Корреляционно-регрессионный анализ

Содержание

1. Введение
2. Виды зависимостей, изучаемых в статистике
3. Основные методы изучения взаимосвязей
4. Проверка на адекватность регрессионной модели
5. Экономическая интерпретация параметров уравнения регрессии
6. Заключение

Введение

Явления, которые в случае событий массового характера отличаются определенной закономерностью, однако не обнаруживаются на основе единичного наблюдения, называются **массовыми явлениями**. Сама такая закономерность называется **статистической закономерностью**.

Статистическая закономерность наблюдается в тех случаях, когда:

а) в исследуемом процессе действует один общий комплекс причин;

б) наряду с этим в каждом отдельном случае действуют особые дополнительные причины, всякий раз иные.

Для исследования интенсивности, вида и формы причинных связей широко применяется **корреляционный** и **регрессионный** анализы.

Теория и методы **корреляционного анализа** используются для выявления связи между случайными переменными и оценки ее тесноты.

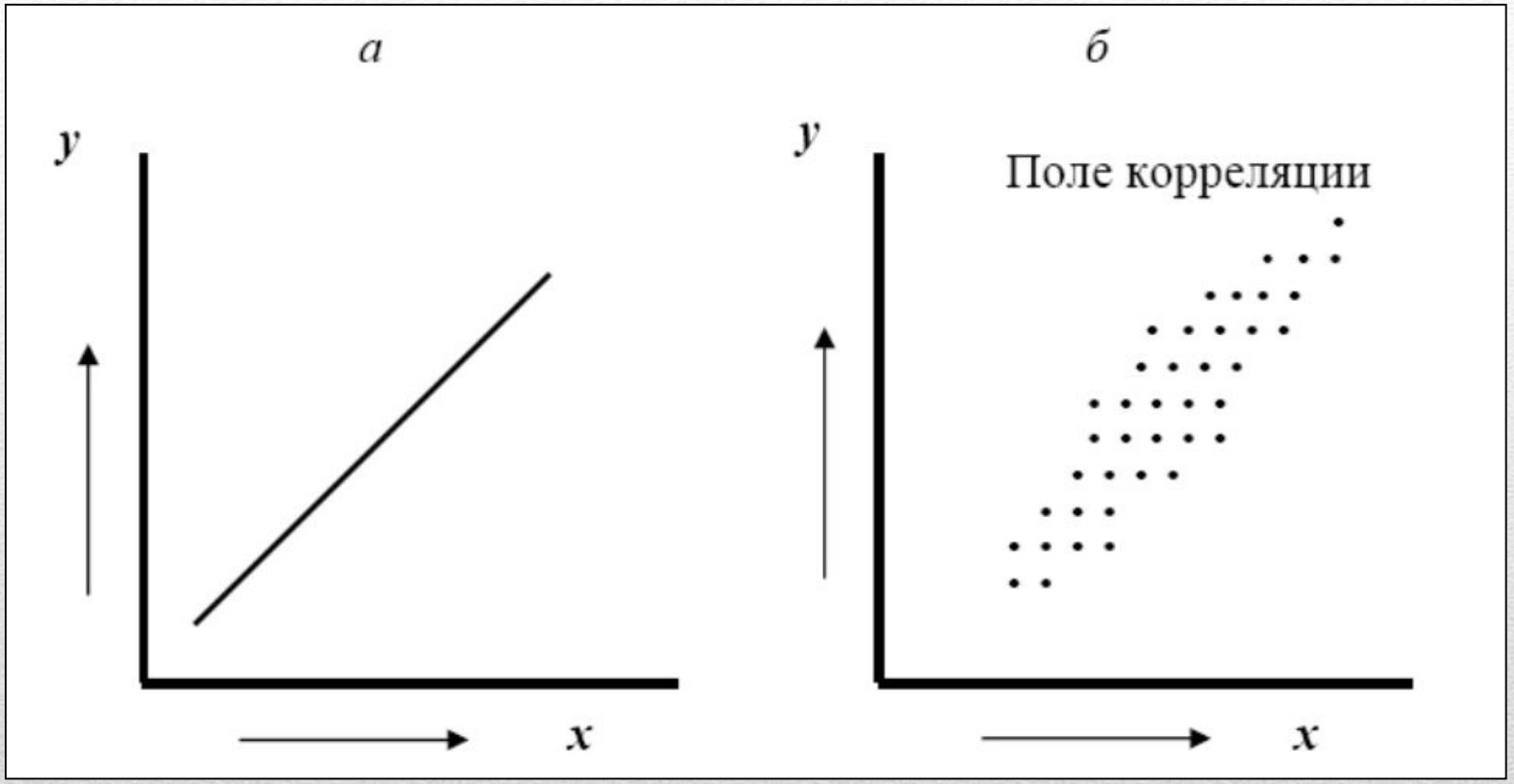
Основной задачей **регрессионного анализа** является установление формы и изучение зависимости между переменными.

