ПАРНАЯ ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ

ГЛАВА 3

Содержание

§ 9. Понятие корреляционной зависимости. Задачи теории корреляции

В новых условиях хозяйственной деятельности предприятий возрастает роль экономико-математических методов для управления производством. Управление производством — это сложный динамический процесс. Поэтому при выработке оптимального решения по управлению производственно-хозяйственной деятельностью предприятия необходимо не только учитывать изменения параметров и характеристик, описывающих эту деятельность, но и уметь их прогнозировать, основываясь на экономических законах, которые наиболее полно отражают взаимосвязи основных показателей предприятия и его подразделений. Математическая формализация этих связей создает условия для экономического обоснования целесообразных объемов производимой продукции,

§ 9. Понятие корреляционной зависимости. Задачи теории корреляции

Для решения этих задач применяют методы корреляционного анализа. При анализе зависимостей между производственными показателями методами корреляционного анализа выделяют два основных типа переменных количественных признаков: независимые переменные (факторные признаки) и зависимые переменные (результативные признаки).

При изучении взаимосвязей между переменными признаками надо, прежде всего, установить, к какому типу зависимостей относится эта связь.

Зависимость между признаками *X* и *Y* называется *корреляционной*, если каждому возможному значению *xi* признака *X* сопоставляется условная средняя соответствующего распределения признака *Y*

§ 9. Понятие корреляционной зависимости. Задачи теории корреляции

Среднее арифметическое значение признака *Y* , вычисленное при условии, что признак *X* принимает фиксированное значение x_i, называется *условным средним*, обозначается через $\bar{y}_{xi}$ и вычисляется по формуле:

$$\bar{y}_{x_i} = \frac{\sum n_{ij}\ y_j}{n_{x_i}},$$

где n_{ij} — частоты, показывающие сколько раз повторяются парные значения x_i, y_j в данной выборке, n_{xi} — частота появления значения xi.

§ 9. Понятие корреляционной зависимости. Задачи теории корреляции

Теория корреляции изучает такую зависимость между признаками X и Y , при которой с изменением одного признака меняется распределение другого. Она применяется для того, чтобы при сложном взаимодействии посторонних факторов выяснить, какова должна быть зависимость между признаками X и Y , если бы посторонние факторы не изменялись и своим изменением не искажали истинную статистическую зависимость.

В теории корреляции решается триединая задача, методологической основой которой является триада:

Модель — Свойства — Адекватность

§ 9. Понятие корреляционной зависимости. Задачи теории корреляции

Первая задача — поиск подходящей модели.

На основе опытных данных выявляется характер корреляционной зависимости между признаками *X* и *Y* . При парной корреляции для ее решения применяют графический метод.

Если в корреляционном поле точки (x_i, y_j) хорошо ложатся на прямую, то можно предположить, что связь между признаками *X* и *Y* носит линейный характер. Если точки не ложатся на прямую, то связь будет нелинейной.

Исходя из геометрических соображений, выбирают уравнение линии, которое называют уравнением регрессии, и находят неизвестные параметры,

§ 9. Понятие корреляционной зависимости. Задачи теории корреляции

Вторая задача — изучение свойств модели.

Определяется теснота связи между признаками, включенными в модель, по коэффициенту r корреляции (в случае линейной корреляции) или по корреляционным отношениям η_{yx}, η_{xy} (в случае криволинейной корреляции).

§ 9. Понятие корреляционной зависимости. Задачи теории корреляции

Третья задача — выявление степени адекватности построенной корреляционной модели (проверяется соответствие полученного уравнения регрессии опытным данным).

Если данная модель оказалась не адекватной, то всё начинается сначала — строят новую модель.

§ 10. Парная линейная корреляция

Предположим, что на основе геометрических, физических или других соображений установлено, что между двумя количественными признаками *X* и *Y* существует линейная корреляционная зависимость. Тогда уравнение регрессии записывают в виде:

$$\hat{y}_x = a_0 + a_1 x .$$

§ 10. Парная линейная корреляция

Пусть опытные данные не сгруппированы в корреляционную таблицу, т. е. заданы в виде табл. 18.

