# Лекция 3

## КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

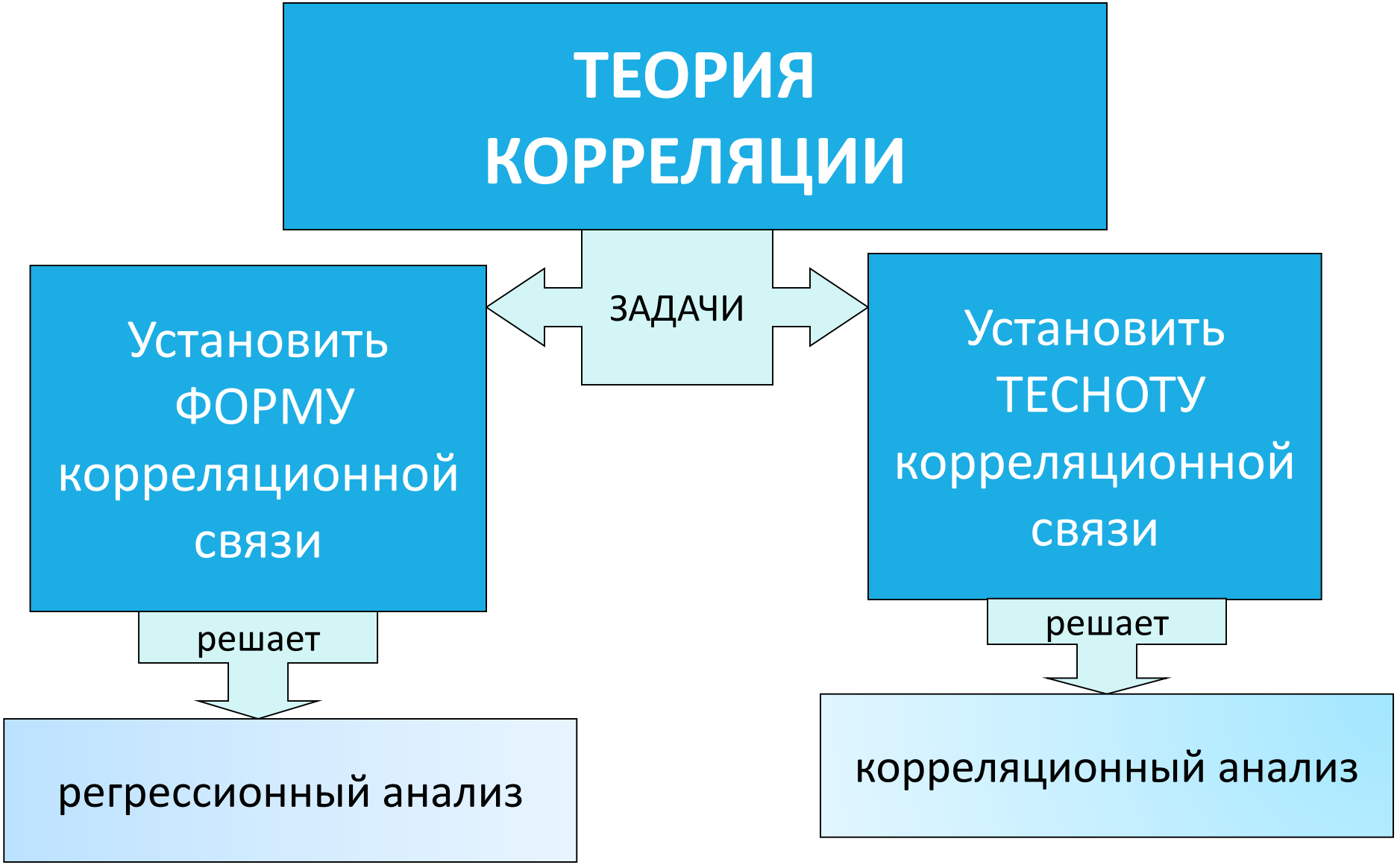
# Понятие корреляционной зависимости

Многие задачи требуют установить и оценить зависимость между двумя или несколькими случайными величинами.

- Зависимость случайных величин называют ***статистической***, если изменение одной величины влечет изменение распределения другой величины.
- Статистическая зависимость называется ***корреляционной***, если при изменении одной величины изменяется среднее значение другой.

Если случайная величина представляет некоторый признак (например, статистические наблюдения некой экономической величины), то под **корреляцией** понимают – меру согласованности одного признака с другим, или с несколькими, либо взаимную согласованность группы признаков.

Функциональная зависимость предполагает взаимно однозначное соответствие аргумента $x$ и функции $y=f(x)$, вероятностная же зависимость допускает некий условный диапазон, в который предположительно (с такой-то долей вероятности) попадает значение признака $y_i$ при значении $x_i$ признака **x**.


ТЕОРИЯ КОРРЕЛЯЦИИ
ЗАДАЧИ
Установить ФОРМУ корреляционной связи
Установить ТЕСНОТУ корреляционной связи
решает
решает
регрессионный анализ
корреляционный анализ

# Корреляционный анализ

**Корреляционный анализ** – один из методов исследования взаимосвязи между двумя или более переменными.

Для применения линейного корреляционного анализа величины, образующие пары, должны быть распределены нормально.

Корреляционная зависимость характеризуется ***формой и теснотой связи***.

**Функция регрессии** определяет форму связи при изучении статистических зависимостей, а тесноту связи определяют с помощью коэффициента корреляции.

# Корреляционный анализ

**В качестве числовой характеристики вероятностной связи используют коэффициенты корреляции**, значения которых изменяются в диапазоне от –1 до +1. После проведения расчетов исследователь, как правило, отбирает только наиболее сильные корреляции, которые в дальнейшем интерпретируются

***Критерием для отбора «достаточно сильных» корреляций*** может быть как абсолютное значение самого коэффициента корреляции (от 0,7 до 1), так и относительная величина этого коэффициента, определяемая по уровню статистической значимости (от 0,01 до 0,1), зависящему от размера выборки.

