Практика 2. Элементы корреляционно-регрессионного анализа.

О связях функциональных, статистических, корреляционных. Сглаживание экспериментальных зависимостей по методу наименьших квадратов. Линейная регрессия. Выборочный коэффициент корреляции. Нелинейная регрессия. Выборочное корреляционное отношение.

В математическом анализе рассматривается связь между величинами, которую называют функциональной. В этом случае величина y определена вполне значениями $x, z,..., u$, т.е. $y = f(x, z,...u)$. Функциональная связь может существовать и между случайными величинами. Но между случайными величинами может существовать связь и другого ряда, заключающаяся в том, что одна из них реагирует на изменение другой изменениями своего закона распределения. Такую связь называют стохастической или вероятностной.

Таким образом, X и Y связаны вероятностной зависимостью, то зная значение одной случайной величины нельзя точно указать, какое значение примет другая величина, а можно указать только закон ее распределения, зависящий от другой случайной величины.

Вероятностная зависимость может быть более или менее тесной; при увеличении степени вероятностной связи она все более и более приближается к функциональной. Функциональную зависимость можно рассматривать как предельный, крайний случай вероятностной зависимости. Другой крайний – полная независимость случайных величин. Между этими двумя «полюсами» находятся все степени вероятностной зависимости – от самой слабой до самой сильной.

Наиболее простым и имеющим важное практическое значение видом вероятностной зависимости является корреляционная зависимость.

Корреляционная зависимость между двумя случайными величинами выражается в том, что на изменения одной случайной величины другая случайная величина реагирует изменениями своего математического ожидания:

$$M(Y/X = x) = f(x); \quad (1)$$

или

$$M(Y/X = y) = f(y); \quad (2)$$

Уравнение (1) называют уравнением случайной величины Y относительно X или уравнением регрессии Y на X. Соответственно уравнение (2) есть уравнение регрессии X и Y.

Таким образом, чтобы изучить корреляционную связь, нужно знать условное математическое ожидание случайной величины. В свою очередь для этого необходимо знать аналитический вид двумерного распределения $(X;Y)$, который зачастую неизвестен. Поэтому идут на упрощение и переходят от

условного математического ожидания случайной величины к условному среднему значению, то есть принимают, что:

$$M(Y/X = x) = \overline{y_x}; \quad (3)$$

или

$$M(Y/X = y) = \overline{x_y}; \quad (4)$$

Тогда из формул (1) и (3) называемое эмпирическое уравнение(эмпирическую функцию) регрессии Y на X:

$$f(x) = \overline{y_x}; \quad (5)$$

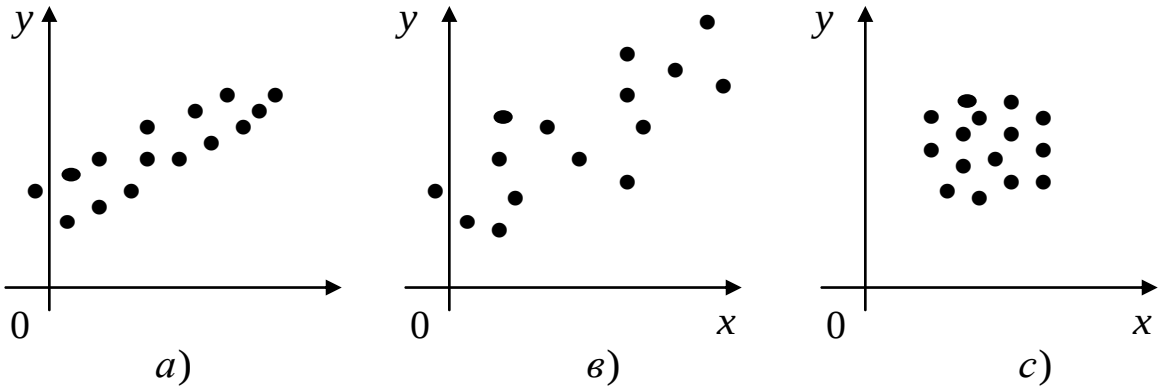
Аналогично из (2) и (4) имеем эмпирическую функцию регрессии Y на X:

$$f(x) = \overline{x_x}; \quad (6)$$

Вопрос в том, что принять за зависимую переменную, а что за независимую, следует решать применительно к каждому конкретному случаю

При изучении корреляционных связей возникает три основных вопроса: наличие связи, форма связи и сила связи.

Допустим, что проведено n испытаний и при каждом отмечались значения двух случайных величин. В результате получатся n пар выборочных значений $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Для наглядности эти пары значений можно рассматривать как координаты точек на плоскости. Образовавшуюся совокупность точек обычно называют полем корреляции. Поле корреляции дает представление о силе корреляции дает представление о силе корреляцию на рис



а) *в*) *с*)

8 приведены примеры совокупностей точек, соответствующих сильной (а), слабой (в) корреляции и полному ее отсутствию (с).