Виды зависимостей, изучаемых в статистике

Зависимость одной случайной величины от значений, которые принимает другая случайная величина (физическая характеристика), в статистике называется **регрессией**.

Рассматривая зависимости между признаками, необходимо выделить прежде всего две категории связи:

1. Функциональные – характеризуются полным соответствием между изменением факторного признака и изменением результативной величины.
2. Корреляционные (статистические) - рассматриваются как признак, указывающий на взаимосвязь ряда числовых последовательностей.



Функциональная (а) и статистическая (б) зависимости

функциональная и статистическая зависимости

Аналитически **функциональная** зависимость представляется в следующем виде: $y = f(x)$.

Графически **статистическая** зависимость двух признаков может быть представлена с помощью поля корреляции, при построении которого на оси абсцисс откладывается значение факторного признака X, а по оси ординат – результирующего Y.

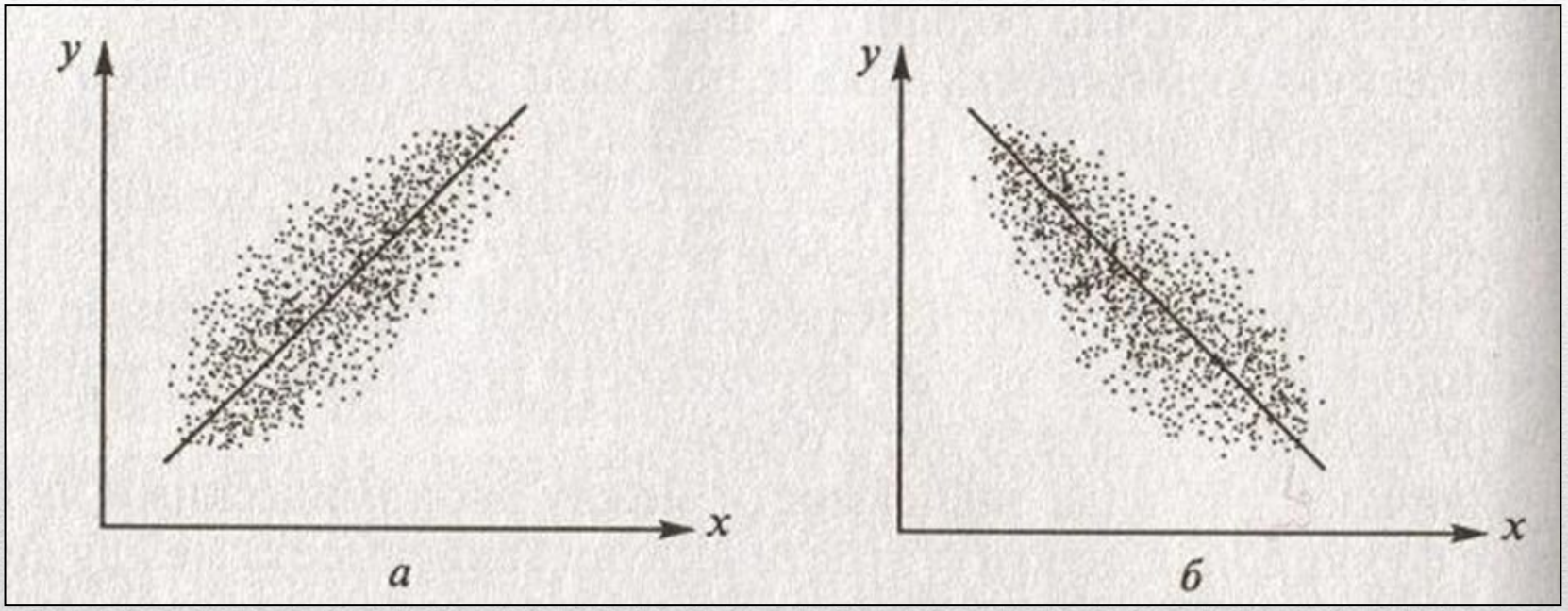


Рисунок 2. Поле корреляции

При исследовании корреляционных зависимостей между признаками решаются следующие задачи:

1. Предварительный анализ свойств моделируемой совокупности единиц.
2. Установление наличия связи в фактическом материале, определение ее направления и формы.
3. Измерение степени тесноты связи между признаками, т. е. степени приближения ее к функциональной зависимости.
4. Построение регрессионной модели, ее экономическая интерпретация и практическое использование.

Основные методы изучения взаимосвязей

Корреляцию и **регрессию** принято рассматривать как совокупный процесс статистического исследования, поэтому их использование в статистике часто именуют корреляционно-регрессионным анализом.

