7 Корреляционный и регрессионный анализ

1. Корреляционный анализ статистических данных.
2. Регрессионный анализ статистических данных.

Статистические связи между переменными можно изучать методами дисперсионного, корреляционного и регрессионного анализа. Методами дисперсионного анализа устанавливается наличие влияния заданного фактора на изучаемый процесс. Корреляционный анализ позволяет оценить силу такой связи, а методами регрессионного анализа можно выбрать конкретную математическую модель и оценить ее адекватность.

Корреляционная связь – это согласованное изменение признаков, отражающее тот факт, что изменчивость одного признака находится в соответствии с изменчивостью другого. Парная корреляция изучает взаимосвязи между двумя случайными величинами, множественная – между большим числом величин.

Основная задача корреляционного анализа – выявление и оценка связи между случайными величинами, основная задача регрессионного анализа – установление формы и изучение зависимости между случайными величинами (рисунок 3).

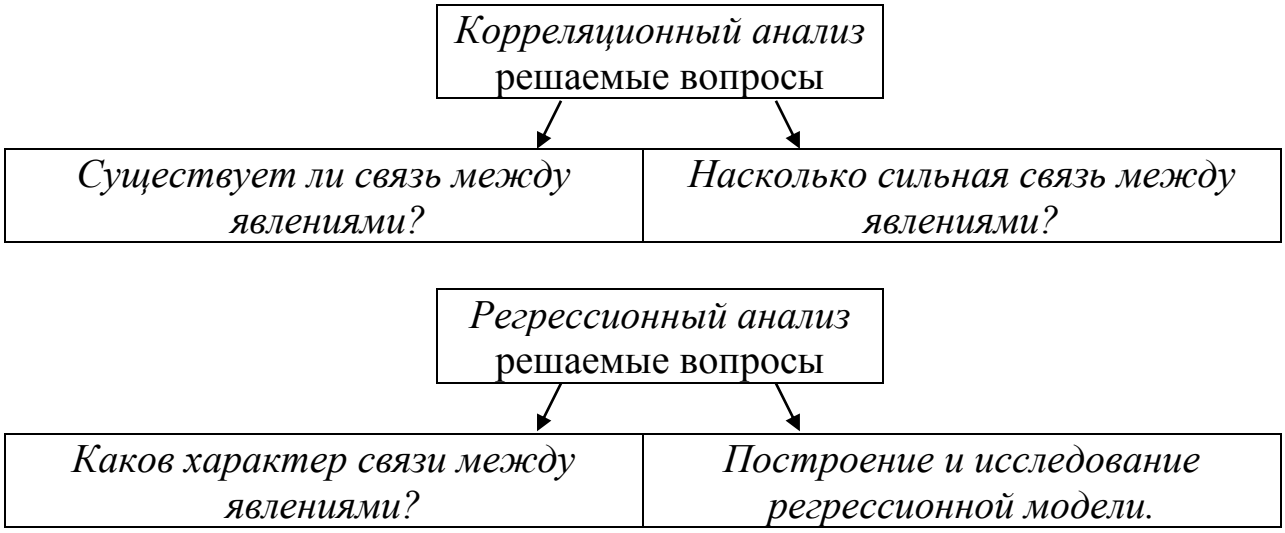


Рисунок 3

Элементы корреляционного анализа. Пусть $(x_1, y_1),...,(x_n, y_n)$ – выборка объема n из наблюдений случайной величины (ξ, η), имеющей двумерное нормальное распределение. Изображая элементы выборки точками в декартовой системе координат, получим *диаграмму рассеивания или корреляционное поле*. Иногда по виду корреляционного

поля можно сделать предположение о наличии и характере связи между случайными величинами ξ и η.

