СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА СВЯЗИ

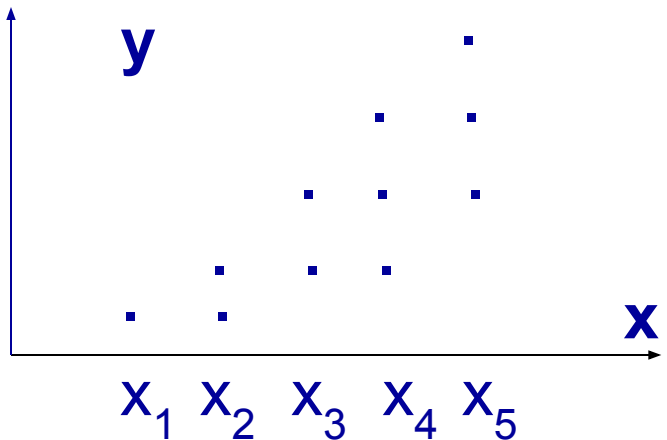
Признаки, которыми характеризуются единицы совокупности, могут быть взаимосвязанными. Взаимосвязанные признаки могут выступать в одной из ролей:

- роли ***признака-результата*** (аналог зависимой переменной (**Y**) в математике);
- роли ***признака-фактора***, (аналог независимой переменной (**X**) в математике). Значение признака-фактора определяют значение признака-результата

Связи в статистике классифицируют **по степени тесноты**, **направлению, форме**, **числу факторов**.

По ***степени тесноты*** связи делят на *статистические* и *функциональные*.

Статистическая (стохастическая) связь – это такая связь между признаками, при которой для каждого значения признака-фактора X признак-результат Y может в определенных пределах принимать любые значения с некоторыми вероятностями; при этом его статистические (обобщающие) характеристики (например, среднее значение) изменяются по определенному закону.



Модель стохастической связи может быть представлена в общем виде уравнением: $y=f(x)+u$, где $f(x)$ - часть результативного признака, сформировавшаяся под воздействием фактора x;

u - случайная составляющая, часть результативного признака, являющаяся результатом действия прочих (неучтенных) факторов, а также ошибок измерения признаков.

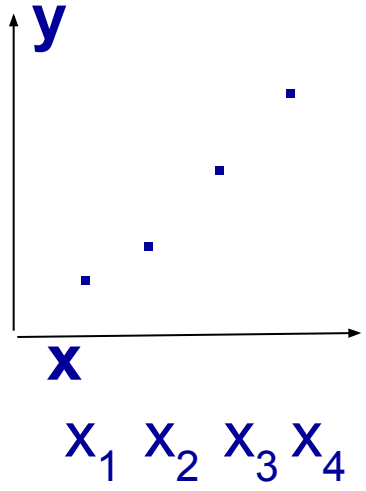
Корреляционная связь частный случай стохастической связи. При корреляционной связи с изменением значения признака X среднее значение признака Y закономерно (функционально) изменяется.

Функциональная связь – такая связь, при которой для каждого значения признака-фактора

признак-результат принимает одно (иногда несколько) строго определенных значений.

Она имеет место, когда все факторы, действующие на результативный признак, известны и учтены в модели и ошибки измерения отсутствуют.

Модель функциональной связи может быть представлена как: y=f(x).



По направлению связи делят на *прямые* и *обратные* связи.

При **прямой связи** направление изменения результата совпадает с направлением изменения признака-фактора.

При **обратной связи** направление изменения результата противоположно направлению изменения признака-фактора.

Например, чем выше квалификация рабочего, тем выше уровень производительности его труда (прямая связь). Чем выше производительность труда, тем ниже себестоимость единицы продукции (обратная связь).

По ***форме связи*** (виду функции f) связи делят на *линейные* (прямолинейные) и *нелинейные* (криволинейные) связи.

Линейная связь отображается прямой линией; **криволинейная** отображается кривой (параболой, гиперболой и т.п.).

При линейной связи с увеличением на единицу значения признака-фактора происходит равномерное возрастание (убывание) значения признака-результата.

При криволинейной связи с увеличением на единицу значения признака-фактора возрастание (убывание) признака-результата происходит неравномерно (гиперболическая форма связи) или же меняется направление связи (параболическая форма связи).

По ***количеству факторов***, действующих на результат, связи подразделяют на **однофакторные** (парные) и **многофакторные** связи.

Порядок изучения парной статистической связи:

1. Качественный (содержательный) анализ связи. На этом этапе производят предварительный анализ направления и формы связи.
2. Сбор данных (статистическое наблюдение).
3. Эмпирический анализ связи.
4. Количественная оценка тесноты связи (корреляционный анализ).
5. Установление аналитической зависимости между признаками (регрессионный анализ):

5.1. выбор формы связи (вида аналитической зависимости);

5.2. оценка параметров уравнения регрессии;

5.3. оценка качества уравнения регрессии.

