Регрессионный анализ

Содержание

Регрессионный анализ. Основные понятия

Это группа методов, направленных на выявление и математическое выражение тех изменений и зависимостей, которые имеют место в системе случайных величин. Если такая система моделирует педагогическую, психологическую, биологическую и т. п., то, следовательно, путем регрессионного анализа выявляются и математически выражаются явления научного эксперимента и зависимости между ними [3–10, 13, 14, 18, 24, 39, 42, 48–50]. Характеристики этих явлений измеряются в разных шкалах, что накладывает ограничения на способы математического выражения изменений и зависимостей, которые изучаются исследователем.

Методы регрессионного анализа рассчитаны, главным образом, на случай устойчивого нормального распределения, в котором изменения от опыта к опыту проявляются лишь в виде независимых испытаний.

Выделяются различные формальные задачи регрессионного анализа. Они могут быть простыми или сложными по формулировкам, по математическим средствам и трудоемкости. Перечислим и рассмотрим на примерах те из них, которые представляются основными.

Первая задача – выявить факт изменчивости изучаемого явления при определенных, но не всегда четко фиксированных условиях.

Вторая задача – выявить тенденцию как периодическое изменение признака.

Третья задача – это выявление закономерности, выраженной в виде корреляционного уравнения (регрессии).

Линейная модель парной регрессии

Регрессия – функция, позволяющая по величине одного коррелируемого признака определить среднюю величину другого признака.

Выделим основные этапы регрессионного анализа. ***Первый этап***. Предположение. На этом этапе происходит выбор формы связи между переменными (модель). ***Второй этап***. Параметризация – происходит оценка значений параметра в выбранной формуле статистической связи. Форма связи (функция) линейная, нелинейная. ***Третий этап.*** Проверка надёжности полученных оценок. На этом этапе осуществляются следующие тесты: *F*-тест (проверка статистической значимости выбранной формы связи), *t*-тест (проверка статистической значимости найденных числовых значений параметра). В результате анализа статистических данных, выбора и построения модели последовательно выполняются все три этапа.

Замечание. На одну переменную, входящую в модель, должно приходиться не менее 6–7 объектов из рассматриваемой выборки.

Рассмотрим простейшую модель регрессии – ***линейную регрессию***. Линейная регрессия для *x* и *y* сводится к нахождению уравнения вида

$$y_x = a + b \cdot x, \text{ или } y = a + b \cdot x + \varepsilon. \qquad (1)$$

Уравнение вида $y_x = a + b \cdot x$ позволяет по заданным значениям фактора x находить теоретические значения результативного признака, подставляя в него фактические значения фактора x.

Построение линейной регрессии сводится к оценке её параметров – a и b. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК), который позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических значений y_x минимальна, то есть:

$$\sum_{i=1}^{n}\left(y_i - y_{x_i}\right)^2 = \sum_{i=1}^{n}\varepsilon_i^2 \;\rightarrow\; \min \cdot$$

Таким образом, из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной (рис. 1):

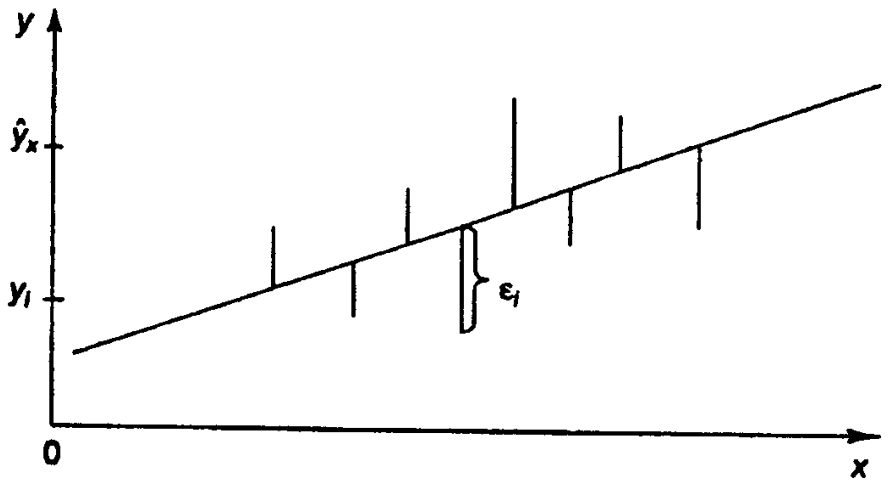

Рис. 1. Линия регрессии с минимальной дисперсией остатков

Как известно из курса математического анализа, чтобы найти минимум функции (1), необходимо вычислить частные производные по каждому из параметров a и b, приравнять их к нулю.

Обозначим $\sum_i \varepsilon_i^2$ через $S(a,\, b)$, тогда: $S(a,\, b) = \sum\left(y - a - b \cdot x\right)^2$.