***В малых выборках*** для дальнейшей интерпретации корректнее отбирать сильные корреляции на основании уровня статистической значимости.

Для исследований, которые проведены на больших выборках, лучше использовать абсолютные значения коэффициентов корреляции.

# Корреляционный анализ. Подготовка данных

**Измерение** - процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу.

В процессе подготовки данных измеряется не сам объект, а его характеристики.

***Шкала*** - правило, в соответствии с которым объектам присваиваются числа.

Переменные могут являться **числовыми** данными либо **символьными**.

Числовые данные, в свою очередь, могут быть дискретными и непрерывными.

Существует пять типов шкал измерений: номинальная, порядковая, интервальная, относительная и дихотомическая.

# Корреляционный анализ. Коэффициенты корреляции

| Значение | 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 |

**В настоящее время разработано множество различных коэффициентов корреляции.** Наиболее применяемыми являются r-Пирсона, r-Спирмена и τ-Кендалла.

Современные компьютерные статистические программы предлагают именно эти три коэффициента, а для решения других исследовательских задач предлагаются методы сравнения групп.

Выбор метода вычисления коэффициента корреляции зависит от типа шкалы, к которой относятся переменные

| 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 $\varphi$, четырехполевая корреляция |
| Дихотомическая | Ранговая | Рангово-бисериальный коэффициент |
| Дихотомическая | Интервальная или отношений | Бисериальный коэффициент |
| Интервальная | Ранговая | Не разработан |

# Корреляционный анализ. Подготовка данных. Типы шкал

| Название | Содержание | Пример |
|---|---|---|
| **Номинальная шкала (nominal scale)** | шкала, содержащая только категории; данные в ней не могут упорядочиваться, с ними не могут быть произведены никакие арифметические действия. | профессии, город проживания, семейное положение |
| **Порядковая (ранговая) шкала (ordinal scale):** | шкала, в которой числа присваивают объектам для обозначения относительной позиции объектов, но не величины различий между ними. | место (1, 2, 3-е), которое команда получила на соревнованиях, номер студента в рейтинге успеваемости (1-й, 23-й, и т.д.), при этом неизвестно, насколько один студент успешней другого, известен лишь его номер в рейтинге |
| **Интервальная шкала (interval scale):** | шкала, разности между значениями которой могут быть вычислены, однако их отношения не имеют смысла. | температура воды в море утром - 19 градусов, вечером - 24, т.е. вечерняя на 5 градусов выше, но нельзя сказать, что она в 1,26 раз выше |
| **Относительная шкала (ratio scale) или шкала отношений:** | шкала, в которой есть определенная точка отсчета и возможны отношения между значениями шкалы. Является числовой | вес новорожденного ребенка (4 кг и 3 кг). Первый в 1,33 раза тяжелее |
| **Дихотомическая шкала (dichotomous scale):** | шкала, содержащая только две категории. | пол (мужской и женский) |

# Коэффициент корреляции Пирсона

Наиболее часто используемый **коэффициент корреляции Пирсона** *r* измеряет степень линейных связей между переменными.

Числовое значение коэффициента корреляции Пирсона определяется формулой:

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

# Коэффициент корреляции Пирсона

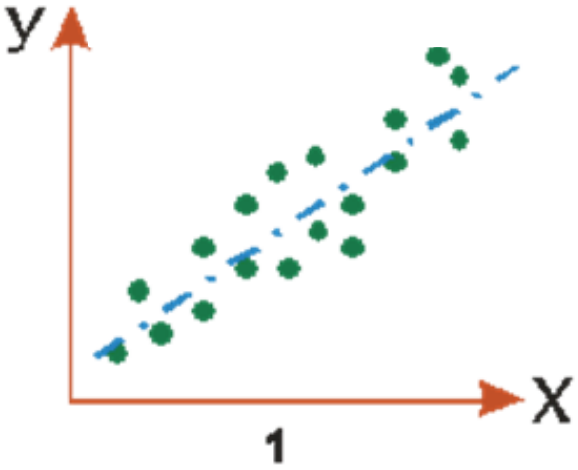
- Значение коэффициента корреляции $r$ лежит в интервале [-1;1].
- При $r=0$ корреляция отсутствует.
- При $|r|=1$ корреляция является полной или абсолютной.
- Чем ближе $|r|$ к 1, тем теснее связь между переменными.
- Отрицательное значение коэффициента корреляции свидетельствует об обратной зависимости между переменными, положительное значение – о прямой.

# Коэффициент корреляции Пирсона

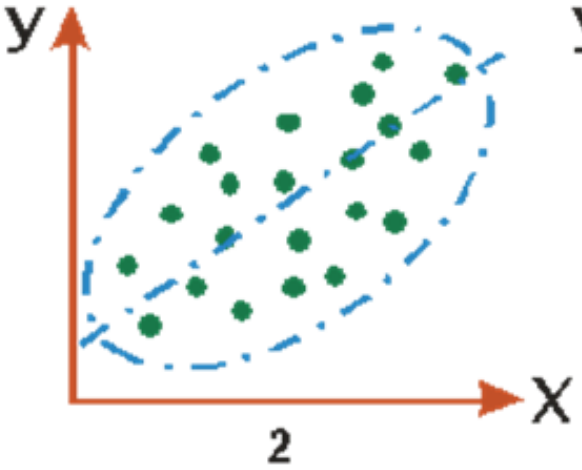
Чем больше разбросанность точек по всему корреляционному полю, тем слабее зависимость между переменными.

Если на графике зависимость можно представить прямой линией (с положительным или отрицательным углом наклона), то корреляция между переменными будет высокая.

