Глава четвертая

Многомерная статистика и проблема измерения

1. Постановка проблемы измерения в многомерной статистике

С развитием многомерной статистики многие ее методы начали в целях измерения с успехом использоваться в социальных исследованиях: социологии, демографии, психологии. Исторически первым таким методом является факторный анализ. Первоначально он был развит в применении к количественным данным, получаемым в психологическом тестировании. Спустя несколько десятилетий, в 40-х годах был развит так называемый латентный анализ в применении к качественным данным социологии и социальной психологии. В последнее время для реализации вероятностной классификации в социологии начала использоваться методика распознавания образов.

Современный аппарат многомерной статистики позволяет выработать единый подход к проблеме измерения данных любой природы — количественных и качественных. Предположим, мы хотим измерить отношение к труду посредством анкеты. Ответы на вопросы будут представлять собой некоторые значения эмпирической переменной. Изучаемое отношение к труду можно рассматривать как некоторую гипотетическую (латентную) переменную, причем, и это существенно, как в данном случае, одномерную переменную. Если анкета имеет n вопросов, то эмпирическая переменная будет n-мерной величиной (n-мерным вектором), а исследуемая латентная переменная — одномерным вектором. В общем случае латентная переменная может быть представлена m-мерпым вектором. Большая трудность связана с характером компонент эмпирического и латентного векторов. В шкалах Лайкерта и Терстона латентная переменная представлялась порядковой переменной, а эмпирические переменные n-мерным

вектором (по числу вопросов в вопроснике), причем каждая компонента векторов была количественной переменной. В принципе компоненты обоих векторов могут быть величинами любой природы.

В общем случае обозначим эмпирическую переменную, состоящую из я компонентов, *x,* а латентную переменную, состоящую из *m* компонентов, — *y.* Когда индивид отвечает на вопросы анкеты, то это означает, что он, обладая определенным значением латентной переменной *y,* реализует определенное значение эмпирической переменной, т. е. можно предположить, что существует условное распределение *x* и *y* [1]:

F (x|y).

Нам неизвестно распределение латентной переменной *y*— *L(y),* но из данных ответов мы получаем безусловное распределение *x—H(x).* Эти три функции распределения — *F(x|y), L(y), H (x)* —связаны известным соотношением:

$$H(x)=\int F(x|y)dL(y)$$

Если бы нам были известны функции *F* и *L,* то проблема оценки латентной переменной *y* из наблюдаемой (эмпирической) переменной *x* сводилась бы к проблеме Бейеса. Однако обычно *F* и *L* неизвестны. В общем виде предложенное интегральное уравнение не решается. Для того чтобы получить его решение и, следовательно, решить проблему измерения латентной переменной *y* через посредство эмпирической переменной ***x,*** необходимо наложить на *F* и *L* определенные ограничения. Т. Андерсон вводит два ограничения: предположение об условной независимости и предположение о линейной регрессии. Предположение об условной независимости можно записать таким образом:

$$F(x\,|\,y)=\prod_{i=1}^{n}F_i(x_i\,|\,y),$$

и оно означает, что эмпирические переменные x_i распределены независимо при данном значении латентной переменной *y.* В переводе на простой язык это говорит о том, что определенный ответ на какой-то вопрос анкеты не влияет на ответы на другие вопросы, предполагая, что индивид в момент ответа обладает присущим ему, но неизвестным значением исследуемой латент-

[1] *Anderson T. W.* Some scaling models and estimation proceduces in the latent class model.— In: Probability and Statistics. U. Grenander (Ed.). Stockholm — New York, 1959.

ной переменной. Это предположение используем при определении моментов распределения $F(x|y)$.
По определению, первый момент:
$E(x|y)=\mu(y)$
Второй момент:
$E\{[x-\mu(y)][x-\mu(y)']|y\}=D(y)$.

В силу предположения условной независимости матрица $D(y)$ диагональная ($d_{ii}=0$, $i\neq j$) и выражение для второго момента принимает вид:

$E(xx'|y)=D(y)+\mu(y)\,\mu'(y)$,

Второе предположение о линейности регрессии записывается в виде

$E(x-y)=\mu(y)=\Lambda(y)+\mu$

что означает, что среднее x при данном y представляет собой линейную функцию от y, где Λ — матрица размерности nm.
Без потери общности можно принять, что $Ey=0$, $Eyy'=M$.
Тогда
$E(x-\mu)(x-\mu)'=\Psi=D+\Lambda M\Lambda'$
Если положить M=J, то
$\Psi=D+\Lambda M\Lambda'$

Таким образом, получаем модель факторного анализа: из известной ковариационной матрицы ψ определяем матрицу факторных нагрузок Λ (при выполнении второго предположения о линейности регрессии).

Можно показать, что если x считать дихотомической переменной и функция $F(x|y)$ определяет вероятность положительного ответа на x при данном y, то получается модель латентно-структурного анализа. В этом случае обозначим

$P(x_i=1|y)=\pi_i(y)$

и функцию $F\{x|y)$ заменим на $\pi(x_i\ldots x_n|y)$.

Теперь наша задача — рассмотреть более детально использование моделей факторного и латентного анализа в социологии. Также мы остановимся на специальном варианте регрессионного анализа, который получил в литературе название причинного анализа.