4. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

4.1. Задачи и проблемы корреляционного анализа

Главной задачей корреляционного анализа является оценка взаимосвязи между переменными величинами на основе выборочных данных.

Различают два вида зависимостей между экономическими явлениями: функциональную и стохастическую. При функциональной зависимости имеет место однозначность отображения множества значений изучаемых величин, т.е. существует правило ***y=f(x)*** - соответствия независимой переменной ***x*** и зависимой переменной ***y***. В экономике примером функциональной связи может служить зависимость производительности труда от объема произведенной продукции и затрат рабочего времени.

При изучении массовых явлений зависимость между наблюдаемыми величинами проявляется часто лишь в случае, когда число единиц изучаемой совокупности достаточно велико. При этом каждому фиксированному значению аргумента соответствует определенный закон распределения значений функции и, наоборот, заданному значению зависимой переменной соответствует закон распределения объясняющей переменной. Например, при изучении потребления электроэнергии у в зависимости от объема производства ***x*** каждому значению ***x*** соответствует множество значений ***y*** и наоборот. В этом случае можно констатировать наличие стохастической (корреляционной) связи между переменными.

Множественность результатов при анализе связи ***x*** и ***y*** объясняется прежде всего тем, что зависимая переменная y испытывает влияние не только фактора ***x***, но и целого ряда других факторов, которые не учитываются. Кроме того, влияние выделенного фактора может быть не прямым, а проявляется через цепочку других факторов.

При изучении **корреляционной зависимости** между переменными возникают следующие задачи:

1. Измерение силы (тесноты) связи.
2. Отбор факторов, оказывающих наиболее существенное влияние на результативный признак.
3. Обнаружение неизвестных причин связей.
4. Построение корреляционной модели и оценка ее параметров.
5. Проверка значимости параметров связи.
6. Интервальное оценивание параметров связи.

Пусть из генеральной совокупности, которую образуют “***k***” признаков, являющихся случайными величинами, сделана выборка объемом ***n***, тогда выборка будет представлять собой ***n*** независимо наблюдаемых ***k***-мерных точек (векторов):

(x_{i1}, x_{i2},…,x_{ij},…x_{ik}), где ***i=1÷n***, а каждая координата x_{ij} наблюдаемой точки является вариантом соответствующего признака x_j ***(j=1÷k)*** генеральной совокупности, изучаемой с точки зрения взаимозависимости ***k*** признаков.

В настоящее время при построении корреляционных моделей исходят из условия нормальности многомерного закона распределения генеральной совокупности. Эти условия обеспечивают линейный характер связи между изучаемыми признаками, что делает правомерным использование в качестве показателей тесноты связи: парного, частного и множественного коэффициентов корреляции.

На практике не всегда строго соблюдаются предпосылки корреляционного анализа: один из признаков оказывается величиной не случайной или признаки не имеют совместного нормального распределения. Для изучения связи между признаками в этом случае существует общий показатель зависимости признаков, который называется корреляционным отношением.

В практике статистического анализа возможны случаи, когда с помощью корреляционных моделей обнаруживают достаточно сильную “зависимость” признаков, в действительности не имеющих причинной связи друг с другом. Такие корреляции называют ложными.

4.2. Двумерная корреляционная модель

Рассмотрим случай изучения корреляционной зависимости между двумя признаками Y и X. Построение двумерной корреляционной модели предполагает, что закон распределения двумерной случайной величины в генеральной совокупности является нормальным, а выборка репрезентативной.

Плотность двумерного нормального закона распределения образуется формулой:

$$\varphi(x,y)=\frac{1}{\sigma_x\sigma_y 2\pi\sqrt{1-\rho^2}}\cdot\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}-2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\}$$

и определяется пятью параметрами:

$MX=\mu_x$ - математическое ожидание X;

$MY=\mu_y$ - математическое ожидание Y;

$DX=\sigma^2_x$ - дисперсия X;

$DY=\sigma^2_y$ - дисперсия Y;

$\rho=M\left[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Y-\mu_y}{\sigma_y}\right]$ - парный коэффициент корреляции, характеризует тесноту линейной связи между величинами X и Y.

В двумерной корреляционной модели используется так же, как мера тесноты связи, ρ^2 - коэффициент детерминации, указывающий долю дисперсии одной случайной величины, обусловленную вариацией другой.

Для получения точечных оценок параметров двумерной корреляционной модели обычно используют метод моментов, т.е. в качестве точечных оценок неизвестных начальных моментов первого и второго порядков генеральной совокупности берутся соответствующие выборочные моменты, и расчеты производят в соответствии со следующими формулами:

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ - оценка для μ_x;

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ - оценка для μ_y;

$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}$ - оценка для $M(X^2)$;

$\overline{y^2} = \frac{\sum_{i=1}^{n} y_i^2}{n}$ - оценка для $M(Y^2)$;

$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ - оценка для M(XY);

$S_x^2 = \overline{x^2} - (\bar{x})^2$ - оценка для σ_x^2;

$S_y^2 = \overline{y^2} - (\bar{y})^2$ - оценка для σ_y^2;

$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y}$ - оценка для ρ.

Полученные оценки являются состоятельными, а $\bar{x}$ и $\bar{y}$ также обладают свойствами несмещенности и эффективности. Следует отметить, что в корреляционной модели распределение выборочных средних $\bar{x}$ и $\bar{y}$ не зависит от законов распределения S_x^2, S_y^2, r.

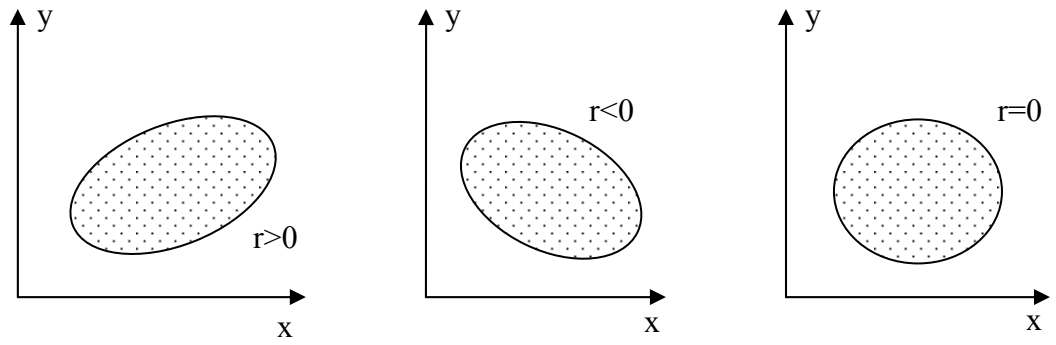
Парный коэффициент корреляции ρ в силу своих свойств является одним из самых распространенных способов измерения связи между случайными величинами в генеральной совокупности; для выборочных данных используется эмпирическая мера связи r.

