$N(x_1)=7$, $N(x_2)=8$, $N=15$, $\bar{y}_1=155{,}14$, $\bar{y}_2=163{,}25$, $\sigma_y=9{,}31$, $r_{rb}=0{,}42$

Формула (II,7,10) представляет собой алгебраическое упрощение коэффициента *r* для случая, когда X – дихотомическая переменная, поэтому все расчеты можно было бы проводить и по формулам для *r*, например, (II,5,1) или (II,5,3). Обобщения этих коэффициентов (полисериальные коэффициенты) не получили широкого распространения.

[140]

Глава III

РЕГРЕССИИ

1. Основные понятия. Прямая регрессия. Криволинейные связи. Корреляционное отношение

Как отмечалось, при исследовании связи между двумя признаками находят распределение совокупности в виде корреляционной таблицы $\{N_{ij}\}$; тесноту связи характеризуют с помощью коэффициентов корреляции (глава II), а форму – с помощью уравнений регрессии, к рассмотрению которых мы и переходим.

Напомним, что каждому значению x_i, соответствует распределение *y:* y_j , N_{ij}, где $j=\overline{1,l}$. Такие распределения называют условными, условными называют и соответствующие средние

$$\overline{y_i}=\frac{\sum_{j=1}^{l} y_j N_{ij}}{N(x_i)},(i=\overline{1,k}) \qquad \text{(III,1,1)}$$

Полную среднюю $\bar{y}$ можно рассматривать как взвешенную сумму условных средних $\overline{y_i}$.

Упражнение 71. Показать, что $\bar{y}$, равное, по определению, $\frac{1}{N}\sum_{j=1}^{l} y_i N(y_j)$ равно $\frac{1}{N}\sum_{i=1}^{k}\overline{y_i}N(x_i)$.

Далее мы будем изучать связь $\overline{y_i}$, с x_i. Если ее можно представить в виде $\bar{y}=f(x_i)$, где $f(x)$ – некоторая известная функция, то уравнение $\overline{y_i}=f(x)$, следуя Гальтону, называют *уравнением регрессии У на X*, а соответствующую ему кривую – *кривой регрессии*[1]. *С* таким уравнением мы уже встречались в примере 42 (§1 главы II).

[141]

Аналогично (III,1,1) определяется условная средняя

$$\overline{x_i}=\frac{\sum_{i=1}^{i} x_i N_{ij}}{N(y_j)}, \qquad \text{(III,1,2)}$$

соответствующая y_j (III, 1,2).

[1] Индекс *x* показывает, что речь идет об условном среднем.

Упражнение 72. Показать, что $\overline{x}$ является взвешенной суммой условных средних $\overline{x_j}$; т.е. что

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{i} N(y_j)\overline{x_j}, \qquad \text{(III,1,3)}$$

Уравнение $\overline{x_y} = \varphi(y)$ называется уравнением регрессии *X* на *Y*. Подчеркнем, что, вообще говоря, обе регрессии – *Y* на *X* и *X* на *Y* – различны; влияния *X* на *Y* и *Y* на *X* не одинаковы. Следовательно, функции *f* и φ не являются взаимно обратными.

Пример 27. В ряде случаев связь удается представить в виде линейной зависимости типа $\overline{y_x} = ax + b$ и соответственно $\overline{x_y} = cy + d$.

Рассмотрим такую корреляционную таблицу для признаков *X* и *Y*.

X	Y					$N(x_i)$	$\overline{y_x}$
	y_1 =20	y_2 =30	y_3=40	y_4=50	y_5=60		
x_1=10	38	37	42	0	0	117	30,3
x_2=20	0	47	40	48	0	135	40,1
x_3=30	0	0	41	28	39	108	49,8
$N(y_j)$	38	84	123	76	39	360	---
$\overline{x_y}$	10,0	15,6	19,9	23,7	30,0	---	---

Упражнение 73. Вычислить $\overline{y_x}$ и $\overline{x_y}$ по данным таблицы примера 27 (ответы выписаны в соответствующей колонке и строке этой таблицы). Исходя из значений $\overline{y_x}$ и $\overline{x_y}$, приведенных в крайних маргиналах, можно записать приближенные равенства:

$$\overline{y_x} = x + 20 \qquad \text{(III,1,4)}$$

$$\overline{x_y} = 0{,}5 * y \qquad \text{(III,1,5)}$$

[142]

В дальнейшем мы рассмотрим нахождение уточненных уравнений регрессии, а сейчас подчеркнем, что уравнения (III 1,4) и (III,1,5) существенно различны: из одного нельзя получить другое. В этом, в частности проявляется специфика корреляционных связей. Иное дело – связи функциональные. Получаемые для опытных данных регрессии $\overline{y_x} = f(x)$ и $\overline{x_y} = \varphi(y)$, являющиеся выражением одной и той же функциональной связи, должны быть в случае надежных данных взаимно обратными. (Кстати, взаимная обратность функций *f* и φ является обычно критерием надежности эмпирического материала).

Наша задача заключается в нахождении уравнения регрессии. Как она решается, рассмотрим на примере прямой регрессии общего вида, а затем вернемся к нашему примеру.