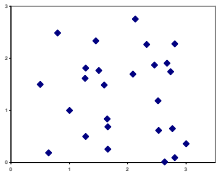
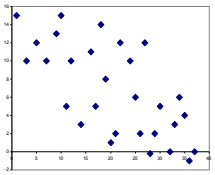
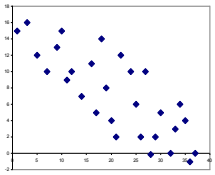
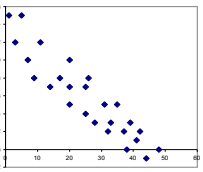
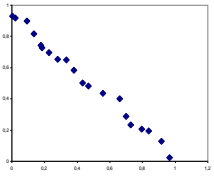
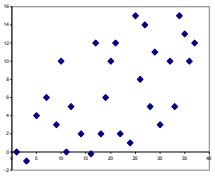
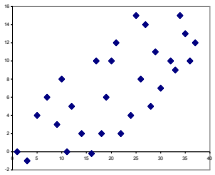
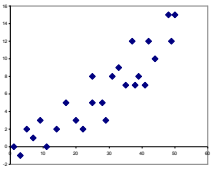
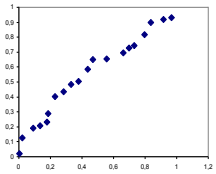
Выборочным *коэффициентом корреляции* называется число

$$r_{\text{выб.}} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}}{\tilde{S}_x \tilde{S}_y}.$$

Можно показать, что $|r_{\text{выб.}}| \le 1$.

В таблице 11 приведены возможные формы корреляционного поля в зависимости от значения выборочного коэффициента корреляции.

Таблица 11

$-0{,}2 < r_{\text{выб.}} < 0{,}2$	$-0{,}5 < r_{\text{выб.}} \le -0{,}2$	$-0{,}75 < r_{\text{выб.}} \le -0{,}5$	$-0{,}95 < r_{\text{выб.}} \le -0{,}75$	$-1 \le r_{\text{выб.}} \le -0{,}95$
				
$0{,}2 \le r_{\text{выб.}} < 0{,}5$	$0{,}5 \le r_{\text{выб.}} < 0{,}75$	$0{,}75 \le r_{\text{выб.}} < 0{,}95$	$0{,}95 \le r_{\text{выб.}} \le 1$	
				

На практике большой интерес представляет задача проверки гипотезы о *значимости* корреляционной связи между случайными величинами, т. е. значимости отклонения коэффициента корреляции от нуля. Пусть $r_{\text{выб.}}$ – выборочный коэффициент корреляции. При заданном уровне значимости α проверяется гипотеза $H_0 : r = 0$ о равенстве нулю теоретического коэффициента корреляции. Если нулевая гипотеза будет отвергнута, то говорят о значимости коэффициента корреляции, а значит о том, что случайные величины ξ и η коррелированы. Если нулевая гипотеза принимается, то коэффициент корреляции незначим, и случайные величины ξ и η некоррелированы.

Статистика критерия имеет вид

$$t_{\text{набл}} = r_{\text{выб.}} \sqrt{\frac{n-2}{1-r_{\text{выб.}}^2}}.$$

Находится $t_{\frac{\alpha}{2};n-2} = t(100\frac{\alpha}{2}\%, n-2)$ – значение процентной точки

распределения Стьюдента с $(n-2)$ степенями свободы.

Схема принятия решения выглядит следующим образом:

– если $|t_{\text{набл}}| = \left| r_{\text{выб.}} \sqrt{\frac{n-2}{1-r_{\text{выб.}}^2}} \right| < t_{\frac{\alpha}{2};n-2}$, то нет оснований отвергать нулевую гипотезу, коэффициент корреляции не значим, а ξ и η некоррелированы;

– если $|t_{\text{набл}}| = \left| r_{\text{выб.}} \sqrt{\frac{n-2}{1-r_{\text{выб.}}^2}} \right| \geq t_{\frac{\alpha}{2};n-2}$, то гипотеза отвергается, и коэффициент корреляции значимо отличается от нуля, а ξ и η коррелированы.

Пример 7.1 Предполагая, что $(x_1, y_1),...,(x_n, y_n)$ – выборка из наблюдений случайной величины (ξ, η), имеющей двумерное нормальное распределение, вычислить выборочный коэффициент корреляции и при заданном уровне значимости $\alpha = 0{,}05$ проверить гипотезу о равенстве нулю теоретического коэффициента корреляции.