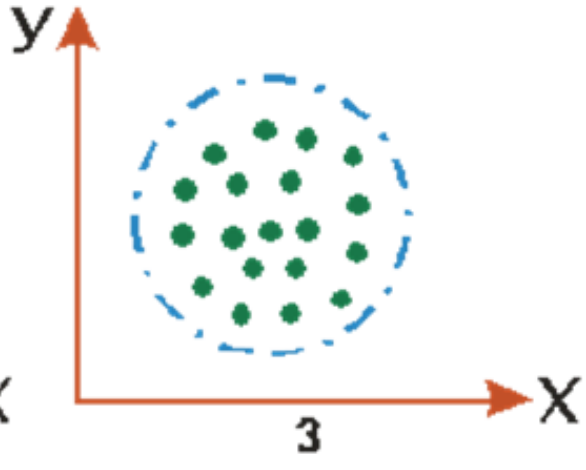
# Примеры корреляционной зависимости




Значимая положительная корреляция (r>0,8)




Имеется некоторая корреляция (0,5 <r<0,6)




Корреляция отсутствует ( r = 0)

Значимая положительная корреляция ( r = +1)
Иначе: функциональная зависимость

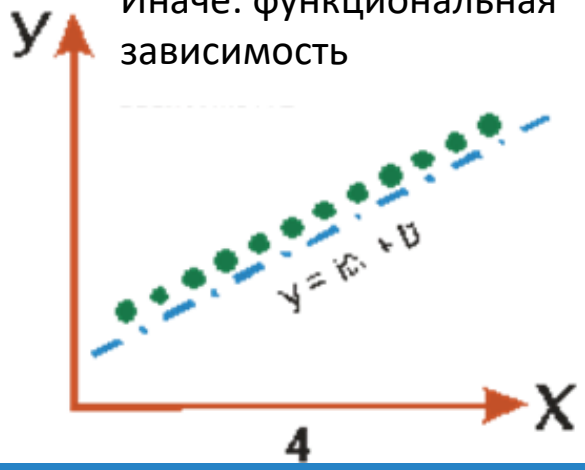
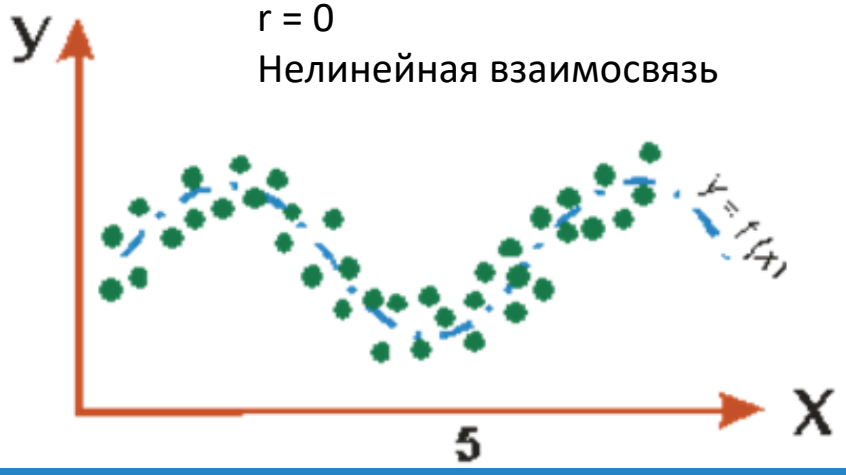



Диаграмма рассеяния показывает однозначное соответствие
r = 0
Нелинейная взаимосвязь



# Проверки гипотезы о независимости наблюдений

Для проверки гипотезы о независимости наблюдений используют **t-критерий Стьюдента**.

Расчетное значение критерия вычисляется по формуле

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

Если критическое значение критерия Стьюдента, соответствующее выбранному уровню значимости и числу степеней свободы, равному

$n$ - 2 (где $n$ – число пар ($x$, $y$)), меньше расчетного ($t_{кр}<t$), то гипотеза о независимости значений $X$ и $Y$ должна быть отвергнута.

# Показатель ранговой корреляции Спирмена

Для определения корреляции порядковых признаков используют **показатель ранговой корреляции Спирмена**.

Расчет такого коэффициента корреляции не требует нормальности распределения и линейной зависимости от переменных, и он может быть применен как к количественным, так и порядковым признакам.

# Показатель ранговой корреляции Спирмена

Идея **коэффициента ранговой корреляции Спирмена** заключается в том, что:

- все данные упорядочиваются по возрастанию переменной, а сами значения заменяются их рангами,

- затем вычисляются разностные ранги, по которым рассчитывается *коэффициент корреляции Спирмена* по формуле

$$r_s = 1 - \frac{6\sum d_i}{n^3 - n},$$

где $d_i$ - разность рангов для каждого члена выборки;

$n$ - число пар значений $x$, $y$.

# Показатель ранговой корреляции Спирмена

Для определения различий между признаками находят критическое значение **коэффициента корреляции Спирмена** для выбранного доверительного уровня и заданного объема выборки. Если объем выборки n>50, то используют критерий Стьюдента

$$t_s = \frac{r_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}}.$$

# Показатель ранговой корреляции Спирмена

Расчетное значение сравнивается с критическим для числа степеней свободы m=n-2 и заданным уровнем значимости.

Если критическое значение критерия Спирмена меньше, чем расчетное (или критическое значение t-критерия Стьюдента меньше, чем расчетное для случая n>50), то различия считаются статистически значимыми.

# Коэффициент корреляции Кенделла

**Коэффициент ранговой корреляции $\tau$-Кендалла** является самостоятельным оригинальным методом, опирающимся на вычисление соотношения пар значений двух выборок, имеющих одинаковые или отличающиеся тенденции (возрастание или убывание значений).

Этот коэффициент называют еще *коэффициентом конкордации*.