Кроме того, по расположению точек на поле корреляции можно в первом приближении сделать предположение о форме и тесноте корреляционной связи.

Пусть сделано предположение о форме корреляционной связи (линенйная, квадратичная, экспоненциальная и т.д.), тем самым можно записать аналитический вид функции $f(x)$ из уравнения (5) пока с неопределенными коэффициентами. Для линейной зависимости будем иметь:

$$\bar{y}(x) = a_0 + a_1 x; \quad (7)$$

Для квадратичной зависимости:

$$\bar{y}(x) = a_0 + a_1 x + a_2 x^2; \quad (8)$$

Для экспоненциальной зависимости:

$$\bar{y}(x) = a_0 e^{a_1 x}; \quad (9)$$

Для обратно пропорциональной зависимости:

$$\bar{y}(x) = a_0 + \frac{a_1}{x}; \quad (10)$$

Во всех уравнениях (9.42) – (9.45) a_0, a_1, a_2 - коэффициенты регрессии; x - независимая случайная переменная.

Неизвестные коэффициенты регрессии находят, исходя из принципа наименьших квадратов. Согласно принципу наименьших квадратов, наилучшее уравнение приближенной регрессии дает та функция из рассматриваемого класса (линейных, квадратичных и т.д.) функций, для которой сумма квадратов:

$$S = \sum_{i=1}^{n} \left[y_i - f(x_i) \right]^2; \quad (11)$$

Имеет наименьшее значение. В формуле (11) функция $f(x)$ записана со всеми неопределенными коэффициентами $a_0, a_1, a_2, \ldots$; y_i - измеренное значение y.

Величину S теперь можно рассматривать как функцию от этих неопределенных коэффициентов. Задача состоит в том, чтобы найти набор коэффициентов $a_0, a_1, a_2, \ldots$, минимизирующих величину S. В математической статистике, как правило, рассматриваются функции $f(x)$, дифференцируемые по всем своим коэффициентам. При этом условие отыскание минимизирующего набора коэффициентов превращается в несложную задачу математического анализа. Как известно, необходимым условием минимума дифференцируемой функции многих переменных $S(a_0, a_1, a_2, \ldots)$ является выполнение равенств: $\frac{\partial S}{\partial a_0} = 0, \frac{\partial S}{\partial a_1} = 0, \frac{\partial S}{\partial a_2} = 0, \ldots$ эти равенства можно рассматривать как уравнения относительно $a_0, a_1, a_2, \ldots$; в математической статистике они называются нормальными уравнениями. Так как $S \geq 0;$ при любых $a_0, a_1, a_2, \ldots$; то у нее обязательно должен существовать хотя бы один минимум. Поэтому если система нормальных уравнений имеет единственное решение, то оно и является минимальным для величины S.

Используя правила дифференцирования, получим систему нормальных уравнений:

$$\begin{cases} \sum_{i=1}^{n} 2[y_i - f(x_i)]\dfrac{\partial f(x_i)}{\partial a_0} = 0; \\ \sum 2[y_i - f(x_i)]\dfrac{\partial f(x_i)}{\partial a_0} = 0; \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

или

$$\begin{cases} \sum_{i=1}^{n} y_i \dfrac{\partial f(x_i)}{\partial a_0} - \sum f(x_i)\dfrac{\partial f(x_i)}{\partial a_0} = 0; \\ \sum_{i=1}^{n} y_i \dfrac{\partial f(x_i)}{\partial a_0} - \sum f(x_i)\dfrac{\partial f(x_i)}{\partial a_1} = 0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

Покажем, как составляются нормальные уравнения для случая линейной регрессии (7). отметим, что линейная форма связи занимает особое место в теории корреляции. Можно показать, что линейная регрессия обуславливается двумерным нормальным законом распределения пары случайных величин $(X;Y)$.Уравнение (11) для случая линейной формы связи между случайными переменными приобретает вид:

$$S = \sum_{i=1}^{n} (y_i - a_0 + a_1 x)^2;$$

Согласно вышеизложенному алгоритму получение системы нормальных уравнений, находим частные производные функции S по a_0 и a_1 и приравниваем их к нулю.

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = -2\sum_{i=1}^{n} (y_i - a_0 - a_i x_i) = 0; \\ \dfrac{\partial S}{\partial a_0} = -2\sum (y_i - a_0 - a_i x_i) x_i = 0; \end{cases}$$

После небольших преобразований получим:

$$\begin{cases} \sum_{i=1}^{n} y_i - \sum a_0 - \sum a_i x_i = 0; \\ \sum_{i=1}^{n} y_i x_i - \sum a_0 x_i - \sum a_i x_i^2 = 0; \end{cases}$$

Величины a_0 и a_1 являются постоянными, поэтому их можно вынести за знак суммы; $\sum_{i=1}^{n} a_0$ есть не что иное, как na_0. В результате имеем:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum y_i; \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum x_i^2 = \sum y_i x_i; \end{cases} \quad (12)$$

Решая систему нормальных уравнений (9.47), получим значения коэффициентов регрессии:

$$a_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}; \quad (13)$$

$$a_1 = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}; \quad (14)$$

Система нормальных уравнений для уравнения регрессии вида (8) согласно (11) получается аналогично:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i; \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i; \\ a_0 \sum_{i=1}^{n} x_i^2 + a_1 \sum_{i=1}^{n} x_i^3 + a_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i; \end{cases}$$

Полученная система линейна относительно неизвестных коэффициентов a_0, a_1, a_2, и ее нетрудно решить, пользуясь известными методами, например, по формулам Крамера или методом Гаусса.