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum\left(y - a - b \cdot x\right) = 0; \\ \dfrac{\partial S}{\partial b} = -2\sum x\left(y - a - b \cdot x\right) = 0. \end{cases}$$

После несложных преобразований получим следующую систему линейных уравнений для оценки параметров a и b:

$$\begin{cases} a \cdot n \quad + b \cdot \sum x = \sum y; \\ a \cdot \sum x + b \cdot \sum x^2 = \sum x \cdot y. \end{cases} \qquad (2)$$

При решении системы уравнений (2) находят искомые оценки параметров a и b. Здесь можно воспользоваться готовыми формулами, которые следуют непосредственно из решения системы:

$$a = \overline{y} - b \cdot \overline{x}, \quad b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2},$$

где $\operatorname{cov}(x, y) = \overline{y \cdot x} - \overline{y} \cdot \overline{x}$ – ковариация признаков x и y, $\sigma_x^2 = \overline{x^2} - \overline{x}^2$ – дисперсия признака x и $\overline{x} = \frac{1}{n}\sum x$, $\overline{y} = \frac{1}{n}\sum y$, $\overline{y \cdot x} = \frac{1}{n}\sum y \cdot x$, $\overline{x^2} = \frac{1}{n}\sum x^2$.

Параметр b называется коэффициентом регрессии, его величина показывает среднее изменение результата с изменением фактора на одну единицу.

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает линейный коэффициент корреляции $r_{xy} = b \cdot \frac{\sigma(x)}{\sigma(y)}$.

После того как найдено уравнение линейной регрессии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров.

Проверить значимость уравнения регрессии – значит установить, соответствует ли аналитическая модель, выражающая зависимость между переменными, экспериментальным данным, и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

Оценка значимости уравнения регрессии в целом производится на основе *F*-критерия Фишера [48, 49].

Эмпирическое значение критерия Фишера находят по формуле:

$$F_{эмп} = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n - 2). \qquad (3)$$

Критическое значение критерия Фишера находят по статистической одноимённой таблице Приложения: $F_{крит}(\alpha;\ k_1;\ k_2)$ при уровне значимости α и степенях свободы $k_1 = m$ и $k_2 = n - m - 1$ Эмпирическое и критическое значения критерия между собой сравниваются: если $F_{эмп} < F_{крит}$, то на уровне значимости α признаётся статистическая незначимость уравнения регрессии в целом.

В парной линейной регрессии оценивается значимость не только уравнения в целом, но и отдельных его параметров.

t-распределение Стьюдента применяется для ***проверки существенности коэффициента регрессии***. Первоначально определяется стандартная ошибка коэффициента регрессии S_b по формуле:

$$S_b = \sqrt{\frac{S_{ост}^2}{\sum (x - \overline{x})^2}} = \frac{S_{ост}}{\sigma_x \cdot \sqrt{n}},$$

где $S_{ост}^2 = \frac{\sum \left(y - y_x\right)^2}{n - 2}$ – остаточная дисперсия на одну степень свободы.

Для оценки существенности коэффициента регрессии его величина сравнивается с его стандартной ошибкой, то есть определяется эмпирическое значение t-критерия Стьюдента:

$$t_b = \frac{b}{S_b}. \qquad (4)$$

Доверительный интервал для коэффициента регрессии определяется по формуле: $b \pm t_{\text{табл}} \cdot S_b$.

Знак коэффициента регрессии указывает на рост результативного признака y при увеличении признака-фактора x ($b > 0$), уменьшение результативного признака при увеличении признака-фактора ($b < 0$) или его независимость от независимой переменной ($b = 0$).

Поэтому границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, например $-1,5 \le b \le 0,8$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть.

Стандартная ошибка параметра a определяется по формуле:

$$S_a = \sqrt{\frac{S_{\text{ост}}^2 \cdot \sum x^2}{n\sum (x - \bar{x})^2}} = \frac{S_{\text{ост}} \cdot \sqrt{\sum x^2}}{\sigma_x \cdot n}.$$

Процедура оценивания существенности данного параметра не отличается от рассмотренной выше для коэффициента регрессии. Вычисляется t-критерий: $t_a = \frac{a}{S_a}$, его величина сравнивается с табличным значением при $n-2$ степенях свободы.

Значимость линейного коэффициента корреляции проверяется на основе величины ошибки коэффициента корреляции m_r: $S_r = \sqrt{\frac{1-r^2}{n-2}}$. Фактическое значение t-критерия Стьюдента определяется как $t_r = \frac{r}{S_r}$.

Критическое значение критерия Стьюдента находится по одноимённой статистической таблице из Приложения на уровне значимости α и числе степеней свободы $k = n - 2$. Эмпирические и критическое значения критерия между собой сравниваются: если $t_b < t_{крит}$, $t_a < t_{крит}$, $t_r < t_{крит}$, то на уровне значимости 0,05 признаётся статистическая незначимость параметров регрессии и показателя тесноты связи.

