Лекция. Корреляционно-регрессионный анализ

1. Линейный коэффициент корреляции
2. Ранговая корреляция
3. Уравнение регрессии

Понятие связи

- Исследование объективно существующих связей между явлениями (признаками) - важнейшая задача статистики.
- Признаки по их значению для изучения взаимосвязи можно разделить на два класса.
- Признаки, обуславливающие изменения других, связанных с ним признаков, называются **факторными**.
- Признаки, изменяющиеся под действием факторных признаков, называются **результативными**.

- Для измерения тесноты связи между двумя количественными признаками *x* и *y* часто используется ***линейный коэффициент корреляции r***.
- Он применим лишь в случае линейной зависимости между нормально распределёнными признаками. Если форма связи между признаками ***x*** и ***y*** ещё не определена, то его рассчитывают для ответа на вопрос, можно ли считать зависимость линейной.

Линейный коэффициент корреляции – это средняя величина из произведений нормированных отклонений для ***x*** и ***y***:

$$r = \frac{\sum \frac{x-\bar{x}}{\sigma_x} \cdot \frac{y-\bar{y}}{\sigma_y}}{n}$$

- Вынося σ_x и σ_y за знак суммы (как постоянные величины), получим другой вид формулы линейного коэффициента корреляции:

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$$

- Путём несложных математических преобразований можно получить другие формулы линейного коэффициента корреляции. Например:

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

- или

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

- Линейный коэффициент корреляции может принимать значения от **-1 до 1**. В случае ***прямой*** зависимости *r* ***положителен***, а в случае ***обратной*** он ***отрицателен***.
- Коэффициент корреляции, равный 1, означает функциональную зависимость между *x* и *y*.
- При $r = 0$ связь между признаками *x* и *y* отсутствует.
- При $0 < |r| < 0,3$ связь слабая.
- При $0,3 \le |r| \le 0,7$ связь средней силы.
- При $0,7 < |r| < 1$ связь сильная.

Пример. Имеются сравнительные показатели связи между средней взвешенной ценой и объёмом продаж облигаций на ММВБ 01.01.2014 г. Вычислить коэффициент корреляции.

Номер серии	Средняя взвешенная цена, тыс. руб. x	Объём продаж, млрд. руб. y
A	84,42	69,5
B	82,46	72,9
C	80,13	71,4
D	63,42	135,1
E	76,17	76,3
F	75,13	74,7
G	74,84	97,4
H	73,03	75,1
I	73,41	75,5
J	71,34	98,2

Номер серии	Средняя взвешенная цена, тыс. руб. x	Объём продаж, млрд.руб. y	x^2	y^2	xy
A	84,42	69,5	7126,74	5867,19	4830,25
B	82,46	72,9	6799,65	6011,33	5314,41
C	80,13	71,4	6420,82	5721,28	5097,96
D	63,42	135,1	4022,10	8568,04	18252,01
E	76,17	76,3	5801,87	5811,77	5821,69
F	75,13	74,7	5644,52	5612,21	5580,09
G	74,84	97,4	5601,03	7289,42	9486,76
H	73,03	75,1	5333,38	5484,55	5640,01
I	73,41	75,5	5389,03	5542,46	5700,25
J	71,34	98,2	5089,40	7005,59	9643,24
$\sum$	754,35	846,1	57228,52	62913,84	75366,67
Средняя величина	75,43	84,61	5722,85	6291,38	7536,67

- Рассчитаем средние квадратические отклонения для признаков *x* и *y*:

$$\sigma_x = \sqrt{\overline{x^2} - \overline{x}^2} = \sqrt{5722,85 - 75,43^2} = 5,69;$$

$$\sigma_y = \sqrt{\overline{y^2} - \overline{y}^2} = \sqrt{6291,38 - 84,61^2} = 19,44$$

- Линейный коэффициент корреляции:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \sigma_y} = \frac{7536,67 - 75,43 \cdot 84,6}{5,69 \cdot 19,44} = -0,82$$

- Полученный результат позволяет сделать вывод о том, что между средней взвешенной ценой акций и объёмом продаж существует ***сильная обратная*** связь.