3 этап – эмпирический анализ связи состоит в построении группировок *(аналитической* или *комбинационной)* и графиков.

Для анализа связи между признаками служат графики: корреляционное поле и эмпирическая линия регрессии.

Корреляционное поле – точечный график, построенный в системе координат X, Y. Число точек равно числу единиц в совокупности. Каждая точка соответствует некоторой единице совокупности и имеет координаты по оси абсцисс – значение признака-фактора X, а по оси ординат – значение признака-результата Y.

Эмпирическая линия регрессия - ломанная линия, построенная по данным аналитической группировки. Число точек ломанной равно числу групп в аналитической группировке. Каждая точка имеет абсциссу равную среднему значению признака-фактора в группе и ординату равную среднему значению признака-результата в этой же группе.

Форма графиков позволяет делать выводы о направлении, форме и тесноты связи.

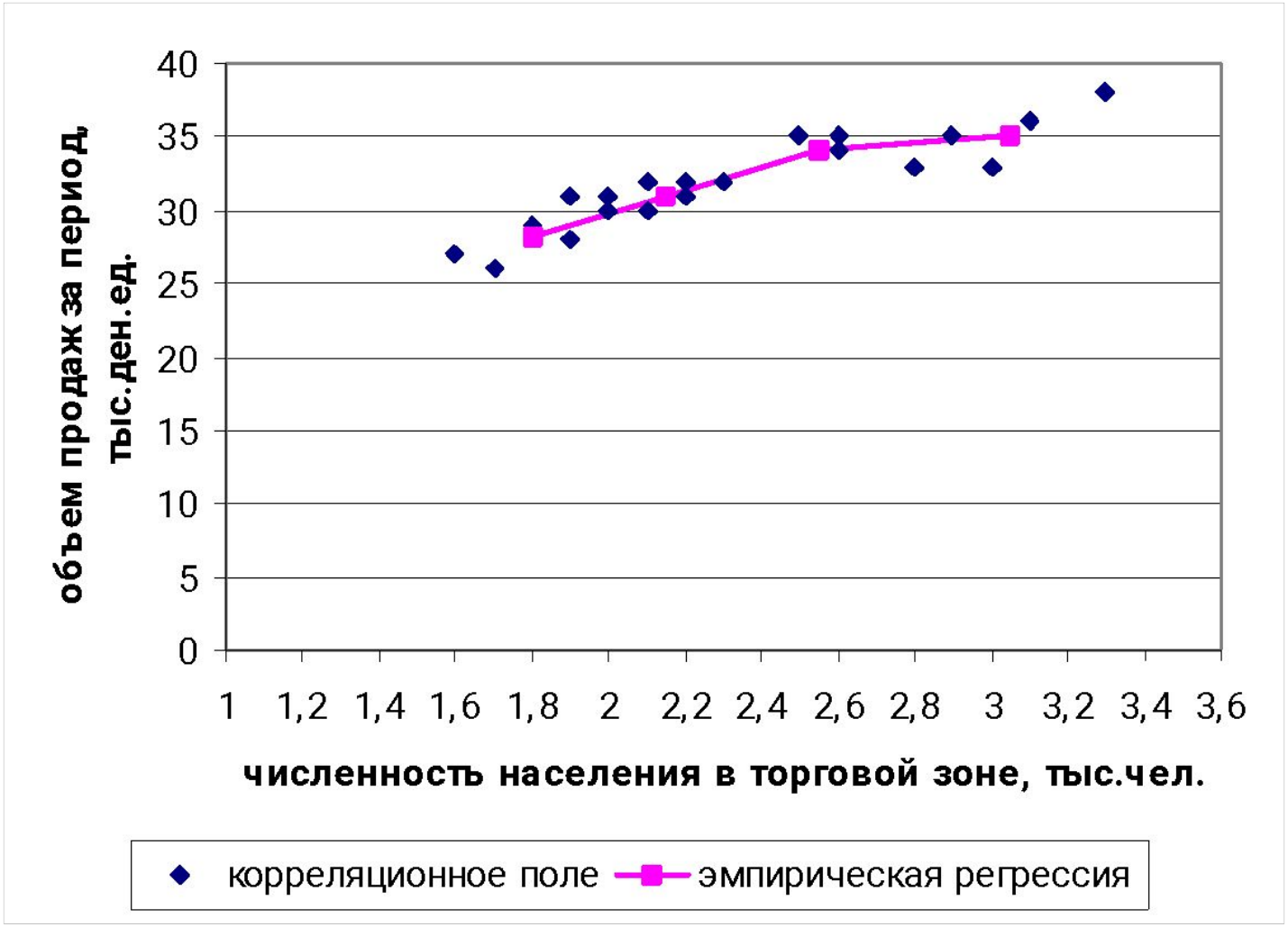
Пример. Имеется совокупность из 20 магазинов розничной торговли. Проведем анализ связи между признаками X- численность населения в торговой зоне, тыс.чел. и Y- объем продаж магазина, тыс.ден.ед. за период.

