Корреляционный и регрессионный анализы

Между различными признаками в природе существуют определенные взаимосвязи. Знание этих связей, зависимости одного признака от другого важно и для агрономической практики. Выращивание сельскохозяйственных культур предполагает знание связей между продуктивностью и обеспеченностью растений элементами питания, теплом и т. п.

Для описания связей между переменными величинами (признаками) применяют понятие функции *f*, которая ставит в соответствие каждому определенному значению независимой переменной *X* определенное значение зависимой переменной *Y*. Такие однозначные связи называются функциональными ($Y = f(x)$). Но в природе они встречаются не всегда. В биологических, сельскохозяйственных науках чаще встречаются такие соотношения между переменными, когда каждому значению признака *X* соответствует не одно, а множество значений *Y*. Например, колосья пшеницы одной и той же длины могут содержать различное число зерен. Причиной такого варьирования является тот факт, что каждый биологический признак представляет собой функцию многих переменных: на него влияют и генетические, и средовые факторы. Поэтому зависимость между такими признаками имеет не функциональный, а стохастический характер. Эти связи обнаруживаются при массовом изучении признаков и называются корреляционными, или корреляцией.

Так как при корреляции разным значениям одной переменной соответствуют различные распределения другой переменной, то форма стохастической связи может быть описана не как зависимость отдельных значений *Y* от величины *X*, а как зависимость частных средних $\overline{Y}_x$ от значений *X*. Изменение функций в зависимости от определенного изменения значений одного или нескольких аргументов называется регрессией. Описанию корреляционных связей служит корреляционно-регрессионный анализ.

Корреляции подразделяют по направлению, форме и числу связей.

По числу связей корреляция бывает простой (зависимость между двумя признаками) и множественной (три и более), по форме – прямолинейной и криволинейной, по направлению – прямой и обратной.

Под прямолинейной корреляцией понимают такую зависимость, которая носит линейный характер и выражается уравнением прямой линии $Y = a + b \cdot X$. Когда при одинаковых приращениях аргумента функция имеет неодинаковые изменения, корреляция называется криволинейной. Если при увеличении аргумента функция возрастает, то корреляция называется положительной или прямой, а если убывает – отрицательной или обратной.

В качестве числового показателя простой линейной корреляции, отражающего тесноту (силу) и направление связи, используют отвлеченное безразмерное число, называемое коэффициентом корреляции и обозначаемое буквой *r*.

Для анализа линейной корреляции между *X* и *Y* проводят *n* независимых парных наблюдений, исходом каждого из которых является пара чисел (X_1; Y_1), (X_2; Y_2), …, (X_n; Y_n). Технику вычисления коэффициента корреляции рассмотрим на примере (табл. 1).

Коэффициент корреляции вычисляют по формуле

$$r = \frac{\sum (X - \bar{x})(Y - \bar{y})}{\sqrt{\sum (X - \bar{x})^2 \cdot \sum (Y - \bar{y})^2}} = \frac{5{,}61}{\sqrt{0{,}3778 \cdot 136}} = \frac{5{,}61}{\sqrt{51{,}3808}} = 0{,}783$$

или, минуя вычисления отклонений и квадратов отклонений, – по формуле

$$r = \frac{\sum XY - (\sum X \cdot \sum Y) \cdot n}{\sqrt{(\sum X^2 - (\sum X)^2 : n)\,(\sum Y^2 - (\sum Y)^2 : n)}},$$

или через средние квадратические отклонения

$$r = \frac{\sum (X - \bar{x})(Y - \bar{y})}{n \cdot S_{\bar{x}} \cdot S_{\bar{y}}},$$

$$S_{\bar{x}} = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}, \qquad S_{\bar{y}} = \sqrt{\frac{\sum (Y - \bar{y})^2}{n - 1}},$$

где $(X - \bar{x})$ и $(Y - \bar{y})$ – отклонения значений X и Y от своих средних значений $\bar{x}$ и $\bar{y}$ в n сопоставимых парах.