Коэффициент корреляции не имеет размерности и, следовательно, его можно сопоставлять для разных статистических рядов. Величина его

лежит в пределах (-1 до +1). Значение ρ=±1 свидетельствует о наличии функциональной зависимости между рассматриваемыми признаками. Если ρ=0, можно сделать вывод, что линейная связь между ***x*** и ***y*** отсутствует, однако это не означает, что они статистически независимы. В этом случае не отрицается возможность существования иной формы зависимости между переменными. Положительный знак коэффициента корреляции указывает на положительную корреляцию, т.е. все данные наблюдения лежат вблизи прямой с положительным углом наклона в плоскости ***xy*** и с увеличением ***x*** растет ***y***. Когда ***x*** уменьшается, то ***y*** уменьшается. Отрицательный знак коэффициента свидетельствует об отрицательной корреляции. Чем ближе значение **|r|** к единице, тем связь теснее, приближение **|r|** к нулю означает ослабление линейной зависимости между переменными. При **|r|**=1 корреляционная связь перерождается в функциональную.

На практике при изучении зависимости между двумя случайными величинами используют «поле корреляции», с помощью которого при минимальных затратах труда и времени можно установить наличие корреляционной зависимости.

Поле корреляции представляет собой диаграмму, на которой изображается совокупность значений двух признаков. Каждая точка этой диаграммы имеет координаты **(x_i, y_i)**, соответствующие размерам признаков в **i**-м наблюдении. Три варианта распределения точек на поле корреляции показаны на рисунках 1.4.1; 1.4.2; 1.4.3. На первом из них основная масса точек укладывается в эллипсе, главная диагональ которого образует положительный угол с осью **X**. Это график положительной корреляции. Второй вариант распределения соответствует отрицательной корреляции. Равномерное распределение точек в пространстве **(XУ)** свидетельствует об отсутствии корреляционной зависимости (рис. 1.4.3.).



Если наблюдаемые значения У и Х представляют собой выборку из двумерного нормального распределения, то формально можно рассматривать два уравнения регрессии:

$$M(Y|X) = \beta_0 + \beta_1 x \text{ и } M(X|Y) = \alpha_0 + \alpha_1 y.$$

В двумерном корреляционном анализе, обычно строят корреляционную таблицу, поле корреляции, рассчитывают точечные оценки параметров корреляционной модели, оценивают уравнения регрессии, проверяют значимость параметров связи и для значимых параметров строят интервальные оценки, не разделяя при этом задачи корреляционного и регрессионного анализа.

Имея оценки параметров модели $\bar{x}, \bar{y}, S_x, S_y^2, r$, можно рассчитать оценки уравнений регрессии в соответствии с формулой для генеральной регрессии:

$$M(y|x) - M(y) = \beta_{yx}[x - M(x)],$$

где $\beta_{yx} = \rho\frac{\sigma_y}{\sigma_x}$ - коэффициент регрессии ***y*** на ***x***, оценка здесь $y|x - \bar{y} = b_{yx}(x - \bar{x})$, где $b_{yx} = r\frac{S_y}{S_x}$ - оценка генерального коэффициента регрессии β_{yx}.

Аналогичные формулы расчета справедливы для оценки уравнения регрессии ***x*** на ***y***:

$M(x|y) - M(x) = \beta_{xy}[y - M(y)]$ - генеральная регрессия ***x*** на ***y***,

где $\beta_{xy} = \rho\frac{\sigma_x}{\sigma_y}$ - коэффициент регрессии ***x*** на ***y***,

$x|y - \bar{x} = b_{xy}(y - \bar{y})$ - где $b_{xy} = r \cdot \frac{S_x}{S_y}$ - оценка генерального коэффициента регрессии β_{xy}.

Можно показать, что формулы $M(y|x) - M(y) = \beta_{yx}[x - M(x)]$ и $M(y|x) = \beta_0 + \beta_1 x$ идентичны. Из формулы:

$$M(y|x) = \beta_{yx}x - \beta_{yx}M(x) + M(y)$$

полагая, что $\beta_{yx} = \beta_1$, а $-\beta_{yx}M(x) + M(y) = \beta_0$, запишем: $M(y|x) = b_0 + b_1 x$.

Аналогично можно показать идентичность формул попарно:

$$M(x|y) - M(x) = \beta_{xy}[y - M(y)] \text{ и } M(x|y) = \alpha_0 + \alpha_1 y;$$
$$y|x - \bar{y} = b_{yx}(x - \bar{x}) \text{ и } y = b_0 + b_1 x;$$
$$x|y - \bar{x} = b_{xy}(y - \bar{y}) \text{ и } x = a_0 + a_1 y$$
$$b_{yx} \cdot b_{xy} = r^2$$
$$b_{yx} / b_{xy} = S_y^2 / S_x^2$$

В двумерной модели параметрами связи являются коэффициент корреляции ρ (или коэффициент детерминации ρ^2) и коэффициенты регрессии β_{yx}, β_{xy}, которые обычно бывают неизвестны.

По результатам выборки рассчитывают их точечные оценки, соответственно ***r***, b_y, b_x, проверяют гипотезу о значимости (существенности) параметров. В двумерной модели достаточно проверить значимость только коэффициента корреляции. Проверяется гипотеза $\mathbf{H_0: \rho=0}$. Если на уровне значимости α гипотеза отвергнется, то коэффициент корреляции считается значимым и рассчитанное по выборке значение ***r*** может быть использовано в качестве его точечной оценки. Если коэффициент корреляции оказывается незначимым, то гипотеза не отвергается и на практике обычно принимают, что ***x*** и ***y*** в генеральной совокупности линейно независимы.

Доказано, что если верна гипотеза $\mathbf{H_0: \rho=0}$, то статистика $t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ имеет распределение Стьюдента с $\boldsymbol{\nu=n-2}$ числом степеней свободы. По таблице распределения Стьюдента были определены значения статистики $\mathbf{t_{табл}(\alpha;\nu=n-2)}$ для α=0,001; 0,01; 0,02; 0,05 и рассчитаны соответственно границы для ***r*** (таблица Фишера-Иейтса). Таким образом, для проверки гипотезы $\mathbf{H_0: \rho=0}$, находят $\mathbf{r_{табл}(\alpha, \nu=n-2)}$ и сравнивают его с $\mathbf{r_{набл}}$, рассчитанным по выборочным данным. Если $\mathbf{|r_{набл.}| \geq r_{табл}}$, то гипотеза $\mathbf{H_0}$ отвергается на уровне значимости α, если $\mathbf{|r_{набл.}| \leq r_{табл}}$, то гипотеза не отвергается.

При $\mathbf{n}$>100, считая распределение статистики нормированным нормальным, проверяют гипотезу $\mathbf{H_0: \rho}$=0, исходя из условия, что при справедливой гипотезе выполняется равенство: $\mathbf{P(|t| \leq t_{табл})=\gamma=\Phi(t_{табл})}$, т.е. если $\mathbf{|t| \leq t_{табл}}$, то гипотеза $\mathbf{H_0}$ не отвергается. Статистика $r\sqrt{n-1}$, если $\mathbf{n}$>100, также имеет нормированный нормальный закон распределения при справедливости $\mathbf{H_0: \rho}$=0 и этим можно пользоваться для проверки значимости коэффициента корреляции.