№	1	2	3	4	5	6	7	8	9	10
X	1,8	2	2,3	1,9	1,7	3,1	2,8	2,1	2,2	1,9
Y	29	30	32	31	26	36	33	30	31	28
№	11	12	13	14	15	16	17	18	19	20
X	1,6	2,2	2,1	3	2,5	2,9	2	3,3	2,6	2,6
Y	27	32	32	33	35	35	31	38	35	34

Для построения **эмпирической линии регрессии** нам потребуются данные аналитической группировки:

Аналитическая группировка магазинов розничной торговли

X_j – численность населения в торговой зоне, тыс.чел.	Число магазинов - N_j	Средний объем продаж - $\overline{y_j}$
[1,6 ; 2,0)	5	28,20
[2,0 ; 2,3)	6	31,00
[2,3 ; 2,8)	4	34,00
[2,8 ; 3,3]	5	35,00
Итого	20	



Вывод: зависимость между признаками прямая (возрастающая) и скорее линейная чем нелинейная

4 этап – количественная оценка тесноты связи (корреляционный анализ) состоит в расчете показателей тесноты связи:

- эмпирического коэффициента детерминации,
- эмпирического корреляционного отношения,
- коэффициента Фехнера,
- коэффициента линейной парной корреляции.

Эмпирический коэффициент детерминации (эмпирическое дисперсионное отношение) - ρ^2.

Данный показатель рассчитывается по данным аналитической группировки, как отношение межгрупповой дисперсии к общей (на основе теоремы о сложении дисперсий):

$$\rho^2 = \frac{\delta_y^2}{\sigma_y^2} = \frac{\sigma_y^2 - \varepsilon_y^2}{\sigma_y^2} = 1 - \frac{\varepsilon_y^2}{\sigma_y^2}$$

Эмпирический коэффициент детерминации показывает процент (долю) вариации признака-результата, обусловленную признаком-фактором, положенным в основу группировки.

Межгрупповая дисперсия рассчитывается по формуле :

$$\delta_y^2 = \frac{\sum_{j=1}^{k} (\overline{y_j} - \overline{y})^2 \cdot N_j}{\sum_{j=1}^{k} N_j}$$

Остаточная дисперсия рассчитывается по формуле:

$$\varepsilon_y^2 = \frac{\sum_{j=1}^{k} \sigma_j^2 \cdot N_j}{\sum_{j=1}^{k} N_j}$$

Где σ^2_j – дисперсия признака Y в j-ой группе

Пример: Рассчитаем эмпирический коэффициент детерминации $\rho^2=\delta^2_y/\sigma^2_y$ для измерения тесноты связи между численностью населения в торговой зоне и объемом продаж магазина розничной торговли по данным аналитической.

Для расчета межгрупповой дисперсии δ^2_y необходимо знать общее среднее арифметическое значение признака Y. Оно в нашем примере равно:

$$\overline{y} = \frac{29+30+32+\ldots+34}{20} = 31{,}9$$

Тогда межгрупповая дисперсия будет равна:

$$\delta_y^2 = \frac{(28{,}2-31{,}9)^2\cdot 5+(31-31{,}9)^2\cdot 6+(34-31{,}89)^2\cdot 4+(35-31{,}9)^2\cdot 5}{20} =$$

$$= 6{,}95$$

Общая дисперсия признака Y для нашего примера будет равна:

$$\sigma_y^2 = \frac{(29-31{,}9)^2 + (30-31{,}9)^2 + (32-31{,}9)^2 + \ldots + (34-31{,}9)^2}{20} =$$

$$= 9{,}09$$

Тогда эмпирический коэффициент детерминации ρ^2=6,95 / 9,09= 0,765

Вывод: 76,5% вариации объема продаж магазина розничной торговли обусловлено численностью населения в торговой зоне.