Для уравнения регрессии вида (10) согласно (11) имеем:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} \frac{1}{x_i} = \sum y_i; \\ a_0 \sum_{i=1}^{n} \frac{1}{x_i} + a_1 \sum \frac{1}{x_i^2} = \sum \frac{y_i}{x_i}; \end{cases}$$

Эта система относительно неизвестных a_0 и a_1 также линейна.

Для суждения о степени тесноты связи между случайными величинами чаще всего используют коэффициент корреляции r или корреляционное отношение η. Возможность измерения тесноты связи между случайными величинами с помощью коэффициента корреляции и корреляционного отношения следует из свойств этих показателей, приведенных ниже:

1. Если коэффициент корреляции $r = \pm 1$, то x и y связаны точной прямолинейной связью вида: $y = a_0 + a_1 x$; или $x = b_0 + b_1 x$;
2. Если $r = 0$ между x и y не существует прямолинейной корреляционной связи, но криволинейная возможна.
3. Чем ближе r к ±1, тем точнее прямолинейная корреляционная связь между x и y. Она ослабевает с приближением r к 0.
4. если корреляционное отношение $\eta_{Y/X} = 0$, то между x и y нет корреляционной связи.
5. Если $\eta_{Y/X} = 1$, то y связано с x однозначной связью, то есть всякому значению x соответствует одно определенное значение y (функциональная связь).
6. Чем ближе $\eta_{Y/X}$ к единице, тем теснее связь между x и y; чем ближе $\eta_{Y/X}$ к нулю, тем слабее эта связь.
7. Если $\eta_{Y/X} = |r|$, то регрессия x по y точно линейна и обратно: если регрессия x по y точно линейна, то $\eta_{Y/X} = |r|$.

Оценка r^* коэффициента корреляции по выборке может быть найдена по формуле: $$r^* = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}; \quad (15)$$

или

$$r^* = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{S_x S_y}; \quad (16)$$

Где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ - средние всех наблюдений x_i и y_i; S_x, S_y - выборочные средние квадратические отклонения случайных величин x и y соответственно.

При малом числе наблюдений r^* удобно вычислять по формуле:

$$r^* = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}; \quad (17)$$

Отметим, что если коэффициент корреляции положительный, то связь между переменными положительная. Это значит, что с ростом значений x увеличивается y. Если коэффициент корреляции имеет отрицательное

значение, то связь между переменными отрицательная, то есть с ростом значений x величина y уменьшается.

Если коэффициент корреляции равен 0, то говорят, что случайные величины не коррелированные. Некоррелированность не следует смешивать с независимостью, независимые случайные величины не коррелированы. Однако обратное утверждение неверно: некоррелированные случайные величины могут быть зависимы и даже функционально.

При отклонении исследуемой зависимости от линейного вида коэффициент корреляции r теряет свой смысл как характеристика степени тесноты связи. Более надежной характеристикой при этом оказывается корреляционное отношение $\eta_{Y/X}$, интерпретация которого не зависит от вида исследуемой зависимости. Выборочное корреляционное отношение $\eta^{*}{}_{Y/X}$ вычисляется по формуле:

$$\eta^{*}{}_{Y/X} = \frac{\frac{1}{n}\sum_{i=1}^{n} m_i \left(\overline{y_i} - \overline{y}\right)^2}{\frac{1}{n}\sum_{i=1}^{k}\sum_{i=1}^{m_i} \left(y_{ij} - \overline{y}\right)^2} = \frac{S^2_{y(x)}}{S^2_y}; \qquad (18)$$

Где числитель $S^2_{\overline{y}(x)}$ характеризует рассеяние частных средних $\overline{y_i} = \frac{1}{m_i}\sum_{i=1}^{m_i} y_{ij}$ около своего общего среднего $\overline{y} = \frac{1}{n}\sum_{i=1}^{k} m_i \overline{y}_i$, а знаменатель – дисперсия S^2_y индивидуальных результатов наблюдений относительно общего среднего $\overline{y}$. Аналогично определяется выборочное значение $\eta^{*}{}_{Y/X}$.