Корреляционно-регрессионный анализ используется для исследования форм связи, устанавливающих количественные соотношения между случайными величинами изучаемого процесса.

Методы

- метод сопоставления
- метод параллельных рядов
- балансовый метод
- графический метод
- методы аналитических группировок
- метод дисперсионного анализа
- метод корреляционного анализа

№ банка	Активы банка, млн.руб	Прибыль, млн.руб
1	80	3,4
2	80	4,2
3	80	3,8
4	82	5,9
5	82	6,0
6	88	5,8
7	88	5,6
8	88	6,3
9	88	6,9
10	90	6,9
11	90	8,2
12	92	7,7
13	92	8,5
14	92	8,0
15	92	9,6
16	92	10,1
17	95	12,0
18	95	11,6
19	99	13,4
20	99	15,8
21	99	16,2
22	105	17,4
23	105	16,5
24	108	19,0
25	108	23,4
26	110	22,6
27	110	20,7
28	110	26,3
29	115	32,0
30	115	31,5

Метод параллельных рядов

Метод параллельных рядов – ряд значений факторного признака и соответствующих ему значений результативного признака (значение признака X располагается в возрастающем порядке, затем прослеживают направление изменения величины результативного признака Y).

Метод аналитических группировок

Сущность этого метода заключается в том, что единицы статистической совокупности группируются, как правило, по факторному признаку и для каждой группы исчисляется средняя или относительная величина по результативному признаку.

Группа банков по активам, млн.руб	Число банков в группе	Средняя прибыль в данных группах банков, млн. руб
80	3	3,8
82	2	5,95
88	4	6,15
90	2	7,55
92	5	8,78
95	2	11,8
99	3	15,13
105	2	16,95
108	2	21,2
110	3	23,2
115	2	31,75
-	30	-

Балансовый метод

Сущность метода заключается в том, что данные взаимосвязанных показателей изображаются в виде таблицы и располагаются таким образом, чтобы итоги между отдельными частями были равны, т.е. чтобы был баланс.