***Основной идеей*** данного метода является то, что о направлении связи можно судить, попарно сравнивая между собой «испытуемых»:

- если у пары «испытуемых» изменение по *X* совпадает по направлению с изменением по *Y*, это свидетельствует о положительной связи,
- если не совпадает – об отрицательной связи, например, при исследовании личностных качеств, имеющих определяющее значение для семейного благополучия.

В этом методе одна переменная представляется в виде монотонной последовательности в порядке возрастания величин; другой переменной присваиваются соответствующие ранговые места.

Количество инверсий (нарушений монотонности по сравнению с первым рядом) используется в формуле для корреляционных коэффициентов.

# Коэффициент корреляции Кенделла

**Коэффициент корреляции $\tau$-Кенделла** (Kendall tau rank correlation coefficient) – мера линейной связи между случайными величинами.

Корреляция Кенделла является ранговой, то есть для оценки силы связи используются не численные значения, а соответствующие им ранги.

Коэффициент инвариантен по отношению к любому монотонному преобразованию шкалы измерения.

Заданы две выборки $x=(x_1,\ldots,x_n),\ y=(y_1,\ldots,y_n)$

Коэффициент корреляции Кенделла вычисляется по формуле:

$$R=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\left[x_i<x_j\right]\neq\left[y_i<y_j\right]\right]$$

где $\tau=1-\frac{4}{n(n-1)}R$ - количество инверсий, образованных величинами $y_i$ расположенными в порядке возрастания соответствующих $x_i$.

# Коэффициент корреляции Кенделла

Коэффициент $\tau$ принимает значения из отрезка [-1, 1].

Равенство $\tau=1$ указывает на строгую прямую линейную зависимость,

$\tau=-1$ на обратную.

Коэффициент $\tau$ (линейно связанный с $R$ ) можно считать мерой неупорядоченности второй последовательности относительно первой.

# Бисериальный коэффициент корреляции

В тех случаях, когда одна переменная измеряется в дихотомической шкале (переменная x), а другая в шкале интервалов или отношений (переменная y), используется так называемый бисериальный коэффициент корреляции. Переменная x, полученная в дихотомической шкале, принимает только два значения (кода) 0 и 1.

$$R_{\text{бис}} = \frac{\bar{x}_1 - \bar{x}_0}{S_y}\sqrt{\frac{n_1 n_0}{N(N-1)}}$$

Несмотря на то что бисериальный коэффициент корреляции принимает значения в диапазоне от -1 до + 1, его знак для интерпретации результатов не играет никакой роли. Это одно из исключений из общего правила. В данном случае речь идет только о наличии или отсутствии значимой связи.

# Коэффициент сопряженности Бравайса

В случае, если данные представлены в номинальной шкале типа «*да*» и «*нет*» (т.е. имеется таблица сопряженности 2х2), то для выяснения тесноты связи используется специальная форма коэффициента корреляции Пирсона, которая носит название **коэффициента сопряженности Бравайса**.

Расчет коэффициента сопряженности Бравайса проводится по формуле

$$C = \frac{ad - bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$$

где $a$, $b$, $c$, $d$ - значения в клетках таблицы 2х2.

# Регрессионный анализ

Регрессионный анализ направлен на получение статистических оценок параметров в уравнении связи изучаемых переменных.

Вид уравнения связи между переменными определяется характером поставленных задач. Чаще всего используется зависимость линейного типа

$$Y=a_0+a_1X,$$

и в качестве статистических оценок определяются коэффициенты $a_0$ и $a_1$ уравнения линейной регрессии.

# Регрессионный анализ

Помимо линейной зависимости существует множество ***нелинейных моделей регрессии*** относительно входящих в них параметров. Приведем примеры некоторых из них.

- Мультипликативная зависимость: $y = a_0 x^{b_0}$.
- Экспоненциальная зависимость: $y = e^{a_0 x + a_1}$.
- Логарифмическая зависимость: $y = a_0 + a_1 x + a_2 \ln x.$
- Степенная зависимость: $y = a_0 x^{a_1}$.
- Показательная зависимость: $y = a_0 e^{a_1 x}$.

# Регрессионный анализ

- Параболическая зависимость: $y = a_0 + a_1 x + a_2 x^2.$

- Логистическая зависимость: $y = \frac{a_1}{1+10^{\gamma+\beta x}} + a_0.$

График логистической функции представляет собой симметричную S-образную кривую, которая является постоянно возрастающей функцией.

# Регрессионный анализ

***Суть простого регрессионного анализа*** выражается в следующем утверждении:

*для любого произвольного или фиксированного значения X соответствующая ему величина Y имеет нормальное распределение относительно некоторого теоретического среднего значения.*

График зависимости этих средних от X отражает основное соотношение между X и Y.

# Регрессионный анализ

Коэффициент при независимой переменной X в уравнении регрессии ($a_1$) называется ***коэффициентом регрессии*** Y на X.

Он определяет угол наклона прямой на графике и служит мерой среднего изменения величины Y при изменении X на единицу.

Коэффициент регрессии может быть положительным и отрицательным, а если X и Y независимы, то он равен нулю.

# Общая задача регрессионного анализа. МНК

Общая задача регрессионного анализа состоит в том, чтобы

- по наблюдениям $x_i$ и $y_i$ оценить параметры модели $a_0$ и $a_1$ «*наилучшим образом*»;
- проверить гипотезу о значимости уравнения и коэффициентов регрессии;
- оценить адекватность полученной зависимости и т.д.

Если под «наилучшим образом» понимать минимальную сумму квадратов расстояний до прямой от наблюдаемых точек, вычисленных вдоль оси ординат, то такой метод построения уравнения регрессии называется ***методом наименьших квадратов***.