Эмпирическое корреляционное отношение - ρ.
Данный показатель представляет собой корень из эмпирического коэффициента детерминации. Он измеряет тесноту связи между фактором (группировочным признаком) и результатом. Область допустимых значений эмпирического корреляционного отношения от 0 до +1.
При достаточно тесной связи между признаками эмпирический коэффициент детерминации стремится к 1.
При слабой связи - к нулю.

$$\rho = \sqrt{\frac{\delta_y^2}{\sigma_y^2}} = \sqrt{\rho^2}$$

В нашем примере:

$$\rho^2 = \sqrt{\frac{\delta_y^2}{\sigma_y^2}} = \sqrt{\rho^2} = \sqrt{0,765} = 0,87 \to 1$$

Следовательно, связь между численностью населения в торговой зоне и объемом продаж достаточно тесная.

Коэффициент Фехнера - Кф служит для измерения тесноты линейной связи. Изменяется в пределах от -1 до +1.

Если | Кф |→1 , то связь близка к линейной функциональной. Если признаки х и у взаимно независимы, то |Кф|→0 .

Но равенство нулю коэффициента корреляции означает отсутствие только линейной связи. Если Кф<0,то связь между признаками обратная. Если Кф>0, то связь - прямая.

$$Кф = \frac{C - H}{C + H}$$

где С – число совпадений, Н – несовпадений знаков отклонений **X** от своего среднего значения и **Y** от своего среднего значения.

Пример: рассчитаем коэффициент Фехнера по данным о 20 магазинах розничной торговли для оценки тесноты связи между численностью населения в торговой зоне и Y- объемом продаж за период. Среднее значение по X = 2,33 тыс.чел.; среднее значение по Y =31,9 тыс.ден.ед. Желтым цветом выделены магазины (единицы), у которых знаки отклонений совпадают.

№	X	Y	X-Xcp	Y-Yср	№	X	Y	X-Xcp	Y-Yср
11	1,6	27	-	-	12	2,2	32	-	+
5	1,7	26	-	-	3	2,3	32	-	+
1	1,8	29	-	-	15	2,5	35	+	+
4	1,9	31	-	-	19	2,6	35	+	+
10	1,9	28	-	-	20	2,6	34	+	+
2	2	30	-	-	7	2,8	33	+	+
17	2	31	-	-	16	2,9	35	+	+
8	2,1	30	-	-	14	3	33	+	+
13	2,1	32	-	+	6	3,1	36	+	+
9	2,2	31	-	-	18	3,3	38	+	+

Таким образом число совпадений С=17, число несовпадений равно Н=3.
Следовательно, Кф= (17 - 3) / (17 + 3) = 0,7.

Вывод: так как значение Кф ближе к 1, то связь можно охарактеризовать как достаточно тесную, а положительное значение Кф свидетельствует о прямой зависимости.

Коэффициент линейной парной корреляции используется для оценки степени тесноты линейной связи:

$$r_{x,y} = \frac{\overline{X \cdot Y} - \overline{X} \cdot \overline{Y}}{\sigma_x \cdot \sigma_y}$$

$$\overline{X \cdot Y} = \frac{\sum_{i=1}^{N} X_i \cdot Y_i}{N}$$

- среднее из произведения

σ_x, σ_y - среднее квадратические отклонения признаков X и Y.

Область допустимых значений линейного коэффициента корреляции от -1 до +1.

Если $| r_{x,y} | \to 1$, то связь близка к линейной функциональной.

Если признаки x и y взаимно независимы, то $| r_{x,y} | \to 0$

! Равенство нулю коэффициента корреляции означает отсутствие только линейной связи. Признаки могут быть связаны тесной нелинейной зависимостью и при этом иметь нулевой коэффициент корреляции (например, в случае параболической формы связи).

Если $r_{x,y} < 0$, то связь между признаками обратная.

Если $r_{x,y} > 0$, то связь - прямая.

Пример: рассчитаем коэффициент линейной парной корреляции между численностью населения в торговой зоне и Y- объемом продаж по данным о 20 магазинах розничной торговли.

$$\overline{X \cdot Y} = \frac{1{,}6 \cdot 27 + 1{,}7 \cdot 26 + \ldots + 3{,}3 \cdot 38}{20} = 75{,}645$$

$$\overline{X} = 2{,}33 \quad \overline{Y} = 31{,}9 \quad \sigma_y = \sqrt{9{,}09} = 3{,}015 \quad \sigma_x = 0{,}4818$$

$$r_{x,y} = \frac{\overline{X \cdot Y} - \overline{X} \cdot \overline{Y}}{\sigma_x \cdot \sigma_y} = \frac{75{,}645 - 2{,}33 \cdot 31{,}9}{0{,}4818 \cdot 3{,}015} = 0{,}907$$

Вывод: зависимость между признаками объем продаж за период и численность населения в торговой зоне можно характеризовать как очень тесную ($r \to 1$) и возрастающую (т.к. $r > 0$).