Проверка коэффициента корреляции на значимость

- Оценка значимости линейного коэффициента корреляции основана на сопоставлении значения ***r*** с его средней квадратической ошибкой: $\frac{|r|}{\sigma_r}$
- Если число наблюдений достаточно велико (n>50), и выборка осуществлена из нормальной совокупности, средняя ошибка коэффициента корреляции рассчитывается по следующей формуле:

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

- Если $\frac{|r|}{\sigma_r} > 3$, то коэффициент считается значимым, а связь реальной.

- При небольшом числе наблюдений (n<30) средняя ошибка коэффициента корреляции определяется как:

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

- Значимость *r* проверяется на основе t-критерия Стьюдента. Для этого вычисляется расчетное значение критерия:

$$t_{расч} = \frac{|r|}{\sigma_r}$$

- и сопоставляется с $t_{табл}$ при заданном уровне значимости α и числе степеней свободы $\nu = n-2$.
- Если $t_{расч} > t_{табл}$, то коэффициент корреляции считается значимым, а связь между *x* и *y* – реальной.

- Если $t_{расч} < t_{табл}$ то коэффициент корреляции считается незначимым, т.е. связь между *x* и *y* отсутствует, и значение *r*, отличное от нуля, получено случайно.
- Пример.

Проверим на значимость линейный коэффициент корреляции, рассчитанный в примере.

Число наблюдений *n*=10, *r*=-0,82. Рассчитаем среднюю ошибку коэффициента корреляции

$$\sigma_r = \frac{\sqrt{1-r^2}}{\sqrt{n-2}} = \frac{\sqrt{1-0,82^2}}{\sqrt{10-2}} = 0,20$$

Отсюда

$$t_{расч} = \frac{|r|}{\sigma_r} = \frac{0,82}{0,20} = 4,1$$

- При $\alpha = 0,05$ и $\nu = n-2 = 10-2 = 8$ $t_{табл} = 2,3060$.
- Полученное значение $t_{расч} = 4,1$ больше $t_{табл} = 2,3060$ Поэтому можно сделать вывод о значимости коэффициента корреляции, подтверждая реальную связь между *x* и *y*.

Ранговая корреляция

- Ранговая корреляция основана на корреляции не самих значений коррелируемых признаков, а их рангов.
- ***Ранг*** – это порядковый номер, присваиваемый каждому индивидуальному значению признаков (отдельно) в ранжированном ряду. Оба признака необходимо ранжировать (нумеровать) в одном и том же порядке: от меньших значений к большим и наоборот.
- Если встречаются несколько одинаковых значений *x* или *y*, используют объединённые (или связные) ранги. Всем связным рангам присваивается один и тот же ранг, равный среднему арифметическому рангов, входящих в данную группу. Например, если в ранжировке объекты, находящиеся на 3-м, 4-м, 5-м и 6-м местах, неразличимы по данному признаку, то каждому из них присваивается ранг, равный

$$\frac{3+4+5+6}{4}=4,5$$

- т.е. мы получаем последовательность 4,5; 4,5; 4,5; 4,5.

Коэффициент корреляции рангов Спирмена

- Для расчета коэффициента Спирмена значения признаков *x* и *y* нумеруют (отдельно) в порядке возрастания от 1 до *n*, т.е. им присваивают определённый ранг (и). Для каждой пары рангов находят их разность

$$d = N_x - N_y$$

и квадраты этой разности суммируют. Коэффициент Спирмена определяется как:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

- Коэффициент Спирмена принимает значения от -1 до 1.
- Если ранги двух признаков совпадают, т.е. для каждого значения $N_x = N_y$ и $\sum d^2 = 0$, то $\rho = 1$. Это указывает на максимально тесную прямую связь между признаками.
- Если ранги двух признаков имеют строго противоположное направление, то $\rho = -1$, что указывает на полную обратную связь.
- Если $\rho = 0$, то связь между признаками отсутствует.
- Во всех остальных случаях коэффициент Спирмена довольно близок к ***r***, так как он представляет собой модификацию одной их формул линейного коэффициента корреляции, где вместо значений признаков *x* и *y* используются их ранги.