Прямая регрессия

О прямой (точнее – прямолинейной) регрессии говорят в том случае, когда точки $(x_i, \overline{y_i})$ располагаются близко к некоторой прямой $y=ax+b$. Уравнение регрессии будет полностью известно, если мы найдем *a* и *b*. Естественным условием их нахождения является минимум отклонений эмпирических точек $(x_i, \overline{y_i})$ от прямой, являющейся линией регрессии.

Мерой отклонения опытных точек от прямой может служить величина дисперсии

$$S=\frac{1}{N}\sum_{i=1}^{k}N(x_i)(\overline{y_i}-y_i)^2, \qquad \text{(III,1,6)}$$

где $y_i=ax_i+b$ – теоретическое значение Y, соответствующее x_i, а $\overline{y}_i$ –эмпирическое среднее, определяемое соотношением (III,1,1).

В S-отклонение $\overline{y}_i$ от y_i входит: 1) в квадрате, так как не должны компенсироваться отклонения разных знаков; 2) со своим «удельным весом» $\frac{N(x_i)}{N}$.

У нас $S=S(a, b)$. Параметры a и b найдем из условия минимума S, т.е. суммы квадратов отклонений (отсюда и название способа – «метод наименьших квадратов»).

Представим уравнение регрессии $y=ax+b$ в виде

$$y-\overline{y}=a(x-\overline{x})+c, \qquad \text{(III,1,7)}$$

где $c=b-\overline{y}+a\overline{x}$.

[143]

Теперь

$$S=S(a,c)=\frac{1}{N}\sum_{i=1}^{k}N(x_i)\left[\overline{y}_i-\overline{y}-a(x_i-\overline{x})-c\right]^2, \qquad \text{(III,1,8)}$$

и задача свелась к нахождению a и c, обеспечивающих минимум S.

S можно рассматривать как взвешенную сумму квадратов отклонений величины $\overline{y}_i-\overline{y}-a(x_i-\overline{x})$ от c. Согласно четвертому свойству дисперсии (§ 3 главы 1) S достигает минимума, когда c равно среднему значению величины $\overline{y}_i-\overline{y}-a(\overline{x}_i-x)$, т.е.

$$c=\frac{1}{N}\sum_{i=1}^{k}N(x_i)[\overline{y}_i-\overline{y}-a(x_i-\overline{x})]=0.$$

Здесь мы использовали соотношения (III,1,1) и определения $\overline{x}$ и $\overline{y}$, Величину a нужно найти из условия минимума

$$S(a,0)=\frac{1}{N}\sum N(x_i)[\overline{y}_i-\overline{y}-a(x_i-\overline{x})]^2 \qquad \text{(III,1,9)}$$

Читатель, знакомый с элементами высшей математики, легко поймет, что условие минимума $\frac{\partial S}{\partial a}=0$ принимает вид

$$\sum N(x_i)[\overline{y}_i-\overline{y}-a(x_i-\overline{x})](x_i-\overline{x})=0 \qquad \text{(III,1,10)}$$

Откуда

$$a=\frac{\sum N(x_i)(\overline{y}_i-\overline{y})(x_i-\overline{x})}{\sum N(x_i)(x_i-\overline{x})^2} \qquad \text{(III,1,11)}$$

Упражнение 74, Убедиться, что при этом $\frac{\partial^2 S}{\partial a^2}>0$, т.е. действительно имеет место минимум.

Для читателя, не знакомого с высшей математикой, заметим, что a можно найти также с помощью соображений, основанных на элементарной математике.

Действительно, перепишем S в виде

$$[\frac{1}{N}\sum N(x_i)(x_i-\overline{x})^2]a^2-2[\frac{1}{N}\sum N(x_i)(\overline{y_i}-\overline{y})(x_i-\overline{x})]a+$$

$$+\frac{1}{N}\sum N(x_i)(\overline{y}_i-\overline{y})^2=Aa^2-2Ba+B=$$

$$=A\left(a-\frac{B}{A}\right)^2B-\frac{B^2}{A}$$

[144]

(где смысл обозначений *A*, *B*, *D* очевиден). Минимальное значение *S*, равное $D-\frac{B^2}{A}$, достигается при

$$a=\frac{B}{A}=\frac{\sum N(x_i)(\overline{y_i}-\overline{y})(x_i-\overline{x})}{\sum N(x_i)(x_i-\overline{x})^2}.$$

Тем самым мы независимо обосновали справидливость (III,1,11).

Это обстоятельство будет использовано в дальнейшем.

Зная *a,* из условия *c*=0 можно найти *b*:

$$b=\overline{y}-a\overline{x}. \qquad \text{(III,1,12)}$$

Тем самым полностью определено уравнение регрессии. Обратим внимание на то, что мы здесь фактически доопределили, уточнили понятие уравнения регрессии. Раньше таковым называлось уравнение $\overline{y_i}=f(x_i)$ (в случае регрессии *Y* на *X*). Теперь мы видим, что уравнение регрессии описывает кривую, отклонение эмпирических точек ($x_i, \overline{y_i}$) от которой минимально. Ясно, что задача отыскания «точной» кривой, на которой лежат эти точки, и очень сложна и нецелесообразна. Доопределенное уравнение регрессии, способ нахождения которого здесь рассмотрен, позволяет сравнительно просто и надежно судить о форме связи между переменными.