Если сравнить значения эмпирического корреляционного отношения (**ρ**) с линейным парным коэффициентом корреляции (**r**), то можно сделать *вывод о форме связи*.

Если разность **ρ - |r| > 0,1**, то связь считают ***нелинейной***.

Если данное неравенство не выполняется, то связь считают ***линейной***.

Пример: так как ρ - |r| =0,87 – 0,907 = -0,03 < 0,1 , то связь между признаками объем продаж за период и численность населения в торговой зоне скорее линейная, чем нелинейная.

5 этап - установление аналитической зависимости между признаками* (*регрессионный анализ)

Регрессия – зависимость среднего значения какой-либо случайной величины от одной или нескольких независимых величин.

Установление аналитической зависимости сводится к построению уравнения регрессии.

Уравнение регрессии – уравнение связи в среднем, а именно, уравнение, описывающее корреляционную зависимость признака-результата **y** (его среднего значения) от значения признака-фактора **x** (или факторов).

Линейное парное (однофакторное) уравнение регрессии имеет вид:

$$\mathbf{M(y_i \mid x=x_i)= f(x_i) = \mathit{a} + \mathit{b}\cdot x_i} ,$$

где $M(y_i \mid x=x_i)$ – условное математическое ожидание зависимой переменной – y при значении независимой переменной x равном x_i;

i – номер единицы совокупности (наблюдения), i=1;n,

n - всего наблюдений.

a,b - параметры (коэффициенты) уравнения регрессии.

При построении уравнения регрессии f(x) мы должны:

1) определить вид уравнения (линейное или нелинейное и какое именно нелинейное: парабола, показательное уравнение или другое);

2) оценить параметры регрессии (a, b) по имеющимся данным наблюдений y_i, x_i.

5.1. Выбор формы связи (вида аналитической зависимости).

Наиболее часто для описания статистической связи признаков используется **линейное уравнение регрессии**.
Внимание к линейной форме связи объясняется четкой экономической интерпретацией параметров линейного уравнения регрессии,
ограниченной вариацией переменных,
и тем, что в большинстве случаев нелинейные формы связи для выполнения расчетов преобразуют (путем логарифмирования или замены переменных) в линейную форму.

Методы выявления формы связи:
- графический (вид корреляционного поля и эмпирической линии регрессии);
- теоретический анализ и опыт предыдущих аналогичных исследований;
- сравнение эмпирического корреляционного отношения с коэффициентом корреляции;
- перебор всевозможных видов функций и выбор наилучшей по показателю качества.

5.2. ***Оценки параметров линейной регрессии*** (***a*** и ***b***) могут быть найдены разными методами: методом наименьших квадратов; методом максимального правдоподобия; методом моментов.

Наиболее распространенным является ***метод наименьших квадратов*** **(МНК)**, который при определенных условиях дает наилучшие оценки.

Суть МНК:

Пусть имеются **n** наблюдений признаков **x** и **y**.

Причем известен вид уравнения регрессии - f(x) (например, прямолинейная зависимость:

$f(x_i)=\boldsymbol{a} + \boldsymbol{b}\cdot x_i$.

Задача состоит в оценке параметров (***a*** и ***b***), которые подбираются таким образом, чтобы минимизировать сумму квадратов отклонений фактических значений признака-результата $\mathbf{y_i}$ от расчетных (теоретических) значений $\mathbf{f(x_i)}$ для всех наблюдений i=1;n :

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

Проиллюстрируем суть данного метода графически. Попытаемся подобрать прямую линию, которая ближе всего расположена к точкам корреляционного поля. Согласно методу наименьших квадратов прямая подбирается так, чтобы сумма квадратов расстояний по вертикали между точками корреляционного поля и этой линией была бы минимальной.

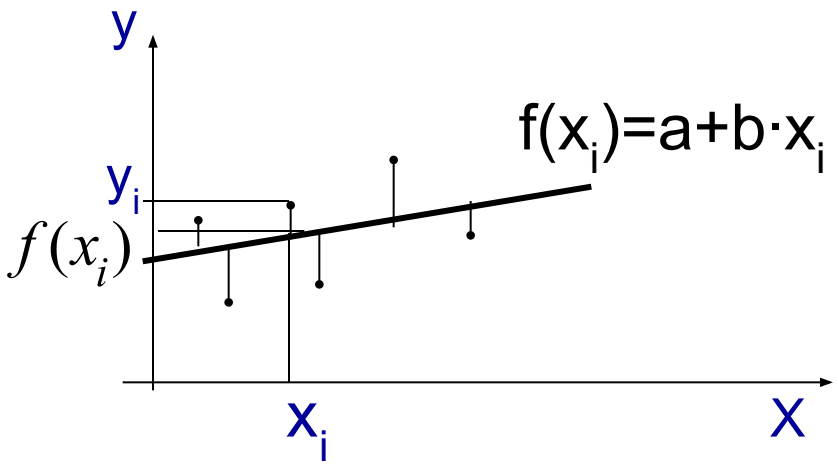


Рис. Линия регрессии с минимальной суммой квадратов отклонений

Значения y_i и x_i i=1;n нам известны, это данные наблюдений. В функции S они представляют собой константы. Переменными в данной функции являются искомые оценки параметров – a и b .