Для двумерной корреляционной модели, если отвергается гипотеза $\mathbf{H_0: \rho}$=0, то параметры связи $\boldsymbol{\rho}$, $\boldsymbol{\beta_{yx}}$, $\boldsymbol{\beta_{xy}}$ считаются значимыми и для них имеет смысл найти интервальные оценки, для чего нужно знать закон распределения выборочных оценок параметров.

Плотность вероятности выборочного коэффициента корреляции имеет сложный вид, поэтому используют специально подобранные функции от выборочного коэффициента корреляции, которые подчиняются хорошо изученным законам, например, нормальному или Стьюдента.

При нахождении доверительного интервала для коэффициента корреляции ρ чаще используют преобразование Фишера:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}.$$

Эта статистика уже при **n**>10 распределена приблизительно нормально, с параметрами $M(z)=\frac{1}{2}\ln\frac{1+\rho}{1-\rho}+\frac{\rho}{2(n-1)}$.

По таблице **z** - преобразования Фишера для выборочного коэффициента ***r*** находят соответствующее ему $\mathbf{z_r}$ и находят интервальную оценку для **M(z)** из условия:

$$P\left(z_r-t_\gamma\sqrt{\frac{1}{n-3}}\le M(z)\le z_r+t_\gamma\sqrt{\frac{1}{n-3}}\right)=\gamma=\Phi(t_\gamma),$$

где $\mathbf{t_\gamma}$ находят по таблице интегральной функции Лапласа:

$$\Phi(t)=\frac{2}{\sqrt{2\pi}}\int_o^t e^{-\frac{t^2}{2}}dt$$ для данного $\gamma=1-\alpha$.

Получив доверительный интервал: $z_{\min}\le M(z)\le z_{\max}$, с помощью таблицы **z** - преобразования Фишера получают интервальную оценку: $r_{\min}\le c\le r_{\max}$, где $\mathbf{r_{min}}$ и $\mathbf{r_{max}}$ выбираются с учетом того, что **z** - функция нечетная, а поправочным членом $\frac{\rho}{2(n-1)}$ пренебрегают.

Для значимых коэффициентов регрессии β_{yx} и β_{xy} с надежностью $\gamma=1-\alpha$, находят интервальные оценки из условия, что статистики:

$$t=(b_{yx}-\beta_{yx})\frac{S_x\sqrt{n-2}}{S_y\sqrt{1-r^2}};$$

$$t=(b_{xy}-\beta_{xy})\frac{S_y\sqrt{n-2}}{S_x\sqrt{1-r^2}}$$

имеют распределение Стьюдента с $\nu=n-2$ степенями свободы и, следовательно, из условия $P(|t|<=t_\alpha)=\gamma$ можно рассчитать интервальные оценки:

$$b_{yx}-t_\alpha\frac{S_y\sqrt{1-r^2}}{S_x\sqrt{n-2}}\le\beta_{yx}\le b_{yx}+t_\alpha\frac{S_y\sqrt{1-r^2}}{S_x\sqrt{n-2}};$$

$$b_{xy}-t_\alpha\frac{S_x\sqrt{1-r^2}}{S_y\sqrt{n-2}}\le\beta_{xy}\le b_{xy}+t_\alpha\frac{S_x\sqrt{1-r^2}}{S_y\sqrt{n-2}}$$

где $\mathbf{t_\alpha}$ определяется по таблице Стьюдента для данного $\alpha=1-\gamma$ и $\nu=n-2$.

Пример 4.1. На основании выборочных данных о производительности труда **(x)** и себестоимости продукции **(y)**, полученных с однотипных предприятий за месяц и представленных в таблице 4.1, найти: а) точеную оценку коэффициента корреляции между **x** и **y**, проверить его значимость при α=0,05 и найти интервальную оценку коэффициента корреляции при γ=0,95; б) оценку уравнения регрессии, характеризующего зависимость себестоимости продукции от производительности труда.

Таблица 4.1

производительность труда ***x***	5	4	3	20	10	15
себестоимость продукции ***y***	7	10	12	2	5	4

Решение

Составим вспомогательную таблицу 4.2

Таблица 4.2

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
	5	7	35	25	49
	4	10	40	16	100
	3	12	36	9	144
	20	2	40	400	4
	10	5	50	100	25
	15	4	60	225	16
Σ	57	40	261	775	338
средние	9,5	6,67	43,5	129,17	56,33

а) Выборочный парный коэффициент корреляции рассчитывается по формуле:

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{S_x S_y} = \frac{43{,}5 - 9{,}5 \cdot 6{,}67}{6{,}24 \cdot 3{,}44} = -0{,}93,$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}57 = 9{,}5$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}40 = 6{,}67$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1}{6}261 = 43{,}5$$

$$S_x = \sqrt{S_x^2} = \sqrt{\overline{x^2} - \left(\bar{x}\right)^2} = \sqrt{129{,}17 - 90{,}25} = \sqrt{38{,}92} = 6{,}24$$

$$S_y = \sqrt{S_y^2} = \sqrt{\overline{y^2} - \left(\bar{y}\right)^2} = \sqrt{56{,}33 - 44{,}49} = \sqrt{11{,}84} = 3{,}44.$$

Для проверки значимости коэффициента корреляции сформулируем статистическую гипотезу $\mathbf{H_0: \rho=0}$. По таблице Фишера-Йейтса находим $\mathbf{r_{табл}\ (\alpha=0{,}05;\ \upsilon=n-2=4)=0{,}811}$. Сравнение $|r_{набл}|=0{,}93$ с $\mathbf{r_{табл}}=0{,}811$ свидетельствует о том, что нулевая гипотеза отвергается и, следовательно, коэффициент корреляции ρ значим.

Интервальную оценку для ρ рассчитаем с помощью **z** - преобразований Фишера. По таблице значений статистики $z = \frac{1}{2}\ln\frac{1+r}{1-r}$

находим $\mathbf{z_r}$=0,93. Из условия, что $\gamma=\Phi(t_\gamma)$=0,95, находим по таблице интегральной функции Лапласа t_γ=1,96. Тогда интегральная оценка для MZ_r определяется: **Z(r=0,93)**=1,6584

$$-1{,}6584-1{,}96\sqrt{\frac{1}{3}} \le MZ_r \le -1{,}6584+1{,}96\sqrt{\frac{1}{3}}$$

$$-2{,}7900 \le MZ_r \le -0{,}5268$$

Воспользовавшись таблицей **z** - преобразования Фишера, перейдем от **z** к ρ и найдем интегральную оценку с надежностью γ=0,95:

$$-0{,}992 \le \rho \le -0{,}48\,.$$

б) Для нахождения оценок уравнения регрессии себестоимости продукции от производительности труда y $=\mathbf{b_0+b_1x}$, воспользуемся формулой:

$$b_1 = b_{yx} = r\frac{S_y}{S_x} = -0{,}93\frac{3{,}44}{6{,}24} = -0{,}51$$

b_0=6,67-(-0,51)9,5=11,52.