Таблица 18

x_i	x_1	x_2	x_3	...	x_k
y_i	y_1	y_2	y_3	...	y_k

§ 10. Парная линейная корреляция

В этом случае значения a_0, a_1, являющиеся оценками истинных величин уравнения регрессии, находят по методу наименьших квадратов, решая систему линейных алгебраических уравнений (СЛАУ) относительно a_0, a_1:

$$\begin{cases} na_0 + [x]a_1 = [y], \\ [x]a_0 + [x^2]a_1 = [xy], \end{cases}$$

где $[x] = \sum_{i=1}^{k} x_i$, $[y] = \sum_{i=1}^{k} y_i$, $[x^2] = \sum_{i=1}^{k} x_i^2$, $[xy] = \sum_{i=1}^{k} x_i y_i$.

§ 10. Парная линейная корреляция

Для нахождения сумм, входящих в систему, составляется табл. 19.

Таблица 19

x_i	y_i	$x_i y_i$	x_i^2
$[x]$	$[y]$	$[xy]$	$[x^2]$

§ 10. Парная линейная корреляция

Если опытные данные сгруппированы в корреляционную таблицу, то значения a_0 и a_1 уравнения регрессии находят по методу наименьших квадратов, решая СЛАУ

$$\begin{cases} na_0 + [n_x x]a_1 = [n_y y], \\ [n_x x]a_0 + [n_x x^2]a_1 = [n_{xy} xy], \end{cases}$$

где n_X и n_Y — частоты признаков X и Y , n_{XY} — частота совместного появления признаков X и Y .

§ 10. Парная линейная корреляция

Для нахождения сумм, входящих в систему, составляется табл. 20.

Таблица 20

y \ x	x_1	x_2	…	x_k	n_y	$n_y y$
y_1			…			
y_2			…			
…	…	…	…	…	…	…
y_m			…			
n_x			…			$[n_y y]$
$n_x x$			…		$[n_x x]$	
$n_x x^2$			…		$[n_x x^2]$	
$n_{xy} xy$			…		$[n_{xy} xy]$	

Суммы $[n_X x]$, $[n_X x^2]$, $[n_{XY} xy]$ в табл. 20 находятся по строкам, а сумма $[n_Y y]$ — по последнему столбцу табл. 20.

§ 10. Парная линейная корреляция

В уравнении регрессии параметр a_0 характеризует усредненное влияние на результативный признак Y неучтенных (не выявленных для исследования) факторных признаков X_j.

Параметр a_1 показывает, на сколько изменяется в среднем значение результативного признака Y при увеличении факторного признака на единицу.

Используя параметр a_1, вычисляют коэффициент эластичности $K_э$ по формуле:

$$K_э = a_1 \frac{\bar{x}}{\bar{y}}.$$

§ 10. Парная линейная корреляция

Коэффициент эластичности $K_Э$ показывает, на сколько процентов изменяется результативный признак *Y* при изменении факторного признака *X* на 1 %.

В случае линейной корреляционной зависимости между признаками *X* и *Y* , уравнения регрессий находят по формулам:

$$\hat{y}_x = \bar{y} + r\frac{S_y}{S_x}(x - \bar{x}),$$

$$\hat{x}_y = \bar{x} + r\frac{S_x}{S_y}(y - \bar{y}),$$

где $\bar{x}$, $\bar{y}$ — выборочные средние признаков X и Y .

§ 10. Парная линейная корреляция

S_X , S_Y — выборочные средние квадратические отклонения признаков X и Y , вычисляемые по формулам:

$$\hat{S}_x = \sqrt{\hat{S}_x^2}\ , \text{где}\ \hat{S}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (n < 50)\ ,$$

$$\hat{S}_y = \sqrt{\hat{S}_y^2}\ , \text{где}\ \hat{S}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (n < 50)\ .$$

§ 10. Парная линейная корреляция

При $n \geq 50$ S_X , S_y находят по формулам:

$$S_x = \sqrt{S_x^2}\text{, где } S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$S_y = \sqrt{S_y^2}\text{, где } S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

