КОРРЕЛЯЦИОННО - РЕГРЕССИОННЫЙ АНАЛИЗ

1. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Установление связи между различными процессами имеет большое значение для понимания функционирования организма в целом. Так, при возрастании физической нагрузки учащаются сердцебиение и частота дыхания, имеется связь между минутным объемом крови и средним давлением в левом предсердии, между ростом и массой человека и т.п.

Именно поэтому естественно использовать наличие связи между изучаемыми процессами для предсказания, прогнозирования изменения одного процесса, определить значение которого затруднительно, через изменения другого связанного с ним процесса, измерения которого не вызывают трудности.

Следует отметить, что определение связи позволяет установить наличие взаимных изменений между характеристиками двух процессов, но не выявить наличие причинно-следственной связи между ними.

Могут быть третьи процессы, которые одновременно влияют на изучаемые процессы и являются причиной их изменений.

Вопросами выявления тесноты, направления и формы связи между двумя переменными занимаются корреляционный анализ и регрессионный анализ.

Определения

Корреляционный анализ — это статистический метод, изучающий связь между явлениями, если одно из них входит в число причин, определяющих другое, или если имеются общие причины, воздействующие на эти явления. Основная задача — выявление связи между случайными переменными.

Регрессионный анализ — это статистический метод, изучающий зависимость между результативным признаков Y и входной переменной X. Основная задача — установление формы связи между переменными и изучение зависимости между ними.

Функциональная и корреляционная зависимости

Существует два вида категории зависимостей между признаками: функциональная и корреляционная.

Функциональная зависимость — это зависимость вида $y = f(x)$, когда каждому возможному значению случайной величины X соответствует одно возможное значение случайной величины Y. Например, площадь круга S однозначно связаны с радиусом окружности R: $S = \pi R^2$.

Все аналитические формулы физики являются примером функциональной зависимости.

Однако в биологических системах, которые относятся к вероятностным системам, такая зависимость встречается редко. Например, зависимость между ростом и длиной стопы человека.

В большинстве случаев более высокий человек имеет больший размер обуви. Однако бывают и исключения. Именно поэтому здесь можно говорить о статистической зависимости между двумя случайными величинами. Размер ноги, в среднем, зависит от роста человека, и между переменными имеется корреляционная связь.

Корреляционная зависимость — это статистическая зависимость, проявляющаяся в том, что при изменении одной из величин изменяется среднее значение другой: $\bar{y} = f(x)$.

Например, рост и масса. При одном и том же росте масса различных индивидуумов может быть разной, но между средними значениями этих показателей имеется определенная зависимость.

Установление взаимосвязи между различными признаками и показателями функционирования организма позволяет по изменениям одних судить о состоянии других.

Для изучения корреляционной связи, данные о статистической зависимости удобно задавать в виде корреляционной таблицы или в виде двумерной выборки (табл. 8.1).

Таблица 8.1. Двумерная выборка

X	x_1	x_2	...	x_n
Y	y_1	y_2	...	y_n

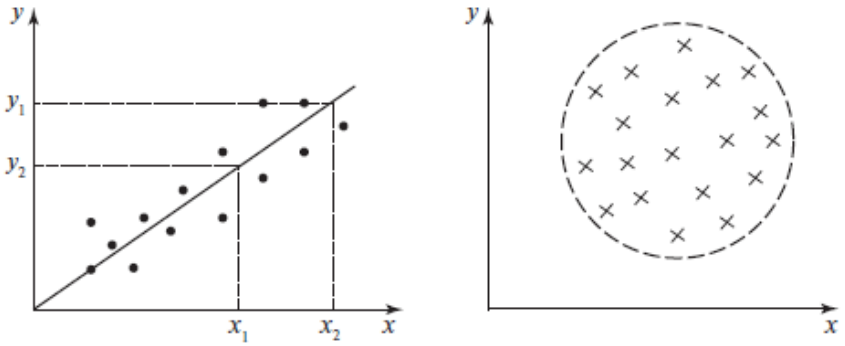
Схема эксперимента следующая: пусть имеется выборка объема n из генеральной совокупности N. На каждом объекте выборки определяют числовые значения признаков, между которыми требуется установить наличие или отсутствие связи. Таким образом, получают два ряда числовых значений.

