3.5 Элементы теории корреляции

В сельскохозяйственных науках, в отличие от точных наук, полные (точные) функциональные связи встречаются редко, так как возможность искусственной изоляции влияния других факторов на изучаемые признаки в большинстве случаев неосуществима.

Например, связь урожайность - удобрения, имеется, но есть еще ряд факторов, влияющих на урожайность (севообороты, семена, предшественники, агротехника - субъективные факторы; метеорологические факторы- объективные).

Поэтому связь урожайность - удобрения неполная функциональная связь. Эту связь называют **корреляционной** (англ. correlation – соотношение, соответствие).

Метод корреляции применяется для того, чтобы при сложном взаимодействии посторонних влияний выяснить, какова была бы зависимость между результатом и фактором, если бы посторонние причины (факторы) не изменялись и своим изменением не искажали основную зависимость.

Первая задача корреляции: выявление на основе наблюдений над большим количеством фактов того, как изменяется в среднем результативный признак в связи с изменением данного фактора (парная корреляция) или группы факторов (множественная корреляция). Эта задача решается нахождением уравнения связи.

Вторая задача корреляции: определение степени влияния искажающих факторов. Эта задача решается при помощи различных показателей тесноты связи: коэффициента корреляции, корреляционного отношения.

Процесс нахождения связи между признаками называется ***выравниванием.***

Выравнивание ведет к нахождению переменной средней $\bar{y}_x$, исчисленной в предположении функциональной зависимости **y** от **x**, т.е. $\bar{y}_x = f(x)$, и называется **уравнением регрессии.**

При изучении влияния одних признаков на другие выделяются два признака - **факториальный** и **результативный**. Выделение этих признаков осуществляется путем логического анализа.

Например, в связи урожайность - осадки, урожайность - результативный признак, а осадки - факториальный.

Графическое изображение связи.

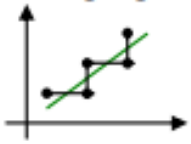
Графическое изображение связи изучаемых явлений позволяет не только установить наличие или отсутствие связи между ними, но и изучить характер этой связи (форму связи и тесноту связи).

Если имеются числовые характеристики факториальных и результативных признаков одного и того же явления, то каждую пару чисел можно изобразить графически, откладывая по оси абсцисс - факториальный признак, по оси ординат - результативный признак.

Ломаная, соединяющая эти точки, называется **ломаной регрессии**.

По форме этой ломаной приближенно определяют вид зависимости.

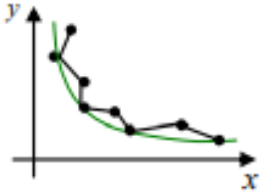
1. Если из графика видно, что связь близка к прямолинейной, то уравнение регрессии пишется в виде:



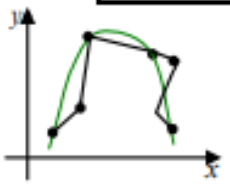
$$\bar{y} = ax + b$$

2. Если экспериментальные данные располагаются так, что через них можно провести гиперболу, то можно ожидать уравнение в виде:

$$\bar{y} = \frac{k}{x};\quad \bar{y} = \frac{a}{x+b},\quad \bar{y} = \frac{a}{x+b} + c$$



3. Если кривая имеет max или min, то зависимость определяется по уравнению: $\bar{y} = ax^2 + bx + c$



Для выявления функциональных зависимостей и определения неизвестных коэффициентов этой зависимости можно воспользоваться методом наименьших квадратов.

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + b \cdot n = \sum_{i=1}^{n} y_i \end{cases} \Rightarrow \quad y = ax + b$$

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 \cdot y \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \cdot y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + c \cdot n = \sum_{i=1}^{n} y_i \end{cases} \Rightarrow \quad y = ax^2 + bx + c$$

Коэффициент корреляции.