Замечание 1. Между t-критерием Стьюдента и F-критерием Фишера существует связь вида: $t_b = t_r = \sqrt{F_{эмп}}$.

Замечание 2. В том случае, когда оба теста выполнятся, построенная модель будет пригодна для дальнейшего анализа и прогнозирования. В прогнозных расчётах по уравнению регрессии определяется предсказываемое y_p значение как точечный прогноз y_x при $x_p = x_k$, то есть путём подстановки в уравнение регрессии $y_x = a + b \cdot x$ соответствующего значения x. Однако точечный прогноз явно не

реален. Поэтому он дополняется расчётом стандартной ошибки y_p и соответственно интервальной оценкой прогнозного значения y_p: $y_p - \Delta_{y_p} \le y_p \le y_p + \Delta_{y_p}$, где $\Delta_{y_p} = m_{y_p} \cdot t_{\text{табл}}$, а m_{y_p} – средняя ошибка прогнозируемого индивидуального значения.

$$m_{y_p} = S_{\text{ост}} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{n \cdot \sigma_x^2}} \cdot$$

Пример. По данным проведённого опроса восьми групп семей известны связи расходов населения на продукты питания *y* (в тыс. руб.) с уровнем доходов семьи *x* (в тыс. руб.) (табл. 1).

Таблица 1

y	0,9	1,2	1,8	2,2	2,6	2,9	3,3	3,8
x	1,2	3,1	5,3	7,4	9,6	11,8	14,5	18,7

Решение. Построим график зависимости между *x* и *y* (рис. 2).

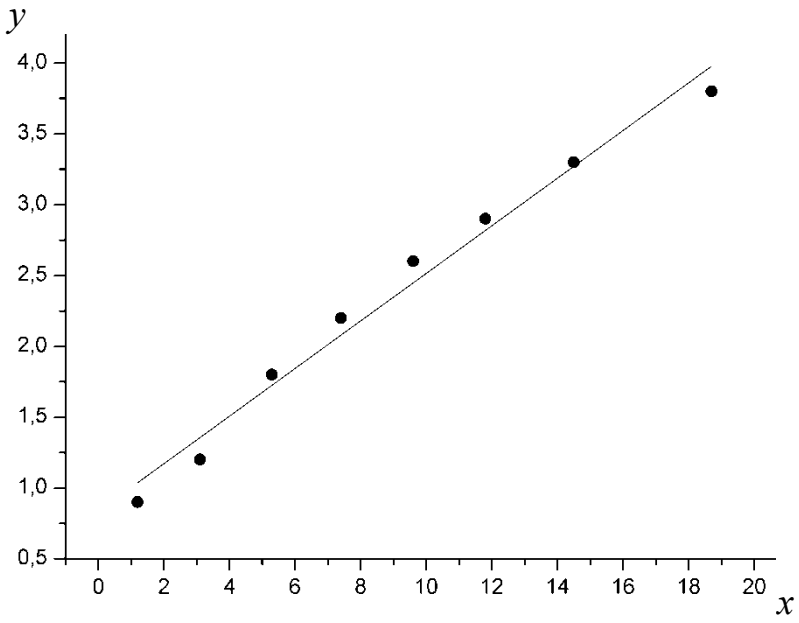

Рис. 2

Предположим, что связь между доходами семьи и расходами на продукты питания линейная, и составим таблицу 2:

Таблица 2

№	*x*	*y*	*xy*	x^2	y^2
1	1,2	0,9	1,08	1,44	0,81
2	3,1	1,2	3,72	9,61	1,44
3	5,3	1,8	9,54	28,09	3,24
4	7,4	2,2	16,28	54,76	4,84
5	9,6	2,6	24,96	92,16	6,76
6	11,8	2,9	34,22	139,24	8,41
7	14,5	3,3	47,85	210,25	10,89

№	x	y	xy	x^2	y^2
8	18,7	3,8	71,06	349,69	14,44
Сумма	71,6	18,7	208,71	885,24	50,83
Среднее значение	8,95	2,34	26,09	110,66	6,35
σ	5,53	0,935	–	–	–
σ^2	30,56	0,874	–	–	–

Рассчитаем параметры линейного уравнения парной регрессии, для этого воспользуемся формулами:

$$b = \frac{\operatorname{cov}(x,\, y)}{\sigma_x^2} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{26,09 - 8,95 \cdot 2,34}{30,56} = 0,168;$$

$$a = \overline{y} - b \cdot \overline{x} = 2,34 - 0,168 \cdot 8,95 = 0,836.$$

Получаем линейное уравнение вида: $y_x = 0,836 + 0,168 \cdot x$, то есть с увеличением дохода семьи на 1000 руб. расходы на питание увеличиваются на 168 руб.