Упражнение 75. По данным корреляционной таблицы примера 27 найти уравнения регрессии *Y* на *X* и *X* на *Y*.

Указание. Использовать формулы (III,1, 11), (III,1,12).

Ответ: $\overline{y_x}=0{,}974\cdot x+20{,}58$ (III,1,4')

$\overline{x_y}=0{,}468\cdot y+1{,}11$ (III,1,5')

Эти соотношения являются уточнением уравнений (III,1,4), (III,1,5), которые были получены «на глазок».

Теперь уравнение регрессии (III,1,7) принимает вид:

$$y-\overline{y}=r\frac{\sigma_y}{\sigma_x}(x-\overline{x}), \qquad \text{(III,1,13)}$$

где

$$r=\frac{1}{N\sigma_x\sigma_y}\sum_i N(x_i)(\overline{y_i}-\overline{y})(x_i-\overline{x}).$$

Упражнение 76. Показать, что:

1. $$r=\frac{1}{N\sigma_x\sigma_y}\sum_i\sum_j N_{ij}(\overline{y_j}-\overline{y})(x_i-\overline{x}) \qquad \text{(III,1,14)}$$

[145]

2. $r = \frac{1}{N\sigma_x\sigma_y}(\overline{xy} - \bar{x}\cdot\bar{y})$ (III,1,15)

3. $r = \frac{1}{N\sigma_x\sigma_y}\sum_j N(y_i)(y_j - \bar{y})(\overline{x_i} - \bar{x}).$ (III,1,16)

Рассмотрим, например,

$$\sum_{i,j} N_{ij}(y_j - \bar{y})(x_i - \bar{x}) = \sum_i (x_i - \bar{x})\sum_j N_{ij}(y_j - \bar{y}) =$$

$$= \sum_i N(x_i)(x_i - \bar{x})(\overline{y_i} - \bar{y}).$$

Таким образом, мы показали, что в уравнение регрессии входит ранее определенная величина r (§ 5 главы II) и тем самым пришли к парному коэффициенту корреляции из теоретических соображений.

Наиболее простую интерпретацию r допускает в так называемых нормальных координатах. Введем $t_x = \frac{x - \bar{x}}{\sigma_x}$ и $t_y = \frac{y - \bar{y}}{\sigma_y}$. Новые переменные безразмерны, имеют нулевые средние и единичные дисперсии. Они не зависят от масштаба.

В этих переменных уравнение регрессии принимает вид:

$t_y = rt_x.$ (III,1,17)

Таким образом, r показывает, на сколько изменяется зависимая переменная при изменении независимой на единицу. Величина $\rho_{yx} = r\frac{\sigma_y}{\sigma_x}$ угловой коэффициент уравнения регрессии Y на X.

Упражнение 77. Показать, что регрессия X на Y имеет вид:

$$x - \bar{x} = r\frac{\sigma_y}{\sigma_x}(y - \bar{y}).$$ (III,1,18)

Теперь $\rho_{xy} = r\frac{\sigma_x}{\sigma_y}$. Ясно, что произведение угловых коэффициентов в уравнениях регрессии Y на X и X на Y равно квадрату коэффициента парной корреляции, а регрессии совпадают только в том случае, когда $|r| = 1$.

При подстановке в уравнение регрессии координат точек $(x_i, \overline{y_i})$ мы получим точное равенство только в том случае, когда все эмпирические точки лежат на одной прямой. На практике этого не бывает и равенство $\overline{y_i} - y = \rho_{yx}(x_i -$

$-\overline{x})$ выполняется приближенно. В качестве меры точности естественно принять среднеквадратическую погрешность, т.е. квадратный корень из отклонения (дисперсии). Мера точности, таким образом $\sqrt{S_{\min}}$, где

$$S_{min} = D - \frac{B^2}{A} = \frac{1}{N}\sum N(x_i)(\overline{y}_i - \bar{y})^2 -$$

$$- \frac{\left[\sum N(x_i)(x_i - \bar{x})(\overline{y_i} - \bar{y})\right]^2}{N\sum N(x_i)(x_i - \bar{x})^2},$$ (III,1,19)

если учесть определения *A*, *B* и *D*.

До сих пор мы рассматривали прямолинейную регрессию, используя метод наименьших квадратов. Этот метод может быть применен и для изучения криволинейной зависимости. В некоторых случаях не потребуется решать криволинейную задачу, ее можно свести к рассмотренной. Для этого используется замена переменных.

Мы приведем интересный социально-демографический пример в форме своеобразного *упражнения* (*№78*): часть выкладок читателю предстоит выполнить самостоятельно. (Впрочем, понять смысл рассматриваемого примера можно и не прибегая к несколько громоздким, хотя и несложным выкладкам, которые предлагаются читателю по ходу изложения материала).

Пример 28. В 1965 г. И. С. Шкловским был установлен гиперболический закон роста численности населения земного шара на материале статистики с 1600 г. по 1960 г. Математически он выглядит так: $y(x)=\frac{A}{B-x}$, где x – календарное время, $y(x)$ – численность населения, а A и B – параметры уравнения. Статистический материал, которым располагал Шкловский[2], приведен в табл. 33.