Найдем теперь оценку неизвестных значений $a_0$ и $a_1$, основанную на имеющейся у нас выборке объема $n$. *Наилучшие* оценки $b_0$ и $b_1$ для $a_0$ и $a_1$ получаются *минимизацией* соответственно по $a_0$ и $a_1$ суммы квадратов отклонений

$$S = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2.$$

Как известно, минимум функции можно найти, приравняв к нулю ее производную.

Далее находим частные производные функции $S$ по $a_0$ и $a_1$ и приравниваем их к нулю.

Решая полученную систему уравнений находим оценки наименьших квадратов:

$$b_0 = \bar{y} - b_1 \bar{x},$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Оценкой уравнения регрессии (или прямой наименьших квадратов) будет

$$\hat{y} = b_0 + b_1 x.$$

Разницей между наблюдаемым и предсказанным значением Y при X=$x_i$ называется отклонением или остатком: $d_i = y_i - \widehat{y}_i$.

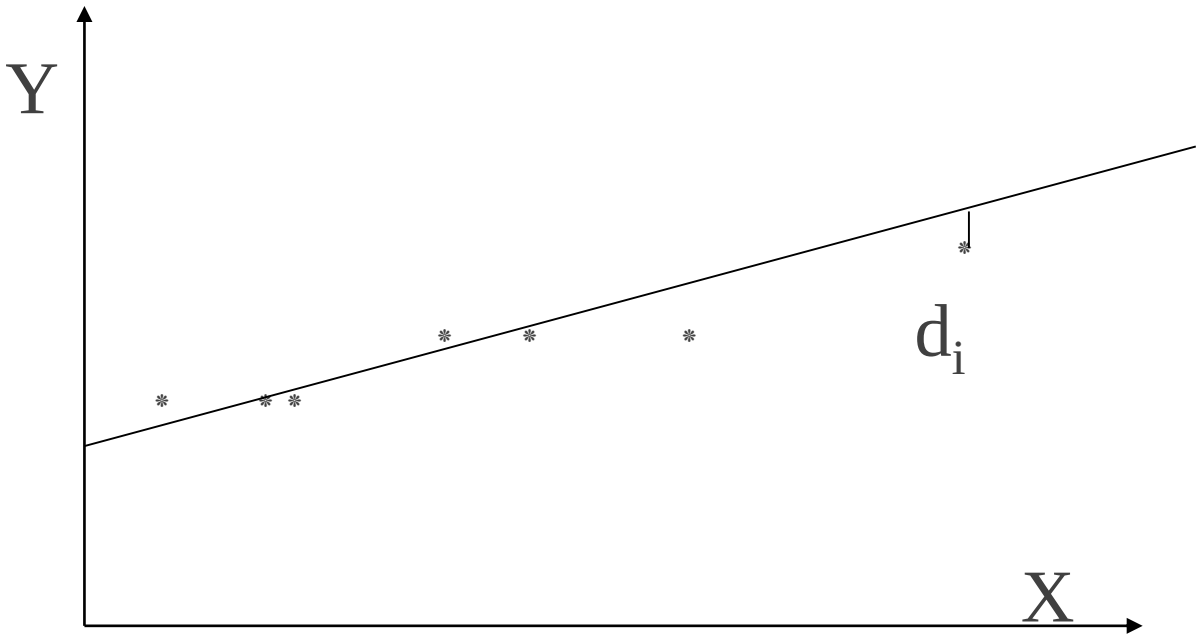



Рис. График прямой наименьших квадратов

Рассмотрим некоторые свойства полученных оценок $b_0$ и $b_1$:

1) $b_0$, $b_1$ независимые случайные величины;

2) оценки подчиняются нормальному закону распределения;

3) математическое ожидание $M(b_0)=a_0$ и $M(b_1)=a_1$;

4) ковариация $cov(b_0,b_1)=0$;

5) дисперсии оценок равны

$$\sigma_{b_0}^2 = \sigma^2 / n, \qquad \sigma_{b_1}^2 = \sigma^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

где $\sigma^2$ – дисперсия ошибок.

Чтобы сделать статистические выводы о $b_0$, $b_1$ и $\hat{y}$, сначала необходимо оценить дисперсию $\sigma^2$. Согласно теории общей линейной модели, обычная несмещенная оценка для $\sigma^2$ определяется через дисперсию оценки

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2}{n-2}.$$

Положительный квадратный корень из этой величины называют ***стандартной ошибкой оценки***.

Для проверки нулевой гипотезы о том, что простая линейная регрессия Y по X отсутствует ($H_0$: $b_1=0$), построим таблицу дисперсионного анализа.

Если полученное значение F-отношения больше табличного с заданным уровнем значимости $\alpha$ и числом степенями свободы (1; n-2), то гипотезу отвергаем ($b_1 \neq 0$), т.е. Y линейно зависит от X.

# Таблица дисперсионного анализа для модели линейной регрессии

| Источник дисперсии | Сумма квадратов | Степени свободы | Средний квадрат | F-отноше-ние |
|---|---|---|---|---|
| Регрессия | $SS_D = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$ | $v_D=1$ | $MS_D=SS_D/v_D$ | $F = MS_D / MS_R$ |
| Отклонение от регрессии | $SS_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ | $v_R=n-2$ | $MS_R=s^2=SS_R/v_R$ | |
| Полная | $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$ | $v_T=n-1$ | | |

**Примечание**. $SS_D$ – обусловленная регрессией сумма квадратов; $SS_R$ – сумма квадратов отклонений от линии регрессии или остаточная сумма квадратов; $SS_T$ – полная сумма квадратов.

# Регрессионный анализ. Коэффициент детерминации

Отношение $SS_D/SS_T$ есть доля вариации Y, объясняемая регрессией Y по X. Это отношение называется ***коэффициентом детерминации*** ($r^2$).

Коэффициент детерминации является мерой качества предсказанных значений зависимой переменной Y моделью линейной регрессии.

Интервал распределения коэффициента детерминации - [0;1].