Найдём показатель тесноты связи – линейный коэффициент корреляции r_{xy}:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 0,168 \cdot \frac{5,53}{0,935} = 0,994 \cdot$$

Близость коэффициента корреляции к 1 указывает на тесную линейную связь между признаками.

Коэффициент детерминации $r_{xy}^2 = 0,987$ показывает, что уравнением регрессии объясняется 98,7% дисперсии результативного признака, а на долю прочих факторов приходится лишь 1,3%.

Оценим качество уравнения регрессии в целом с помощью F-критерия Фишера, найдём эмпирическое значение критерия $F_{эмп}$ по формуле (3):

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n - 2) = \frac{0,987}{1 - 0,987} \cdot 6 = 455,54 \cdot$$

Критическое значение критерия Фишера для $k_1 = 1$, $k_2 = n - 2 = 6$, $\alpha = 0,05$ равно $F_{крит} = 5,99$. Таким образом, $F_{эмп} > F_{крит}$, так как 455,54 > 5,99, и на уровне значимости 0,05 признается статистическая значимость уравнения в целом.

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитаем t-критерий Стьюдента и доверительные интервалы каждого из показателей. Рассчитаем случайные ошибки параметров линейной регрессии и коэффициента корреляции:

$$\left(S_{\text{ост}}^2 = \frac{\sum \left(y - y_x \right)^2}{n - 2} = \frac{0,1257}{8 - 2} = 0,021 \right),\ S_b = \frac{S_{\text{ост}}}{\sigma_x \cdot \sqrt{n}} = \frac{\sqrt{0,021}}{5,53 \cdot \sqrt{8}} = 0,0093,$$

$$S_a = \frac{S_{\text{ост}} \cdot \sqrt{\sum x^2}}{\sigma_x \cdot n} = \frac{\sqrt{0,021 \cdot 885,24}}{5,53 \cdot 8} = 0,0975,\ S_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0,987}{6}} = 0,0465 \cdot$$

Эмпирическое значения t-критерия: $t_b = \frac{0,168}{0,0093} = 18,065$, $t_a = \frac{0,836}{0,0975} = 8,574$, $t_r = \frac{0,994}{0,0465} = 21,376$. Критическое значение t-критерия Стьюдента при $\alpha = 0,05$ и числе степеней свободы $n - 2 = 6$ равно $t_{крит} = 2,447$.

Таким образом, $t_b > t_{крит}$, $t_a > t_{крит}$, $t_r > t_{крит}$, и на уровне значимости 0,05 признаём статистическую значимость параметров регрессии и показателя тесноты связи. Рассчитаем доверительные интервалы для параметров регрессии a и b. Получим, что $a \in (0,597; 1,075)$ и $b \in (0,145; 0,191)$.

На основании того, что оба теста выполнились, можем по построенной модели осуществлять анализ и прогнозирование. Найдём прогнозное значение результативного фактора y_p при значении признака-фактора, составляющем 110% от среднего уровня $x_p = 1,1 \cdot \bar{x} = 1,1 \cdot 8,95 = 9,845$, то есть найдём расходы на питание, если доходы семьи составят 9,85 тыс. руб.

$$y_p = 0,836 + 0,168 \cdot 9,845 = 2,490 \text{ (тыс. руб.)}$$

Значит, если доходы семьи составят 9845 руб., то расходы на питание будут 2490 руб.

Найдем доверительный интервал прогноза. Ошибка прогноза составляет:

$$m_{y_p} = S_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{n \cdot \sigma_x^2}} = \sqrt{0,021 \cdot \left(1 + \frac{1}{8} + \frac{\left(9,845 - 8,95\right)^2}{8 \cdot 30,56}\right)} = 0,154,$$

а доверительный интервал ($y_p - \Delta_{y_p} \le y_p \le y_p + \Delta_{y_p}$): $2,113 < y_p < 2,867$, то есть прогноз является статистически надёжным.

Пример. Изучалась зависимость между массой матерей x_i, измеряемой в начале беременности (кг), и массой новорождённых детёнышей y_i (кг) (табл. 3).

Таблица 3

i	1	2	3	4	5	6	7	8	9	10
x_i	10	10	10,1	10,2	10,8	11	11,1	11,3	11,3	11,4
y_i	0,7	0,7	0,65	0,61	0,73	0,65	0,65	0,75	0,7	0,7

i	11	12	13	14	15	16	17	18	19	20
x_i	11,8	12	12	12,1	12,3	13	13,4	13,5	14,5	15,6
y_i	0,69	0,72	0,6	0,75	0,63	0,8	0,78	0,7	0,7	0,85

Решение. Здесь под независимой переменной x будем понимать массу матерей, а под зависимой переменной y – массу новорожденных детёнышей.

Для расчёта необходимых сумм и произведений составим вспомогательную таблицу 4.