Сделаем замену переменных: перейдем от x к $X'=x-x_0$,, где x_0 – начало отсчета времени, т.е. 1600, и от y к $Y'=\frac{1}{y}$.

Построив график Y'=Y'(X'), видим, что все точки тесно группируются возле прямой линии. (Убедитесь самостоятельно. Именно здесь начинается для читателя само-упражнение. Кстати, постройте график y=y(x), убедитесь, что точки ложатся на гиперболу).

[147]

В силу сказанного, станем искать $Y'(X')$ в виде $-aX'+b$, используя метод наименьших квадратов. (Знак минус показывает, что с ростом X' величина Y' уменьшается – так и должно быть: ведь $Y'=\frac{1}{y}$).

Таблица 33

Численность населения земного шара

Год	Численность (млн. чел)	Рассчитанная численность (млн. чел)	Отклонения (%)
1600	486	481	1,0
1650	545	545	0
1700	617	627	-1,7
1750	728	739	-1,6
1800	906	900	0,6
1850	1171	1150	1,8
1900	1608	1592	1,0
1910	-	1725	-
1920	1861	1882	-1,1
1930	2070	2070	0
1940	2295	2300	0,2
1950	2517	2588	-2,8
1960	3010	2958	1,7

[2] Таблица заимствуется из книги: *Суслов И. П., ГражданниковЕ.*Д. Основы социальной статистики, Новосибирск, 1973 (мы несколько уточнили приведенные авторами расчеты и устранили имеющиеся опечатки).

Теперь

$$a = \frac{\sum X_i' \sum Y_i' - N \sum X_i' Y_i'}{N \sum (X_i')^2 - (\sum X_i')^2}$$

$$b = \frac{\sum (X_i')^2 \sum Y_i' - \sum X_i' \sum X_i Y_i'}{N \sum (X_i')^2 - (\sum X_i')^2}$$

Это несложно показать, если внимательно рассмотреть материал данного параграфа. Для каждого x_i, можно вычислить x'_i и y'_i, и, следовательно, найти a и b (сделайте это),

Теперь $y(x) = \frac{A}{B - x}$, где $A = \frac{1}{a}$, $B = x_0 + \frac{b}{a}$

После соответствующих вычислений получим: A = 207052, B = 2030, т.е. окончательно:

$y(x) = \frac{207052}{2030 - x}$ — закон Шкловского.

Найдем расчетную численность. Эти данные приводятся в таблице (колонка 3). Подсчет относительных отклонений

[148]

показывает, что они не превосходят по абсолютной величине 2,8 (колонка 4).

Итак, получено теоретическое уравнение. Читатель вправе задать вопрос: «Ну и что? Для чего это уравнение? Что оно дает нам? Значения y_i, которые были известны заранее, да и то, как видно из таблицы, приближенно?!»

Попытаемся ответить. Мы установили закономерность, которой подчиняется эмпирический материал, а знание закономерности может стать источником новых сведений. Но экстраполируя данные, полученные с помощью формулы, на прошлое и будущее, нужно помнить, что наши предсказания будут тем надежней, чем меньше выбираемый интервал. Например, из формулы Шкловского следует, что к 2030 г. население должно стать бесконечно большим. Этот результат, конечно, не имеет, как принято говорить, «физического» смысла, что отнюдь не свидетельствует о неправильности формулы. Просто нужно помнить, что обычно закономерности относятся ко вполне определенным условиям, что устанавливаемые формулы имеют границы применимости. Так, мы с достоверностью не можем, зная закон Шкловского, вычислить величину народонаселения, скажем, в 1500 или 2000 году. Расчеты для 1970 и 1980 годов по этой формуле дают 3450 и 4140 млн. человек, что на 5,1 и 6,3% ниже реальной численности (3635 и 4415 млн. соответственно). Хотя ошибка несколько возрастает, формула дает очень хорошее приближение к реальным данным.

Можно предположить, что в ближайшие десятилетия мы станем свидетелями изменения темпов роста населения земного шара – закон перестанет быть гиперболическим. Это, само собой, нисколько не опровергает формулу Шкловского, установленную для рассмотренных временных интервалов. Отметим, что она дает возможность определять численность населения в те годы внутри изученного интервала, для которых статистика отсутствует или ненадежна. Так, в 1910 г. население примерно составляло 1725 млн. человек и т.д.

Корреляционное отношение

Вернемся, однако, к рассмотрению регрессий. В случае *криволинейной* зависимости целесообразно использовать так называемое корреляционное отношение

$$\eta_{нч} = \frac{\sqrt{\frac{1}{N}\sum N(x_i)(\overline{y_i} - \overline{y})^2}}{\sigma_y}, \qquad \text{(III,1,20)}$$

[149]

которое, по определению, представляет собой отношение среднего квадратического отклонения условных средних $\overline{y_i}(\sigma_{\bar{y}})$ к полному среднему квадратическому отклонению (σ_y): $\eta_{yx} = \sigma_{\bar{y}} / \sigma_y$ (см. § 5 главы II).

С учетом (III,1,13), (I1I,1,19), (III,1,20)

$\min S = \sigma_y^2(\eta_{yx}^2 - r^2)$.