Для наглядности полученного материала каждую можно представить в виде точки на координатной плоскости. По оси абсцисс откладывают значения одного вариационного ряда — x_i, а по оси ординат — другого — y_i.

Такое изображение статистической зависимости называется *полем корреляции*, или *корреляционным полем точек*. Оно создает общую картину корреляции.

Если точки группируются вдоль некоторого направления (рис. 8.1, *а*), то это говорит о наличии линейной корреляционной связи между признаками.

Если точки распределены равномерно (рис. 8.1, *б*), то линейная корреляционная связь отсутствует.



а *б*

Рис. 8.1. Корреляционное поле: *а* — наличие линейной корреляционной связи между признаками; *б* — отсутствие линейной корреляционной связи

Коэффициент линейной корреляции и его свойства

На практике исследователя часто может интересовать не сама зависимость одной переменной от другой, а именно характеристика тесноты связи между ними, которую можно было бы выразить одним числом. Эта характеристика была предложена К. Пирсоном и называется *выборочным коэффициентом линейной корреляции* $r_в$.

Требования к корреляционному анализу: корреляционный анализ — это метод, используемый, когда данные можно считать случайными и выбранными из совокупности, распределенной по *нормальному закону*.

Выборочный коэффициент линейной корреляции $r_в$ характеризует тесноту линейной связи между количественными признаками в выборке:

$$r_в = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (8.1)$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ;$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

Если $r > 0$, то корреляционная связь между переменными прямая, при $r = 0$ связь обратная.

Свойства коэффициента корреляции $r_в$ проявляются при достаточно большом объеме выборки n.

1. Коэффициент корреляции принимает значения на отрезке $-1 \le r_в \le 1$. В зависимости от того, насколько $|r_в|$ приближается к 1, различают связи:

- $r_в < 0{,}3$ — слабая связь;
- $r_в = 0{,}3–0{,}5$ — умеренная связь;

- $r_в$ = 0,5–0,7 — заметная (значительная);
- $r_в$ = 0,7–0,8 — достаточно тесная;
- $r_в$ = 0,8–0,9 — тесная (сильная);
- $r_в$ > 0,9 — очень сильная, то есть чем ближе $|r_в|$ к 1, тем теснее связь.

2. При $r_в = 1$ — функциональная зависимость $y = f(x)$.

3. Чем ближе $|r_в|$ к 0, тем слабее связь.

4. При $r_в = 0$ линейная корреляционная связь отсутствует.

5. $r_{xy} = r_{yx}$ — случайные переменные симметричные. x и y могут взаимозаменяться, не влияя на величину $r_в$.

6. Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.

7. Коэффициент корреляции — величина безразмерная.

Пример 8.1. Вычислить коэффициент корреляции между ростом (X) и массой (Y) некоторых животных. Исходные данные приведены в выборке объема $n = 10$ (табл. 8.2).

Таблица 8.2. Исходные данные для примера 8.1

x_i	31	32	33	34	35	35	40	41	42	46
y_i	7,8	8,3	7,6	9,1	9,6	9,8	11,8	12,1	14,7	13,0

Решение. Применим формулу (8.1):

$$r_в = \frac{\sum_{i=1}^{10}(x_i - \overline{x})\cdot(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{10}(x_i - \overline{x})^2 \sum_{i=1}^{10}(y_i - \overline{y})^2}}.$$

Средний рост $\overline{x}$:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{10} x_i}{10}, \quad \overline{x} = \frac{31+32+...+46}{10} = \frac{369}{10} = 36,9.$$

Средняя масса $\overline{y}$:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{10} y_i}{10}, \quad \overline{y} = \frac{7,8+8,3+...+13,0}{10} = 10,38.$$