Таблица 4

Масса матерей, x_i	Масса детёнышей, y_i	$x_i \cdot y_i$	x_i^2	y_i^2
10	0,7	7	100	0,49
10	0,7	7	100	0,49
10,1	0,65	6,565	102,01	0,4225
10,2	0,61	6,222	104,04	0,3721
10,8	0,73	7,884	116,64	0,5329
11	0,65	7,15	121	0,4225
11,1	0,65	7,215	123,21	0,4225
11,3	0,75	8,475	127,69	0,5625
11,3	0,7	7,91	127,69	0,49
11,4	0,7	7,98	129,96	0,49
11,8	0,69	8,142	139,24	0,4761
12	0,72	8,64	144	0,5184
12	0,6	7,2	144	0,36
12,1	0,75	9,075	146,41	0,5625
12,3	0,63	7,749	151,29	0,3969
13	0,8	10,4	169	0,64
13,4	0,78	10,452	179,56	0,6084
13,5	0,7	9,45	182,25	0,49
14,5	0,7	10,15	210,25	0,49
15,6	0,85	13,26	243,36	0,7225
Сумма 237,4	14,06	167,919	2861,6	9,9598

Определим параметры линейной регрессии, используя формулу (1) и ранее найденное значение коэффициента корреляции.

$$b = r_{xy} \cdot \frac{\sigma_y}{\sigma_x} = 0{,}565 \cdot \frac{\sqrt{0{,}003781}}{\sqrt{2{,}1831}} = 0{,}0235.$$

$$a = \bar{y} - k_1 \cdot \bar{x} = \frac{14{,}06}{20} - 0{,}0235 \cdot \frac{237{,}4}{20} = 0{,}424.$$

Следовательно, фактическое уравнение регрессии массы детёнышей (шимпанзе) по значениям массы их матерей имеет вид $\hat{y}_x = 0{,}024 \cdot x + 0{,}424$, то есть при увеличении массы матери на 1 кг у детёныша ожидается увеличение массы на 0,024 кг.

В свою очередь найдём квадрат коэффициента корреляции: $R^2 = (r_{xy})^2 = 0{,}565^2 = 0{,}319$. Коэффициент детерминации показывает, что вариация массы новорождённых детёнышей на 31,9% обусловлена изменчивостью массы матерей.

Оценим качество уравнения регрессии в целом с помощью F-критерия Фишера, найдём эмпирическое значение критерия $F_{эмп}$ по формуле (3):

$$F = \frac{r_{xy}^2}{1-r_{xy}^2}\cdot(n-2) = \frac{0,987}{1-0,987}\cdot 6 = 455,54\cdot$$

Критическое значение критерия Фишера для $k_1 = 1$, $k_2 = n - 2 = 18$, $\alpha = 0,05$ равно $F_{крит} = 4,41$. Таким образом, $F_{эмп} > F_{крит}$, так как 455,54 > 4,41, и на уровне значимости 0,05 признаётся статистическая значимость уравнения в целом.

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитаем t-критерий Стьюдента: $t_b = \sqrt{F_{эмп}} = \sqrt{455,54} = 21,34$. Для уровня значимости α=0,05 найдём критическое значение критерия Стьюдента из Приложения: $t_{крит} = t(\alpha; k) = t(0,05; 18) = 2,10$.

Таким образом, $t_b \geq t_{крит}$, так как 21,34 > 2,1, и на уровне значимости 0,05 делаем вывод о статистической значимости показателя, стоящего перед x.

В силу того, что значение коэффициента корреляции, уравнение регрессии и параметр при x статистически значимы, по найденному уравнению регрессии можем делать статистические прогнозы, так, если масса самки шимпанзе равна 15 кг, то ожидаемая масса новорождённого детёныша будет равна $\hat{y}_x = 0,024\cdot 15 + 0,424 = 0,784$ кг.

Пример. Исследуется зависимость между доходом и размерами помещичьего хозяйства в России на рубеже XIX–XX вв. по сведениям о размерах (в десятинах) и доходах (в тыс. руб.) десяти помещичьих имений

Исходные данные (x – размеры имения в десятинах, y – доход имения в тыс. руб.) Найти уравнение линейной регрессии, описывающее корреляционную связь между размерами и доходом помещичьего имения (табл. 5).