В отличии от коэффициента корреляции корреляционное отношение несимметрично по отношению к исследуемым переменным, то есть $\eta^{*}{}_{Y/X} \neq \eta_{Y/X}$. Отметим, что между $\eta^{*}{}_{Y/X}$ и $\eta_{Y/X}$ нет какой-либо простой зависимости. Некоррелированность Y с X (то есть равенство нулю величины $\eta_{Y/X}$) не влечет за собой непосредственно некоррелированность Y с X.

Величина $\eta^{*}{}_{Y/X}$ - r^{*} используется в качестве отклонения зависимости от линейной, т.к. обычно $(\eta^{*}{}_{Y/X})^2 > r^2$, $(\eta^{*}{}_{X/Y})^2 > r^2$ и лишь в случае линейной зависимости $r^2 = (\eta^{*}{}_{Y/X})^2 = (\eta^{*}{}_{X/Y})^2$.

Замечание. Из теории вероятностей известно, что характеристикой связи (линейной) между случайными величинами Y и X служит коэффициент корреляции:

$$r = \frac{K_{xy}}{\sigma_x \sigma_y};$$

где $K_{xy} = M\left[(X - MX)(Y - MY)\right]$-корреляционный момент, σ_x, σ_y-средние квадратические отклонения случайных величин Y и X соответственно

Тогда очевидно, что числитель в формуле (15) есть оценка корреляционного момента ,т.е. выборочный корреляционный момент K^*_{xy}. Для небольших выборок рекомендуется использовать несмещенную оценку:

$$\tilde{K}^*_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}); \qquad (19)$$

а выборочные средние квадратические отклонения случайных величин Y и X вычислять по формуле.

Числитель в формуле (15) достаточно просто преобразуется к выражению в числителе формулы (16). Чаще используется формула (16) для вычисления величины r^*, поэтому для получения несмещенной оценки выборочного корреляционного момента, стоящего в числителе формулы (16), рекомендуется вычисленный числитель умножить на $\frac{n}{n-1}$; величины S_x и S_y определить по формуле .

При корреляционном анализе необходимо оценить достоверность связи между переменными, то есть выяснить, не объясняется ли величина коэффициента корреляции, полученная по выборочным данным, случайностями выборки. Для этого оценивается значимость (существенность) коэффициента корреляции. Проверяется гипотеза H_0 о том, что r=0, альтернативной является гипотеза H_1 при $r \neq 0$.

В случае совместной нормальной распределенности исследуемых переменных и при достаточно большом объеме выборки n распределение r^* можно считать приближенно нормальным со средним, равным своему теоретическому значению r, и дисперсией $\sigma_r^2 = \frac{\left(1 - r^2\right)^2}{n}$. Оценка для σ_r вычисляется по формуле:

$$\sigma_r \approx S_r = \frac{1 - r^{*2}}{n}; \qquad (20)$$

Можно доказать, что в указанной ситуации величина $t = \frac{r^* - r}{\sigma_r}$ имеет приближенно нормальное распределение с математическим ожиданием, равным нулю, и дисперсией равной единице. Поэтому проверка значимости (или существенности) коэффициента корреляции сводится к следующему:

вычисляется значение $t = \dfrac{|r^*|}{\sigma_r}$, которое затем сравнивается с найденным по табл.

2. Приложение для заданной вероятности $\dfrac{p}{2}$ значением t_p.

Если $t < t_p$, то принимается гипотеза H_0, то есть коэффициент корреляции считать существенным нельзя и его отклонение от нуля обусловлено неизбежными случайными колебаниями выборки. Если $t > t_p$, то гипотеза H_0 отвергается и коэффициент корреляции можно считать существенными, а связь между случайными величинами Y и X достоверной.

Однако следует учитывать, что при малых значениях n и значениях r, близких $k \pm$, это приближение оказывается очень грубым.

Пример 1. Результаты наблюдений случайной величины (X; Y) представлены в табл. 9.

Таблица 9

Y\X	**20**	**25**	**30**	**35**	**40**	**45**	$\mathbf{m_j}$
34	4	2					6
38		5	3				8
42			5	45	5		55
46			2	8	7		17
50				4	7	3	14
$\mathbf{m_i}$	4	7	10	57	19	3	100

Необходимо:

1) вычислить групповые средние $\overline{x_i}$ и $\overline{y_j}$ и построить по ним ломаные эмпирических линию регрессии; 2) предполагая, что между переменными X и Y существует линейная корреляционная зависимость:

а) найти уравнение прямых регрессий и построить их графики на том же чертеже, на котором изображены ломаные по групповым средним;

б) вычислить коэффициент корреляции, на уровне значимости $\alpha = 0{,}05$ оценить его существенность и сделать вывод о тесноте и направлении связи;

в) используя соответствующее уравнение регрессии, определить среднее значение величины Y для $x = 38$.

Пояснения к табл. 9: в последнем столбце таблицы представлены частоты $\mathbf{m_j}$ появлений значений y_j, $j = \overline{1;5}$; в последней строке таблицы представлены частоты $\mathbf{m_i}$ появление значений x_j, $i = \overline{1;6}$; на пересечении строк и столбцов представлены частоты n_{ij} появлений пары (x_i, y_j). Объем выборки, как видно из таблицы, $n = 100$.

