ГЛАВА 2

ДВУХМЕРНЫЙ КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Методы двухмерного корреляционно-регрессионного анализа позволяют определить тесноту и вид зависимостей между парами стереометрических показателей одного или двух микрообъектов изучаемого морфологического и патолого-анатомического объекта. Иерархическая подчиненность микрообъектов в биоструктурах и многомерные взаимозависимости между их метрическими свойствами позволяют завершать стереометрический анализ изучением двухмерных распределений. Такие двухмерные модели будут содержать в себе следы воздействия остальной совокупности факторов. Недостаток метода в данном случае состоит только в том, что на его основе нельзя определить зависимость между двумя изучаемыми показателями и вклад в их взаимосвязи всей остальной совокупности неучтенных свойств объекта.

Двухмерный регрессионный анализ есть частный случай многомерного и входит в него как составная часть. Поэтому знание его также важно и для понимания принципов многомерного анализа. Именно сложность многомерного анализа объясняет сдержанный интерес морфологов к корреляционно-регрессионному анализу вообще.

Двухмерные корреляционные связи можно устанавливать не только между какими-либо одними стереометрическими параметрами объектов, но и между разными наборами пар стереометрических параметров одного объекта. Исключение составляют только стереометрические константы, которые, как правило, безразмерны и в норме неизменны относительно величины биологического объекта. При проведении исследования следует также учитывать и размерность пар признаков, которые принимают за сопряженные. Неодинаковые размерности очень часто могут служить причиной искажения реальных форм связей. Поэтому рекомендуется анализировать величины с одинаковой размерностью.

Таблица 10

Корреляционная решетка, отражающая зависимость между диаметром и длиной сегментов лимфатических капилляров эпикарда собаки (мкм)

Длина сегмента (y)	Диаметр (x)												
	13	18	23	28	33	38	43	48	53	58	63	68	$\sum P_x$
25	3	2											5
35		6	4										10
45		1	13	**5**									19
55		1	2	4	8	1							16
65			1		4	4	2						11
75					2	6	6	2					16
85							1	5					6
59								1	4	1			6
105									2	4	1	1	8
115											1	1	2
125													1
$\sum P_y$	3	10	20	9	14	11	9	8	6	5	2	3	100

Взаимозависимость между метрическими показателями биологических объектов можно показать на любом конкретном примере, в частности на диаметре (x) и длине (y) капиллярных сегментов лимфатического русла. В табл. 10 представлены результаты сопряженных замеров диаметра и длины лимфатических капилляров эпикарда собаки. Между диаметром и длиной устанавливается прямая зависимость таким образом, что с увеличением диаметра капилляра происходит одновременное увеличение его длины. Причем эта зависимость

не функциональная, а корреляционная, так как одному значению диаметра соответствует несколько вариантов значений длины. Так, капилляры диаметром 38 мкм в 1 случае имели длину 55 мкм, в 4—65 мкм и в 6—75 мкм. Из таблицы также видно, что численные значения частот встречаемости пар диаметра и длины капилляров группируются около линии, которую можно провести из верхнего левого угла в нижний правый угол. Таблица показывает не только наличие, но и вид, направление связи. Как мы уже отметили, такое направление отражает прямую зависимость. Если значения частот сопряженных пар признаков группируются вокруг линии, идущей из правого верхнего угла в нижний левый угол, то в таком случае зависимость будет не прямой, а обратной. Группировка значений частоты встречаемости сопряженных пар признаков в центре таблицы по фигуре, близкой к окружности, является отражением отсутствия либо нелинейности корреляционной связи.

Таблицы такого типа являются первым этапом проведения двухмерного корреляционно-регрессионного анализа и называются корреляционными решетками. Их построение осуществляется следующим образом. Проводят замеры сопряженных пар признаков. Например, замеряют диаметр и длину, диаметр и число объектов в единице объема, абсолютную и удельную площадь поверхности и другие пары стереометрических показателей одного микрообъекта (диаметры, длины, число поверхность, объем и другие параметры двух структур), находящегося в тесной структурной зависимости по какому-либо одному или совокупности морфологических признаков. Такими признаками могут быть топографические (капилляр и печеночная балка), структурно-функционные (терминальная и респираторная бронхиолы, респираторная бронхиола и альвеолярный ход) и т.д. После проведения сопряженных замеров формируют таблицу. Одному признаку приписывают строки, а другому колонки. Для наглядности результаты замеров группируют в колонках и строках в порядке возрастания, а в клетках их пересечений вписывают частоты встречаемости сопряженных замеров одного класса величины двух параметров.