Таблица 1. **Вычисление коэффициента корреляции между числом зерен в колосе (*X*) и продуктивностью растений озимой ржи сорта Ясельда (*Y*)**

№ растения	Значение признаков		Отклонение от средней		Квадраты отклонений		Произведение отклонений
	Число зерен в колосе X, шт.	Продуктивность растений Y, г/раст.	$X - \bar{x}$	$Y - \bar{y}$	$(X - x)^2$	$(Y - \bar{y})^2$	$(X - \bar{x}) \cdot (Y - \bar{y})$
1	38	1,74	–3	–0,13	9	0,0169	0,39
2	46	2,06	5	0,19	25	0,0361	0,95
3	38	1,75	–3	–0,12	9	0,0144	0,36
4	42	2,00	1	0,13	1	0,0169	0,13
5	38	1,53	–3	–0,34	9	0,1156	1,02
6	44	1,78	3	–0,09	9	0,0081	–0,27
7	38	1,77	–3	–0,10	9	0,0100	0,30
8	37	1,80	–4	–0,07	16	0,0049	0,28
9	48	2,22	7	0,35	49	0,1225	2,45
10	41	2,05	0	0,18	0	0,0324	0
Сумма	$\sum X = 410$	$\sum Y = 18{,}70$	$\sum (X - \bar{x}) = 0$	$\sum (Y - \bar{y}) = 0$	$\sum (X - \bar{x})^2 == 136$	$\sum (Y - \bar{y}) =^2 = 0{,}3778$	$\sum (X - \bar{x}) \cdot (Y - \bar{y}) = = 5{,}61$
Среднее	$\bar{x} = \frac{\sum X}{n} = = \frac{410}{10} = 41$	$\bar{y} = \frac{\sum Y}{n} = = \frac{18{,}7}{10} = 1{,}8$	–	–	–	–	–

Примечание: $n = 10$.

Значения коэффициента корреляции могут находиться в пределах от +1 при прямой функциональной связи до –1 при обратной функциональной связи. При полном отсутствии

корреляции $r = 0$, при $r < \pm 0{,}3$ корреляционная зависимость слабая, при $r = \pm 0{,}3 \div 0{,}7$ – средняя, а при $r > \pm 0{,}7$ – сильная.

Знак при коэффициенте корреляции указывает направление связи: «+» – прямая зависимость; «–» – связь обратная.

Таким образом, связь между числом зерен в колосьях и продуктивностью растений озимой ржи сорта Ясельда сильная ($r = 0{,}783$).

Степень связи между признаками более точно измеряется коэффициентом детерминации d_{yx}, равным квадрату коэффициента корреляции:

$$d_{yx} = r^2.$$

Он показывает долю (%) тех изменений, которые зависят от изучаемого фактора. В нашем примере $d_{yx} = 0{,}783^2 = 0{,}613$. Таким образом, только 61,3 % изменчивости признака Y (продуктивность растений озимой ржи сорта Ясельда) обусловлено действием факториального признака X (числом зерен в колосе), остальная часть корреляционной связи ($1 – 0{,}613 = 0{,}387$) обусловлена другими факторами.

Коэффициент корреляции выборочных наблюдений подвержен случайным колебаниям, которые зависят от объема выборки и точности проведения наблюдений. Поэтому для оценки надежности выборочного коэффициента корреляции вычисляют его ошибку и критерий существенности.

Стандартную ошибку коэффициента корреляции определяют по формуле

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0{,}783^2}{10 - 2}} = 0{,}220,$$

где S_r – ошибка коэффициента корреляции;

r – коэффициент корреляции;

n – число пар значений выборки.

Чем больше число наблюдений, тем меньше будет ошибка коэффициента корреляции. Значение коэффициента корреляции обычно записывается вместе с его ошибкой:

$$r \pm S_r = 0{,}783 \pm 0{,}220.$$

Критерий существенности коэффициента корреляции вычисляют по формуле

$$t_r = \frac{r}{S_r} = \frac{0{,}783}{0{,}220} = 3{,}56.$$

Сопоставляя фактические и теоретически рассчитанные значения t_r при числе степеней свободы, равном $n – 2$, оценивают существенность корреляционной связи. Если $t_{r_{\text{факт}}} \geq t_{r_{\text{теор}}}$, то корреляционная связь существенна, а при $t_{r_{\text{факт}}} < t_{r_{\text{теор}}}$ – несущественна.