После того, как уравнение регрессии найдено, находят так называемый ***коэффициент корреляции.*** Он используется для оценки тесноты связи между величинами при прямолинейной зависимости. Обозначается буквой *r* и определяется по формуле:

$$r=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}$$, где

$\bar{x}$ - среднее значение факториального (причинного) признака $\bar{x}=\frac{\sum x_i}{n}$

$\bar{y}$ - среднее значение результативного признака $\bar{y}=\frac{\sum y_i}{n}$

Промежуточные вычисления удобно располагать в виде таблицы:

№ наблюдения	x_i	y_i	$x_i-\bar{x}$	$\left(x_i-\bar{x}\right)^2$	$y_i-\bar{y}$	$\left(y_i-\bar{y}\right)^2$	$\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)$
$\sum$	…	…	…	…	…	…	…

Величина коэффициента корреляции находится в пределах $-1\le r\le 1$:

1) Чем ближе $|r|$ к 1, тем теснее связь между факториальным и результативным признаками.

2) при $|r|=1$ получается полная функциональная связь.

3) если $|r|\to 0$, то связь между признаками слабая.

4) при $|r|=0$ связи между признаками нет (линейная зависимость отсутствует).

5) при $r>0$ зависимость между признаками прямая (возрастающая).

6) при $r<0$ зависимость обратная (убывающая).

Если зависимость между признаками прямая, то можно пользоваться уравнением прямой регрессии: $y-\bar{y}=b_{y/x}\left(x-\bar{x}\right)$, где

$b_{y/x}$ - коэффициент регрессии, который определяется по формуле:

$$b_{y/x}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$$

Пример Для 10 петушков леггорнов 15 дневного возраста были получены следующие данные о весе их тела (**x**) в граммах и весе гребня (**y**) (в мг):

x_i	83	72	69	90	90	95	95	91	75	70
y_i	56	42	18	84	56	107	90	68	31	48

Требуется:

1) найти коэффициент корреляции и сделать вывод о тесноте и направлении линейной корреляционной связи между признаками;

2) составить уравнение прямой регрессии;

3) нанести на чертеж исходные данные и построить прямую регрессии.

Решение: Составим вспомогательную таблицу

№	x_i	y_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	83	56	0	0	-4	16	0
2	72	42	-11	121	-18	324	198
3	69	18	-14	186	-42	1764	588
4	90	84	7	49	24	576	168
5	90	56	7	49	-4	16	-28
6	95	107	12	144	47	2209	564
7	95	90	12	144	30	900	360
8	91	68	8	64	8	64	64
9	75	31	-8	64	-29	841	232
10	70	48	-13	169	12	144	156
$\sum$	30	600	0	990	0	6854	2302

Вычисляем средние:

$$\bar{x}=\frac{\sum x_i}{n}=\frac{830}{10}=83 \qquad \bar{y}=\frac{\sum y_i}{n}=\frac{600}{10}=60$$

1) найдем коэффициент корреляции:

$$r=\frac{\sum(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\cdot\sqrt{\sum(y_i-\bar{y})^2}} \qquad r=\frac{2302}{\sqrt{990\cdot 6854}}=0,88$$

Вывод: между весом тела **x** и весом гребня **y** у 15- дневных петушков существует тесная положительная линейная корреляционная связь.

2) найдем коэффициент регрессии:

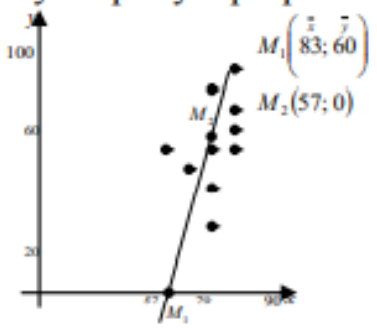
$$b_{y/x}=\frac{\sum(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \qquad b_{y/x}=\frac{2302}{990}\approx 2,32$$

Подставим в уравнение прямой регрессии:

$$y-\bar{y}=b_{y/x}(x-\bar{x}) \quad y-60=2,32(x-83)$$

$$\underline{y=2,32x-132,56}$$

3) наносим исходные данные на координатную плоскость и строим найденную прямую регрессии



Задания для решения в аудитории

Дана таблица значений *x* и *y*

x	2,8	3,4	3,7	3,4	2,8	1,5	4,9	7,2	1,7	3,4
y	1,3	2,0	4,4	3,0	2,2	1,8	5,0	2,8	9,1	4,4

Требуется:

1. найти коэффициент корреляции и сделать вывод о тесноте и направлении линейной корреляционной связи между признаками;
2. составить уравнение прямой регрессии;
3. нанести на чертеж исходные данные и построить прямую регрессии.