§ 10. Парная линейная корреляция

Коэффициент линейной корреляции r находят по формуле:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x \cdot S_y},$$

где $\overline{xy}$ — средняя произведения значений признаков X и Y ,

$\bar{x}$, $\bar{y}$ — средние значения признаков X и Y ,

S_X , S_Y — выборочные средние квадратические отклонения признаков X и Y, вычисленные по вышеприведенным формулам

§ 10. Парная линейная корреляция

$$\hat{y}_x = \bar{y} + r\frac{S_y}{S_x}(x - \bar{x})$$ — уравнение регрессии y на x

$$\hat{x}_y = \bar{x} + r\frac{S_x}{S_y}(y - \bar{y}),$$ — уравнение регрессии x на y .

§ 10. Парная линейная корреляция

Если данные выборки для признаков *X* и *Y* заданы в виде корреляционной таблицы и объем выборки $n>30$, то для нахождения величин, входящих в уравнения линий регрессий, переходят к вспомогательному распределению с условными вариантами u_i и v_j , вычисляемых по формулам

$$u_i = \frac{x_i - C_1}{h_1},$$

$$v_j = \frac{y_j - C_2}{h_2},$$

где $C_1 = M_0X$, $C_2 = M_0Y$, h_1 и h_2 — шаги значений признаков X и Y.

§ 10. Парная линейная корреляция

Выборочный коэффициент линейной корреляции r в этом случае находят по формуле

$$r = \frac{\sum n_{uv}\, uv - n\bar{u}\,\bar{v}}{nS_u S_v},$$

где

$$S_u = \sqrt{\overline{u^2} - \bar{u}^2},\ S_v = \sqrt{\overline{v^2} - \bar{v}^2}.$$

§ 10. Парная линейная корреляция

Для нахождения суммы $\sum n_{uv}uv$ составляется расчетная табл. 21.

Таблица 21

u \ v	v_1	v_2	…	v_k	n_u
u_1	u_1v_1 / $n_{u_1v_1}$	u_1v_2 / $n_{u_1v_2}$	…	u_1v_k / $n_{u_1v_k}$	n_{u_1}
u_2	u_2v_1 / $n_{u_2v_1}$	u_2v_2 / $n_{u_2v_2}$	…	u_2v_k / $n_{u_2v_k}$	n_{u_2}
…	…	…	…	…	…
u_n	u_nv_1 / $n_{u_nv_1}$	u_nv_2 / $n_{u_nv_2}$	…	u_nv_k / $n_{u_nv_k}$	n_{u_n}
n_v	n_{v_1}	n_{v_2}	…	n_{v_k}	$\sum n_{uv}uv$

§ 10. Парная линейная корреляция

Статистики $\bar{x}$, $\bar{y}$, S_X, S_Y находят по формулам

$$\bar{x} = \bar{u}h_1 + C_1,\ \bar{y} = \bar{v}h_2 + C_2,\ S_x = S_u h_1,\ S_y = S_v h_2 .$$

§ 11. Коэффициент корреляции, его свойства и значимость

После выбора функции как формы корреляционной зависимости между признаками X и Y решается задача, состоящая в определении тесноты связи между ними, в оценке рассеяния относительно линии регрессии значений одного признака для различных значений другого.

Для этого используют выборочный коэффициент r корреляции,

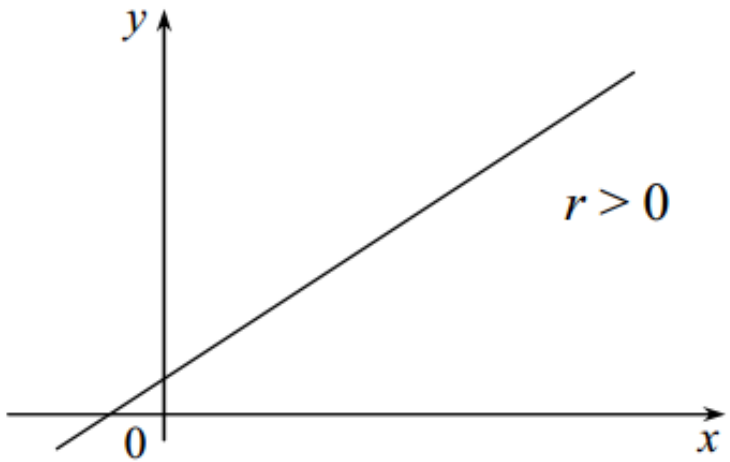

Рис. 6.