Так как, по определению, $S \geq 0$, то $\eta_{yx}^2 \geq r^2$ или $\eta_{yx} \geq |r|$. Итак, *minS=0,* если все (x_i, $\overline{y_i}$) лежат на одной прямой, т.е. регрессия *Y* на *X* прямолинейная. Таким образом, равенство является условием того, что регрессия прямолинейная. Во всех остальных случаях (криволинейная зависимость!)

$\eta_{yx} > |r|$.

Мы видели (§ 4 главы III), что $0 \leq \eta \leq 1$. Можно аналогично показать, что $-1 \leq r \leq 1$. Доказательство справедливости этого утверждения составит содержание следующего *упражнения.*

Упражнение 79.

Указание. Использовать очевидное неравенство

$$\sum_{i,j} N_{ij}\left[y_i - \overline{y} - r\frac{\sigma_y}{\sigma_x}(x_i - \overline{x})\right]^2 \geq 0,$$

преобразуя его к виду $\sigma_y^2(1 - r^2)$. Тогда $1 - r^2 \geq 0$, т.е. $|r| \leq 1$.

Итак, мы нашли диапазон возможных значений, принимаемых *r* и *η*, выяснили условие того, что регрессия прямолинейная и нашли меру криволинейной связи (*η*). Так как обычно связи криволинейные, следует обратить особое внимание на корреляционное отношение.

К сожалению, в социологической литературе, как уже отмечалось, наблюдается злоупотребление коэффициентом *r*, который вычисляется без обоснования правомерности его использования. Лишь в редких случаях исследователи применяют *η*, хотя ситуация должна быть обратной.

Упражнение 80. Показать, что в случае корреляционной таблицы:

$$\eta_{yx}^2 = \frac{N\sum\frac{1}{N(x_i)}(\sum N_{ij}y_j)^2 - \left[\frac{1}{N}\sum N(y_i)y_i\right]^2}{N\sum N(y_j)y_j^2 - \left[N^{-1}\sum N(y_j)y_j\right]^2} \qquad \text{(III,1,21)}$$

[150]

Вернемся к рассмотрению $\eta_{yx} = \sigma_{\bar{y}} / \sigma_y$. Стоящая в числителе величина $\sigma_{\bar{y}}$ описывает колеблемость *Y* под влиянием *X*. σ_y описывает полную колеблемость величины *Y* под влиянием всех условий. Следовательно, η_{yx} показывает, какую часть общей изменчивости *Y* обусловливает влияние *X*. Это отношение выявляет степень воздействия *X* на *Y*.

Таблица 34

Пример расчета корреляционного отношения

	Выполнение нормы выработки, % (*Y*)	$N(x_i)$

Возраст, лет (X)	Выполнение нормы выработки, % (Y)					$N(x_i)$
	95-100	100-105	105-110	110-115	115-120	
19-22	5	7	2	4	4	22
22-25	1	7	2	3	12	21
25-28	3	2	2	8	13	28
28-31	1	1	3	1	5	11
31-34	0	0	3	5	3	11
34-37	0	0	1	2	5	8
37-40	0	0	3	2	4	9
40-43	0	0	0	0	0	0
43-46	0	0	0	0	1	1
46-49	0	0	0	0	0	0
49-52	0	1	0	0	0	1
$N(y_j)$	10	14	16	25	47	112

Аналогично η_{yx} может быть определена величина η_{xy} , которая характеризует воздействие Y на X:

$$\eta_{xy} = \frac{\sigma_{\bar{x}}}{\sigma_x} \qquad \text{(III,1,22)}$$

Вообще говоря $\eta_{xy} \neq \eta_{yx}$, ибо воздействия X на Y и Y на X неравнозначны. Поэтому целесообразно вычислять оба корреляционных отношения, если они имеют содержательный смысл. Для Y – производительности труда рабочих, а X – стажа значение η_{yx} можно рассматривать как степень влияния стажа на производительность, корреляционное отношение η_{xy} в данном случае интерпретировать нельзя.

Упражнение 81. Записать выражение для η_{xy}.

Упражнение 82. Для таблицы 34 рассчитать корреляционное отношение[3]. Указание: Для вычислений удобно

[151]

перейти к $x' = \frac{x-a}{\alpha_x}$ и $y' = \frac{x-b}{\alpha_y}$, полагая a=35,5; b=107,5; α_x=3, α_y=5 (убедиться, что η при этом не изменится!)

В новых переменных x'_i, y'_j корреляционное отношение

$$\eta^2_{yx} = \frac{\sum_{i=1}^{11}\left[N(x_i)\right]^{-1}\left(\sum_{j=1}^{5} y'_j N_{ij}\right)^2 - \frac{1}{N}\left[\sum_{j=1}^{5} y'_j N(y_j)\right]^2}{\sum_{j=1}^{5} y'^2_j N(y_j) - \frac{1}{N}\left[\sum_{j=1}^{5} y'_j N(y_j)\right]^2}$$

С учетом данных таблицы имеем:

η_{yx}=0,41.

(Читатель, испытывающий затруднения при вычислении этого коэффициента, может обратиться к с.150 – 151 «Методики и техники…», где найдет подробные выкладки.