1) Вычисляем групповые средние $\overline{x_i}$ и $\overline{y_j}$.

Для x = 20: $\overline{y_1} = \frac{34 \cdot 4}{4} = 34;$ для x = 25: $\overline{y_2} = \frac{34 \cdot 2 + 38 \cdot 5}{7} = 36{,}86;$ для x = 30: $\overline{y_3} = \frac{38 \cdot 3 + 42 \cdot 5 + 46 \cdot 2}{10} = 41{,}6;$ для x = 35: $\overline{y_4} = 43{,}12;$ для x = 40: $\overline{y_5} = 46{,}42;$ для x = 45: $\overline{y_6} = 50.$

Составляем таблицу:

x	**20**	**25**	**30**	**35**	**40**	**45**
$\overline{y_j}$	34	36,86	41,6	43,12	46,42	50

На рис. 9 представлена ломаная эмпирической линии регрессии Y по X. Так как объем выборки велик, то эта ломаная более наглядно представляет тенденцию изменения значений Y при изменении значений X, чем корреляционное поле.

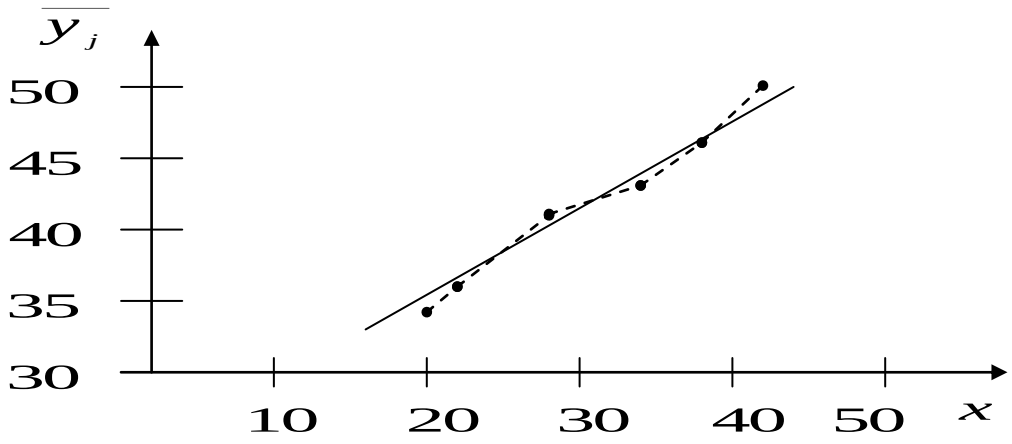


Рис. 9 «Ломаная» линия регрессии по групповым средним; график уравнения регрессии Y по X.

По виду ломаной можно предположить наличие линейной корреляционной зависимости между переменными X и Y.

2) Уравнение регрессии ищем в виде $\overline{y_x} = a_0 + a_1 x$. Коэффициенты a_0, a_1. найдем из системы нормальных уравнений. С учетом частностей появления значений переменных систем (9.47) принимает вид.

$$\begin{cases} n a_0 + a_1 \sum_i x_i m_i = \sum_j y_j m_j; \\ a_0 \sum_i x_i m_i + a_1 \sum_i m_i x_i^2 = \sum_{ij} m_{ij} y_j x_i; \end{cases}$$

Составляем расчетную таблицу для определения коэффициентов при неизвестных a_0, a_1. в системе нормальных уравнений.

Таблица 9

Y\X	**20**	**25**	**30**	**35**	**40**	**45**	$\mathbf{m_j}$	$\mathbf{y_j m_i}$	$\mathbf{y_i^2 m_i}$	$\sum_i x_i y_j m_{ij}$;
34	4	2					6	204	6936	4420
38		5	3				8	304	11552	8170
42			5	45	5		55	2310	97020	80850
46			2	8	7		17	782	35972	28520
50				4	7	3	14	700	35000	27750
$\mathbf{m_i}$	4	7	10	57	19	3	100	4300	186480	149710
$\mathbf{x_i m_i}$	80	175	300	1995	760	135	3445			
$\mathbf{x_i^2 m_i}$	1600	4375	9000	69825	30400	6075	121275			

Пояснение к табл. 9:

$$\sum_{j=1}^{5} y_j m_j = y_1 m_1 + ... + y_5 m_5 = 34 \cdot 6 + 38 \cdot 8 + 42 \cdot 55 + 46 \cdot 17 + 50 \cdot 14 = 204 + 304 + 2310 + 782 + 700 = 4300$$

$$\sum_{j=1}^{6} y_j m_j = y^2{}_1 m_1 + ... + y^2{}_5 m_5 = 34^2 \cdot 6 + 38^2 \cdot 8 + 42^2 \cdot 55 + 46^2 \cdot 17 + 50^2 \cdot 14 =$$

$$= 6936 + 11552 + 97020 + 35972 + 3500 = 186480;$$

$$\sum_{i=1}^{6} x_i m_j = 3445; \sum x^2{}_i m_j = 121275.$$

Для первой строчки последнего столбца: $\sum_{i=1} x_i y_1 n_{i1} = 34 \cdot 20 \cdot 4 + 34 \cdot 25 \cdot 2 = 4420$;

аналогично для строк со второй по пятую включительно: $\sum_{j=1}^{5} \sum_{i=1}^{6} x_i y_j n_{ij} = 149710$.