Находим:

$$\sum_{i=1}^{10}(x_i - \overline{x})\cdot(y_i - \overline{y}) =$$

$$= (31-36,9)\cdot(7,8-10,38)+...+(46-36,9)\cdot(13-10,38) = 99,9;$$

$$\sum_{i=1}^{10}(x_i - \overline{x})^2 = (31-36,9)^2+(32-36,9)^2+...+(46-36,9)^2 = 224,8;$$

$$\sum_{i=1}^{10}(y_i - \overline{y})^2 = (7,8-10,38)^2+(8,3-10,38)^2+...+(13,0-10,38)^2 = 51,9.$$

Подставим полученные значения в формулу для $r_в$:

$$r_в = \frac{99,9}{\sqrt{224,8\cdot 51,9}} = 0,925.$$

Величина $r_в$ близка к 1, что говорит о тесной связи роста и массы.◁

Следует отметить, что коэффициент корреляции Пирсона характеризует лишь связи, имеющие линейную или близкую к линейной зависимости. Именно поэтому если коэффициент корреляции не слишком отличается от нуля, то это говорит об отсутствии линейной связи, хотя нелинейная связь может иметь место. В этом случае необходимо обратить внимание на корреляционное поле точек. Если точки находятся вблизи воображаемой прямой (см. рис. 8.1, *а*), то в этом случае имеет смысл искать тесноту линейной связи между признаками.

Проверка гипотезы о значимости выборочного коэффициента линейной корреляции

Это ответ на вопрос: существует ли вообще эта связь? Эмпирический коэффициент корреляции, как и $r_в$ или любой другой выборочный показатель, служит оценкой своего генерального параметра. Выборочный коэффициент линейной корреляции $r_в$ — величина *случайная*, так как он вычисляется по значениям переменных, случайно попавших в выборку из генеральной совокупности, а значит, как и любая случайная величина имеет ошибку m_r.

Для того чтобы выяснить, находятся ли случайные величины X и Y генеральной совокупности в линейной корреляционной зависимости, надо проверить значимость $r_в$. Для этого проверяют нулевую гипотезу о равенстве нулю коэффициента корреляции генеральной совокупности H_0: $r_{ген} = 0$, то есть предположение «линейная корреляционная связь между признаками X и Y случайна». Выдвигают альтернативную гипотезу H_1: $r_{ген} \neq 0$, то есть эта линейная корреляционная связь имеется. Задают уровень значимости, например $\alpha \leq 0{,}05$.

Критерием для проверки нулевой гипотезы является отношение выборочного коэффициента корреляции к своей ошибке

$$t_{набл} = \frac{r_в}{m_r},$$

где m_r — ошибка коэффициента корреляции.

Если объем выборки $n < 100$, то $m_r = \sqrt{\frac{1-r_в^2}{n-2}}$.

Если объем выборки $n > 100$, то $m_r = \frac{1-r_в^2}{\sqrt{n}}$.

Число степеней свободы для проверки критерия равно $f = n - 2$. Гипотезу проверяют по таблицам распределения Стьюдента в соответствии с выбранным уровнем значимости.

По таблице распределения Стьюдента (см. табл. П1 в приложении) находим $t_{крит}(\alpha; t)$, определенное на уровне значимости $\alpha \leq 0{,}05$ при числе степеней свободы $f = n - 2$, где n — объем двумерной выборки.

Если $t_{набл} > t_{крит} \Rightarrow H_1$ — отвергают нулевую гипотезу и принимают альтернативную: $r_{ген} \neq 0$, имеется линейная корреляционная связь между признаками.

Если $t_{набл} < t_{крит}$, то нет оснований отвергать нулевую гипотезу, а $r_в$ статистически незначим. Эта связь случайна.

Пример 8.2. Изучали зависимость между систолическим давлением (мм рт.ст.) у мужчин в начальной стадии шока и возрастом X (годы). Результаты наблюдений приведены в виде двумерной выборки объема 11 (табл. 8.3). Вычисленный коэффициент корреляции равен $r_в = -0{,}61$. Проверить значимость выборочного коэффициента линейной корреляции.

