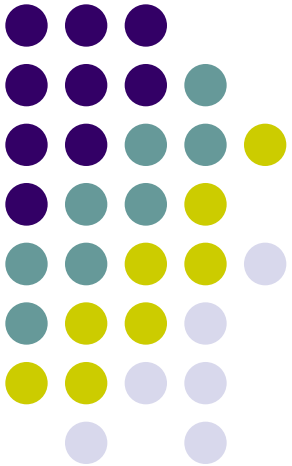
ЛЕКЦИЯ 2
Методы и модели корреляционно-регрессионного анализа

Вопрос 1. Общие сведения

Вопрос 2. Исходные предпосылки регрессионного анализа и свойства оценок

Вопрос 3. Этапы построения многофакторной корреляционно-регрессионной модели



Виды зависимостей между экономическими явлениями и процессами

Функциональная

- имеется однозначное отображение множества *A* на множество *B.*
- Множество *A* называют областью определения функции, а множество *B* - множеством значений функции.
- Функциональная зависимость встречается редко.

Стохастическая (вероятностная, статистическая)

- В большинстве случаев функция *(Y)* или аргумент (X) — случайные величины.
- Если на случайную величину X действуют факторы Z1, **Z2,** ..., **V1**, V2, а на У— Z0, **Z2, V1**, V3 ..., то наличие двух общих факторов *Z2* и V1 позволит говорить о вероятностной или статистической зависимости между *X и Y.*

Стохастическая зависимость: статистическая

- ***Статистической*** называется зависимость между случайными величинами, при которой изменение одной из величин влечет за собой изменение закона распределения другой величины.
- В частном случае статистическая зависимость проявляется в том, что при изменении одной из величин изменяется математическое ожидание другой. **В этом случае говорят о корреляции или корреляционной зависимости.**
- Статистическая зависимость проявляется только в массовом процессе, при большом числе единиц совокупности.

Стохастическая зависимость: вероятностная

- При ***стохастической*** закономерности для заданных значений зависимой переменной можно указать ряд значений объясняющей переменной, случайно рассеянных в интервале.
- Каждому фиксированному значению аргумента соответствует определенное статистическое распределение значений функции.
- Это обусловливается тем, что зависимая переменная, кроме выделенной переменной, подвержена влиянию ряда неконтролируемых или неучтенных факторов.
- Поскольку значения зависимой переменной подвержены случайному разбросу, они не могут быть предсказаны с достаточной точностью, а только **указаны с определенной вероятностью**.
- Односторонняя вероятностная зависимость между случайными величинами, устанавливающая соответствие между этими величинами есть **регрессия**.
- Односторонняя стохастическая зависимость выражается с помощью функции, которая называется **регрессией**.

ВИДЫ РЕГРЕССИЙ

1. Регрессия относительно числа переменных:

- простая регрессия — регрессия между двумя переменными;
- множественная регрессия — регрессия между зависимой переменной y и несколькими объясняющими переменными x1, x2, ..., xm.

Множественная линейная регрессия имеет следующий вид:

$$Y = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m \quad ,$$

где y - функция регрессии;

- x1, x2, ..., xm — независимые переменные;
- a1, a2, ..., am —коэффициенты регрессии;
- a0 — свободный член уравнения;
- m — число факторов, включаемых в модель.

ВИДЫ РЕГРЕССИЙ (продолжение)

2. Регрессия относительно формы зависимости:

- линейная регрессия - регрессия выражаемая линейной функцией;
- нелинейная регрессия - регрессия выражаемая нелинейной функцией.

3. В зависимости от характера регрессии:

- положительная регрессия: она имеет место, если с увеличением (уменьшением) объясняющей переменной значения зависимой переменной также соответственно увеличиваются (уменьшаются);
- отрицательная регрессия: в этом случае с увеличением или уменьшением объясняющей переменной зависимая переменная уменьшается или увеличивается.

4. Относительно типа соединения явлений:

- непосредственная регрессия: в этом случае зависимая и объясняющая переменные связаны непосредственно друг с другом;
- косвенная регрессия: в этом случае объясняющая переменная действует на зависимую через ряд других переменных;
- ложная регрессия: она возникает при формальном подходе к исследуемым явлениям без уяснения того, какие причины обусловливают данную связь.

КОРРЕЛЯЦИЯ

Корреляция в широком смысле слова означает связь, соотношение между объективно существующими явлениями.

- Связи между явлениями могут быть различны по силе.
- При измерении тесноты связи говорят о корреляции в узком смысле слова.
- Если случайные переменные причинно обусловлены и можно в вероятностном смысле высказаться об их связи, то имеется корреляция.
- Понятия «корреляция» и «регрессия» тесно связаны между собой.
- В корреляционном анализе оценивается сила связи, а в регрессионном анализе исследуется ее форма.