Середина интервала, y	5,785	10,565	15,345	20,125	24,905	29,685	f_x	$\bar{y}_j$
Гр по y / Гр по x	3,4-8,17	8,18-12,95	12,96-17,73	17,74-22,51	22,52-27,29	27,3-32,07		
80	3						3	5,785
82	2						2	5,785
88	4						4	5,785
90	1	1					2	8,175
92	2	3					5	8,653
95		2					2	10,565
99			3				3	15,345
105			2				2	15,345
108				1	1		2	22,515
110				1	2		3	23,312
115						2	2	29,685
f_y	12	6	5	2	3	2	30	-

Аналитический метод

Изучение корреляционных зависимостей основывается на исследовании таких связей между переменными, при которых значение одной переменной можно принять за зависимую переменную, которая «в среднем» изменяется в зависимости от того, какие значения принимает другая переменная, рассматриваемая как причина изменения зависимой переменной.

Теоретической линией регрессии – называется линия, вокруг которой группируются точки корреляционного поля и которая указывает основное направление, основную тенденцию связи.

Важным этапом регрессионного анализа является определение **типа функции**, с помощью которой характеризуется зависимость между признаками.

Тип уравнения выбирается на основе теоретического анализа и исследования фактических данных. В большинстве случаев связи в общественных явлениях изучают по уравнению прямой, вида $y=a+bx$, где a и b – параметры искомой прямой.

Параметры уравнения a и b находятся выравниванием по способу наименьших квадратов, которые приводят к системе двух нормальных уравнений:

$$\begin{cases} na + b\Sigma x = \Sigma y \\ a\Sigma x + b\Sigma x^2 = \Sigma yx \end{cases}$$

$$a = \frac{\Sigma y \Sigma x^2 - \Sigma yx \Sigma x}{n\Sigma x^2 - (\Sigma x)^2}$$

$$b = \frac{n\Sigma yx - \Sigma y \Sigma x}{n\Sigma x^2 - (\Sigma x)^2}$$

Для измерения тесноты линейной зависимости рассчитывают линейный коэффициент корреляции (r).

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}$$

или

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\left[\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2\right]}}$$

Линейный коэффициент корреляции может принимать любые значения в пределах от -1 до +1. Чем ближе коэффициент корреляции по абсолютной величине к 1, тем теснее связь между признаками. Знак при линейном коэффициенте корреляции указывает на направление связи: прямой зависимости соответствует знак «+», а обратной зависимости – знак «-».

Зная коэффициент корреляции, можно дать качественно-количественную оценку тесноты связи. Используются, например, специальные табличные соотношения (так называемая шкала Чеддока).

Качественная оценка тесноты связи

Величина коэффициента парной корреляции	Характеристика силы связи
До 0,3	Практически отсутствует
0,3–0,5	Слабая
0,5–0,7	Заметная
0,7–0,9	Сильная
0,9–0,99	Очень сильная

	x	y	x*y	x^2	y^2	$\hat{Y}$
1	80,000	3,400	272,000	6 400,000	11,560	1,482
2	80,000	4,200	336,000	6 400,000	17,640	1,482
3	80,000	3,800	304,000	6 400,000	14,440	1,482
4	82,000	5,900	483,800	6 724,000	34,810	2,936
5	82,000	6,000	492,000	6 724,000	36,000	2,936
6	88,000	5,800	510,400	7 744,000	33,640	7,296
7	88,000	5,600	492,800	7 744,000	31,360	7,296
8	88,000	6,300	554,400	7 744,000	39,690	7,296
9	88,000	6,900	607,200	7 744,000	47,610	7,296
10	90,000	6,900	621,000	8 100,000	47,610	8,749
11	90,000	8,200	738,000	8 100,000	67,240	8,749
12	92,000	7,700	708,400	8 464,000	59,290	10,203
13	92,000	8,500	782,000	8 464,000	72,250	10,203
14	92,000	8,000	736,000	8 464,000	64,000	10,203
15	92,000	9,600	883,200	8 464,000	92,160	10,203
16	92,000	10,100	929,200	8 464,000	102,010	10,203
17	95,000	12,000	1 140,000	9 025,000	144,000	12,383
18	95,000	11,600	1 102,000	9 025,000	134,560	12,383
19	99,000	13,400	1 326,600	9 801,000	179,560	15,290
20	99,000	15,800	1 564,200	9 801,000	249,640	15,290
21	99,000	16,200	1 603,800	9 801,000	262,440	15,290
22	105,000	17,400	1 827,000	11 025,000	302,760	19,650
23	105,000	16,500	1 732,500	11 025,000	272,250	19,650
24	108,000	19,000	2 052,000	11 664,000	361,000	21,830
25	108,000	23,400	2 527,200	11 664,000	547,560	21,830
26	110,000	22,600	2 486,000	12 100,000	510,760	23,284
27	110,000	20,700	2 277,000	12 100,000	428,490	23,284
28	110,000	26,300	2 893,000	12 100,000	691,690	23,284
29	115,000	32,000	3 680,000	13 225,000	1 024,000	26,918
30	115,000	31,500	3 622,500	13 225,000	992,250	26,918
Сумм	2 869,000	385,300	39 284,200	277 725,000	6 872,270	385,300