Подставляя данные из табл. 9 в систему нормальных уравнений получим:

$$\begin{cases} 100 a_0 + 3445 a_1 = 4300; \\ 3445 a_0 + 121275 a_1 = 149710. \end{cases}$$

Решая эту систему, находим $a_0 = 22{,}089, a_1 = 0{,}607$, тогда уравнение регрессии Y по X имеет вид:

$$\overline{y_x} = 22{,}089 + 0{,}607 x; \qquad (21)$$

Строим график этой прямой по двум точкам:

При $x = 25$ $\overline{y_x} = 22{,}089 + 15{,}175 = 37{,}264;$

При $x = 45$ $\overline{y_x} = 22{,}089 + 27{,}315 = 49{,}404.$

Уравнение регрессии (21) дает возможность прогнозировать значение среднее переменной Y в предположении, что независимая переменная X примет определенное значение. Например, для $x = 38$ из уравнения (21) получим $\overline{y} = 45{,}155$.

Данные, приведенные в табл. 9, позволяет определить уравнение регрессии X и Y. Находим аналогично предыдущему групповые средние:

для y = 34: $\overline{x_1} = \dfrac{20 \cdot 4 + 25 \cdot 2}{6} = 21{,}67$; для y = 38: $\overline{x_2} = \dfrac{25 \cdot 5 + 30 \cdot 3}{8} = 26{,}875$; для y = 42: $\overline{x_3} = 35$; для y = 46: $\overline{x_4} = 36{,}47$; для y = 50: $\overline{x_5} = 39{,}64$. Составляем таблицу:

y	**34**	**38**	**42**	**46**	**50**
$\overline{x_j}$	21,67	26,87	35	36,47	39,64

Строим ломаную эмперической линии прогрессии (рис. 10).

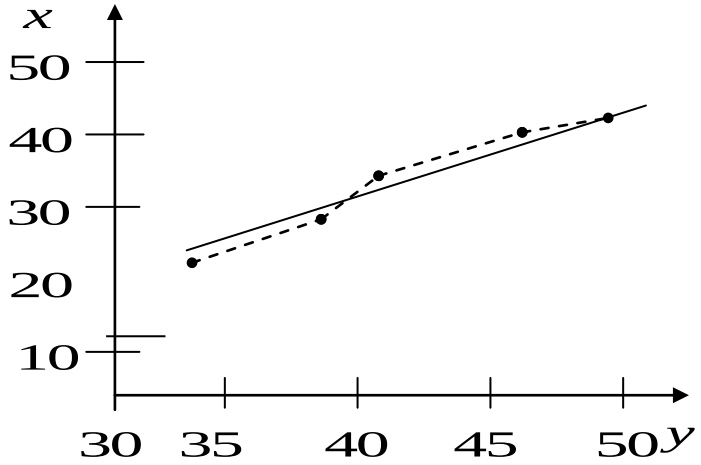


Рис. 10 Ломаная линии регрессии по групповым средним; график уравнения регрессии X и Y

Уравнение регрессии для зависимости $\overline{x_y} = g(y)$ ищем в виде: $\overline{x_y} = b_0 + b_1 y$.

Система нормальных уравнений имеет вид:

$$\begin{cases} nb_0 + b_1 \sum_j y_j m_i = \sum_i x_i m_i; \\ b_0 \sum_j y_j m_j + b_1 \sum_j m_j x_j^2 = \sum_{ij} m_{ij} y_j x_i \ . \end{cases}$$

Исследуя данные таблицы, имеем:

$$\begin{cases} 100 b_0 + 4300 b_1 = 3445; \\ 4300 b_0 + 186480 b_1 = 149710 \ . \end{cases}$$

Решая систему, получаем $b_0 = -8{,}414, b_1 = 0{,}9968$ уравнение регрессии X по Y принимает вид: $\overline{x_y} = -8{,}414 + 0{,}9968$.

График этой функции строим по двум точкам:

При $y = 40$ $\overline{x_y} = -8{,}414 + 39{,}872 = 31{,}458$.

При $y = 50$ $\overline{x_y} = -8{,}414 + 49{,}84 = 41{,}426$.