Чтобы найти минимум функции 2-ух переменных необходимо вычислить частные производные данной функции по каждому из параметров и приравнять их нулю,

т.е. $\partial S/\partial a = 0$ и $\partial S/\partial b = 0$.

В результате получим систему из 2-ух нормальных линейных уравнений:

$$\begin{cases} \sum_{i=1}^{n} y_i = a \cdot n + b \sum_{i=1}^{n} x_i & \dfrac{\partial S}{\partial a} = 0 \\ \sum_{i=1}^{n} y_i \cdot x_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^{\,2} & \dfrac{\partial S}{\partial b} = 0 \end{cases}$$

Решая данную систему, найдем искомые оценки параметров.

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

Оценка параметра b может быть рассчитана также через коэффициент корреляции:

$$b = r_{x,y}\cdot\frac{\sigma_y}{\sigma_x}$$

Знак коэффициента регрессии b указывает направление связи (если $b>0$, связь прямая, если $b<0$, то связь обратная).

Величина b показывает на сколько единиц изменится в среднем признак-результат -y при изменении признака-фактора - x на 1 единицу своего измерения.

Формально значение параметра a – среднее значение признака-результата y при нулевом значении x. Если признак-фактор не имеет или не может иметь нулевого значения, то интерпретация параметра a не имеет смысла.

Пример: построим линейное уравнение регрессии объема продаж магазина (y) от значений фактора x– численности населения в торговой зоне:

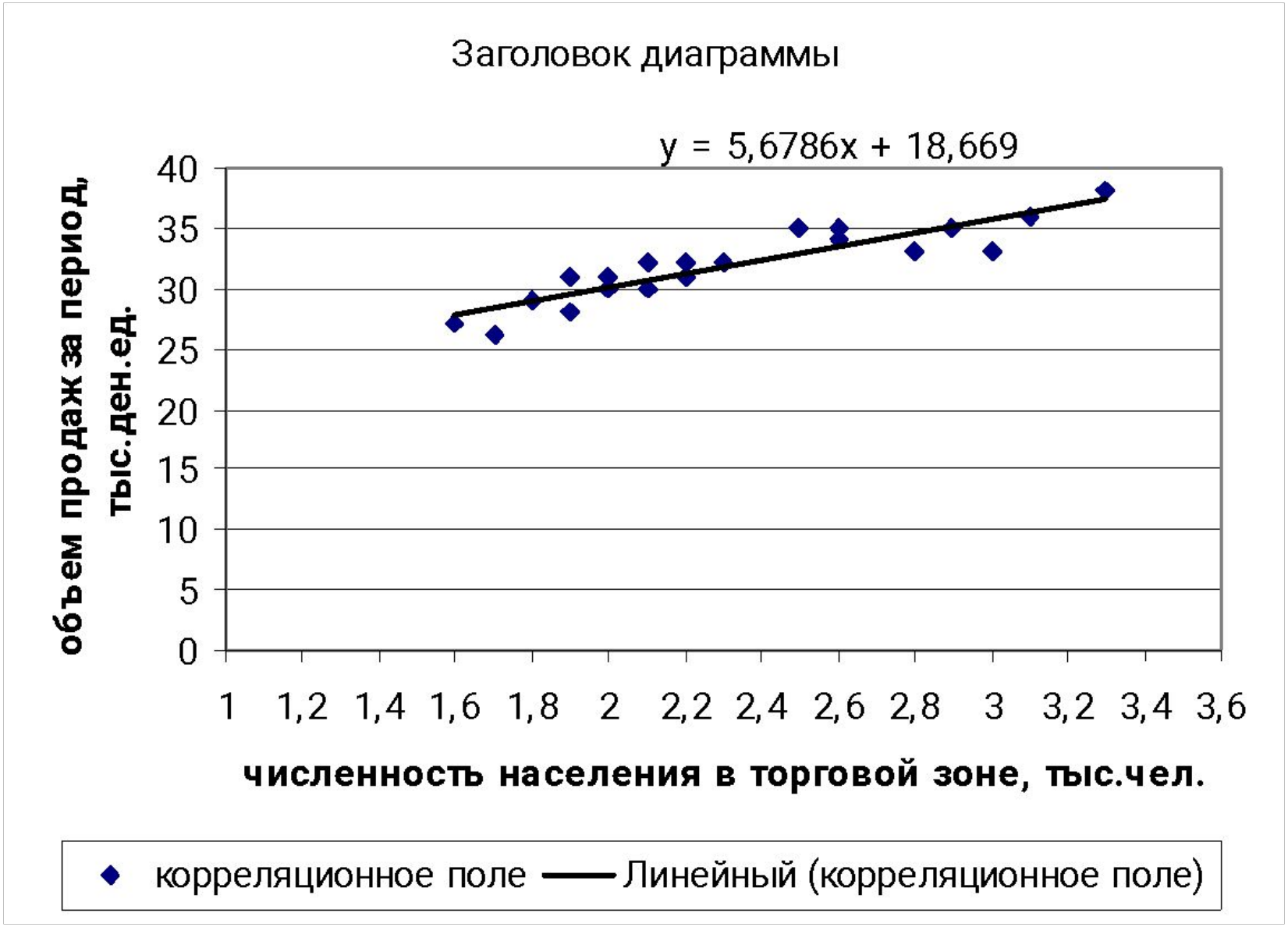
$f(x_i)=a+b\cdot x_i$, $f(x_i)$-расчетное значение признака y.