Тогда используя $y-\bar{y}=b_1(x-\bar{x})$, находим:

$$\hat{y} = 11{,}52 - 0{,}51x\,.$$

4.3. Трехмерная корреляционная модель

На примере трехмерной генеральной совокупности достаточно четко можно продемонстрировать основные задачи и особенности многомерного корреляционного анализа.

Пусть признаки X, Y, Z образуют трехмерную нормально распределенную генеральную совокупность, которая определяется девятью параметрами:

- тремя математическими ожиданиями

$$M(X)=\mu_x \quad \mathrm{M(Y)}=\mu_y \quad \mathrm{M(Z)}=\mu_z \tag{4.9.}$$

- тремя дисперсиями

$$D(X)=\sigma_x^2 \quad \mathrm{D(Y)}=\sigma_y^2 \quad \mathrm{D(Z)}=\sigma_z^2 \tag{4.10.}$$

- тремя парными коэффициентами корреляции

$$\rho_{xy}=M\left[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Y-\mu_y}{\sigma_y}\right];\quad \rho_{xz}=M\left[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Z-\mu_z}{\sigma_z}\right];\quad \rho_{yz}=M\left[\frac{Y-\mu_y}{\sigma_y}\cdot\frac{Z-\mu_z}{\sigma_z}\right].$$

Следует отметить, что частные одномерные (X, Y, Z) и двумерные [(X, Y), (X,Z), (Y, Z)] распределения компонент, а также условные

распределения при фиксированных одной [(X,Y)|Z; (X,Z)|Y; (Y,Z)|X] и двух [X|(Y,Z); Y|(X,Z); Z|(X,Y)] переменных являются нормальными. Поэтому поверхности и линии регрессии являются плоскостями и прямыми, соответственно.

Для изучения разнообразия связей между тремя случайными величинами рассчитывают не только парные, но частные и множественные коэффициенты корреляции (детерминации).

Частные коэффициенты корреляции между двумя случайными величинами при фиксированной третьей (в силу их независимости от фиксированных переменных) характеризуют тесноту связи между этими двумя величинами при исключении из рассмотрения фиксированной третьей величины. Поэтому, если при прямой связи парный коэффициент корреляции между теми же двумя случайными величинами оказался больше соответствующего частного коэффициента, то можно сделать вывод о том, что третья фиксированная величина усиливает взаимосвязь между изучаемыми величинами, т.е. более высокое значение парного коэффициента обусловлено присутствием третьей величины. Уменьшение значения парного коэффициента корреляции, в сравнении с соответствующим частным, свидетельствует об ослаблении связи между изучаемыми величинами действием фиксируемой величины.

Частный коэффициент корреляции обладает всеми свойствами парного коэффициента корреляции, т.к. он является коэффициентом корреляции условного двумерного распределения.

Для трехмерной модели можно рассчитать три частных коэффициента корреляции:

$$\rho_{xy|z} = \frac{\rho_{xy} - \rho_{xz} \cdot \rho_{yz}}{\sqrt{\left(1-\rho_{xz}^2\right)\cdot\left(1-\rho_{yz}^2\right)}};$$

$$\rho_{xz|y} = \frac{\rho_{xz} - \rho_{xy} \cdot \rho_{zy}}{\sqrt{\left(1-\rho_{xy}^2\right)\cdot\left(1-\rho_{zy}^2\right)}};$$

$$\rho_{yz|x} = \frac{\rho_{yz} - \rho_{yx} \cdot \rho_{zx}}{\sqrt{\left(1-\rho_{yx}^2\right)\cdot\left(1-\rho_{zx}^2\right)}}.$$

(4.11.)

Множественный коэффициент корреляции в трехмерной нормальной совокупности служит мерой связи между одной случайной величиной и совместным действием двух остальных. Для трехмерной корреляционной модели можно рассчитать три множественных коэффициента корреляции:

$$R_x = R_{x|yz} = \sqrt{\frac{\rho_{xy}^2 + \rho_{xz}^2 - 2\rho_{xy}\rho_{xz}\rho_{yz}}{1-\rho_{yz}^2}};$$

$$R_y = R_{y|xz} = \sqrt{\frac{\rho_{yx}^2 + \rho_{yz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}}{1-\rho_{xz}^2}};$$

$$R_z = R_{z|xy} = \sqrt{\frac{\rho_{zx}^2 + \rho_{zy}^2 - 2\rho_{zx}\rho_{zy}\rho_{xy}}{1-\rho_{xy}^2}}.$$

(4.12.)

По величине множественный коэффициент корреляции заключен между нулем и единицей. Если R_x=1, то связь между величинами X и (Y, Z) является функциональной, линейной: точки (x, y, z) расположены в плоскости регрессии X на (Y, Z). Если R_x=0, то одномерная случайная величина X и двумерная случайная величина (Y, Z) являются независимыми (в силу нормальности распределения). Множественный коэффициент детерминации R_x^2 показывает долю дисперсии случайной величины X, обусловленную изменением величины (Y, Z).

Множественный коэффициент корреляции может увеличиваться при введении в модель дополнительных признаков и не увеличится при исключении некоторых признаков из модели. Наибольшему множественному коэффициенту детерминации соответствует большие частные коэффициенты детерминации. Например, если $R_x^2 > R_z^2$ и $R_x^2 > R_y^2$, то:

$$\rho_{xz|y} > \rho_{zy/x}$$

$$\rho_{xy|z}^2 > \rho_{zy/x}^2.$$

При фиксировании одной случайной величины трехмерное нормальное распределение превращается в двумерное нормальное распределение, определяемое пятью параметрами. Если фиксирована случайная величина Z, то двумерное нормальное распределение (X,Y|Z) характеризуется следующими параметрами:

$$\mu_{x|z} = \mu_x + \rho_{zx}\frac{\sigma_x}{\sigma_z}(z - \mu_z)$$

$$\mu_{y|z} = \mu_z + \rho_{zy}\frac{\sigma_y}{\sigma_x}(z - \mu_z)$$

$$\sigma^2_{x|z} = \sigma^2_x\left(1 - \rho^2_{zx}\right)$$

$$\sigma^2_{y|z} = \sigma^2_y\left(1 - \rho^2_{zy}\right)$$

$$\rho_{xy|z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{\left(1 - \rho^2_{xz}\right)\left(1 - \rho^2_{yz}\right)}}$$

(4.13.)

Зависимость между величинами **X** и **У** при фиксированном значении случайной величины **Z** выражается прямыми регрессиями в плоскости **Z=z**:

$$M\left(Y|X\right)|Z - \mu_{y|z} = \beta_{yx|z}\left(X - \mu_{x|z}\right),$$
$$M\left(X|Y\right)|Z - \mu_{x|z} = \beta_{xy|z}\left(Y - \mu_{y|z}\right).$$

(4.14.)

Коэффициенты частной регрессии рассчитывают в соответствии с формулами:

$$\beta_{yx|z} = \rho_{xy|z}\frac{\sigma_{y|z}}{\sigma_{x|z}} = \frac{\beta_{yx} - \beta_{yz}\beta_{zx}}{1 - \beta_{xz}\beta_{zx}};$$

$$\beta_{xy|z} = \rho_{xy|z}\frac{\sigma_{x|z}}{\sigma_{y|z}} = \frac{\beta_{xy} - \beta_{xz}\beta_{zy}}{1 - \beta_{yz}\beta_{zy}},$$

(4.15.)