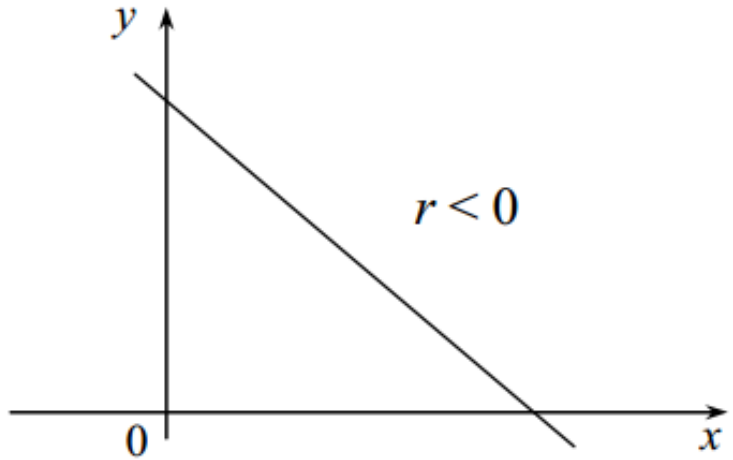

Рис. 7.

§ 11. Коэффициент корреляции, его свойства и значимость

Линейный коэффициент корреляции изменяется на отрезке $[-1; 1]$, то есть $|\ r\ | \leq 1$.

Если $r = \pm 1$, то корреляционная зависимость становится функциональной.

В случае $r > 0$ говорят о *положительной корреляции* величин *X*, *Y* (рис. 6); например, вес и рост человека связаны положительной корреляцией; в случае $r < 0$ — об *отрицательной корреляции* (рис. 7).

Положительная корреляция между случайными величинами означает, что при возрастании одной из них другая имеет тенденцию в среднем возрастать; отрицательная корреляция означает, что при возрастании одной из случайных величин другая имеет тенденцию в среднем убывать.

§ 11. Коэффициент корреляции, его свойства и значимость

Если $r = 0$, то линейная связь между признаками X и Y отсутствует, но может существовать криволинейная корреляционная связь или нелинейная функциональная. Оценку тесноты линейной корреляционной связи определяют, пользуясь табл. 22.

Таблица 22

Теснота связи	Величина r	
	Положительная	Отрицательная
Линейной связи нет	$0 \div 0{,}2$	$0 \div -0{,}2$
Слабая	$0{,}2 \div 0{,}5$	$-0{,}2 \div -0{,}5$
Средняя	$0{,}5 \div 0{,}75$	$-0{,}5 \div -0{,}75$
Сильная	$0{,}75 \div 0{,}95$	$-0{,}75 \div -0{,}95$
Функциональная	$0{,}95 \div 1$	$-0{,}95 \div -1$

§ 11. Коэффициент корреляции, его свойства и значимость

Значимость выборочного коэффициента корреляции проверяют по критерию Стьюдента. По опытным данным вычисляют расчетную статистику t_p, пользуясь формулой:

$$t_p = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}.$$

§ 11. Коэффициент корреляции, его свойства и значимость

Затем по таблице критических точек распределения Стьюдента (приложение 6) по заданному уровню значимости α и числу степеней свободы $k = n - 2$ находят табличное значение $t_{кр}$ двусторонней критической области.

Если $t_р < t_{кр}$, то коэффициент корреляции r — незначимый (мало отличается от нуля) и признаки *X* и *Y* некоррелированные.

Если $t_р > t_{кр}$, то приходят к выводу о наличии линейной корреляционной связи.