Если $r^2=1$, то наблюдаемые точки в точности лежат на линии регрессии.

Если $r^2=0$, то Y не зависит от X.

Если, например, $r^2$=0,95 , это означает, что 95% отклонений от среднего значения зависимой переменной объясняет построенная регрессия, а 5% отклонений остаются необъясненными.

Далее в регрессионном анализе для проверки адекватности полученной модели проводят анализ остатков. Для этого строят график $d_i$ в зависимости от $x_i$ или $\hat{y}_i$, i=1,…n.

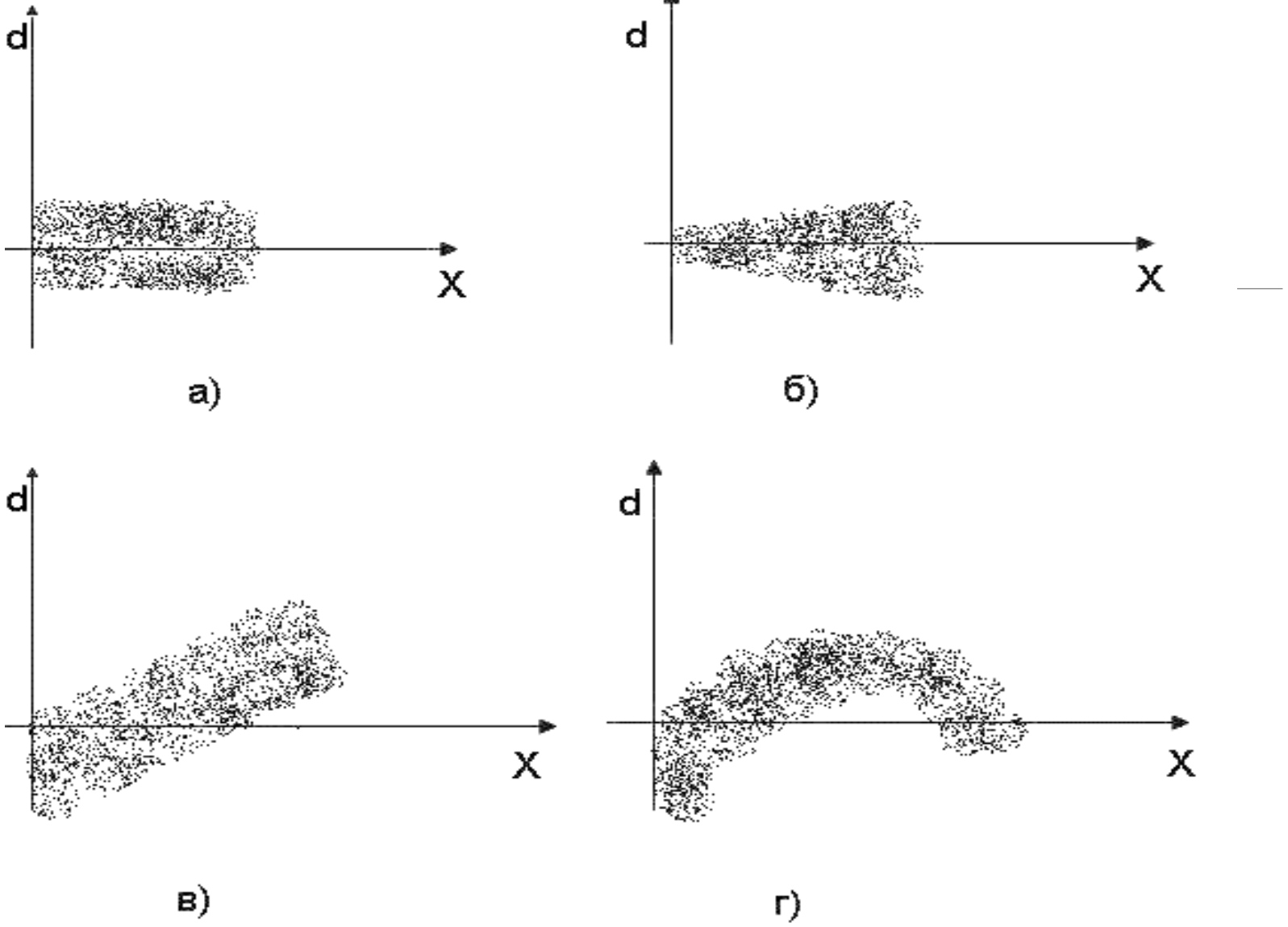



Рис. Примеры графиков остатков

- Если остатки попадают в горизонтальную полосу с центром на оси абсцисс, модель можно рассматривать как ***адекватную*** (*рис.а).*
- Если полоса расширяется, когда x или $\hat{y}$ возрастает (рис. б), это указывает на ***гетероскедастичность*** (т.е. на отсутствие постоянства дисперсии $\sigma^2$).
- В частности, $\sigma$ может быть функцией $\beta_0+\beta_1 x$, что делает необходимым преобразование переменной Y.
- График, показывающий ***линейный тренд*** (рис. в), дает основание для введения в модель дополнительной независимой переменной.
- График вида, представленного на рисунке г), указывает, что в модель должен быть добавлен линейный или квадратный член.

Если предсказанная регрессия удовлетворительно описывает истинную зависимость между Y и X, то остатки должны быть независимыми нормально распределенными случайными величинами с нулевым средним, и в значениях $d_i$ должен отсутствовать тренд.

**Независимость остатков** может быть проверена при помощи ***коэффициента Дарбина-Ватсона***, имеющего вид:

$$D = \sum_{i=2}^{n} (d_i - d_{i-1})^2 \Big/ \sum_{i=1}^{n} d_i^2 .$$

Если $D>D_1$, то с достоверностью α принимается гипотеза о наличии соответственно отрицательной или положительной корреляции остатков. Если

$D_2(\alpha)>D>D_1(\alpha)$, то критерий не позволяет принять решение по гипотезе о наличии или отсутствии корреляции остатков.