Упражнение 83. По данным последней таблицы рассчитать r. Для этой цели удобно использовать формулу (11,5,4). Ответ: 0,21.

[3] Данные заимствованы из «Методики и техники...», с.150.

Итак, $r<\eta$. Связь нелинейная[4]. Для установления ее формы целесообразно построить эмпирическую кривую регрессии по точкам $(x_i, \overline{y_i})$. Эта работа составит содержание *упражнения 84*.

2. Частная корреляция. Случай трех признаков

Наличие статистической связи между двумя величинами может быть следствием связи обеих с некоторой третьей (либо совокупностью некоторых величин). Следовательно, возникает необходимость устранить влияние «третьих» величин. Заметим, что в простейшем случае этого можно достичь, изучая связи между двумя данными величинами в совокупности однородных объектов (при фиксированном «третьем» признаке). Однако для такой процедуры необходимы большие общности, особенно если устраняется влияние не одного, а нескольких признаков. Для изучения связи в таких ситуациях служит специальный аппарат частной корреляции. Рассмотрим принципиальную схему метода. Если корреляция данных признаков уменьшается при устранении неко-

торого признака, то взаимозависимость выделенных признаков определяется, в частности, и этим признаком. В предельном случае, когда устранение обращает коэффициент корреляции в нуль, можно считать, что этот признак обусловливает изучаемую связь.

Если при устранении коэффициент корреляции увеличивается, то данный признак ослабляет связь. Если же коэффициент корреляции практически не меняется, то соответствующий признак на связь не влияет.

Рассмотрим одну содержательную задачу. При изучении связи между производительностью труда и возрастом было установлено наличие прямой корреляции. Но на производительность влияет и стаж работы, который тоже оказывается в прямой корреляции с возрастом и с производительностью. Чтобы выяснить, прямая или обратная связь между производительностью и собственно возрастом, нужно, очевидно, устранить влияние стажа. Решить этот вопрос, непосредственно сопоставляя между собой три полученных парных коэффициента корреляции, невозможно. (Забегая вперед, укажем, что связь между производительностью и возрастом при устранении стажа оказалась отрицательной, а между производительностью и стажем при устранении возраста положительной, но более тесной).

Перейдем к рассмотрению техники частной корреляции, ограничившись для простоты выкладок случаем трех признаков. (Рассмотрение общего случая не потребует новых идей, хотя и оказывается значительно более громоздким).

Допустим, что изучаемая совокупность из N объектов может быть описана с помощью количественных признаков Y, X_1, и X_2. (Во избежание недоразумений подчеркнем, что теперь X_i, – не i-ое значение признака, как было раньше, а сам i-ый признак $(i=1, 2)$, который может, в свою очередь, принимать ряд различных значений). Если признак Y принимает m различных значений y_g $(g=\overline{1,m})$, то $\overline{y} = \frac{1}{N}\sum_{g=1}^{m} N_g y_g$, где N_g – число индивидов, у которых $Y=y_g$. Обозначим через $\overline{x_{ig}}$ среднее значение признака X_i, у индивидов с $Y=y_g$, тогда $\overline{x_i} = \frac{1}{N}\sum_{g=1}^{m} N_g \overline{x_{ig}}$.

[4] Значимость отклонения от линейности определяется с помощью критерия Фишера *(Закс* Л. Статистическое оценивание. М., 1976, с. 401).

Если, например, Y – квалификация, а X_1 – возраст рабочих некоторого коллектива из N индивидов, то $y_g = g\ (\ g = \overline{1,6}\ )$ – тарифно-квалификационный разряд (для

[153]

определенности, предполагается, что сетка имеет 6 разрядов). N_g – число рабочих, у которых разряд g, $\overline{x_{1g}}$ – средний возраст рабочих с разрядом g, а $\overline{x_1}$ – средний возраст рабочих данного коллектива. Аналогично интерпретируется $\overline{x_{2g}}, \overline{x_2}$, если X_2, скажем, стаж работы и т.д. Найдем линейную зависимость Y от X_i (i=1, 2), которая удовлетворяет принципу наименьших квадратов. Для этого введем величину $\delta y = y - \overline{y}$ и $\delta x = x - \overline{x}$. Теперь указанная выше зависимость, по аналогии с предыдущим, может быть представлена в виде $\delta y = a_1 \delta x_1 + a_2 \delta x_2$.

Найдем a_i (i=1, 2) из условия минимума суммы квадратов отклонений.

$$S = S(a_1 a_2) = \sum_g N_g (\delta y_g - a_1 \delta x_{1g} - a_2 \delta x_{2g})^2,$$

условия минимума по аналогии с (III,1,10) принимают вид:

$$\begin{cases} \sum N_g (\delta y_g - a_1 \delta x_{1g} - a_2 \delta x_{2g}) \delta x_{1g} = 0 \\ \sum N_g (\delta y_g - a_1 \delta x_{1g} - a_2 \delta x_{2g}) \delta x_{2g} = 0 \end{cases}$$

Или:

$$\begin{cases} a_1 \sum N_g \delta x_{1g}^2 + a_2 \sum N_g \delta x_{1g} \delta x_{2g} = \sum N_g \delta x_{1g} \delta y_g \\ a_1 \sum N_g \delta x_{1g} \delta x_{2g} + a_2 \sum N_g \delta x_{2g}^2 = \sum N_g \delta x_{2g} \delta y_g \end{cases}$$