причем:

$$\rho^2_{xy|z} = \beta_{xy|z}\beta_{yx|z};$$

для расчета условных средних квадратических отклонений используют формулы:

$$\sigma_{z|zx} = \sigma_{y|z}\sqrt{1-\rho^2_{xy|z}} = \sigma_{y|x}\sqrt{1-\rho_{yz|x}{}^2};$$

$$\sigma_{z|yz} = \sigma_{x|z}\sqrt{1-\rho^2_{xy|z}} = \sigma_{x|y}\sqrt{1-\rho^2_{xz|y}}.$$

(4.16.)

Условное распределение при фиксировании величины (X, У) будет одномерным Z|(X,Y), которое характеризуется условным математическим ожиданием:

$$M_z\left|(X,Y) = M\left(Z\middle|X\right)\right|y = M\left(Z\middle|Y\right)\Big|x\,.$$

и условной дисперсией $D_z\left|(X,Y) = \sigma^2_{z|xy}\right.$.

Плоскость регрессии Z на (X, У) будет получена при изменении точки (X, У):

$$MZ\left|(X,Y) - \mu_z = \beta_{zx|y}\left(X - \mu_x\right) + \beta_{zy|x}\left(Y - \mu_y\right).\right. \qquad (4.17)$$

Остаточная дисперсия относительно плоскости регрессии рассчитывается в соответствии с формулой:

$$\sigma^2_{z|xy} = \sigma^2_{z|y}\left(1-\rho^2_{zx|y}\right) = \sigma^2_{z|x}\left(1-\rho^2_{yz|x}\right).$$

(4.18)

Для оценки девяти параметров трехмерной корреляционной модели используют следующие формулы:

$$\mu_x \to \bar{x} = \frac{\sum x}{n};\ \sigma^2_{\text{x}} \to S^2_x = \overline{x^2} - \left(\bar{x}\right)^2;\ \ \rho_{\text{xy}} \to r_{xy} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{S_x S_y};$$

$$\mu_y \to \bar{y} = \frac{\sum y}{n};\ \sigma^2_{\text{y}} \to S^2_y = \overline{y^2} - \left(\bar{y}\right)^2;\ \ \rho_{\text{xz}} \to r_{xz} = \frac{\overline{xz} - \bar{x}\cdot\bar{z}}{S_x S_z}; \qquad (4.19)$$

$$\mu_z \to \bar{z} = \frac{\sum z}{n};\ \sigma^2_{\text{z}} \to S^2_z = \overline{z^2} - \left(\bar{z}\right)^2;\ \ \rho_{\text{yz}} \to r_{yz} = \frac{\overline{yz} - \bar{y}\cdot\bar{z}}{S_y S_z}.$$

Оценки условных средних квадратических отклонений при фиксировании одной переменной, частных коэффициентов корреляции, условных средних квадратических отклонений при двух фиксированных переменных и множественных коэффициентов корреляции рассчитываются в соответствии со следующими формулами:

$$S_{x|y} = S_x\sqrt{1-r_{xy}^2};\ \ \mathrm{S_{x|z}} = S_x\sqrt{1-r_{xz}^2};\ \ \mathrm{S_{y|z}} = S_y\sqrt{1-r_{yz}^2};$$

$$S_{y|x} = S_y\sqrt{1-r_{xy}^2};\ \ \mathrm{S_{z|x}} = S_z\sqrt{1-r_{xz}^2};\ \ \mathrm{S_{z|y}} = S_z\sqrt{1-r_{yz}^2};$$

$$r_{xy|z} = \frac{r_{xy}-r_{xz}\cdot r_{yz}}{\sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)}};\ \ r_{xz|y} = \frac{r_{xz}-r_{xy}\cdot r_{yz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{yz}^2\right)}};\ \ r_{yz|x} = \frac{r_{yz}-r_{xy}\cdot r_{xz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{xz}^2\right)}}; \qquad (4.20)$$

$$S_{z|yz} = S_{x|y}\sqrt{1-r_{xz|y}^2};\ \ \mathrm{S_{y|zx}} = S_{y|z}\sqrt{1-r_{yx|z}^2};\ \ \mathrm{S_{z|xy}} = S_{z|x}\sqrt{1-r_{yz|x}^2};$$

$$r_{x|yz}^2 = 1-\frac{S_{x|yz}^2}{S_x^2};\ \ \mathrm{r_{y|xz}^2} = 1-\frac{S_{y|xz}^2}{S_y^2};\ \ \mathrm{r_{z|xy}^2} = 1-\frac{S_{z|xy}^2}{S_z^2},$$

причем: $1-r_{z|xy}^2 = (1-r_{zx}^2)(1-r_{zy|x}^2) = (1-r_{zy}^2)(1-r_{zx|y}^2)$.

Для проверки значимости параметров связи трехмерной корреляционной модели формулируется нулевая гипотеза о равенстве нулю проверяемого параметра. Если на уровне значимости α гипотеза отвергается, то с надежностью $\boldsymbol{\gamma=1-\alpha}$ можно утверждать, что параметр значимо отличается от нуля. Если же гипотеза принимается, то параметр связи незначим.

В трехмерном корреляционном анализе проверяется значимость только частных и множественных коэффициентов корреляции или коэффициентов детерминации. Коэффициенты регрессии одновременно равны нулю или отличны от нуля вместе с соответствующими коэффициентами корреляции (детерминации).

Проверка значимости парных коэффициентов корреляции для трехмерной модели обычно не проводится. Чтобы установить значимость частного коэффициента корреляции, необходимо на выбранном уровне значимости α проверить гипотезу $\mathbf{H_0:\ \rho_{частн}=0}$. В основе критерия используемого для проверки этой гипотезы, лежит статистика:

$$t = \frac{r_{частн}}{\sqrt{1-r_{частн}^{\ 2}}}\sqrt{n-3}, \qquad (4.21)$$

которая при справедливости нулевой гипотезы подчиняется распределению Стьюдента с числом степеней свободы $\boldsymbol{\nu=n-3}$. Для упрощения процедуры проверки значимости разработаны таблицы, где табулирован $\mathbf{r_{табл}\ (\alpha,\ \nu=n-3)}$ в соответствии с перечисленными условиями. Если $\mathbf{|r_{частн}|>r_{табл}(\alpha,\ \nu)}$, то $\boldsymbol{\rho_{частн}}$ считается значимым на

уровне α. В противоположном случае , когда $\mathbf{H_0}$**:** $\rho_{\text{частн}}=0$ не отвергнется, следует считать, что между соответствующими признаками связь отсутствует, либо провести анализ на основе другой выборки.