3) Коэффициент корреляции удобно вычислить по формуле (9.52), так все необходимые суммы получены в расчетной табл. 9:

При $x = 25$ $\overline{y_x} = 22{,}089 + 15{,}175 = 37{,}264;$

$$\sum_{ij} x_i y_j m_{ij} = 149710, \sum_i x_i m_i = 3445, \sum_j y_j - m_j = 4300, \sum_i x_i^2 m_i = 121275,$$

$$\sum_i y_j^2 m_j = 186480, n = 100, \text{ тогда}$$

$$r^* = \frac{100 \cdot 149710 - 3445 \cdot 4300}{\sqrt{(100 \cdot 121275 - 11868025)}\sqrt{(100 \cdot 186480 - 1849000)}} = 0{,}778.$$

По формуле (9.55) найдем оценку среднеквадратического отклонения коэффициента корреляции:

$$\sigma_r \approx \frac{1 - 0{,}6058}{10} = 0{,}039.$$

Зададимся доверительной вероятностью $p = 0{,}95$ (уровень значимости $\alpha = 1 - p = 0{,}05$), и по табл. 2. приложения найдем значение t_p: $t_{0{,}95} = 1{,}96.$

Вычисляем величину $t = \frac{|r^*|}{\sigma_r} \approx \frac{0{,}778}{0{,}039} = 19{,}95.$

Величина $t_p >> t_{0{,}95}$ (знак «>>» означает «значительно больше»), поэтому можно сделать вывод о том, что корреляционная зависимость. Так как $r = 0{,}778,$ то есть достаточно близок единицы, то эта зависимость может считаться вполне достаточно тесной; положительный знак коэффициента корреляции указывает на прямо пропорциональную зависимость, то есть с возрастанием значений, например, X значения Y также будут возрастать. Графики уравнений регрессии также подтверждают этот вывод.

При наибольшем объеме выборки ($n \le 50$) величина выборочного коэффициента корреляции r^* считается значимо отличной от нуля, если выполняется неравенство:

$$r^{*2} > \left[1 + (n-2)/t_\alpha^2\right]^{-1}$$

Где t_α есть критическое значение t -распределение Стьюдента с $(n-2)$ степенями свободы, соответствующее выбранному уровню значимости α. Поэтому для проверки значимости выборочного коэффициента корреляцтии вычисляется величина:

$$t = r^* \sqrt{\frac{n-2}{1 - r^{*2}}}; \qquad (22)$$

Для проверки нулевой гипотезы находят по табл. 6 распределения Стьюдента по фиксированному уровню значимости α и числу степеней свободы $(n-2)$ критическое значение $t_{\alpha;n-2}$, удовлетворяющее условию $p(|t| \ge t_{\alpha;n-2}) = \alpha$. Если для $t_{набл.}$ (значения t), вычисленного по формуле (22) выполняется $|t_{набл.}| \ge t_{\alpha;n-2}$, то нулевую гипотезу об отсутствии линейной зависимости между переменными X и Y следует отвергнуть. Если же $t_{набл.} < t_{\alpha;n-2}$, то нет оснований отвергать нулевую гипотезу о некоррелированности переменных X и Y.

Если же известно, что $r \ne 0$, то необходимо воспользоваться Z-преобразованием Фишера (независящим от r, n):

$$Z = \frac{1}{2}\ln\frac{1+r^*}{1-r^*}$$

Все вышеприведенные рассуждения и формулы, если подходить достаточно строго, справедливы в предложении, что двумерное распределение исследуемых переменных (X,Y) в генеральной совокупности предполагается нормальным или близким к нему.

Пример 2. В результате наблюдений получена выборка:

X	70	110	85	65	100	90	120	80	130	110
Y	2,8	3,5	2,4	2,1	3,4	3,2	3,6	2,5	4,1	3,3

Требуется построить корреляционное поле найти уравнение регрессии, сделать вывод о тесное связи между переменными (показателями) X и Y, оценить ожидаемое срежднее значение Y при $x = 80.$

На рис. 10 построено корреляционное поле.

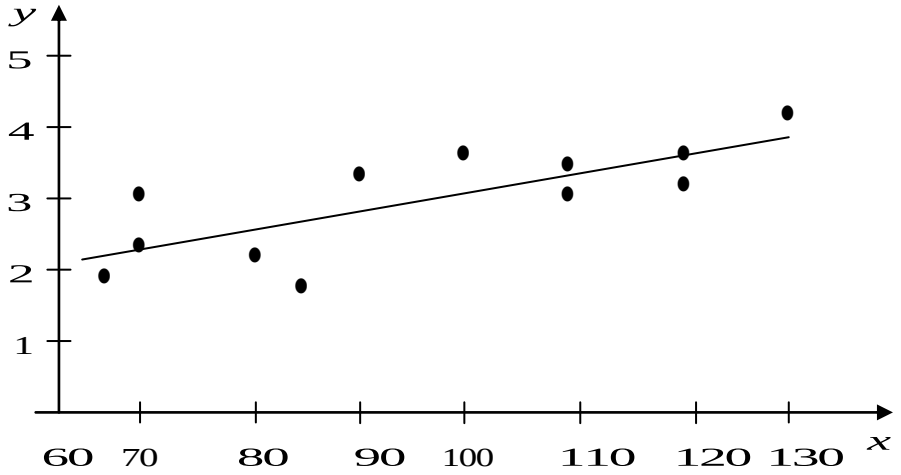


Рис. 10 Корреляционное поле и прямая регрессия

Расположение точек корреляционного поля позволяет высказать предположение о линейном виде корреляционной зависимости между переменными X и Y. Найдем коэффициенты уравнения регрессии $\overline{y_x} = a_0 + a_1 x.$

Составим расчетную табл. 10.