$$b=\frac{\overline{x\cdot y}-\overline{x}\cdot\overline{y}}{\sigma_x^2}=r_{x,y}\cdot\frac{\sigma_y}{\sigma_x}=0{,}907\cdot\frac{3{,}015}{0{,}4818}=5{,}68$$

$$a=\overline{y}-b\cdot\overline{x}=31{,}9-2{,}33\cdot5{,}68=18{,}67$$

Величина ***b*** в нашем примере показывает, что при увеличении численности населения в торговой зоне на 1 тыс.чел. объем продаж магазина за период в среднем возрастает (т.к. b>0) на 5,68 тыс.ден.ед. Значение параметра *a* не интерпретируется, т.к. нет среди исходных данных значений x равных нулю.

Нанесем график уравнения на корреляционное поле.



5.3. - Оценка качества уравнения регрессии.

Под качеством (адекватностью) уравнения регрессии понимается степень близости (соответствия) рассчитанных по данному уравнению значений признака-результата f(x) фактическим (наблюдаемым) значениям y.

Для оценки качества (адекватности) полученного уравнения регрессии используется ряд показателей:

- теоретический коэффициент детерминации;
- среднеквадратическую ошибка уравнения регрессии;
- средняя ошибка аппроксимации.

Наиболее широкое применение из них получил **теоретический коэффициент детерминации – R^2**. Данный показатель рассчитывается, как отношение объясненной уравнением дисперсии признака-результата - **δ^{*2}**, к общей дисперсии признака-результата **σ^2_y** :

$$R^2_{yx} = \frac{\delta^{*2}}{\sigma^2_y} = \frac{\delta^{*2}}{\delta^{*2} + \varepsilon^{*2}}$$

$$\delta^{*2} = \frac{\sum_{i=1}^{n} (f(x_i) - \bar{y})^2}{n}$$

Объясненная уравнением регрессии дисперсия y

$$\varepsilon^{*2} = \frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{n}$$

Необъясненная уравнением регрессии дисперсия y

В регрессионном анализе также действует теорема о сложении дисперсий, согласно которой общая дисперсия признака-результата равна сумме объясненной уравнением регрессии дисперсии - δ^{*2} и остаточной (необъясненной) дисперсии - ε^{*2} :

$\sigma^2_y = \delta^{*2} + \varepsilon^{*2}$. Поэтому коэффициент детерминации может быть также рассчитан через остаточную и общую дисперсии:

$$R^2 = 1 - \frac{\varepsilon^{*2}}{\sigma_y^2} = 1 - \frac{\sum_{i=1}^{n} (f(x_i) - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Данный показатель (R^2) характеризует долю вариации (дисперсии) признака-результата y, объясняемую уравнением регрессии (a, следовательно, и фактором x), в общей вариации (дисперсии) y.

Коэффициент детерминации R^2 принимает значения от 0 до 1.

Соответственно величина ($1 - R^2$) характеризует долю дисперсии y, вызванную влиянием прочих неучтенных в уравнении факторов и ошибками измерений.

!! При парной линейной регрессии R^2 можно рассчитать по упрощенной формуле: $R^2 = r_{yx}^2$.