Теоретическое значение критерия находят по таблице Стьюдента (прил. 1), принимая 5%-ные или 1%-ные уровни значимости. Так, t_{05} равен значению 2,45 при $n – 2 = 8$ степенях свободы. Значит, коэффициент корреляции в нашем случае существен.

При малых выборках и значениях *r*, близких к единице, распределение выборочных коэффициентов корреляции заведомо отличается от нормального и оценка существенности коэффициента корреляции по критерию Стьюдента становится ненадежной. Р. Фишер в этих случаях предлагает преобразовывать коэффициент корреляции в величину *z*, используя специальные таблицы. Тогда

$$S_z = \frac{1}{\sqrt{n-3}};$$

$$t_r = \frac{z}{S_z};$$

$$z \pm t \cdot S_z.$$

Определив коэффициент корреляции, мы выясняем направление и степень сопряженности в изменчивости признаков. Однако он не позволяет узнать, как количественно изменяется результативный признак (в нашем примере продуктивность растений) при изменении факториального (в нашем примере число зерен в колосе) на единицу измерения (т. е. на 1 шт.). Это решается с помощью регрессионного анализа. Его основная задача – определить формулу корреляционной зависимости. Различают регрессию простую и множественную, а по форме – прямо- и криволинейную.

Сущность регрессионного анализа состоит в том, чтобы построить линию, которая наиболее точно выражала бы зависимость одного признака от другого.

Зависимость между признаками может быть выражена коэффициентом регрессии, показывающим, в каком направлении и на какую величину изменяется в среднем один признак *Y* (функция) при изменении другого *X* (аргумент) на единицу измерения. Коэффициентов регрессии столько, сколько признаков, они имеют знак коэффициента корреляции и вычисляются по следующим формулам:

$$b_{yx} = \frac{\sum(X-\bar{x})(Y-\bar{y})}{\sum(X-\bar{x})^2} = \frac{5{,}61}{136} = 0{,}0413;$$

$$b_{xy} = \frac{\sum(X-\bar{x})\,(Y-\bar{y})}{\sum(Y-\bar{y})^2} = \frac{5{,}61}{0{,}3778} = 14{,}8491.$$

Их можно вычислить и через средние квадратические отклонения:

$$b_{yx} = r \cdot \frac{S_{\bar{y}}}{S_{\bar{x}}};$$

$$b_{xy} = r \cdot \frac{S_{\bar{x}}}{S_{\bar{y}}}.$$

Произведение коэффициентов регрессии равно квадрату коэффициента корреляции:

$$b_{yx} \cdot b_{xy} = r^2.$$

$$r = \sqrt{0{,}04125 \cdot 14{,}8491} = \sqrt{0{,}6125} = 0{,}783.$$

Ошибку коэффициентов регрессии вычисляют по формулам

$$S_{b_{yx}} = S_t \cdot \sqrt{\frac{\sum (Y - \bar{y})^2}{\sum (X - \bar{x})^2}} = 0{,}220 \cdot \sqrt{\frac{0{,}3778}{136}} = 0{,}220 \cdot 0{,}0527 = 0{,}0116;$$

$$S_{b_{xy}} = S_t \cdot \sqrt{\frac{\sum (X - \bar{x})^2}{\sum (Y - \bar{y})^2}} = 0{,}220 \cdot \sqrt{\frac{136}{0{,}3778}} = 0{,}220 \cdot 18{,}97 = 4{,}1741.$$

Критерий существенности коэффициентов регрессии ($S_{b_{yx}}$ и $S_{b_{xy}}$) определяется по формулам

$$t_{b_{xy}} = \frac{b_{xy}}{S_{b_{xy}}} = \frac{0{,}0413}{0{,}0116} = 3{,}56;$$

$$t_{b_{yx}} = \frac{b_{yx}}{S_{b_{yx}}} = \frac{14{,}8491}{4{,}1741} = 3{,}56.$$

Критерии существенности коэффициентов регрессии равны критерию существенности коэффициента корреляции: $t_b = t_r$.