Пример. Рассмотрим расчет коэффициента ранговой корреляции Спирмена для средней взвешенной цены и объёма продаж облигаций на ММВБ 01.01.2004 г.

№ п/п	Средняя взвешенная цена, тыс. руб. x	Объём продаж, млрд.руб. y	Ранги		Разность рангов $d = N_x - N_y$	d^2
			N_x	N_y		
1	84,42	69,5	1	10	-9	81
2	82,46	72,9	2	8	-6	36
3	80,13	71,4	3	9	-6	36
4	63,42	135,1	10	1	9	81
5	76,17	76,3	4	4	0	0
6	75,13	74,7	5	7	-2	4
7	74,84	97,4	6	3	3	9
8	73,03	75,1	8	6	2	4
9	73,41	75,5	7	5	2	4
10	71,34	98,2	9	2	7	49
n=10					$\sum$	304

- Рассчитаем коэффициент Спирмена:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \cdot 304}{10 \cdot (100-1)} = -0,84$$

- Полученное значение $\rho = -0,84$ свидетельствует о сильной обратной связи между признаками *x* и *y*.
- Этот вывод согласуется с полученным значением линейного коэффициента корреляции $r = -0,82$.

Уравнение регрессии

- Этот термин в статистике впервые был использован Френсисом Гальтоном (1886) в связи с исследованием вопросов наследования физических характеристик человека.
- В качестве одной из характеристик был взят рост человека; при этом было обнаружено, что в целом сыновья высоких отцов, что не удивительно, оказались более высокими, чем сыновья отцов с низким ростом.
- Более интересным было то, что разброс в росте сыновей был меньшим, чем разброс в росте отцов. Так проявлялась тенденция возвращения роста сыновей к среднему (*regression to mediocrity*), то есть «регресс». Этот факт был продемонстрирован вычислением среднего роста сыновей отцов, рост которых равен 56 дюймам, вычислением среднего роста сыновей отцов, рост которых равен 58 дюймам, и т. д.
- После этого результаты были изображены на плоскости, по оси ординат которой откладывались значения среднего роста сыновей, а по оси абсцисс — значения среднего роста отцов. Точки (приближённо) легли на прямую с положительным углом наклона меньше 45°; важно, что регрессия была линейной.

Уравнение регрессии

- Найти ***уравнение регрессии*** – значит по эмпирическим данным математически описать изменения взаимно коррелируемых величин.
- Уравнение регрессии определяет, каким будет **среднее значение** результативного признака ***у*** при произвольном значении факторного признака ***х***, если остальные факторы не учитывать.
- Уравнение регрессии называется также **теоретической линией регрессии**. Рассчитанные по уравнению регрессии значения результативного признака называются теоретическими.
- Выбор уравнения регрессии осуществляется на основе исходных эмпирических данных и теоретического обоснования рабочей гипотезы о взаимодействии признаков.

- Для аналитической связи между **x** и **y** могут использоваться следующие простые виды уравнений:
- а) $\bar{y}_x = a_0 + a_1 x$ (прямая);
- б) $\bar{y}_x = a_0 + a_1 x + a_2 x^2$ (парабола);
- в) $\bar{y}_x = a_0 + a_1 \dfrac{1}{x}$ (гипербола);
- г) $\bar{y}_x = a_0 a_1^{\,x}$ (показательная функция);
- д) $\bar{y}_x = a_0 + a_1 \ln x$ (логарифмическая функция);
- е) $\bar{y}_x = \dfrac{d}{1 + e^{a_0 + a_1 x}}$ (логистическая функция) и др.
- Зависимость, выражаемую уравнением прямой, называется линейной, а все остальные – нелинейными.