Если $D_2(\alpha)<D<4 - D_2(\alpha)$ , то гипотеза о корреляции остатков отклоняется. Критические значения $D_1(\alpha)$ и $D_2(\alpha)$ для различных α берутся из табличных данных.

Итак, определим **основные этапы регрессионного анализа** :

1) нахождение коэффициентов регрессии, построение модели;

2) проверка гипотезы о существовании линейной зависимости между переменными;

3) анализ остатков.

# Множественная регрессия

*Уравнение регрессии — это зависимость случайной величины Y от неслучайных факторов* ***X****, т. е. зависимость «следствия» Y от «причин»* ***X****:*

$$Y=\eta\,(\boldsymbol{X},\boldsymbol{\beta})+\boldsymbol{\varepsilon}, \qquad (2.1)$$

где $\boldsymbol{X}=\{x_1, x_2,\ldots,x_j,\ldots,x_k\}$ - вектор факторов, $j= 1,2,\ldots, k$;

$\boldsymbol{\beta}=\{\beta_1,\beta_2,\ldots,\beta_j,\ldots,\beta_d\}$ — вектор параметров модели; $\eta(\boldsymbol{X},\boldsymbol{\beta})$ -функция регрессии (или функция отклика) случайной величины $\boldsymbol{Y}$ на неслучайные $\boldsymbol{X}$; $\boldsymbol{\varepsilon}=\{\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_i,\ldots,\varepsilon_n\}$ - вектор ошибок наблюдений. При многократном однотипном воздействии $\boldsymbol{X}$ на входе получаем на выходе объекта различные значения $Y$.

# Множественная регрессия

Уравнение регрессии вида (2.1) описывает только статику объекта, т. е. предполагается, что взаимосвязь показателя ***Y*** и факторов ***X***, установленная в определенный момент (интервал) времени, от времени не зависит.

***Регрессионные модели в зависимости от рассматриваемых факторов могут быть использованы в целях***: объяснения сути явления (предсказательная модель), прогнозирования (прогнозная модель), управления.

Если установлена зависимость *Y* только от управляемых факторов ***X'***, то это уравнение теоретически может быть использовано в целях управления объектом.

# Множественная регрессия

При построении уравнения по результатам пассивного эксперимента ошибка в управлении может быть неприемлемой.

Функциональная модель и модель для прогнозирования содержат все группы факторов $\boldsymbol{X'}$, $\boldsymbol{Z}$ :

$$Y=\eta(\boldsymbol{X'}, \boldsymbol{Z}, \boldsymbol{\beta})+\varepsilon.$$

Обычно функциональная модель более сложная, чем предсказательная.

В целях построения $\eta(\boldsymbol{X},\boldsymbol{\beta})$ обычно предполагают, что это - гладкая функция в области допустимых значений: $\boldsymbol{X} \in \boldsymbol{X}_{доп}$.

В этом случае возможно ее разложение в ряд Тейлора в окрестности некоторой точки, например, точки, соответствующие «центру» эксперимента, - среднему значению $\bar{X}$.

# Множественная регрессия

В результате получаем полином степени $p$ вида:

$$Y = \beta_0 + \sum_j \beta_j x_j + \sum_{uj} \beta_{uj} x_u x_j + \sum_j \beta_{jj} x_j^2 + \cdots + \varepsilon, \qquad (2.2)$$

где $\sum_j -$ сумма по $j = 1,2,\ldots,k$, $\sum_{uj} -$ сумма парных взаимодействий $x_u x_j$, $u,j = 1,2,\ldots,k$, $u \neq j$, $k$ - число факторов; $\beta_{uj}$ - коэффициент парного взаимодействия, $\beta_{jj}$ - коэффициент при квадрате переменной и т. д.; в формуле (2.2) степень полинома p = 2.

Обычно разложение ограничивают конечным числом членов ряда. Например:

$$Y = \beta_0 + \sum_j \beta_j x_j \qquad (2.3)$$

# Множественная регрессия

По результатам эксперимента могут быть определены не «истинные» коэффициенты регрессии **β**, соответствующие генеральной совокупности, а лишь их оценки

$\boldsymbol{B} = (b_0, b_1, \ldots, b_j, \ldots, b_d)$, вычисленные по выборке объемом **n**.

В этом случае уравнение регрессии в векторной форме имеет вид:

$$\hat{Y} = \eta(\boldsymbol{X}, \boldsymbol{B}), \quad (2.4)$$

где $\hat{Y}$ – предсказанные (прогнозируемые) значения выходной величины.

При выводе и использовании формул регрессионного анализа удобнее пользоваться векторной формой представления уравнений регрессии:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}; \hat{\boldsymbol{Y}} = \boldsymbol{XB}, \qquad (2.5)$$

# Множественная регрессия

***Y*** – вектор наблюдений; ***X***- матрица значений независимых переменных;

***β, B*** - векторы коэффициентов и их оценок соответственно;

**ε** - вектор ошибок:

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ y_i \\ \cdot \\ y_n \end{bmatrix}, X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ik} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nk} \end{bmatrix}, B=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \beta_i \\ \cdot \\ \beta_d \end{bmatrix}, \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \varepsilon_i \\ \cdot \\ \varepsilon_n \end{bmatrix}.$$

Первый столбец матрицы ***X*** содержит фиктивную переменную $x_{0i}=1, i=1,2,...,n$.

# Множественная регрессия

В общем случае, когда $d > k$ (число коэффициентов регрессии больше числа анализируемых факторов $k$), можно записать уравнение регрессии в следующей векторной форме:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{FB}, \qquad (2.6)$$

где $F[f_{iq}\,(X)]_{nd}$ - матрица известных функций $f_{iq}$, от независимых переменных.