При формировании двухмерной выборочной совокупности число интервалов не следует брать большим, так как это уменьшает наглядность представлений о реальной зависимости между изучаемыми факторами. В приведенном примере величина интервалов для диаметра и длины капилляров принята в 5 мкм, хотя для каждого фактора она может быть выбрана иной. Итоговый столбец таблицы ΣP_X образует ряд распределения, такой же ряд образует и итоговая строка ΣP_Y. Из таблицы следует, что между диаметром и длиной лимфатических капилляров имеется пропорциональная зависимость, которая может быть описана линейной функцией.

Кроме табличной интерпретации данных, довольно часто используют графический метод, с помощью которого строят так называемое поле корреляции, по которому устанавливают существующий тип зависимости между факторами x и y (рис. 56). В рассматриваемом примере использовано графическое изображение зависимости между диаметром (D) и длиной капиллярных сегментов *(L)*.

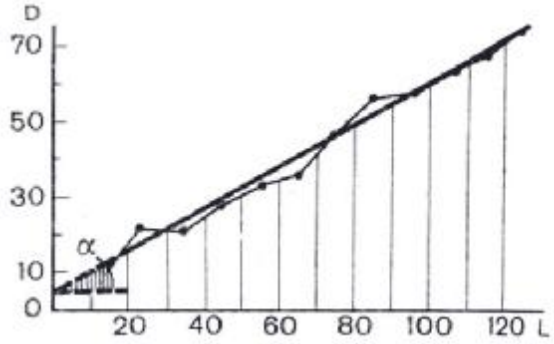

Рис. 56. Изображение результатов сопряженных замеров диаметра (ось ординат) и длины (ось абсцисс) лимфатических капилляров в двухмерной системе координат. Экспериментальная зависимость (ломанная кривая) с узловыми точками хорошо апроксимируется линейной функцией с углом наклона а. Объяснение в тексте.

При изучении корреляционной зависимости устанавливают, в каком направлении происходит смещение рядов распределения. Диаметр капилляров и их длина увеличивается одновременно. Если число замеров пар диаметров и длин капилляров было бы достаточно большим, то линия регрессии приняла бы нормальный вид. За теоретическую линию регрессии принимают то предельное значение эмпирической зависимости, к которому она приближается в пределе при неограниченном увеличении числа наблюдений и одновременном уменьшении величины интервалов аргумента. В данном примере это прямая линия, идущая под углом к оси абсцисс. Исходя из этого, при планировании корреляционного анализа число наблюдений должно быть выбрано таким, чтобы полученную зависимость молено было выразить теоретической функцией.

Прежде чем определить теоретическую зависимость регрессии, например между диаметром и длиной капилляров, выбирают тип уравнения, зависимости, которой будут аппроксимировать реальную функцию. Эта задача является одной из наиболее трудных, так как существует большое число разных теоретических функций, в виде которых с малой погрешностью можно представить результаты опыта. При решении этой задачи прибегают к получению разнообразной дополнительной информации и проводят логический анализ изучаемой связи между явлениями. Если теоретически нельзя выбрать класс функции, пригодной для решения поставленной задачи, эти вопросы решают эмпирически, прибегая к построению разных типов экспериментальных функций. После того как подобрана теоретическая функция, которой приближенно выражают экспериментальную зависимость, определяют параметры уравнения регрессии. Одним из наиболее общепринятых способов оценки параметров регрессии является метод наименьших квадратов.

При выборе линейной зависимости ее задают в виде уравнения:

$$y = f(x) - ax + b, \tag{123}$$

где a и b — параметры уравнения регрессии, подлежащие определению. Параметры уравнения регрессии устанавливают из системы уравнений:

$$\begin{cases} a\sum\limits_{i=1}^{p} x_i^2 + b\sum\limits_{i=1}^{p} y_i = \sum\limits_{i=1}^{p} x_i y_i \\ a\sum\limits_{i=1}^{p} y_i + pb = \sum\limits_{i=1}^{p} x_i \end{cases} \tag{124}$$

где p — количество пар наблюдений функции y и аргумента x в зависимости $y = ax + b$; a — значения аргумента p_i; a и b — параметры, подлежащие определению.

Пример расчета покажем на установлении уравнения регрессии между диаметром и длиной капилляров, приняв за теоретические функции: прямую $y = ax + b$; обратную $x = a_1 y + b_1$.

Для вычислений используем данные, приведенные в табл. 0.