Линейная зависимость y=a+bx

a=	-	56,656
b=		0,73

y=-56,656+0,73x

Проверка на адекватность регрессионной модели

Для практического использования моделей регрессии очень важна их адекватность, т.е. соответствие фактическим данным.

Проверка значимости (существенности) осуществляется с помощью t-критерия Стьюдента. При этом рассчитываются значения t-критерия:

- для параметра a: $t_a = \frac{a}{m_{ост}}\sqrt{n-2}$

- для параметра b: $t_b = \frac{a}{m_{ост}}\sqrt{n-2}\,\sigma_x$

$\sigma_{ост}$ – остаточное среднее квадратическое отклонение результативного признака y от выровненных значений y_x

$$\sigma_{ост} = \sqrt{\frac{\sum (y - y_x)^2}{n}} = \sqrt{\frac{152,9456}{30}} = 2,2579$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2} = \sqrt{\frac{277725}{30} - \left(\frac{2869}{30}\right)^2}$$

$$= \sqrt{9257,5 - 9145,097} = \sqrt{112,403} = 10,6$$

$$y_x = a - 56,656 \cdot \frac{\sqrt{30-2}}{2,2579} = 132,58$$

$$y_x = a0,73 \cdot \frac{\sqrt{30-2}}{2,2579} \cdot 10,6 = 18,107$$

Для линейной однофакторной связи используется формула

$$t_{расч} = r\sqrt{\frac{n-2}{1-r^2}} = 0{,}959\sqrt{\frac{28}{1-0{,}959^2}} = 0{,}959 \cdot 18{,}67 = 17{,}91$$

Экономическая интерпретация параметров уравнения регрессии

После проверки адекватности, установления точности и надежности построенной модели (уравнения регрессии), ее необходимо проанализировать. Прежде всего нужно проверить, согласуются ли знаки параметров с теоретическими представлениями и соображениями о направлении влияния признака-фактора на результативный признак (показатель).

$$V = \frac{\sigma}{\bar{x}} \cdot 100\%,$$

$$\bar{x} = \frac{\sigma}{n} = \frac{2869}{30} = 95{,}63 \text{ млн. руб}$$

$$\sigma = \frac{\sigma}{n} = \frac{385{,}3}{30} = 12{,}84 \text{ млн. руб}$$

$$V = \frac{\sigma}{\bar{x}} \cdot 100\% = 0{,}73 \cdot \frac{95{,}63}{12{,}84} = 5{,}44\ \%$$

Заключение

Полученное уравнение $\hat{y}$=-56,656+0,73x позволяет проиллюстрировать зависимость размера прибыли банков от размера их активов.

А также проведена проверка данной модели на адекватность по критерию Стьюдента, результат оказался положительным (модель адекватна, т.е. ее можно применять), а затем дана экономическая оценка этой модели - при увеличении активов банка увеличивается и прибыль.

Спасибо за внимание