С использованием определения a получаем:

$$\begin{cases} a_1 \sigma_1^2 + a_2 \sigma_1 \sigma_2 \dfrac{\sum N_g \delta x_{1g} \delta x_{2g}}{N \sigma_1 \sigma_2} = \sigma_0 \sigma_1 \dfrac{\sum N_g \delta x_{1g} \delta y_g}{N \sigma_0 \sigma_1} \\ a_1 \sigma_1 \sigma_2 \dfrac{\sum N_g \delta x_{2g} \delta x_{1g}}{N \sigma_1 \sigma_2} + a_2 \sigma_2^2 = \sigma_0 \sigma_2 \dfrac{\sum N_g \delta x_{2g} \delta y_g}{N \sigma_0 \sigma_2} \end{cases}$$

Но $\dfrac{\sum N_g \delta x_{1g} \delta x_{2g}}{N \sigma_1 \sigma_2} = r_{12}$ – коэффициент корреляции признаков X_1 и X_2, а $\dfrac{\sum N_g \delta y_g \delta x_{ig}}{N \sigma_0 \sigma_i} = r_{0i}$ – признаков Y и X_i (индекс 0 соответствует Y).

Имеем линейную систему:

$$\begin{cases} a_1 \sigma_1 + a_2 \sigma_2 r_{12} = \sigma_0 r_{01} \\ a_1 \sigma_1 r_{12} + a_2 \sigma_2 = \sigma_0 r_{02} \end{cases}$$

[154]

которая решается очень просто:

$$a_1 = \frac{\sigma_0}{\sigma_1} \frac{r_{01} - r_{02} r_{12}}{1 - r_{12}^2}, \qquad \text{(III,2,1)}$$

$$a_2 = \frac{\sigma_0}{\sigma_2} \frac{r_{02} - r_{01} r_{12}}{1 - r_{12}^2} \qquad \text{(III,2,2)}$$

Теперь полная регрессия Y на X_1 и X_2 имеет вид:

$$y-\overline{y}=\frac{\sigma_0}{\sigma_1}\frac{r_{01}-r_{02}r_{12}}{1-r_{12}^2}(x-\overline{x_1})+\frac{\sigma_0}{\sigma_2}\frac{r_{02}-r_{01}r_{12}}{1-r_{12}^2}(x-\overline{x_2})$$

Допустим, что X_2 фиксировано; обозначая новые средние через $\overline{y'}$ и $\overline{x'}$, получим:

$$y-\overline{y'}=\frac{\sigma_0}{\sigma_1}\frac{r_{01}-r_{02}r_{12}}{1-r_{12}^2}(x-\overline{x_1'})=A(x-\overline{x_1'})$$

Аналогично для регрессии X на Y:

$$x_1-\overline{x_1'}=\frac{\sigma_1}{\sigma_0}\frac{r_{01}-r_{02}r_{12}}{1-r_{12}^2}(y-\overline{y'})=B(y-\overline{y'})$$

Упражнение 85. Вывести уравнение регрессии X на Y.

Так как произведение коэффициентов регрессии равно квадрату коэффициента корреляции (§ 1, глава III), то, обозначая коэффициент корреляции Y и X_2 при фиксировании X_2 через $r_{01\cdot2}$, получим: $r^2_{01\cdot2}=AB$

Отсюда с учетом очевидных обозначений А и В имеем:

$$r_{01\cdot2}=\frac{r_{01}-r_{02}\cdot r_{12}}{\sqrt{\left(1-r_{02}^2\right)\left(1-r_{12}^2\right)}} \qquad \text{(III,2,3)}$$

Обратим внимание на то, что если связи между X_2 и X_1, с одной стороны, и X_2 и Y, с другой, нет, то $r_{01\cdot2}=r_{01}$, как и следовало ожидать. Полученное выражение является, таким образом, обобщением коэффициента корреляции между двумя признаками (X_1, на Y), если на них влияет третий (X_2). Соотношение (II1,2,3) позволяет определить корреляцию между признаками Y и X_1 при устранении влияния X_2.

Аналогично:

$$r_{02\cdot1}=\frac{r_{02}-r_{01}\cdot r_{12}}{\sqrt{\left(1-r_{01}^2\right)\left(1-r_{12}^2\right)}} \qquad \text{(III,2,4)}$$

$$r_{12\cdot0}=\frac{r_{12}-r_{01}\cdot r_{02}}{\sqrt{\left(1-r_{01}^2\right)\left(1-r_{02}^2\right)}} \qquad \text{(III,2,5)}$$

[155]

Рассмотренные коэффициенты называются коэффициентами корреляции первого порядка (устраняется один признак).

В случае четырех признаков: Y, X_1, X_2, X_3 наряду с коэффициентами рассмотренных типов (r_{01}, r_{12}, $r_{01\cdot2}$ и т.д.) появляются коэффициенты корреляции второго порядка: например, $r_{01\cdot23}$ – коэффициент частной корреляции признаков Y и X_1 при устранении влияния X_2 и X_3.