Рассчитаем значение параметров прямого уравнения регрессии, т.е. свободного члена b и коэффициента a при x.

Сначала найдем оценку параметра b, как

$$b = \frac{\sum\limits_{i=1}^{p} x_i \sum\limits_{i=1}^{p} x_i^2 - \sum\limits_{i=1}^{p} y \sum\limits_{i=1}^{p} xy}{p\sum\limits_{i=1}^{p} x^2 - (\sum\limits_{i=1}^{p} y)^2}, \tag{125}$$

после чего определим величину параметра a_1 из зависимости:

$$a = \frac{\sum_{i=1}^{p} x_i - pb}{\sum_{i=1}^{p} y_i}. \quad (126)$$

Подставляя найденные значения a = 0,552 и b = —0,178 в прямое уравнение регрессии, запишем его в виде y = 0,552x—0,178.

Таким же образом проведем расчеты и параметров обратного уравнения регрессии, при этом величина b окажется равной 5,616, а величина a равной 1,661, что позволит определить x = 1,6610 + 5,616.

Угол, образованный обеими прямыми, представляет собой графическую оценку степени тесноты связи между двумя изученными факторами и называется коэффициентом корреляции. Чем меньше такой угол, тем больше его тангенс, т.е. тем больше коэффициент корреляции и тем более тесная связь между изученными факторами.

По прямому и обратному уравнениям регрессии коэффициент корреляции можно найти как

$$r_{x,y} = \pm\sqrt{a_1 a_2} \quad (127)$$

В этом случае он окажется равным:

$$r_{x,y} = \pm\sqrt{1{,}661 \cdot 0{,}562} = 0{,}917.$$

Когда устанавливаются нелинейные связи между изучаемыми факторами, способ вычисления параметров таких зависимостей не отличается от рассмотренного и принцип наименьших квадратов сохраняется.

Коэффициент корреляции можно рассчитать и без определения прямого и обратного уравнений регрессии, используя формулу

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \quad (128)$$

где x и y— средние арифметические значения x_1 и y_1. Если обе случайные величины независимы друг от друга, коэффициент корреляции равен нулю.

Кроме указанных способов оценки величины коэффициента корреляции, можно использовать также формулу:

$$r_{x,y} = a\frac{s_y}{s_x} \quad (129)$$

где a — параметр при x в уравнении линейной регрессии; σ_x и σ_y — средние квадратические отклонения для аргумента и функции.

В нашем примере параметр a для прямого уравнения регрессии равен 0,552, среднее квадратическое отклонение диаметра капилляров составило 14,52, а длины — соответственно 23,85, что дает оценку коэффициента корреляции, равную

$$r_{x,y} = 0{,}552 \cdot \frac{23{,}85}{14{,}53} = 0{,}907.$$

которая близка к установленной первым способом.

Коэффициент корреляции может изменяться в пределах между — 1 и +1. (Приложение II, табл. 6). Если коэффициент корреляции не равен нулю, то он указывает на наличие зави-

симости между системой, состоящей из двух случайных величин. Причем, чем больше абсолютная оценка коэффициента корреляции, т.е. чем ближе он по абсолютной величине к единице, тем более тесная корреляционная зависимость между изучаемыми случайными величинами. При коэффициенте корреляции, равном единице, корреляционная зависимость переходит в функциональную. Знак коэффициента корреляции показывает вид зависимости. Если он отрицателен, то зависимость обратная. При положительном коэффициенте корреляции имеется прямая зависимость между изменением аргумента и функции.

При стереометрических исследованиях объем выборки бывает сравнительно небольшим и поэтому необходимо проводить проверку коэффициента корреляции выборочной совокупности на репрезентативность для коэффициента корреляции генеральной совокупности. Ошибку коэффициента корреляции называют также коэффициентом надежности, так как она характеризует надежность выборочного коэффициента. Например, при выборке объемом более 100 параметров для определения коэффициента корреляции используют формулу:

$$m_r = \pm\frac{1-r^2}{\sqrt{N}}. \qquad (130)$$

Ожидаемые значения коэффициента корреляции приведены в приложении II (табл. 6).

При объеме выборки меньше 100 расчет ведут по формуле:

$$m_r = \pm\sqrt{\frac{1-r^2}{N-2}}. \qquad (131)$$

При оценке величины коэффициента корреляции следует иметь в виду, что его нулевое значение не указывает на отсутствие связи между анализируемыми распределениями, так как он может быть равен нулю при наличии нелинейных зависимостей. Изложенные этапы двухмерного корреляционного анализа позволяют определить наличие связи, ее линейность или нелинейность.

