностью приготовления образцов для градуировки θ_p, то есть выполняется условие $\theta_p > 0,4S_b$), то ее необходимо учитывать, и в этом случае

$$\Delta'_{\text{гр}} = \sqrt{\Delta^2_{\text{гр}} + \theta^2_p}.$$

Как следует из выражений (3.3) и (3.4) вклад градуировки в погрешность результата анализа уменьшается при увеличении угла наклона градуировочного графика b, при уменьшении рассеяния точек относительно графика S_o, при увеличении числа образцов для градуировки m и уменьшении их погрешности приготовления θ_p.

Требования к выборке для проведения регрессионного анализа

1. Ожидаемая зависимость переменной Y от X должна быть линейной.
2. Для любого значения X_i Y должна иметь нормальное распределение.
3. Для любого значения X_i выборки для Y должны иметь одинаковую дисперсию.
4. Для любого значения X_i выборки для Y должны быть независимы друг от друга.
5. Размер выборки должен более чем в 10 раз превосходить число переменных в анализе (лучше – в 20 раз).
6. Следует исключить аутлаеры (промохи)