Таблица 5

№	x	y	$x\cdot y$	x^2	y^2
1	240	1,5	360	57600	2,25
2	255	1,25	318,75	65025	1,5625
3	265	1,55	410,75	70225	2,4025
4	270	1,4	378	72900	1,96
5	285	1,45	413,25	81225	2,1025
6	295	1,6	472	87025	2,56
7	310	1,8	558	96100	3,24
8	320	1,8	576	102400	3,24
9	325	1,85	601,25	105625	3,4225
10	330	1,9	627	108900	3,61
Сумма	2895	16,1	4715	847025	26,35
Средние	289,5	1,61	471,5	84702,5	2,635

Решение. Вычислим параметры a и b по формулам:

$$b = \frac{471,5 - 289,5\cdot 1,61}{\sqrt{(84702,5 - 289,5^2)\cdot(2,635 - 1,61^2)}} = 0,006,\ a = 1,61 - 0,006\cdot 289,5 = 0,144.$$

Уравнение линейной регрессии примет вид: y=0,006·x–0,144. Коэффициент регрессии в этом уравнении, равный 0,006, означает, что при возрастании размеров

имения на единицу, то есть на 1 десятину, доход имения возрастает на 0,006 тыс. рублей, или на 6 рублей. С помощью уравнения регрессии можно предсказать примерный доход имения любых размеров.

Графическая интерпретация регрессии показывает тенденцию в изменении дохода имения в зависимости от его размеров. Здесь была рассмотрена наиболее простая форма связи между двумя признаками – линейная. Между тем, во-первых, зависимости между признаками могут принимать самые разнообразные формы, а, во-вторых, при более полном анализе взаимосвязей необходимо учитывать, что на результативный признак обычно влияет не один фактор, а несколько. Выявить форму связи между результативным признаком и несколькими факторными признаками позволяет множественный регрессионный анализ.

Нелинейные модели парной регрессии

Если между изучаемыми явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций.

Различают два класса нелинейных регрессий:

1. Регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, например:

- полиномы различных степеней $y_x = a + b \cdot x + c \cdot x^2$, $y_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$;
- равносторонняя гипербола $y_x = a + b/x$;
- полулогарифмическая функция $y_x = a + b \cdot \ln x$.

2. Регрессии, нелинейные по оцениваемым параметрам, например:

- степенная $y_x = a \cdot x^b$;
- показательная $y_x = a \cdot b^x$;
- экспоненциальная $y_x = \mathrm{e}^{a+b \cdot x}$.

Регрессии, нелинейные по включенным переменным, приводятся к линейному виду простой заменой переменных, а дальнейшая оценка параметров производится с помощью метода наименьших квадратов. Рассмотрим некоторые функции.

Парабола второй степени $y_x = a + b \cdot x + c \cdot x^2$ приводится к линейному виду с помощью замены: $x = x_1$, $x^2 = x_2$. Парабола второй степени обычно применяется в случаях, когда для определённого интервала значений фактора меняется характер связи рассматриваемых признаков: прямая связь меняется на обратную, или обратная – на прямую.

Равносторонняя гипербола $\hat{y}_x = a + b \cdot \frac{1}{x}$ может быть использована для характеристики связи, например кривые Филлипса, Энгеля и в других случаях.

Гипербола приводится к линейному виду простой заменой: $z = \frac{1}{x}$; $y_x = a + b \cdot \ln x$ приводится к линейному виду заменой: $z = \ln x$; $y_x = a + b \cdot \sqrt{x} - z = \sqrt{x}$.

Несколько иначе обстоит дело с регрессиями, нелинейными по оцениваемым параметрам, которые делятся на два типа: нелинейные модели внутренне линейные (приводятся к линейному виду с помощью соответствующих преобразований, например, логарифмированием) и нелинейные модели внутренне нелинейные (к линейному виду не приводятся).

К внутренне линейным моделям относятся, например, степенная функция – $y_x = a \cdot x^b$, показательная – $y_x = a \cdot b^x$, экспоненциальная – $y_x = \mathrm{e}^{a+b \cdot x}$, логистическая – $y_x = \frac{a}{1 + b \cdot \mathrm{e}^{-c \cdot x}}$, обратная – $y_x = \frac{1}{a + b \cdot x}$.

К внутренне нелинейным моделям можно, например, отнести следующие модели: $y_x = a + b \cdot x^c$, $y_x = a \cdot \left(1 - \frac{1}{1 - x^b}\right)$.

Покажем приведение к линейному виду путём логарифмирования на примере уравнения $\hat{y}_x = a \cdot b^x$ и $\hat{y}_x = a \cdot e^{bx}$ (эти уравнения применяются в том случае, когда основная тенденция ряда следует или оказывается близкой к закону геометрической прогрессии).

$\hat{y}_x = a \cdot b^x$	$\hat{y}_x = a \cdot e^{bx}$
$\lg y = \lg(a \cdot b^x) =$ $= \lg a + \lg b^x = \lg a + x \lg b$	$\ln y = \ln a e^{bx} = \ln a + \ln e^{bx} =$ $= \ln a + bx \ln e = \ln a + bx$
$\ln y = \ln a + x \ln b$	$\lg y = \lg a + bx \lg e = \lg a + 0{,}43 \cdot bx$

После выбора аналитической формы связи определяют значения коэффициентов уравнения регрессии, для чего решают специальные системы уравнений (табл. 6).