# Множественная регрессия

Например, пусть $k = 2$, а $d = 5$, т. е. необходимо вычислить пять коэффициентов: $b_0, b_1, b_2, b_3, b_4$.

Тогда:

$F = (f_{i1}, f_{i2}, f_{i3}, f_{i4}, f_{i5})$, где $f_{i0} = 1 = x_{i0}, f_{i1} = x_{i1}, f_{i2} = x_{i2}, f_{i3} = x_{i1}x_{i2}, f_{i4} = x_{i1}, f_{i5} = x_{i2},$

т. е. введены переобозначения, и вместо уравнения

$$\hat{Y} = b_0 + b_1x_1 + b_2x_2\, b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2$$

имеем

$$\hat{Y} = b_0 + b_1f_1 + b_2\, f_2 + b_3f_3 + b_4\, f_4 + b_5\, f_5 \quad (2.7)$$

После вычисления коэффициентов регрессии нужно вернуться к первоначальным обозначениям, для того чтобы облегчить интерпретацию результатов.

# Задачи регрессионного анализа:

- вычисление коэффициентов регрессии;
- проверка значимости коэффициентов регрессии;
- проверка адекватности модели;
- выбор «лучшей» регрессии;
- вычисление стандартных ошибок.

Вычисление коэффициентов регрессии осуществляется методом наименьших квадратов (МНК-метод).

Проверка значимости коэффициентов регрессии основана на методах проверки «гипотез о средних».

Проверка адекватности модели основана на методах дисперсионного анализа.

Вычисление стандартных ошибок, по которым можно судить о точности предсказаний, осуществляется по обычным формулам расчета средних квадратичных отклонений.

# Постулаты регрессионного анализа

***Первое условие.*** **Результаты эксперимента должны быть свободны от систематических ошибок, т. е. ожидание *M{Y}* величины У должно быть равно действительному значению $\tilde{Y}$** т.е.:

$M\{Y\} = \tilde{Y}$

$\tilde{Y} = (\tilde{Y}_1, \dots, \tilde{Y}_i, \dots, \tilde{Y}_N)$.

Следовательно, математическое ожидание ошибки $\varepsilon$ будет равно нулю

$M\{\varepsilon\} = M\{Y - \tilde{Y}\} = 0,$

или если действительным значением считать предсказанное по уравнению регрессии значение $\hat{Y}$, то:

$M\{\varepsilon\} = M\{Y - \hat{Y}\} = 0,$

Рассмотрим отдельный опыт в точке $x_i$. Пусть для «истинной модели» (2.1) величина среднего $M\{\ddot{Y}_i\}$ значения $\tilde{y}_i$ равна:

$M\{\tilde{Y}_i\} = \tilde{y}_i$

Определим ошибку $\varepsilon_i$ i-го опыта:

$\varepsilon_i = Y_i - \hat{Y} = \left[\left(Y_i - \hat{Y}_i\right) - \left(\tilde{Y}_i - M\{\tilde{Y}_i\}\right)\right] + [\tilde{Y}_i - M\{\tilde{Y}_i\}] = A_i + B_i,$

где $A_i = (Y_i - \hat{Y}_i) - (\ddot{Y}_i - M\{\hat{Y}_i\})$ - случайная переменная с нулевым средним;

$B_i = [\tilde{Y}_I - M\{\tilde{Y}_I\}]$ - ошибка смещения.

Если построенная модель верна (корректна), то ошибка смещения равна нулю, и первое условие соблюдено.

# Постулаты регрессионного анализа

***Второе условие*** **- дисперсия результатов наблюдения во всех лоточках одинакова**, т. е.:

$D\{Y_i\} = \sigma^2, D\{\varepsilon_i\} = \sigma^2$ для $\forall i$.

***Третье условие*** **- результаты наблюдений в точке $x_i$ не зависят от результатов наблюдений в предыдущей точке $x_{i-1}$**, т.е. $Y_{i-1}$ и $Y_i$ - не коррелированны, так что ковариации равны нулю:

3) $Cov\{Y_{i-1}, Y_i\} = M\{(Y_{i-1} - \hat{Y}_{i-1})(Y_i - \hat{Y}_i)\} = 0;$

$Cov\{\varepsilon_{i-1}, \varepsilon_i\} = M\{\varepsilon_{i-1}, \varepsilon_i\} = 0;$

Поэтому для уравнения регрессии имеем, например:

$M\{Y_i\} = \beta_0 + \sum_j \beta_j x_{ij} + \sum_{uj} \beta_{uj} x_{iu} x_{ij} + \sum_j \beta_{jj} x_{ij}^2 + \cdots + \varepsilon_i,$

***Четвертое условие:*** **$Y_i$, $\varepsilon_i$ - случайные величины, подчиненные нормальному закону распределения** со средними

$M\{\tilde{Y}_i\}$ *и дисперсиями* $D\{\tilde{Y}_i\} = \sigma^2$,*т. е.*

$Y_i \approx N(\tilde{Y}_i, \sigma^2);$

$\varepsilon_i \approx N(0, \sigma^2),$

где $N$ – обозначение нормальных распределений наблюдаемой величины $Y$ и ее ошибки $\varepsilon$.

# Постулаты регрессионного анализа

Представленные условия формулируются в виде следующих *постулатов*.

1. Случайная величина $Y$ и ее ошибка $\varepsilon$ подчинены нормальному закону распределения.
2. Дисперсия выходной величины $Y$ постоянна и не зависит от величины $Y_i$, $i$ =1,2,...n.
3. Результаты наблюдений $Y_i$ в разных точках эксперимента независимы и не коррелированны.

К этим постулатам добавляют еще один, который практически в большой степени обеспечивает выполнение первых трех.

4. Входные переменные $X_j$ - независимы, неслучайны, измеряются без ошибок.