По значениям коэффициентов корреляции и их ошибок судят о достоверности, используя, одно из соотношений:

$$t = \frac{r}{m} \quad (132) \qquad t = r\frac{\sqrt{N}}{\sqrt{1-r^2}} \quad (133) \quad \text{или} \quad t = r\frac{\sqrt{N-2}}{\sqrt{1-r^2}} \qquad (134)$$

Далее полученные значения t сравнивают с $t_{табл.}$, определяемым из таблицы Стьюдента (см. Приложение II, табл. 1), и обычным способом приходят к заключению о достоверности или недостоверности найденного коэффициента корреляции.

Продолжим анализ производимого примера. Последовательно найдем значения ошибки коэффициента корреляции

$$m_r = \frac{1-(0{,}917)^2}{100} = 0{,}0159$$

$$t = \frac{0{,}917}{0{,}159} = 57{,}63.$$

Учитывая, что для N=100 $t_{теор.}$= 1,97 и $t_{теор.}$, принимаем полученный выборочный коэффициент корреляции достоверным.

О достоверности коэффициента корреляции судят не только по его ошибке, но и по так называемому Z-преобразованию Фишера. Это преобразование используют и для расчета

должного объема сопряженной выборки с целью получения достоверного коэффициента корреляции. Находят значение Z из формулы:

$$Z = 0{,}51\ \ 1n\frac{1+r}{1-r} \tag{135}$$

и соответствующую ошибку из формулы

$$m_z = \pm\frac{\Delta}{\sqrt{N-3}}. \tag{136}$$

Достоверность определяется так же, как и для коэффициента корреляции, т.е.

$$t = \frac{Z}{m_z}. \tag{137}$$

По значению Z и t с использованием формулы (Плохинский Н.А., 1970):

$$N = 3 + \frac{t^2}{Z^2} \tag{138}$$

определяют минимальный объем выборки сопряженных пар признаков, достаточной для получения достоверного выборочного коэффициента корреляции. В приводимом примере:

$$Z = 0{,}51 \cdot 1n\frac{1+0{,}917}{1-0{,}917} = 1{,}691;$$

$$m_z = \pm\cdot\frac{1}{\sqrt{100-3}} = \pm 0{,}102;$$

$$t = \frac{1{,}601}{0.102} = 15{,}696$$

$$N = \frac{(15{,}768)^2}{(1{,}601)^2} + 3 \approx 100{,}12$$

Следовательно, должное значение выборки сопряженных пар признаков соответствует измеренному числу диаметров и длин лимфатических капилляров эпикарда сердца собаки.

Объем выборки, достаточной для получения репрезентативного выборочного коэффициента корреляции, можно контролировать и по статистикам распределения изучаемых пар признаков.

Для этих целей отбирают тот признак, у которого отношение среднего квадратического отклонения к самому среднему — коэффициент вариации будет больше. По статистикам этого параметра рассчитывают должный объем сопряженных пар измерений. Методы планирования объема выборочных распределений одномерных признаков изложены в разделе, посвященном статистическому обеспечению стереометрического анализа.

Кроме определения параметров регрессии и расчета коэффициента корреляции в ряде случаев для установления возможных колебаний вариант вокруг теоретического уравнения в принятом 95% доверительном интервале еще используют среднее квадратическое отклоне-

ние уравнения регрессии

$$s = \sqrt{\frac{\sum (y_0 - y_T)_2}{n_1 + n_2 - 2}}, \qquad (139)$$

где y_0 — опытные и $y_т$ — теоретические значения функции; n_1 и n_2 — объем выборки для аргумента и функции.

Среднее квадратическое отклонение уравнения регрессии позволяет установить доверительный интервал для значений функции на всем интервале ее изменений, что очень важно в проблеме установления существенности в изменениях стереометрических параметров при патологических процессах.

Описание парных зависимостей между стереометрическими параметрами с установлением среднего квадратического отклонения для выбранной и приближенной функции представляет интерес в случае, когда нельзя получить достаточный по объему материал для характеристики патологического процесса. В таких случаях на конкретном патолого-анатомическом объекте путем измерений и расчета по стереометрическим формулам устанавливают параметры метрических свойств структуры, по одному параметру, с использованием установленной для нормы корреляционной функции, находят должное значение другого параметра и сопоставляют его с фактическим. При таком подходе удается избежать влияния размерных характеристик объекта, как, например, объема и массы, значительно усложняющих процесс сопоставления и требующих увеличения объема наблюдений.