Таблица 8.3. Исходные данные для примера 8.2

x_i	68	37	50	53	75	66	52	65	74	65	54
y_i	114	149	146	141	114	112	124	105	141	120	124

Решение. Проверим нулевую гипотезу об отсутствии линейной корреляционной связи между переменными X и Y в генеральной совокупности H_0: $r_{ген} = 0$.

При справедливости этой гипотезы

$$t_{набл} = \frac{r_в}{m_r},$$

где ошибка коэффициента корреляции

$$m_r = \sqrt{\frac{1-r_в^2}{n-2}}$$

имеет распределение Стьюдента с $f = n - 2$ степенями свободы.

$$m_r = \sqrt{\frac{1-0,61^2}{11-2}} = \sqrt{\frac{1-0,37}{9}} = 0,26;$$

$$t_{\text{набл}} = \frac{r_{\text{в}}}{m_r} = \frac{0,61}{0,26} = 2,34.$$

По таблице Стьюдента (см. табл. П1 в приложении) находим табличное значение $t_{\text{крит}}$, определенное на уровне значимости $\alpha \leq 0,05$ и при $f = 11 - 2 = 9$.

$$t_{\text{крит}}(0,05; 9) = 2,26.$$

Поскольку $t_{\text{набл}} > t_{\text{крит}}$, то справедлива гипотеза H_1 и коэффициент корреляции значимо отличается от нуля с вероятностью 0,95. ◁

2. ВЫБОРОЧНОЕ УРАВНЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ

Метод наименьших квадратов

При проведении современных клинических исследований обычно нет недостатка в информации: каждому пациенту соответствует целое множество различных клинических показателей и данных. В них могут быть завуалированы некоторые соотношения, основные черты которых и позволяют выявлять методы регрессионного анализа. При этом задача регрессионного анализа состоит в подборе упрощенной аппроксимации этой связи с помощью математической модели.

Регрессионный анализ имеет в своем распоряжении специальные процедуры проверки, является ли выбранная математическая модель *адекватной* для описания имеющихся данных.

Чаще всего регрессионный анализ используется для *прогноза*, то есть предсказания значений ряда зависимых переменных по известным значениям других переменных.

Выше указывалось, что результаты наблюдений, приведенные в двумерной выборке (табл. 8.7), можно представить в виде корреляционного поля точек (рис. 8.2), где каждая точка соответствует отдельным значениям x и y.

Таблица 8.7. Пример записи двумерной выборки

x_i	x_1	x_2	x_3	x_4	x_5
y_i	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$	$\bar{y}_5$

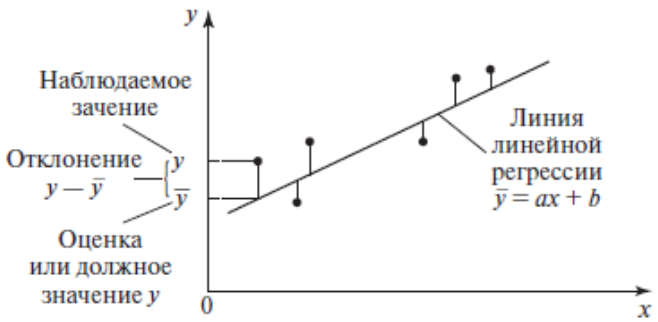


Рис. 8.2. Метод наименьших квадратов

В результате получается диаграмма рассеяния, позволяющая судить о форме и тесноте связи между варьирующими признаками. Довольно часто эта связь может быть аппроксимирована прямой линией (см. рис. 8.2).

Определение

Регрессия — это функция, позволяющая по величине одного признака X находить среднее ожидаемое (должное) значение другого признака Y, корреляционно связанного с X.