Основу критерия оценки значимости множественного коэффициента детерминации, а также и корреляции $\mathbf{r_{мн}}$ составляет статистика:

$$F_{набл.}=\frac{r_{мн}^2/2}{\left(1-r_{мн}^2\right)/(n-3)}, \tag{4.22}$$

которая при справедливости нулевой гипотезы $\mathbf{H_o}$**:** $\mathbf{R^2}=0$ имеет распределение Фишера. По таблице распределения Фишера определяют $\mathbf{F_{табл}(\alpha;\ \nu_1=2;\ \nu_2=n-3)}$ и сравнивают с $\mathbf{F_{набл}}$. Если $\mathbf{F_{набл}}>\mathbf{F_{табл}}$, то гипотеза отвергается и, следовательно, $\mathbf{R^2}$ значимо отличается от нуля.

Осуществляя проверку значимости коэффициентов связи трехмерной корреляционной модели, следует учитывать, что если, например, $\mathbf{R_z}$ незначим, то коэффициенты $\rho_{zx|y}$ и $\rho_{zy|x}$ становятся незначимыми. Или, если $\rho_{zx|y}$ незначим, то множественный коэффициент корреляции незначимо отличается от абсолютной величины парного коэффициента корреляции $\mathbf{R_z=|\rho_{zy}|}$**.**

Для значимых множественных коэффициентов корреляции можно получить оценки уравнения регрессии. Для значимого $\mathbf{R_z}$ оценкой соответствующего уравнения регрессии будет:

$$\overline{z|(x,y)}-\bar{z}=b_{zx|y}\left(x-\bar{x}\right)+b_{xy|z}\left(y-\bar{y}\right), \tag{4.23}$$

где $b_{zx|y}=r_{zx|y}\dfrac{S_{z|y}}{S_{x|y}}$ и $b_{zy|x}=r_{zy|x}\dfrac{S_{z|x}}{S_{y|x}}$ – частные коэффициенты регрессии, причем $b_{zx|y}\cdot b_{zy|x}=r_{zx|y}^2$.

Для значимых параметров связи имеет смысл определить границы доверительного интервала с надежностью $\gamma=1-\alpha$. Исходным равенством интервального оценивания $\rho_{\text{частн}}$ служит:

$$P(r_{\text{частн min}}\le\rho_{\text{частн}}\le r_{\text{частн max}})=\gamma.$$

Для получения более точных значений доверительного интервала в данном случае используется z-преобразование Фишера, так как статистика $z=\dfrac{1}{2}\ln\dfrac{1+r}{1-r}$ имеет приблизительно нормальный закон распределения с параметрами $M_z\approx\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ и $D_z\cong\dfrac{1}{n-4}$. Поэтому первоначально определяют границы доверительного интервала для $\mathbf{M_z}$ исходя из равенства:

$$P\left(Z_{r\ частн}-t_\gamma\frac{1}{\sqrt{n-4}}\le M_z\le Z_{r\ частн}+t_\gamma\frac{1}{\sqrt{n-4}}\right)=\gamma=\Phi\left(t_\gamma\right), \tag{4.24}$$

где $Z_{\text{г част}}$ определяется по таблице **z** - преобразования Фишера для рассчитанного $\mathbf{r_{частн}}$,

$\mathbf{t_\gamma}$ находится по таблице интегральной функции Лапласа для заданного значения γ.

Зная границы интервальной оценки для $\mathbf{M_z}$ по таблице **z** - преобразования Фишера получают доверительные границы для $\rho_{\text{частн}}$.

Для значимых частных и множественных коэффициентов детерминации существуют более предпочтительные точечные оценки, чем выборочные коэффициенты:

$$\frac{(n-2)r^2_{\text{частн}} - 1}{n-3} \quad \text{- оценка для } \rho^2_{\text{частн}};$$

$$\frac{(n-1)r_{мн}^{\,2} - 2}{n-3} \quad \text{- оценка для } \rho^2_{\text{мн}}. \qquad (4.25)$$

Интервальные оценки для коэффициента плоскости регрессии можно найти решением относительно оцениваемого коэффициента регрессии неравенства $|t| \le t(\alpha;\ \nu=n-3)$,
где:

$$t = \frac{\left(b_{zx|y} - \beta_{zx|y}\right) S_{x|y}\sqrt{n-3}}{S_{z|y}\sqrt{1-r^2_{zx|y}}}$$

$$t = \frac{\left(b_{zy|x} - \beta_{zy|x}\right) S_{y|x}\sqrt{n-3}}{S_{z|x}\sqrt{1-r^2_{zy|x}}} \qquad (4.26)$$

- статистики, подчиняющиеся распределению Стьюдента с числом степеней свободы $\nu=n-3$;

$\mathbf{t(\alpha;\ \nu=n-3)}$ - определяют по таблице Стьюдента.

Пример 4.2

С целью изучения эффективности производства продукции была отобрана группа 25 однотипных предприятий. На основании полученной выборки для трех показателей (X - производительность труда, У - фондоотдача, Z - материалоемкость продукции) были вычислены величины:

$\bar{x}$=6,06 $\bar{y}$=2,052 $\bar{z}$=24,32

S_x=0,7782 S_y=0,7925 S_z=3,7086

r_{xy}=0,9016392 r_{xz}=-0,8770319 r_{yz}=-0,8899999

Требуется рассчитать оценки частных и множественных коэффициентов корреляции, проверить на уровне α=0,05 их значимость,

для значимых частных коэффициентов корреляции рассчитать интервальные оценки с надежностью γ=0,95.

Решение. Для расчета частных коэффициентов корреляции воспользуемся формулами:

$$r_{xy|z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1-r_{xz}^2\right)\left(1-r_{yz}^2\right)}} = \frac{0{,}9016392 - 0{,}8770319 \cdot 0{,}8899999}{0{,}203078 \cdot 0{,}2079002} = 0{,}5526811;$$

$$r_{xz|y} = \frac{r_{xz} - r_{xy}r_{yz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{yz}^2\right)}} = -0{,}3782736;$$

$$r_{yz|x} = \frac{r_{yz} - r_{xy}r_{xz}}{\sqrt{\left(1-r_{xy}^2\right)\left(1-r_{xz}^2\right)}} = -0{,}4775473.$$

Множественные коэффициенты корреляции можно вычислить по формулам через парные коэффициенты, например:

$r_{x|yz} = \sqrt{\dfrac{r_{xy}^2 + r_{xz}^2 - 2r_{xy}r_{xz}r_{yz}}{1 - r_{yz}^{\ 2}}} = 0{,}9163587,$ или через коэффициенты детерминации в соответствии с формулами:

$$r_{x|yz}^2 = 1 - \frac{S_{x|yz}^2}{S_x^2} = 1 - \frac{0{,}0970695}{0{,}6056} = 0{,}8397136;$$

$$r_{x|yz} = 0{,}9163588; \quad r_{y|xz} = 0{,}9249889; \quad r_{z|xy} = 0{,}9065585,$$

где $S_{x|yz}^2$ рассчитана в соответствии с формулами (4.20).