2. **Средняя квадратическая ошибка уравнения** регрессии представляет собой среднее квадратическое отклонение наблюдаемых значений признака - результата от теоретических значений, рассчитанных по уравнению, т.е.:

$$s_u = \sqrt{\frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{n-2}} = \sqrt{\frac{(1-R^2)\cdot n\cdot \sigma_y^2}{n-2}}$$

Показатели качества (адекватности) используют также для решения задачи *выбора вида функциональной зависимости*. Выбор может быть осуществлен путем сравнения величин показателя качества (R^2 или s_u), рассчитанных для разных функциональных зависимостей. Чем больше величина коэффициента детерминации R^2 (или чем меньше величина среднеквадратической ошибки s_u), тем уравнение лучше.

Если показатели адекватности оказываются примерно одинаковыми для нескольких функций, то предпочтение отдается более простым видам функций, т.к. они лучше интерпретируются и требуют меньшего объема наблюдений для оценки параметров.

Пример: рассчитаем показатель качества - коэффициент детерминации для уравнения:

$f(xi)=18,67 + 5,68 \cdot xi$

$R^2=r^2_{yx}=0,907^2=0,82.$

То есть 82 % вариации объема продаж за период обусловлено влиянием фактора X – численностью населения в торговой зоне. Соответственно, 18 % (100% - 82%) вариации объема продаж обусловлено влиянием прочих неучтенных факторов.

Если значение коэффициента детерминации существенно отличается от нуля, то уравнение регрессии можно признать качественным.

Прогнозирование по уравнению регрессии означает построение доверительного интервала для ожидаемого (прогнозируемого) значения признака-результата Y при заданном значении признака-фактора X ($X_{прогноз}$).

Заранее задают уровень доверительной вероятности Рдов.

Доверительный интервал прогноза определяется так:

$(Y_{прогноз} - \Delta_{прогноз};\ Y_{прогноз} + \Delta_{прогноз})$,

где $Y_{прогноз}$ – значение Y, полученное по уравнению регрессии: $Y_{прогноз} = f(X_{прогноз})$;

$\Delta_{прогноз}$ – предельная ошибка прогноза.

$\Delta_{прогноз} = \mu_{прогноза} \cdot t$,

где t – коэффициент доверия, определяемый по таблицам распределения Стьюдента, в зависимости от α=1-Рдов и числа степеней свободы=n-2.

$\mu_{прогноза}$ – средняя ошибка прогноза определяется в случае линейной парной регрессии по формуле:

$$\mu_{прогноз} = \sqrt{s_u^2 \left(1 + \frac{1}{n} + \frac{(X_{прогн} - \bar{x})^2}{n \cdot \sigma_x^2} \right)}$$

где s^2_u – средняя ошибка регрессии;

$X_{прогн}$ – значение признака фактора X, для которого выполняется прогноз.

Средняя ошибка регрессии может быть определена по формуле:

$$s_u^2 = \frac{n \cdot (1 - R^2) \cdot \sigma_y^2}{n - 2}$$

Пример: требуется построитьдоверительный интервал для ожидаемого (прогнозируемого) значения Y, если X примет значение равное 105% от своего среднего уровня. (Уровень доверительной вероятности Рдов взять равным 0,95).

Решение:

$X_{прогнозное}$=1,05·2,33=2,4465.

$Y_{прогнозное}$=18,67 + 5,68·2,4465=32,56.

Для расчета предельной ошибки определим коэффициент доверия и среднюю ошибку прогноза.

t – коэффициент доверия, определяется по таблицам распределения Стьюдента. В нашем примере t (α=1-Рдов=0,05; число степеней свободы= n-2=20-2=18) = 2,1.

Для расчета средней ошибки прогноза определим среднюю ошибку регрессии по формуле:

$$s_u^2 = \frac{n \cdot (1 - R^2) \cdot \sigma_y^2}{n - 2} = \frac{20 \cdot (1 - 0{,}82) \cdot 9{,}09}{20 - 2} = 1{,}784$$

Тогда средняя ошибка прогноза будет равна:

$$\mu_{прогноза} = \sqrt{1{,}784\left(1 + \frac{1}{20} + \frac{(2{,}4465 - 2{,}33)^2}{20 \cdot 0{,}2321}\right)} = 1{,}37$$

Тогда $\Delta_{прогноз} = \mu_{прогноза} \cdot t = 1{,}37 \cdot 2{,}1 = 2{,}88$.

Интервал прогноза будет:

(32,56 – 2,88; 32,56 + 2,88) или (29,68; 35,44).

Вывод: с доверительной вероятностью 95% можно утверждать, что при численности населения в торговой зоне, составляющей 105% от среднего уровня, объем продаж магазина не выйдет за пределы от 29,68 до 35,44 тыс.ден.ед.