$x_{(i)}$	1,37	0,11	1,56	–0,11	0,23	–0,76	–0,13	–0,64	–0,46	–0,88
$y_{(i)}$	0,08	0,64	1,59	1,75	0,74	0,89	1,44	0,72	1,26	0,03

–0,56	1,28	1,16	–0,30	–0,31	1,13	–0,17	0,60	–1,16	2,65	1,55
0,92	0,51	0,52	0,43	1,06	0,47	0,11	1,27	1,33	0,32	1,48

0,29	–2,16	–0,77	0,93	0,01	–1,56	1,59	–1,13	–1,74
1,61	1,47	0,98	0,54	0,59	0,34	0,17	0,20	0,27

Выборочный коэффициент корреляции:

$$r_{\text{выб.}} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x} \cdot \bar{y}}{\tilde{S}_x \tilde{S}_y} \approx \frac{\frac{1}{30}(1{,}37 \cdot 0{,}08 + ... + (-1{,}74) \cdot 0{,}27) - 0{,}05 \cdot 0{,}79}{1{,}12 \cdot 0{,}52} \approx -0{,}08.$$

Проверяется гипотеза $H_0 : r = 0$.

Статистика критерия имеет вид

$$t_{\text{набл}} = r_{\text{выб.}}\sqrt{\frac{n-2}{1-r_{\text{выб.}}^2}} \approx -0{,}08 \cdot \sqrt{\frac{30-2}{1-(-0{,}08)^2}} \approx -0{,}4.$$

Находим значение процентной точки распределения Стьюдента $t_{\frac{0{,}05}{2};30-2} = t(2{,}5\,\%, 28) \approx 2{,}048$.

Поскольку $|t_{\text{набл}}| \approx |-0{,}4| < t_{0{,}025;28} \approx 2{,}048$, то нет оснований отвергать нулевую гипотезу H_0, и коэффициент корреляции не значим, а ξ и η некоррелированы.

Линейный регрессионный анализ. Часто требуется определить, как зависит наблюдаемая случайная величина от одной или нескольких других величин. Регрессионный анализ – раздел математической статистики, изучающий связь между зависимой переменной и одной или несколькими независимыми переменными.

Наблюдаются значения $(x_1, y_1),...,(x_n, y_n)$ двумерной случайной величины ξ, η. Исследуется зависимость случайной величины η от случайной величины ξ.

В общем случае регрессионная модель имеет вид

$$y = f(x, \beta_0, \beta_1,..., \beta_k).$$

Параметры $\beta_0, \beta_1,..., \beta_k$ называются *коэффициентами регрессии*.

Одна из задач регрессионного анализа – оценка коэффициентов регрессии. Для оценки коэффициентов регрессии, как правило, используется *метод наименьших квадратов*: в качестве оценок принимаются такие значения параметров, которые минимизируют сумму квадратов отклонений наблюдаемых значений y_i от $\tilde{y}_i = f(x_i, \beta_0, \beta_1,..., \beta_k)$, ($i = 1,..., n$), т. е. метод наименьших квадратов основан на минимизации суммы квадратов:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \to \min.$$

Если предположить, что связь между переменными линейна, то соответствующая регрессионная модель имеет вид

$$y = \beta_0 + \beta_1 x,$$

где β_0 и β_1 – коэффициенты линейной регрессии.

Для линейной модели регрессии задача минимизации имеет вид:

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)^2\to\min.$$

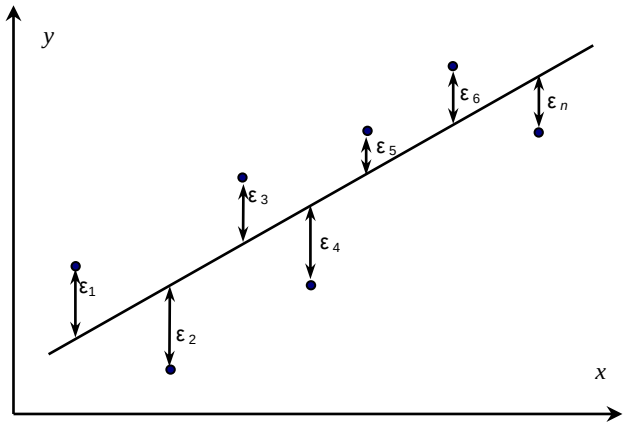


Рисунок 4

На рисунке 4 изображены отклонения $\varepsilon_i=y_i-\tilde{y}_i,\ i=1,...,n$.

