**§4. Регрессионый анализ**

Для выяснения механизмов биологических процессов чаще всего недостаточно установить наличие связи между признаками. Желательно но знать каков вид этой связи, то есть, как конкретно изменяется данный признак при изменении величины другого признака. В математической статистике термин регрессия употребляется для описания зависимости среднего значения одного признака от величины другого признака на основе какой-либо математической модели. Например, линейная регрессия – это описание зависимости одного признака от дру­гого с помощью прямой.

Экспериментатор измеряет попарно значения двух признаков, например, рост и вес у мужчин. Если в выборке ***п*** объектов, то будет получено ***п*** пар значений, которые можно нанести на плоскость в виде ***п*** точек в координатах рост-вес. Через эти точки можно провести, вообще говоря, множество разных кривых, с помощью которых исследователь пытается отразить зависимость одного признака от другого. Задачей регрессионного анализа является отыскание такой кривой, которая наилучшим способом удовлетворяет экспериментальным точкам. Вид этой кривой, правда, зависит от математической модели, связывающей эти два признака. Математическая модель строится на основании некоторых допущений, исходных предположений о связи между признаками. Например, можно предположить, что вес человека прямо пропорционален его росту. В этом случае необходимо будет провести через экспериментальные точки прямую. Но ведь обычно экспериментальные точки не лежат на прямой, как же провести прямую? На этот вопрос и отвечает математическая статистика.

Конечно, можно спросить, почему именно линейная зависимость предполагается между ростом и весом, а не какая-либо другая. Как уже указывалось, все зависит от модели. Если человека рассматривать как цилиндр, высотой ***h*** (рост), то объем этого цилиндра, а значит и вес, будет пропорционален высоте. На самом деле форма тела человека значительно сложней, и модель цилиндра лишь приближенно описывает ее. Учет зависимости ширины плеч, талии и так далее от роста позволит создать более адекватную модель, которая может оказаться нелинейной. Тогда придется проводить через экспериментальные точки не прямую, а более сложную кривую. Но и в этом случае экспериментальные точки не обязательно будут лежать на теоретической кривой. Поэтому провести теоретическую кривую через экспериментальные точки можно многими способами.

Здесь мы рассмотрим случай линейной регрессии. Общий вид линейной регрессии

$y=a+bx,$ (1)

здесь $x$ ***–*** значения одного из признаков, ***у –*** средние значения другого признака; $a,b$ – к коэффициенты.

Задачей линейного регрессионного анализа является нахождение коэффициентов $a$ и $b$, так как эти два коэффициента полностью определяют положение прямой на плоскости.

Для нахождения коэффициентов $a$ и $b$, обычно используется метод наименьших квадратов. В чем суть этого метода? Суть метода наименьших квадратов заключается в том, что сумма квадратов отклонения экспериментальных точек от теоретической линии регрессии должна быть минимальна. В математической форме это утверждение можно записать так:

 $S=\sum\_{}^{}\left(y\_{i}-y\_{i}^{'}\right)^{2}=min$, (2)

где S – сумма квадратов отклонений экспериментальных точек от линии регрессии; $y\_{i}$. - экспериментальные значения признака, $y\_{i}^{'}$ теоретические значения признака, соответствующие теоретической линии регрессии.

В случае линейной регрессии удовлетворение требования (2) приводит к системе из двух уравнений с двумя неизвестными $a и b.$ Решая эту систему, находят а и b. Эта система имеет вид:

 $na+\sum\_{}^{}x\_{i}b=\sum\_{}^{}y\_{i}$

$\sum\_{}^{}x\_{i}a+\sum\_{}^{}x\_{i}^{2}b=\sum\_{}^{}x\_{i}y\_{i} $ (3)

**Пример. Имеется выборка:**

**x : 1 2 3 4 5 6 7 8**

у : **3 3 5 4 7 7 1 0 9**

Провести через эти точки прямую методом наименьших квадратов.

Решение. Поскольку придется решать систему (3) удобно составить таблицу:

|  |  |  |  |
| --- | --- | --- | --- |
| *х* | *у* | *х2* | *у2* |
| 1 | 3 | 1 | 3 |
| 2 | 3 | 4 | 6 |
| 3 | 5 | 9 | 15 |
| 4 | 4 | 16 | 16 |
| 5 | 7 | 25 | 35 |
| 6 | 7 | 36 | 42 |
| 7 | 10 | 49 | 70 |
| 8 | 9 | 64 | 72 |
| $\sum\_{}^{}x\_{i}=36$  | $\sum\_{}^{}y\_{i}=48$  | $\sum\_{}^{}x\_{i}^{2}=204$  | $\sum\_{}^{}x\_{i}y\_{i}=259$  |

Найдя суммы, подставляем их в (2):

 $8a+36b=48$

 $36a+204b=259$

Здесь п=8 потому, что всего у нас 8 точек. Для решения удобно избавиться от одного из неизвестных, приравняв коэффициенты при одном из них. Избавимся от **а**. Для этого умножим первое уравнение на 4.5 и вычтем из него нижнее уравнение почленно. Имеем:

$$36f+162b=216$$

 -

$$36a+204b=259$$

$$-42b=-43$$

$$b=\frac{43}{42}=1.024$$

Теперь найдем ***а*,** подставляя известное значение ***b*** в одно из уравнений, например:

$$8a+36∙1.024=48$$

8a=11.136

$$a=\frac{11.136}{8}=1.392$$

Подставляем значения *а* и *b* в (1)

$$y=1.392+1.024x$$

Это и есть уравнение линейной регрессии. Чтобы проверить близость теоретических точек к экспериментальным, можно вычислить соответ­ствующие значения. Например:

$$y\_{1}^{'}=1.392+1.024∙2.416$$

$$y\_{2}^{'}=1.392+1.024∙3.440$$

. . . . . . . . . . . . . . . . . . . . . . .

$$y\_{8}^{'}=1.392+1.024∙9.584$$

Видно, что теоретические значения близки к эксперименталь­ным. Такая проверка исключает грубые ошибки в вычислениях.

**Контрольные вопросы и задания**

1. Что такое регрессионный анализ?
2. Что такое линейная регрессия?
3. В чем суть метода наименьших квадратов?
4. Как находятся коэффициенты линейной регрессии?
5. Найти коэффициенты уравнения линейной регрессии для следующих выборок:
6. Продолжительность вегетационного периода (х, дни) и вес 1000 семян ячменя (*у*, г):

х: 90 85 80 75 70 65 60

у: 47,50 46,75 45,75 42,85 44,76 41,44 37,00

2) Общая длина в см (х) и вес в г (у) человеческого эмбриона:

x: 15 23 30 35 40 46 50

у: 120 300 640 1210 1700 2240 3250

1. Pост листа валлиснерии в см (*у*) по часам (*х*):

х: 6 16 42 54 65 77 88

у: 0,3 1,7 12,6 15,4 16,1 16,7 17,1

1. Вес растения кукурузы в г (*у*) и дни роста (х):

х: 6 18 30 39 46 53 60 74 93

у: 1 4 9 17 26 42 62 71 74

1. Максимальное (*у*) и минимальное (*х*) артериальное давление у человека:

х: 63 68 73 78 83 88 93 95 98

y: 107 112 119 122 123 130 130 129 133

1. Возраст в годах (*х*) и средний вес в г (*у*) у осетра:

х: 1 2 3 4 5 6 7 8

у: 49 82 119 150 206 275 357 375