§ 12. Определение надежности (доверительного интервала) коэффициента корреляции

Коэффициент корреляции, как правило, рассчитывается по данным выборки. Чтобы полученный результат распространить на генеральную совокупность, возможно возникновение некоторой ошибки, которую оценивают с помощью средней квадратичной ошибки σ_r.

С помощью σ_r производят оценку надежности коэффициента корреляции, построив доверительные интервалы для различных объемов выборки.

§ 12. Определение надежности (доверительного интервала) коэффициента корреляции

Пусть число n наблюдений пар чисел (x; y) меньше 50 ($n < 50$). В этом случае средняя квадратическая ошибка σ_r вычисляется по формуле

$$\sigma_r = \frac{1-r^2}{\sqrt{n-2}},$$

где r — коэффициент парной линейной корреляции, n — объем выборки.

Доверительный интервал для оценки r находят по формуле

$$r - t_\gamma \sigma_r \le \hat{r} \le r + t_\gamma \sigma_r,$$

Где t_γ находят по таблице значений функции Лапласа F(x) (приложение 2).

§ 12. Определение надежности (доверительного интервала) коэффициента корреляции

Пример. Если задать надежность γ=0,95, то F(x)=$\frac{\gamma}{2}$=0,475 и t_{γ} =1,96

§ 12. Определение надежности (доверительного интервала) коэффициента корреляции

Если объем выборки $n > 50$, то погрешность σ_r для коэффициента корреляции r находят также по указанной формуле (см. слайд 32).

Затем вычисляют отношение r / σ_r .

Если это отношение больше 3, то можно считать, что найденный коэффициент корреляции r отражает истинную зависимость между признаками *X* и *Y*.

Величина $r - 3\sigma_r$ является, как правило, *гарантийным минимумом*, а величина $r + 3\sigma_r$ — *гарантийным максимумом* коэффициента корреляции r и доверительный интервал для оценки r запишется в виде

$$r - 3\sigma_r \le \hat{r} \le r + 3\sigma_r$$

§ 13. Коэффициент детерминации

Линейный коэффициент корреляции оценивает тесноту взаимосвязи между признаками и показывает, является ли эта корреляция положительной или отрицательной.

Однако понятия тесноты взаимосвязи бывает недостаточно при содержательном анализе взаимосвязей. В частности коэффициент корреляции не показывает степень воздействия факторного признака *X* на результативный *Y*.

Таким показателем является коэффициент детерминации.

§ 13. Коэффициент детерминации

Пусть по опытным данным для признаков X и Y получены уравнения регрессий $\tilde{y}_x = a_0 + a_1 x$ и $\tilde{x}_y = b_0 + b_1 y$.

Величину $a_1 b_1 = r^2$ называют *коэффициентом детерминации*.

Этот коэффициент детерминации можно находить и по формуле

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_{x_i})^2}{\sum (y_i - \bar{y})^2},$$

где y_i — опытные значения признака Y , $\tilde{y}_{xi}$ — значения y , найденные по уравнению регрессии, $\bar{y}$ — средняя признака Y .

Данной формулой пользуется тогда, когда общее число значений y_i равно числу значений x_i признака X .

§ 13. Коэффициент детерминации

Коэффициент детерминации используется,

во-первых, для контроля вычислений, проводимых при получении уравнений регрессий ($r^2 = a_1 b_1$)

во-вторых, он показывает, какую часть рассеяния результативного признака *Y* можно объяснить принятой регрессионной моделью.

§ 14. Проверка адекватности модели

Для проверки соответствия уравнения регрессии $\tilde{y}_x = a_0 + a_1 x$ опытным данным применяют критерий Фишера — Снедекора. Вычисляют статистику $F_н$ по формуле:

$$F_н = \frac{R^2(n-2)}{1-R^2},$$

где R^2 — коэффициент детерминации, n — объем выборки.

§ 14. Проверка адекватности модели

Чем ближе значение где R^2 к единице, тем лучше модель согласуется с опытными данными. Затем при заданном уровне значимости α и числах степеней свободы k_1 =1, k_2= n - 2 находят по таблице критических точек распределения Фишера — Снедекора (приложение 7) $F_T = F_{\alpha:k1:k2}$.