Таблица 6

Форма связи	Уравнение регрессии	Система уравнений
Линейная	$\hat{y}_x = a + bx$	$\begin{cases} a \cdot n + b \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i \cdot y_i) \end{cases}$
Парабола	$\hat{y}_x = a + bx + cx^2$	$\begin{cases} a \cdot n + b \cdot \sum_{i=1}^{n} x_i + c \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \\ a \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} x_i^2 + c \cdot \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} (x_i \cdot y_i) \\ a \cdot \sum_{i=1}^{n} x_i^2 + b \cdot \sum_{i=1}^{n} x_i^3 + c \cdot \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} (x_i^2 \cdot y_i) \end{cases}$

Форма связи	Уравнение регрессии	Система уравнений
Гипербола	$\hat{y}_x = a + \frac{b}{x}$	$\begin{cases} a \cdot n + b \cdot \sum\limits_{i=1}^{n} \frac{1}{x_i} = \sum\limits_{i=1}^{n} y_i \\ a \cdot \sum\limits_{i=1}^{n} \frac{1}{x_i} + b \cdot \sum\limits_{i=1}^{n} \frac{1}{x_i^2} = \sum\limits_{i=1}^{n} \frac{y_i}{x_i} \end{cases}$
Степенная	$\hat{y}_x = a \cdot x^b$	$\begin{cases} n \cdot \lg a + b \cdot \sum\limits_{i=1}^{n} \lg x_i = \sum\limits_{i=1}^{n} (\lg y_i) \\ \lg a \cdot \sum\limits_{i=1}^{n} \lg x_i + b \cdot \sum\limits_{i=1}^{n} (\lg x_i)^2 = \sum\limits_{i=1}^{n} (\lg x_i \cdot \lg y_i) \end{cases}$
Экспоненциальная	$\hat{y}_x = e^{a+bx}$	$\begin{cases} n \cdot a + b \cdot \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} (\lg y_i) \\ a \cdot \sum\limits_{i=1}^{n} x_i + b \cdot \sum\limits_{i=1}^{n} (x_i)^2 = \sum\limits_{i=1}^{n} (x_i \cdot \lg y_i) \end{cases}$
Показательная	$\hat{y}_x = a \cdot b^x$	$\begin{cases} n \cdot \lg a + \lg b \cdot \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} (\lg y_i) \\ \lg a \cdot \sum\limits_{i=1}^{n} x_i + \lg b \cdot \sum\limits_{i=1}^{n} (x_i)^2 = \sum\limits_{i=1}^{n} (x_i \cdot \lg y_i) \end{cases}$
Полулогарифмическая	$\hat{y}_x = a + b \cdot \lg x$	$\begin{cases} a \cdot n + b \cdot \sum\limits_{i=1}^{n} \lg x_i = \sum\limits_{i=1}^{n} y_i \\ a \cdot \sum\limits_{i=1}^{n} \lg x_i + b \cdot \sum\limits_{i=1}^{n} (\lg x_i)^2 = \sum\limits_{i=1}^{n} (y_i \cdot \lg x_i) \end{cases}$

Замечание 1. Переменная $\hat{y}_x$ обозначает теоретическое значение фактического значения результирующего признака *y*.

Замечание 2. Для оценки тесноты связи нелинейной регрессии находят индекс корреляции:

$$\rho_{xy} = \sqrt{\frac{\sum\limits_{i=1}^{n} (\hat{y}_{x_i} - \bar{y})^2}{\sum\limits_{i=1}^{n} (y_i - \bar{y})^2}}. \qquad (5)$$

Пример. Исследовалась зависимость урожайности зерновых культур *y* (ц/га) от количества осадков *x* (см), выпавших в вегетационный период (в период роста и развития растений) (табл. 7).

Таблица 7

x	25	27	30	35	36	38	39	41	42	45	46	47	50	52	53
y	23	24	27	27	32	31	33	35	34	32	29	28	25	24	25

Решение. Построим поле корреляции (рис. 3).

Рис. 3

Учитывая расположение точек корреляционного поля, можем предположить, что наиболее подходящим уравнением регрессии будет уравнение параболы $\hat{y}_x = a + bx + cx^2$, его параметры находят из решения следующей системы:

$$\begin{cases} a \cdot n + b \cdot \sum_{i=1}^{n} x_i + c \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i, \\ a \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} x_i^2 + c \cdot \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} (x_i \cdot y_i), \\ a \cdot \sum_{i=1}^{n} x_i^2 + b \cdot \sum_{i=1}^{n} x_i^3 + c \cdot \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} (x_i^2 \cdot y_i). \end{cases}$$

Для расчёта необходимых сумм и произведений составим вспомогательную таблицу 8:

Таблица 8

x	y	x^2	x^3	x^4	$x\,y$	$y\,x^2$
25	23	625	15625	390625	575	14375
27	24	729	19683	531441	648	17496
30	27	900	27000	810000	810	24300
35	27	1225	42875	1500625	945	33075
36	32	1296	46656	1679616	1152	41472
38	31	1444	54872	2085136	1178	44764
39	33	1521	59319	2313441	1287	50193
41	35	1681	68921	2825761	1435	58835
42	34	1764	74088	3111696	1428	59976
45	32	2025	91125	4100625	1440	64800
46	29	2116	97336	4477456	1334	61364
47	28	2209	103823	4879681	1316	61852
50	25	2500	125000	6250000	1250	62500
52	24	2704	140608	7311616	1248	64896
53	25	2809	148877	7890481	1325	70225
Сумма 606	429	25548	1115808	50158200	17371	730123

Теперь система примет вид:

$$\begin{cases} a \cdot 15 + b \cdot 606 + c \cdot 25548 = 429, \\ a \cdot 606 + b \cdot 25548 + c \cdot 1115808 = 17371, \\ a \cdot 25548 + b \cdot 1115808 + c \cdot 50158200 = 730123. \end{cases}$$

В результате решения этой системы, получим значение a= –43,932; b=3,834 и c= –0,048, то есть уравнение регрессии будет иметь вид:

$$\hat{y}_x = -43{,}93 + 3{,}834 \cdot x - 0{,}048 \cdot x^2.$$

Для оценки тесноты связи вычислим индекс корреляции. Для этого составим следующую вспомогательную табл. 2, при нахождении $\hat{y}_x$ будем подставлять конкретные значения x_i (с учётом найденного уравнения регрессии) (табл. 9):

Таблица 9

i	x_i	y_i	$\hat{y}_x$	$(\hat{y}_x - \bar{y})^2$	$(y_i - \bar{y})^2$
1	25	23	21,7	47,619	31,36
2	27	24	24,34	18,163	21,16
3	30	27	27,57	1,0586	2,56
4	35	27	31,02	5,8795	2,56
5	36	32	31,43	7,9826	11,56
6	38	31	31,94	11,131	5,76
7	39	33	32,05	11,88	19,36
8	41	35	31,98	11,407	40,96
9	42	34	31,8	10,225	29,16
10	45	32	30,68	4,318	11,56
11	46	29	30,11	2,2841	0,16
12	47	28	29,45	0,719	0,36
13	50	25	26,88	2,967	12,96
14	52	24	24,68	15,364	21,16
15	53	25	23,44	26,661	12,96
Сумма	606	429	429	177,66	223,6

Тогда индекс корреляции:

$$\rho_{xy} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_{x_i} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} = \sqrt{\frac{77{,}66}{223{,}6}} = 0{,}891.$$

Таким образом, полученная зависимость весьма тесная.

В свою очередь коэффициент детерминации $R^2 = \left(\rho_{xy}\right)^2 = 0{,}891^2 \approx 0{,}79$ показывает, что вариация урожайности зерновых культур на 79% обусловлена регрессией, или изменчивостью количества выпавших в вегетационный период осадков.

Эмпирическое значение критерия Фишера найдём по формуле:

$$F_{эмп} = \frac{\rho_{xy}^2}{1 - \rho_{xy}^2} \cdot (n-2) = \frac{0{,}79}{1 - 0{,}79} \cdot 18 = 67{,}71.$$

Критическое значение критерия Фишера находят по статистической одноимённой таблице Приложения: $F_{крит}(0{,}05;\ k_1 = 1;\ k_2 = 18) = 4{,}41$.

Таким образом, $F_{эмп} > F_{крит}$, так как 67,71 > 4,41, и на уровне значимости 0,05 признаётся статистическая значимость уравнения регрессии в целом.

Литература

1. Гласс Дж. Статистические методы в педагогике и психологии / Дж. Гласс, Дж. Стенли. – М.: Прогресс, 1976. – 496 с.

2. Гланц С. Медико-биологическая статистика / С. Гланц. - М.: Практика, 1998. - 459 с.

3. Гмурман В. Е. Теория вероятностей и математическая статистика: учебное пособие для вузов / В. Е. Гмурман. – М.: Высш. шк., 2003. – 479 с.

4. Новиков Д. А. Статистические методы в педагогических исследованиях (типовые случаи) / Д. А. Новиков. – М.: МЗ-Пресс, 2004. – 67 с.

5. Новиков Д. А. Статистические методы в медико-биологическом эксперименте (типовые случаи) / Д. А. Новиков, В. В. Новочадов. – Волгоград: Изд-во ВГМУ, 2005. – 84 с.

6. Шилова З. В. Теория вероятностей и математическая статистика: учебное пособие / З. В. Шилова, О. И. Шилов. – Киров: Изд-во ВГГУ, 2015. – 158 с.