- Для определения параметров уравнения регрессии наиболее часто используется метод наименьших квадратов (МНК).
- В соответствии с этим методом параметры уравнения прямой находятся из системы линейных уравнений:

$$\begin{cases} na_0 + a_1 \sum x = \sum y \\ a_0 \sum x + a_1 \sum x^2 = \sum xy \end{cases}$$

Здесь *x* и *y* – индивидуальные значения признаков;

n – количество пар значений признаков *x* и *y*.

- Решая систему методом определителей, находим параметры:

$$a_0 = \frac{\sum y \sum x^2 - \sum xy \sum x}{n\sum x^2 - \left(\sum x\right)^2} \qquad a_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

- Для определения параметров параболического уравнения регрессии решается система уравнений:

$$\begin{cases} na_0 + a_1\sum x + a_2\sum x^2 = \sum y \\ a_0\sum x + a_1\sum x^2 + a_2\sum x^3 = \sum xy \\ a_0\sum x^2 + a_1\sum x^3 + a_2\sum x^4 = \sum x^2 y \end{cases}$$

- В случае гиперболической функции в систему вместо значений *x* подставляются $\frac{1}{x}$, логарифмической – $\ln x$.
- Показательную функцию можно привести к линейному виду, предварительно её прологарифмировав: $\ln y_x = \ln a_0 + x \ln a_1$.
- Параметр в уравнении линейной регрессии называется ***коэффициентом регрессии***. Он показывает, на сколько изменяется значение результативного признака ***y*** при изменении факторного признака ***x*** на единицу.

- В экономическом анализе также используется ***коэффициент эластичности***, определяемый на основе уравнения регрессии:

$$Э = \frac{d\overline{y}_x}{dx}\frac{x}{\overline{y}_x}$$

- Для линейной зависимости:

$$Э = a_1 \frac{x}{a_0 + a_1 x}$$

- Коэффициент эластичности изменяется с изменением факторного признака. Коэффициент эластичности показывает, на сколько процентов в среднем изменяется результативный признак *y* при изменении факторного признака *x* на 1%.

- Средняя ошибка аппроксимации определяется следующим образом:

$$\overline{\varepsilon} = \frac{\sum \left| y_i - \overline{y}_x(x_i) \right|}{\sum y_i}$$

Теоретическое корреляционное отношение и теоретический коэффициент детерминации

- ***Корреляционное отношение*** является универсальным показателем тесноты связи, применимым ко всем случаям корреляционной зависимости независимо от формы этой связи.
- ***Эмпирическое корреляционное отношение***, рассмотренное выше, рассчитывается по аналитической группировке.
- ***Теоретическое корреляционное отношение*** определяется на основе теоретических значений результативного признака $\overline{y}_x$, рассчитанных по уравнению регрессии.

- Общая дисперсия результативного признака характеризует общую вариацию признака ***y*** под влиянием всех факторов, от которых он зависит:

$$\sigma_y^2 = \frac{\sum \left(y_i - \bar{y}\right)^2}{n}$$

- ***Факторная дисперсия*** характеризует вариацию признака ***y*** только за счёт признака ***x***:

$$\delta_{фактор}^2 = \frac{\sum \left(\bar{y}_x - \bar{y}\right)^2}{n}$$

- Сравнивая эти дисперсии, получим ***теоретический коэффициент детерминации***:
- $$\eta_{теор}^2 = \frac{\delta_{фактор}^2}{\sigma_y^2}$$ или $$\eta_{теор}^2 = \frac{\sum \left(\bar{y}_x - \bar{y}\right)^2}{\sum \left(y_i - \bar{y}\right)^2}$$