В линейной математической модели уравнение *линейной регрессии* имеет вид:

$$\bar{y} = ax + b,$$

где a и b — параметры линейной регрессии; a — коэффициент регрессии, показывающий, насколько в среднем величина одного признака Y изменяется при изменении на единицу меры другого признака X, корреляционно связанного с Y (чем больше α — угловой коэффициент прямой $a = \operatorname{tg}\alpha$, тем круче прямая, то есть быстрее изменяется Y); b — свободный член в уравнении, определяет $\bar{y}$ при $x = 0$; $\bar{y}$ — предсказанное (должное) значение Y для данного x при определенных значениях регрессионных параметров.

Параметры линейной регрессии определяют *методом наименьших квадратов*. Это способ подбора параметров регрессионной модели, согласно которому сумма квадратов отклонений вариант от линии регрессии должна быть минимальной:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \Rightarrow \min . \qquad (8.3)$$

Необходимо подобрать параметры a и b так, чтобы точки, построенные по результатам наблюдений, находились как можно ближе к прямой линии регрессии.

Следовательно необходимо найти значения a и b, при которых функция $F(a, b)$ будет иметь минимум:

$$F(a, b) = \sum_{i=1}^{n}\left(\bar{y}_i - y_i\right)^2 \Rightarrow \min$$

или

$$F(a, b) = \sum_{i=1}^{n}\left(ax_i + b - y_i\right)^2 \Rightarrow \min .$$

Для отыскания минимума необходимо приравнять нулю частные производные:

$$\begin{cases} \dfrac{\partial F}{\partial a} = 2\sum\limits_{i=1}^{n}\left(ax_i + b - y_i\right) = 0; \\ \dfrac{\partial F}{\partial b} = 2\sum\limits_{i=1}^{n}\left(ax_i + b - y_i\right) = 0. \end{cases}$$

Выполнив преобразование, получаем систему двух линейных уравнений относительно a и b:

$$\begin{cases} \left(\sum x^2\right)a + \left(\sum x\right)b = \sum xy; \\ \left(\sum x\right)a + nb = \sum y. \end{cases}$$

Решая эту систему, найдем значения a и b:

$$a = \frac{n\left(\sum xy - \sum x \sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}; \qquad (8.4)$$

$$b = \frac{\left(\sum x^2\right)\left(\sum y\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}. \qquad (8.5)$$

Рассмотрим построение линии линейной регрессии на примере.

Пример 8.4. Исследовали зависимость между содержанием препарата в биологической ткани X и его концентрацией в крови пациента Y. Результаты наблюдений представлены в виде двумерной выборки (табл. 8.8).

Таблица 8.8. Исходные данные для примера 8.4

x_i	1,15	1,9	3,0	5,34	5,4	7,7	7,9	9,03	9,37	10,18
y_i	0,99	0,98	2,6	5,92	4,33	7,68	9,8	9,47	10,64	12,39

Решение. Результаты промежуточных вычислений представлены в табл. 8.9.

Таблица 8.9. Результаты промежуточных вычислений для примера 8.4

i	x_i	y_i	x_i^2	$x_i y_i$	$\bar{y}_i$	$y_i - \bar{y}_i$	$(y_i - \bar{y}_i)^2$
1	1,15	0,99	1,32	1,14	0,27	0,72	0,52
2	1,9	0,98	3,61	1,86	1,21	–0,23	0,05
3	3,0	2,6	9,0	7,80	2,58	0,02	0,00
4	5,34	5,92	28,51	31,61	5,50	0,42	0,18
5	5,4	4,33	29,16	23,38	5,58	–1,25	1,56
6	7,7	7,68	59,29	59,14	8,45	–0,77	0,59
7	7,9	9,8	62,41	77,42	8,71	1,09	1,19
8	9,03	9,47	81,54	85,51	10,12	–0,65	0,42
9	9,37	10,64	87,79	99,69	10,54	0,1	0,01
10	10,18	12,39	103,63	126,13	11,56	0,83	0,69
Σ	60,97	64,8	466,26	513,68		0,28	5,21

Используя формулы (8.4) и (8.5), находим:

$$a = \frac{10 \cdot 513,68 - 60,97 \cdot 64,8}{10 \cdot 466,26 - (60,97)^2} = 1,25;$$

$$b = \frac{466,26 - 64,8 - 60,97 \cdot 513,68}{10 \cdot 466,26 - (60,97)^2} = -1,17.$$

Уравнение линейной регрессии имеет вид:

$$\bar{y}_x = 1,25x - 1,17.$$

Построим корреляционное поле точек (рис. 8.3).