Необходимым условием минимума функции двух переменных β_0 и β_1 является равенство ее частных производных по β_0 и β_1 нулю:

$$\begin{cases}\sum_{i=1}^{n}x_i(y_i-\beta_1x_i-\beta_0)=0,\\ \sum_{i=1}^{n}(y_i-\beta_1x_i-\beta_0)=0.\end{cases}$$

Решение системы дает искомые оценки коэффициентов линейной регрессии:

$$b_0=\bar{y}-r_{\text{выб.}}\frac{S_y}{S_x}\bar{x};$$

$$b_1=r_{\text{выб.}}\frac{S_y}{S_x}.$$

Здесь b_0 и b_1 – оценки β_0 и β_1 соответственно.

Пример 7.2 По данным наблюдений двумерной случайной величины ξ,η построить выборочное уравнение линейной регрессии η на ξ и выборочное уравнение линейной регрессии ξ на η.

1) Выборочное уравнение линейной регрессии η на ξ имеет вид:

$$y=b_1x+b_0,$$

где $b_0=\bar{y}-r_{\text{выб.}}\frac{S_y}{S_x}\bar{x}\approx0{,}79-(-0{,}08)\cdot\frac{0{,}51}{1{,}12}\cdot0{,}05\approx0{,}79;$

$$b_1 = r_{\text{выб.}} \frac{S_y}{S_x} \approx (-0{,}08) \cdot \frac{0{,}51}{1{,}12} \approx -0{,}03 .$$

Таким образом, искомое уравнение:

$$y = -0{,}03x + 0{,}79 .$$

2) Выборочное уравнение линейной регрессии ξ на η имеет вид:

$$x = b_1^* y + b_0^* ,$$

где $b_0^* = \bar{x} - r_{\text{выб.}} \frac{S_x}{S_y} \bar{y} \approx 0{,}05 - (-0{,}08) \cdot \frac{1{,}12}{0{,}51} \cdot 0{,}79 \approx 0{,}19 ;$

$$b_1^* = r_{\text{выб.}} \frac{S_x}{S_y} \approx (-0{,}08) \cdot \frac{1{,}12}{0{,}51} \approx -0{,}17 .$$

Таким образом, искомое уравнение:

$$x = -0{,}17 y + 0{,}19 .$$

На рисунке 5 приводятся графики уравнений линейной регрессии.

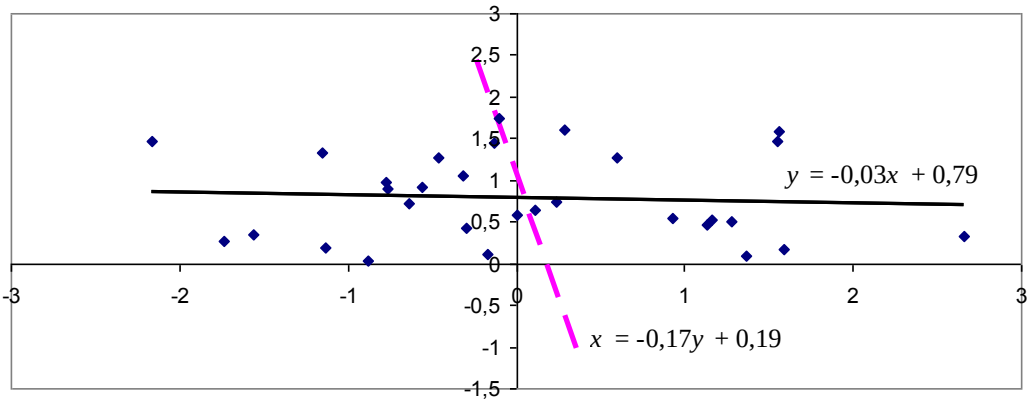


Рисунок 5

Вопросы для самоконтроля

1. Какие задачи решает корреляционный анализ?
2. Как найти выборочный коэффициент корреляции?
3. Какие задачи решает регрессионный анализ?
4. Как построить выборочные уравнения линейной регрессии?