- Коэффициент детерминации показывает, какую долю в общей дисперсии результативного признака занимает дисперсия, выражающая влияние вариации фактора ***x*** на вариацию *y*.
- ***Теоретическое корреляционное отношение*** равно квадратному корню из детерминации:

$$\eta_{теор} = \sqrt{\frac{\delta^2_{фактор}}{\sigma^2_y}} \quad \text{или} \quad \eta_{теор} = \sqrt{\frac{\sum\left(\bar{y}_x - \bar{y}\right)^2}{\sum\left(y_i - \bar{y}\right)^2}}$$

Остаточная дисперсия отражает влияние на вариацию результативного признака всех остальных признаков (кроме *x*), не учтенных в уравнении регрессии:

$$\sigma^2_{ост} = \frac{\sum\left(y_i - \bar{y}_x\right)^2}{n}$$

- Общая дисперсия признака ***y*** равна сумме факторной и остаточной дисперсий:

$$\sigma_y^2 = \delta_{фактор}^2 + \sigma_{ост}^2$$

- Если результативный признак полностью зависит от фактора ***x***, то связь функциональная. В этом случае коэффициент корреляции $\eta = 1$.
- Если фактор **x** не оказывает ни какого влияния на признак **y**, т.е. связь отсутствует, то $\eta = 0$.
- Таким образом, чем η ближе к 1, тем теснее связь между признаками *x* и *y*. Чем ближе η к 0, тем зависимость слабее. При $\eta < 0,3$ говорят о слабой зависимости между коррелируемыми величинами, при $0,3 < \eta < 0,6$ – о средней, при $0,6 < \eta < 0,8$ – о зависимости выше средней, при $\eta > 0,8$ – о сильной зависимости.

Проверка регрессионной модели на адекватность

- Значимость параметров уравнения регрессии проверяется путём их сопоставления со средней ошибкой. ***Средние ошибки параметров*** определяются соответственно:

$$\mu_{a_0} = \frac{\sigma_{ост}}{\sqrt{n-2}}; \qquad \mu_{a_1} = \frac{\sigma_{ост}}{\sigma_x \sqrt{n-2}}.$$

- где

$$\sigma_{ост} = \sqrt{\frac{\sum \left(y - \bar{y}_x \right)^2}{n}}$$

- Отношение коэффициента к его средней ошибке обозначим через *t*:

$$t_{a_0} = \frac{a_0}{\mu_{a_0}}; \qquad t_{a_1} = \frac{a_1}{\mu_{a_1}}.$$

- По значению *t* судят о значимости параметра.
- При большом числе наблюдений (*n*>30) параметр a_i считается значимым, если $t_{a_i} > 3$.

- Если выборка малая (n<30), фактическое (расчетное) t сопоставляется с табличным (критическим) t-критерием Стьюдента, определяемым для числа степеней свободы $\nu = n - 2$ и заданного уровня значимости α (0,05 или 0,01).
- Если $t_{факт} > t_{табл}$, то параметр считается значимым.

- Проверка значимости уравнения регрессии в целом, т.е. проверка адекватности модели осуществляется с помощью F-критерия Фишера.
- F-критерий представляет собой отношение факторной дисперсии результативного признака к остаточной, каждая из которых рассчитана на одну степень свободы:

$$F = \frac{\sigma^2_{факт} \big/ (m-1)}{\sigma^2_{ост} \big/ (n-m)}$$

- где **_m_** – число параметров в уравнении регрессии;

  (m-1) – число степеней свободы для факторной дисперсии;

  n – число наблюдений;

  (n-m) – число степеней свободы для факторной дисперсии.
- Расчетное F сопоставляется с сопоставляется с табличным, определяемым для числа степеней свободы
- $\nu_1 = m-1$ и $\nu_2 = n-m$. Если $F_{расч} > F_{табл}$, то уравнение значимо.