Если окажется, что $F_н > F_T$, то полученное уравнение линейной регрессии согласуется с данными опыта.

§ 14. Проверка адекватности модели

Чем ближе значение где R^2 к единице, тем лучше модель согласуется с опытными данными. Затем при заданном уровне значимости α и числах степеней свободы k_1 =1, k_2= n - 2 находят по таблице критических точек распределения Фишера — Снедекора (приложение 7) $F_T = F_{\alpha:k1:k2}$.

Если окажется, что $F_н > F_T$, то полученное уравнение линейной регрессии согласуется с данными опыта.

Замечание. Формулой для вычисления $F_н$ пользуются тогда, когда исходные данные заданы не в виде корреляционной таблицы.

§ 14. Проверка адекватности модели

Если опытные данные заданы в виде корреляционной таблицы, то проверку модели на адекватность можно выполнить тогда, когда общее число значений y_i больше числа значений x_j. В этом случае находят остаточную сумму квадратов Q_e, характеризующую влияние неучтенных в модели факторов, по формуле

$$Q_e = Q - Q_R,$$

где

$Q = \sum(y_i - \bar{y})^2$ — сумма квадратов отклонений значений y_i от средней $\bar{y}$

$Q_R = \sum(\bar{y}_{xi} - \bar{y})^2$ — сумма квадратов отклонений условных средних $\bar{y}_{xi}$ от средней $\bar{y}$

§ 14. Проверка адекватности модели

Затем вычисляется статистика $F_{\text{Н}}$ по формуле

$$F_{\text{Н}} = \frac{Q_R(n-2)}{Q_e}.$$

По таблице критических точек распределения Фишера – Снедекора (приложение 7) при заданном уровне значимости α и числах степеней свободы k_1 =1, k_2= n - 2 находят по таблице критических точек распределения Фишера — Снедекора (приложение 7) $F_T = F_{\alpha:k1:k2}$.

Если окажется, что $F_{\text{Н}} > F_T$, то модельное уравнение регрессии значимо описывает опытные данные, в противном случае если $F_{\text{Н}} < F_T$ — нет.

§ 15. Оценка величины погрешности

После проверки модельного уравнения линейной корреляции на адекватность находят относительную погрешность уравнения по формуле:

$$\delta = \frac{\sigma_u}{\bar{y}} \cdot 100\%,$$

где $\sigma_u = \sqrt{Du}$, σ_u — стандартная ошибка уравнения регрессии,

$D_u = \frac{\sum (u_i - \bar{u})^2}{n-2}$ — остаточная дисперсия,

$u_i = y_i - \hat{y}_{x_i}$, y_i — опытные значения y,

$\hat{y}_{x_i}$ — значения y, полученные по уравнению регрессии,

$\bar{u} = \frac{1}{n}\sum (y_i - \hat{y}_{x_i})$ — среднее значение u_i,

n — объем выборки.

§ 15. Оценка величины погрешности

Если величина δ мала, то прогнозные качества оцененного регрессионного уравнения высоки.

Одновременно производят оценку коэффициентов уравнения регрессии $\tilde{\tilde{y}}_x = a_0 + a_1 \mathrm{x}$.

Пусть S_{a0} и S_{a1} — стандартные ошибки соответственно коэффициентов a_0 и a_1 уравнения регрессии. Их вычисление производят по формулам:

$$S_{a_0} = S_y \sqrt{1 - r^2} \cdot \sqrt{\frac{[x^2]}{n[x^2] - [x]^2}},$$

$$S_{a_1} = S_y \sqrt{1 - r^2} \cdot \sqrt{\frac{n}{n[x^2] - [x]^2}}.$$

§ 15. Оценка величины погрешности

Коэффициенты a_0 и a_1 считаются значимыми, если $2S_{ai} < |a_i|$. Если же коэффициенты a_0 и a_1 незначимы, то ситуацию можно поправить путем увеличения объема выборки n , увеличения числа факторов, включаемых в модель или изменения формы уравнения связи.

Спасибо за внимание!