Проверку значимости множественных коэффициентов корреляции сделаем с помощью статистики:

$$F_{набл(x)} = \frac{r_{x|yz}^2 / 2}{\left(1-r_{x|yz}^2\right)/(n-3)} = \frac{0{,}8397136 \cdot 22}{2 \cdot 0{,}1602864} = 57{,}627157;$$

$$F_{набл(x)} = \frac{r_{y|xz}^2 / 2}{\left(1-r_{y|zz}^2\right)/(n-3)} = \frac{0{,}8556046 \cdot 22}{2 \cdot 0{,}1443954} = 65{,}1797121;$$

$$F_{набл(x)} = \frac{r_{z|xy}^2 / 2}{\left(1-r_{z|xy}^2\right)/(n-3)} = 50{,}745165.$$

Сравнивая $\mathbf{F_{набл}}$ с $\mathbf{F_{кр}(0{,}05;\ 2;\ 22)=3{,}44}$, найденым по таблице распределения Фишера для α=0,05, ν_1=2; ν_2=**n-3**=22, делаем вывод, что все множественные коэффициенты корреляции $\mathbf{r_{x|yz}}$, $\mathbf{r_{y|xz}}$, $\mathbf{r_{z|xy}}$ генеральной совокупности существенно отличаются от нуля.

Для проверки значимости частных коэффициентов корреляции по таблице Фишера-Иейтса находим $\mathbf{r_{кр}}(0{,}05;\ 25)=0{,}381$ и $\mathbf{r_{кр}}(0{,}05;20)=0{,}423$, тогда с помощью линейной интерполяции рассчитаем:

$$\mathrm{r_{кр}}(0{,}05;\ 22) = 0{,}381 + \frac{0{,}423 - 0{,}381}{25 - 20} \cdot (25 - 22) = 0{,}4062.$$

Так как наблюдаемые значения $|\mathbf{r_{xy|z}}|$ и $|\mathbf{r_{yz|x}}|$ больше, чем $\mathbf{r_{кр}}(0,05;22)$, то с вероятностью ошибки 0,05 гипотеза о равенстве нулю генеральных частных коэффициентов корреляции $\rho_{xy|z}$ и $\rho_{yz|x}$ отвергается. Для частного коэффициента корреляции $\rho_{xz|y}$ гипотеза **H:** $\rho_{xz|y}=0$ не отвергается, т.к. $\mathbf{r_{xz|y}}$= - 0,378 меньше по модулю, чем $\mathbf{r_{кр}}(0,05; 22)=0,4062$. Для значимых частных коэффициентов корреляции $\rho_{xy|z}$ и $\rho_{yz|x}$ с надежностью $\gamma=0,95$ найдем интервальные оценки с помощью **z** - преобразования Фишера. По таблице значений статистики $z = \frac{1}{2}\ln\frac{1+r}{1-r}$ находим для $\mathbf{r_{xy|z}}=0,55$ соответствующее ему $\mathbf{z_r}=0,6184$, тогда:

$$P\left(0,6184 - t_\gamma\sqrt{\frac{1}{n-4}} \le Mz(r_{yz|x}) \le 0,6184 + t_\gamma\sqrt{\frac{1}{n-4}}\right) = 0,95,$$

где $\mathbf{t_\gamma}=1,96$ найдено по таблице значений интегральной функции Лапласа для $\mathbf{\Phi(t_\gamma)}=0,95$;

$$\sqrt{\frac{1}{n-4}} = \sqrt{\frac{1}{21}} = 0,2182 \text{ , } 1,96\cdot 0,2182=0,4277;$$

следовательно,

$$0,1907 \le MZ_{xy|z} \le 1,0461$$

По таблице **z** - преобразования совершим переход к интервальным оценкам ρ:

$$0,19 \le \rho_{xy|z} \le 0,78 \quad (\textit{значим}),$$

Аналогично, для $\mathbf{r_H}$= -0,378≈ -0,38 находим **z (r)=** **z** (-0,38)= - **Z** (0,38) = -0,4001. Тогда $\mathbf{Z_{min}}$=-0,4001-0,4277=-0,8278 и $\mathbf{Z_{max}}$=-0,4001+0,4277=+0,0276. Откуда $\mathbf{r_{min}}$=-0,68 $\mathbf{r_{max}}$=+0,03, т.е. $-0,68 \le \rho_{xz|y} \le 0,03$ (не значим).

На основании полученных расчетов можно сделать вывод, что существует тесная взаимосвязь каждого из исследуемых показателей эффективности работы с другими, т.е. все множественные коэффициенты детерминации значимы и превышают 0,8.

Особенно тесная связь между фондоотдачей и двумя остальными показателями. Изменение фондоотдачи, в среднем, на 85,6% объясняется изменением производительности труда и материалоемкости. При увеличении производительности труда на 1 тыс. руб. фондоотдача увеличивается, в среднем, на 0,55 руб. на рубль основных производственных фондов; при уменьшении материалоемкости на 1% фондоотдача увеличивается, в среднем, на 0,48 руб.

Взаимосвязь между материалоемкостью и производительностью труда не доказана (без учета фондоотдачи). Однако можно сказать, что

фондоотдача усиливает связь между материалоемкостью и производительностью труда, т.к. $|r_{xz}|>|r_{xz|y}|$.

4.4. Методы оценки параметров корреляционных моделей

Для оценки параметров корреляционных моделей в основном используют три метода: моментов, максимального правдоподобия и наименьших квадратов.

Метод моментов был предложен К.Пирсоном. В соответствии с ним первые q моментов случайной величины X приравниваются q выборочным моментам, полученным по экспериментальным данным. Теоретическим обоснованием метода моментов служит закон больших чисел, согласно которому для рассматриваемого случая при большом объеме выборки выборочные моменты близки к моментам генеральной совокупности.

Для двумерной корреляционной модели согласно методу моментов неизвестное ожидание оценивается средним арифметическим (выборочным начальным моментом первого порядка), а дисперсия - выборочной дисперсией (выборочным центральным моментом второго порядка). Коэффициент корреляции ρ оценивается выборочным коэффициентом r, который является функцией выборочных начальных моментов первого порядка самих случайных величин и их произведения.

Метод моментов дает возможность получать состоятельные оценки, т.е. надежность выводов, сделанных при его использовании, зависит от количества наблюдений. Использование метода моментов на практике приводит к сравнительно простым вычислениям.

Метод максимального правдоподобия, предложенный английским математиком Р.А.Фишером, часто приводит к более сложным вычислениям, чем метод моментов, однако оценки, получаемые с его помощью, как правило, оказываются более надежными и особенно предпочтительными в случае малого числа наблюдений.

Метод максимального правдоподобия для оценки математического ожидания предполагает использование средней арифметической, которая обладает свойствами несмещенности, состоятельности и эффективности.

Дисперсию генеральной совокупности согласно методу максимального правдоподобия, рекомендуется оценивать выборочной дисперсией, которая удовлетворяет лишь условию состоятельности. Использование исправленной дисперсии позволяет иметь оценку дисперсии, удовлетворяющую условиям несмещенности и состоятельности.

Применение метода максимального правдоподобия часто приводит к решению сложных систем уравнений, поэтому метод наименьших квадратов, использование которого связано с более

простыми выкладками, имеет большое практическое применение. Основоположниками этого метода являются Гаусс, Лежандр.