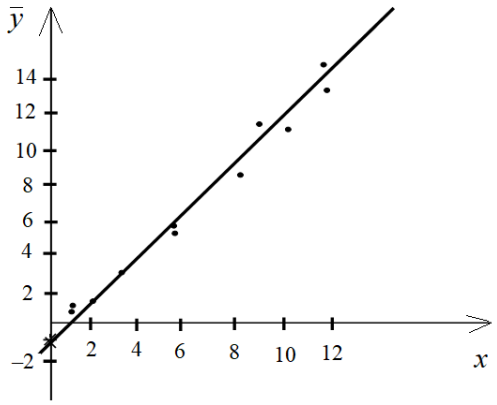


Рис. 8.3. График решения задачи

Рассчитаем должные величины:

- при $x = 0$, $\bar{y}_x = -1,17$;
- при $x = 10$, $\bar{y}_x = 11,33$.

Нанесем линию регрессии на график (см. рис. 8.3).

Из графика видно, что экспериментальные точки достаточно близко расположены к линии регрессии.

Нелинейная регрессия

Если график регрессии $\bar{y} = f(x)$ изображается кривой линией, то это *нелинейная регрессия*.

Вид уравнения регрессии выбирают на основании опыта предыдущих исследований, литературных источников, профессионального мнения и визуального наблюдения расположения точек корреляционного поля. Этот очень важный этап анализа называется *спецификацией*.

Наиболее часто встречаются следующие виды уравнений нелинейной регрессии:

- полиномиальное уравнение:

$$\bar{y} = a_0 + a_1 x + ... + a_n x^n;$$

- уравнение параболы второго порядка:

$$\bar{y} = ax^2 + bx + c;$$

- уравнение параболы третьего порядка:

$$\bar{y} = ax^3 + bx^2 + cx + d;$$

- гиперболическое уравнение:

$$\bar{y} = \frac{a}{x} + b.$$

Для определения неизвестных параметров регрессии используется метод наименьших квадратов.

Пример 8.5. По данным табл. 8.9 исследовать зависимость урожайности зерновых культур *Y* (кг/га) от количества осадков *X* (см), выпавших в вегетационный период (*n* = 15).

Построить корреляционное поле точек и предположить наиболее подходящий вид уравнения регрессии.

Таблица 8.9. Исходные данные для примера 8.5

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	25	27	30	35	36	38	39	41	42	45	46	47	50	52	53
y_i	23	24	27	27	32	31	33	35	34	32	29	28	25	24	25

Решение. Увеличение количества выпавших осадков приведет к увеличению урожайности до некоторого предела, после чего урожайность будет снижаться. Учитывая расположение точек корреляционного поля, можно предположить, что наиболее подходящим уравнением регрессии будет уравнение параболы (рис. 8.4).◁

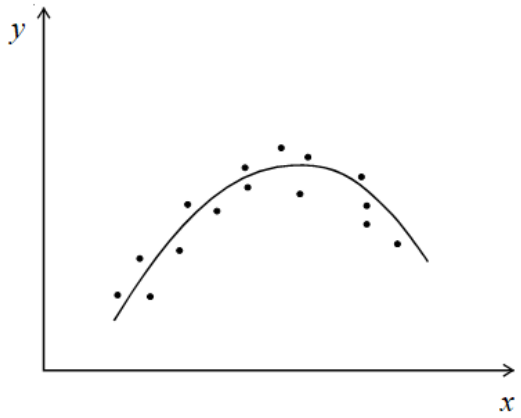


Рис. 8.4. Нелинейная регрессия