В зависимости от того, между какими признаками рассматривается связь, не всегда имеет смысл вычислять все коэффициенты регрессии.

Корреляция может быть изображена графически в виде линии регрессии. Линию регрессии можно построить двумя способами – графическим и аналитическим.

При графическом способе по оси абсцисс откладывают значения признака X, по оси ординат – значения признака Y. Каждое наблюдение под двумя переменными отличается точкой с координатами (X; Y). Такой график называют точечной диаграммой, или корреляционным полем.

На точечной диаграмме с помощью прозрачной линейки проводят линию на глаз так, чтобы она располагалась как можно ближе ко всем точкам, и сумма расстояний этой линии от эмпирических точек была наименьшей. Данный способ приблизителен, так как дает возможность выявить лишь общую тенденцию, поэтому лучше пользоваться аналитическим способом.

При аналитическом способе используют уравнение прямой линии (для линейной регрессии)

$$Y = a + bx.$$

По исходным наблюдениям вычисляют $\bar{x}$, $\bar{y}$, и b_{yx} и подставляют в уравнение линейной регрессии

$$a = Y - bx;$$

$$b = b_{yx},$$

имеющее следующий вид:

$$Y = \bar{y} + b_{yx}\,(X - \bar{x}).$$

$$Y = 1{,}87 + 0{,}0413\,(X - 41) = 1{,}87 + 0{,}0413 \cdot X - 0{,}0413 \cdot 41.$$

$$Y = 0{,}17875 + 0{,}04125 \cdot X.$$

По уравнению находят теоретически усредненные значения Y для крайних (min и max) значений ряда X. Найденные точки ($X_{\min}$, $Y_{\min}$ и $X_{\max}$, $Y_{\max}$) наносят на график и соединяют прямой. Это и будет теоретическая линия регрессии Y по X.

В нашем случае при $X_{\min} = 37$ $Y_{\min}$ составит 1,7048, а при $X_{\max} = 48$ $Y_{\max}$ – 2,1592.

$$Y_{\min} = 0{,}1767 + 0{,}0413 \cdot 37 = 1{,}7048.$$

$$Y_{\max} = 0{,}1767 + 0{,}0413 \cdot 48 = 2{,}1592.$$

Найденные точки (37; 1,7048 и 48; 2,1592) наносим на график и соединяем прямой. Получаем теоретическую линию регрессии Y и X, которая показывает, что при увеличении числа зерен в колосе (на 1 шт.) продуктивность растений увеличивается в среднем на 0,0413 г/раст. Судя по коэффициенту детерминации $d_{yx} = r^2 = (0{,}783)^2 = 0{,}613$, примерно 61,3 % изменений продуктивности колоса у ржи обусловлено изменениями озерненности колоса, 38,7 % изменений связано с другими факторами.

На графике целесообразно указать уравнения регрессии, коэффициент регрессии и корреляции, доверительную зону для истинной линии регрессии в совокупности (рис. 1).

Для установления доверительной зоны необходимо вверх и вниз от теоретической линии регрессии отложить величину одной (68%-ная зона) или двух (95%-ная зона) ошибок отклонений от регрессии, т. е. $\pm S_{yx}$ или $\pm 2S_{yx}$, и соединить найденные точки пунктирными линиями. Область, заключенная между этими линиями, и называется доверительной зоной регрессии.

Вывод. Корреляция между продуктивностью растений озимой ржи сорта Ясельда и числом зерен в колосе прямая, сильная ($r = 0{,}783$). Продуктивность растений на 61,3 % зависит от числа зерен в колосе. При увеличении числа зерен в колосе на 1 шт. продуктивность растений озимой ржи сорта Ясельда увеличивается на 0,0413 г. Данная зависимость выражается уравнением $Y = 0{,}1767 + 0{,}0413 \cdot X$.