X	**Y**	$\mathbf{X^2}$	$\mathbf{Y^2}$	**XY**
70	2,8	4900	7,84	196
11	3,5	12100	12,25	385
85	2,4	7225	5,76	204
65	2,1	4225	4,41	136,5
100	3,4	10000	11,56	340
90	3,2	8100	10,24	288
120	3,6	14400	12,96	432
80	2,5	6400	6,25	200
130	4,1	16900	16,81	533
110	3,3	12100	10,89	363
960	30,9	96350	98,97	3077,5

В последней строке табл. 10 получены значения: $\sum_{i=1}^{10} x_i = 960; \sum_{i=1}^{10} y_i = 30{,}9;$ $\sum_{i=1}^{10} x^2{}_i = 96350; \sum_{i=1}^{10} y^2{}_i = 98{,}97; \sum_{i=1}^{10} x_i y_i = 3077{,}5.$

Получим (см (9.47)) систему нормальных уравнений:

$$\begin{cases} 10a_0 + 960a_1 = 30{,}9; \\ 960a_0 + 96350a_1 = 3077{,}5. \end{cases}$$

Из которй следует $a_0 = 0{,}5445$, $a_1 = 0{,}0265$. (см. (13), (14)).

Тогда уравнение регрессии имеет вид: $\overline{y_x} = 0{,}0265x + 0{,}5445$. строим пряму регрессии по точкам: при При $x = 70$ $\overline{y_x} = 2{,}3995$; $x = 120$ $\overline{y_x} = 3{,}7245$.

Найдем по формуле (15) выборочный коэффициент корреляции r^*. Предварительно вычислим, учитывая (9.6), $\overline{x} = \frac{960}{10} = 96$, $\overline{y} = \frac{30{,}9}{10} = 3{,}09$.

Используя формулу, найдем оценки выборочных дисперсий $S^2{}_x$ и $S^2{}_y$ по формуле:

$$S^2 = \frac{1}{n}\sum x^2{}_i - \frac{1}{n^2}(\sum x_i)^2$$

Тогда $S_x^2 = \frac{1}{10}96350 - \frac{1}{100}960^2 = 419,$ аналогично $S_y^2 = \frac{1}{10}98{,}97 - \frac{1}{100}30{,}9^2 = 0{,}3489$. Выборочный корреляционный момент $K^*{}_{xy}$ найдем по формуле (19): $K^*{}_{xy} = \frac{1}{10}3077{,}5 - 96 \cdot 3{,}09 = 11{,}11$.

Учитывая что объем выборки небольшой, найдем несмещенные оценки выборочных дисперсий и корреляционного момента, умножив их вычисленные значения на величину: $\frac{n}{n-1} = \frac{10}{9}$, тогда $\overline{S}_x^2 = \frac{10}{9} \cdot 419 = 465{,}5555,$ $S_y^2 = \frac{10}{9} \cdot 0{,}3489 = 0{,}3877$, $\tilde{K}^*{}_{xy} = \frac{10}{9} \cdot 11{,}11 = 12{,}3444$. Тогда по формуле (9.50), числитель которой есть выборочный корреляционный момент получим: $r^* = \frac{12{,}3444}{\sqrt{465{,}5555 \cdot 0{,}3877}} = 0{,}9187$.

Величина выборочного коэффициента корреляции говорит о достаточно тесной линейной зависимости между переменными X и Y. Те не менее проверим нулевую гипотезу. По формуле (22) найдем величину $t = 0{,}9187\sqrt{\frac{10-2}{1-0{,}844}} = 6{,}579$. Зададимся уровнем значимости $\alpha = 0{,}05,$ доверительная вероятность $p = 1 - \alpha = 0{,}95,$ число степеней свободы $n - 2 = 10 - 2 = 8,$ и по табл. 6 находим критическое значение $t_{0{,}05;8} = 2{,}31$. Поскольку $t_{табл.} > t_{0{,}05;8}$, то нулевую гипотезу об отсутствии линейной зависимости надо отвергнуть и признать наличие достаточно близкой линейной

корреляционной связи между переменными X и Y. Прогноз среднего значения переменной Y при X=80 составит $\overline{y_x} = 0{,}0265 \cdot 80 + 0{,}5445 \approx 2{,}66.$