Устраним сперва влияние X_3, вычислив $r_{01\cdot3}$, $r_{02\cdot3}$, $r_{12\cdot3}$, а затем влияние X_2, по ранее рассмотренной схеме. Тогда получим

$$r_{01\cdot23}=\frac{r_{01\cdot3}-r_{02\cdot3}\cdot r_{12\cdot3}}{\sqrt{\left(1-r_{02\cdot3}^2\right)\left(1-r_{12\cdot3}^2\right)}} \qquad \text{(III,2,6)}$$

Можно было сперва устранить влияние X_2 , а затем X_3.

Упражнение 86. 1. Записать $r_{01\cdot23}$ в этом случае. 2. Записать $r_{02\cdot13}$ и $r_{12\cdot03}$.

В случае пяти признаков порядок рассмотрения сохраняется, число коэффициентов резко увеличивается. В заключение напомним еще раз, что речь идет о количественных признаках, и все рассмотрение проводилось в предположении линейности связей, а это существенно сужает область применимости данных коэффициентов.

Техника частных корреляций оказывается неприменимой для коэффициентов взаимной сопряженности, ранговой корреляции Спирмена. Однако установлено, что имеет смысл расчет частных корреляций для коэффициента Кендэла. Любопытно, что формулы

элиминирования оказываются аналогичными полученным для *r*. Так, чтобы исключить влияние X_2 на взаимодействие X_1 с Y (случай трех признаков), достаточно рассчитать

$$\tau_{01\cdot 2} = \frac{\tau_{01} - \tau_{02} \cdot \tau_{12}}{\sqrt{\left(1-\tau_{02}^2\right)\left(1-\tau_{12}^2\right)}} \qquad \text{(III,2,7)}$$

и т.д.

3. Множественная регрессия. Случай трех признаков

Частные коэффициенты корреляции, рассмотренные в предыдущем параграфе, выражают связь между результативным признаком («зависимая» переменная) и одним из

[156]

факторов («независимая» переменная) в случае, когда остальные факторы остаются неизменными.

Представляет интерес выявление влияния нескольких признаков (факторов) на результативный. В общем случае это очень сложная задача, которая имеет относительно простое решение, если зависимости линейные.

Рассмотрим для простоты случай трех признаков, который, однако, позволяет понять принцип анализа множественной регрессии в общем случае. Как и в предыдущем параграфе, станем рассматривать признаки Y – результирующий – и факторные X1 и X2. Будем исследовать корреляцию между Y и $U=a_1X_1+a_2X_2$, т.е. признаком, который представляет собой линейную комбинацию факторных. Для этого введем

$$\delta u_g = u_g - \bar{u}, \qquad \text{(III,3,1)}$$

где $g=\overline{1,m}$, как и ранее,

$$u_g = a_1 x_{1g} + a_2 x_{2g}, \qquad \text{(III,3,2)}$$

$$\bar{u} = a_1 \bar{x}_{1g} + a_2 \bar{x}_2 \qquad \text{(III,3,3)}$$

По определению дисперсии

$$\sigma_u^2 = \frac{1}{N} \sum_g N_g (\delta u_g)^2 \qquad \text{(III,3,4)}$$

Естественно определить коэффициент корреляции между Y и U:

$$R = \frac{\sum_g N_g \delta y_g \delta u_g}{N \sigma_0 \sigma_u} \qquad \text{(III,3,5)}$$

Найдем связь между R и r_{ih} *(i, h* =1, 2*)*. Из (III,3,2) и (III,3,3)

$$\delta u_g = a_1 \delta x_{1g} + a_2 \delta x_{2g} \qquad \text{(III,3,6)}$$

Теперь числитель R равен

$$a_1 \sum_g N_g \delta x_{1g} \delta y_g =$$

$$= \frac{N \sigma_0^2}{1 - r_{12}^2} \left(r_{01}^2 + r_{02}^2 - 2 r_{01} r_{02} r_{12} \right),$$

если использовать (Ш,2,1), (111,3,2).

[157]

Далее. С учетом (III,3,4) и (III,3,6), а затем (III,2,1) и (III,2,2):

$$\sigma_u = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 - 2a_1a_2\sigma_1\sigma_2 r_{12} = \frac{\sigma_0^2}{1-r_{12}^2}(r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12})$$

Теперь

$$R = \sqrt{\frac{r_{01}^2 + r_{02}^2 - 2r_{01}r_{02}r_{12}}{1-r_{12}^2}}$$

Вспоминая, что

$$r_{02\cdot 1} = \frac{r_{02} - r_{01}r_{12}}{\sqrt{(1-r_{01}^2)(1-r_{12}^2)}}$$

Мы можем переписать R в виде

$$R = R_{01\cdot 2} = \sqrt{1-(1-r_{01}^2)(1-r_{02\cdot 1}^2)} \qquad \text{(III,3,7)}$$

Индекс у R означает, что коэффициент описывает суммарное влияние признаков X_1 и X_2 на Y.

В случае четырех признаков

$$R_{0\cdot 123} = \sqrt{1-(1-r_{01}^2)(1-r_{02\cdot 1}^2)(1-r_{03\cdot 12}^2)}$$

Заметим, что возможности применения R крайне ограничены, так как линейность встречается в социологии очень редко.

[158]