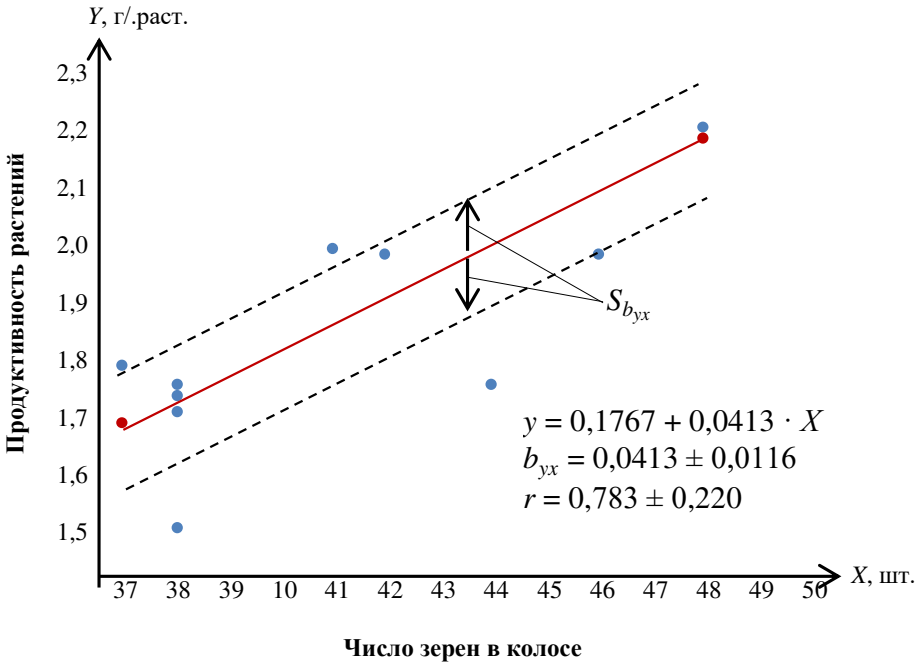

Рис. 1. Точечный график и теоретическая линия регрессии при прямой корреляции между продуктивностью растений и числом зерен в колосе

Задание. Вычислить коэффициенты прямолинейной корреляции и регрессии, рассчитать уравнение регрессии, представить полученные данные в виде графика и сделать выводы.

Исходные данные приведены в табл. 2.

Таблица 2. **Независимые переменные**

Номер растения	Урожайность Y, ц/га	Высота растений X_1, см	Продуктивная кустистость X_2, шт.	Длина колоса X_3, см	Число колосков в колосе X_4, шт.	Число зерен в колосе X_5, шт.	Число зерен с растения X_6, шт.	Масса 1000 семян X_7, г	Продуктивность растения X_8, г	Густота стеблестоя X_9, шт./м^2
1	32	91	3	8,5	17	33	67	32	1,90	402
2	36	99	2	9,3	18	35	69	40	1,97	431
3	48	102	6	10,6	19	47	75	43	2,30	456
4	51	110	7	11,7	19	50	83	50	2,63	520
5	33	98	4	8,0	17	32	73	36	1,88	411
6	40	103	3	9,4	17	35	69	40	2,05	438
7	42	100	5	9,0	18	37	90	38	2,12	426
8	55	112	6	12,0	19	46	85	52	2,77	451
9	54	110	7	10,8	19	49	76	49	2,83	480
10	49	105	7	10,7	17	46	74	46	2,65	465
11	40	102	4	9,4	18	38	75	46	1,99	431
12	43	101	2	9,0	17	44	83	37	2,07	463
13	47	112	5	10,7	19	37	96	34	2,53	472
14	33	93	2	8,8	17	33	80	36	1,80	428
15	39	99	2	8,7	18	37	77	43	1,88	433
16	35	96	3	8,5	18	35	72	40	1,73	418
17	40	105	5	10,3	19	46	89	47	2,12	453
18	43	112	4	11,2	17	49	92	44	2,49	460
19	52	108	7	11,7	19	47	108	50	2,69	489
20	58	102	6	11,7	18	54	117	53	2,80	493
21	35	97	2	9,2	17	32	63	38	1,87	427
22	33	92	3	8,7	17	33	87	33	1,92	465
23	48	105	5	10,1	18	36	92	47	2,12	483
24	41	100	3	9,8	18	39	85	43	2,00	470
25	53	98	4	11,3	21	40	112	50	2,68	459