Основная идея метода наименьших квадратов сводится к тому, чтобы в качестве оценки неизвестного параметра принимать значение, которое минимизирует сумму квадратов отклонений между оценкой и параметром для всех наблюдений.

Так как нормальный закон распределения генеральной совокупности является исходной предпосылкой построения корреляционных моделей, метод наименьших квадратов и метод максимального правдоподобия дают одинаковые результаты.

В анализе двумерной корреляционной модели обычно оценку уравнения регрессии производят с помощью метода наименьших квадратов.

4.5. Ранговая корреляция

Для изучения взаимосвязи признаков, не поддающихся количественному измерению, используются различные показатели ранговой корреляции. В этом случае элементы совокупности располагают в определенном порядке в соответствии с некоторыми признаками (качественным и количественным), т.е. производят ранжирование. При этом каждому объекту присваивается порядковый номер, называемый рангом. Например, элементу с наименьшим значением признака присваивается ранг 1, следующему за ним элементу - ранг 2 и т.д. Элементы можно располагать также в порядке убывания значений признака. Если объекты ранжированы по двум признакам, то можно изменить силу связи между признаками, основываясь на значениях рангов.

Коэффициент ранговой корреляции Спирмэна является парным, и его использование не связано с предпосылкой нормальности распределения исходных данных:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)},$$

где **d** - разность значений рангов, расположенных в двух рядах у одного и того же объекта.

Величина $\mathbf{r_s}$ для двух рядов, состоящих из **n** рангов, зависит только от $\mathbf{\Sigma d^2}$. Если два ряда полностью совпадают, то $\mathbf{\Sigma d^2=0}$ и, следовательно $\mathbf{r_s=1}$, т.е. при полной прямой связи: $\mathbf{r_s=1}$. При полной обратной связи, когда ранги двух рядов расположены в обратном порядке, $\mathbf{r_s=-1}$. При отсутствии корреляции между рангами $\mathbf{r_s=0}$.

Пример 4.3. При ранжировании оценок на вступительных экзаменах и средних баллов за первую экзаменационную сессию одних и тех же лиц получены следующие ранги:

Таблица 4.3

Ранг	студент	А	Б	В	Г	Д	Е	Ж	З	И	К
	вступит. экзамен	2	5	6	1	4	10	7	8	3	9
	экзамен. сессия	3	6	4	1	2	7	8	10	5	9
d		-1	-1	2	0	2	3	-1	-2	-2	0
d^2		1	1	4	0	4	9	1	4	4	0

Из данных таблицы 4.3 следует: $\mathbf{\Sigma d^2}$=28; $\mathbf{r_s}$=1- $\frac{6 \cdot 28}{10(10-1)}$ =0,83, что свидетельствует о достаточно высокой связи между изучаемыми признаками.

Для измерения тесноты связи между признаками, не поддающимися точной количественной оценке, используются и другие коэффициенты, например, коэффициент Кэндела, конкордации, ассоциации, контингенции и др.

4.6. Нелинейная парная корреляция

Для изучения связи между признаками, которая выражается нелинейной функцией, используется более общий, чем коэффициент корреляции, показатель тесноты связи - корреляционное отношение.

Нелинейная (или криволинейная) связь между двумя величинами - это такая связь, при которой равномерным изменениям одной величины соответствует неравномерные изменения другой, причем эта неравномерность имеет определенный закономерный характер.

Использование корреляционного отношения основано на разложении общей дисперсии зависимой переменной на составляющие: дисперсию, характеризующую влияние объясняющей переменной, и дисперсию, характеризующую влияние неучтенных и случайных факторов:

$$S_y^2 = S_{y|x}^2 + S_{ост}^2,$$

где

S^2_y - общая дисперсия зависимой переменной, т.е. дисперсия относительно среднего значения;

$S^2_{y|x}$ - дисперсия функции регрессии относительно среднего значения зависимой переменной, характеризующая влияние объясняющей переменной;

$S^2_{ост}$ - дисперсия зависимой переменной ***y*** относительно функции регрессии, т.е. остаточная дисперсия.

Корреляционной отношение выборочных данных определяется по формуле:

$$\eta_{yx} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

или $\eta_{yx} = \sqrt{1 - \frac{S^2_{ост}}{S^2_y}}$.

Влияние корреляционного отношения заключено в пределах:

$$0 \le \eta_{yx} \le 1.$$

Если дисперсия $\mathbf{S^2_{y|x}}$, обусловленная зависимостью величины y от объясняющей переменной ***x***, равно общей дисперсии $\mathbf{S^2_y}$ (а это возможно лишь при наличии функциональной связи), то $\boldsymbol{\eta_{yx}}$=1. Если же остаточная (т.е. необъясненная) дисперсия $\mathbf{S^2_{ост}}$ равна общей дисперсии S^2_y, то $\boldsymbol{\eta_{yx}}$=0, т.е. корреляционная связь отсутствует.

В предыдущей главе было отмечено, что линейный коэффициент парной корреляции является симметричной функцией относительно ***x*** и ***y***. Следует подчеркнуть, что этим свойством не обладает корреляционное отношение, т.е. $\boldsymbol{\eta_{xy} \neq \eta_{yx}}$. Для линейной связи $\boldsymbol{\eta_{xy}=\eta_{yx}=r_{xy}}$ Поэтому величину $\boldsymbol{\eta^2 - r^2}$ можно использовать для характеристики нелинейности связи между переменными.

В качестве одного из самых простых критериев оценки нелинейности связи можно использовать следующий:

$$K_н = \frac{\sqrt{n}}{0{,}67449} \cdot \frac{1}{2}\sqrt{\eta^2_{yx} - r^2_{yx}}.$$

Если значение $K_н$>2,5, то корреляционную связь можно считать нелинейной.

Проверка и построение доверительных интервалов для корреляционного отношения генеральной совокупности осуществляются так же, как аналогичные процедуры для линейного коэффициента парной корреляции:

$$t_н = \frac{\eta_{yx}}{\sqrt{1 - \eta^2_{yx}}}\sqrt{n-2};$$

$\mathbf{t_{кр}}$ находится по таблице распределения Стьюдента из условия:

$$\left.\begin{matrix} \alpha \\ \nu = n-2 \end{matrix}\right\} \rightarrow t_{кр}$$

доверительный интервал имеет вид:

$$\eta - t_\gamma \sqrt{\frac{1-\eta^2}{n-3}} \le \eta \le \eta + t_\gamma \sqrt{\frac{1-\eta^2}{n-3}},$$

где $\mathbf{t_\gamma}$ находится по таблице интегральной функции Лапласа с учетом уровня доверительной вероятности $\boldsymbol{\gamma}$.

Следует обратить внимание на то, что использование корреляционного отношения $\boldsymbol{\eta}$ имеет смысл только для функций криволинейных, но линейных относительно параметров. Для функций, нелинейных относительно параметров [типа $\tilde{y} = f(x)$], корреляционное отношение не может служить точным измерителем тесноты связи.

Вернуться к руководству