Виды корреляции

1) относительно характера:

- Положительная (прямая);
- Отрицательная (обратная);

2) относительно числа переменных:

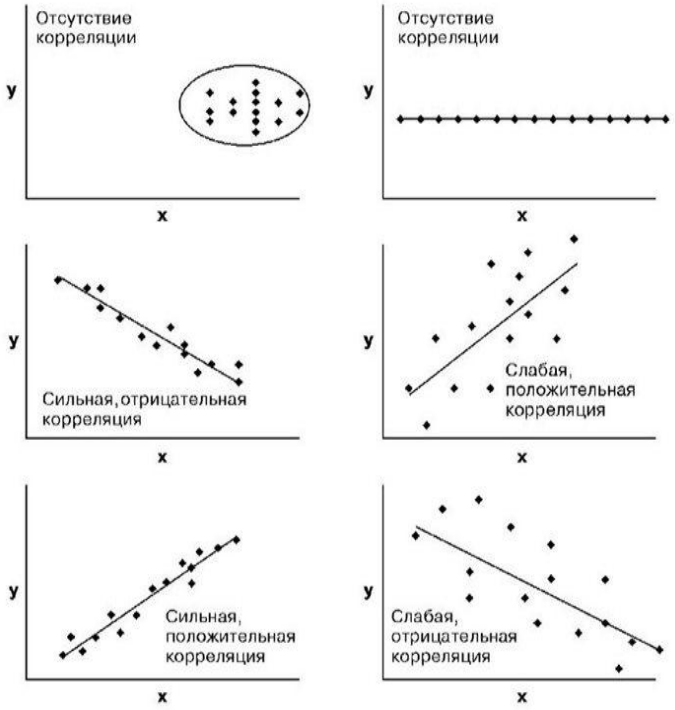
- простая;
- множественная;
- частная;

3) относительно формы связи:

- Линейная (прямолинейная);
- Нелинейная (криволинейная);

4) относительно типа соединения:

- непосредственная;
- косвенная;
- ложная;



Задачи корреляционного анализа

- Измерение степени связности (тесноты, силы) двух и более явлений:
  - Здесь речь идет в основном о подтверждении уже известных связей.
- Отбор факторов, оказывающих наиболее существенное влияние на результативный признак на основе измерения тесноты связи между явлениями.
- Обнаружение неизвестных причинных связей.
  - Корреляция непосредственно не выявляет причинных связей между явлениями, но устанавливает степень необходимости этих связей и достоверность суждений об их наличии.
  - Причинный характер связей выясняется с помощью логически-профессиональных рассуждений, раскрывающих механизм связей.

Задачи регрессионного анализа

- Установление формы зависимости (линейная или нелинейная; положительная или отрицательная и т. д.).
- Определение функции регрессии и установление влияния факторов на зависимую переменную:

  Важно:
  - определить форму регрессии,
  - казать общую тенденцию изменения зависимой переменной
  - выяснить, каково было бы действие на зависимую переменную главных факторов, если бы прочие не изменялись и если бы были исключены случайные элементы. Для этого определяют функцию регрессии в виде математического уравнения того или иного типа.
- Оценка неизвестных значений зависимой переменной, т. е. решение задач экстраполяции и интерполяции.
  - В ходе экстраполяции распространяются тенденции, установленные в прошлом, на будущий период.
  - В ходе интерполяции определяют недостающие значения, соответствующие моментам времени между известными моментами, т.е. определяют значения зависимой переменной внутри интервала заданных значений факторов.

Выборочные уравнения регрессии

- Условное математическое ожидание случайной величины *Y: M(У/X)* есть функция от *X,* которая называется *функцией регрессии* и равна f(x), т. е.

$$M(У/X) = f(x); \qquad (5.2)$$

Аналогично

$$M(X/Y) = \varphi(y). \qquad (5.3)$$

Графическое изображение f(x), или (y) *называется линией регрессии,* а записанные уравнения (5.2) и (5.3) — *уравнениями регрессии.*

Поскольку условное математическое ожидание *M* случайной величины *У* есть функция от (x), то его оценка $\bar{y}$, т. е. условная средняя, также является функцией от *X.*

Обозначим эту функцию через:

$$\overline{y_x} = f^*(x) \qquad (5.4)$$

Уравнение (5.4) определяет выборочное уравнение регрессии ***y*** на ***x***.

Сама функция называется *выборочной регрессией У* на *X,* а график У*(x) — *выборочной регрессией.*

- Аналогично определяется для случайных величин *X:*

$$\overline{x_y} = \varphi^*(y). \qquad (5.5)$$

- Функция регрессии необратима, так как речь идет о средних величинах для некоторого конкретного значения фактора.
- Функция регрессии формально устанавливает соответствие между переменными *X и У,* хотя такой зависимости может и не быть в экономике (ложная регрессия).

Линейная регрессия

- Пусть задана система случайных величин X и *У и* случайные величины *X и У* зависимы.
- Представим одну из случайных величин как линейную функцию другой случайной величины *X:*

$$Y = g(x) = \alpha + \beta x, \qquad (5.6)$$

где α, β — параметры, которые подлежат определению.

В общем случае эти параметры могут быть определены различными способами, наиболее часто используется метод наименьших квадратов (МНК).

Применение метода наименьших квадратов (МНК)

Функцию g(x) называют *наилучшим приближением* в смысле МНК, если математическое ожидание $M[Y - g(x)]^2$ принимает наименьшее возможное значение.

Применение МНК (продолжение)

Рассмотрим определение параметров выбранного уравнения прямой линии средней квадратической регрессии по несгруппированным данным.

Применение МНК (продолжение)

Итак, требуется найти:

$$y = \rho x + b \quad . \qquad (5.10)$$

Очевидно, параметры *p u b* нужно подобрать так, чтобы точки (x_1, y_1), (x_2, y_2), … (x_n, y_n), построенные по исходным данным, лежали как можно ближе к прямой (5.10) (рис. 5.1).

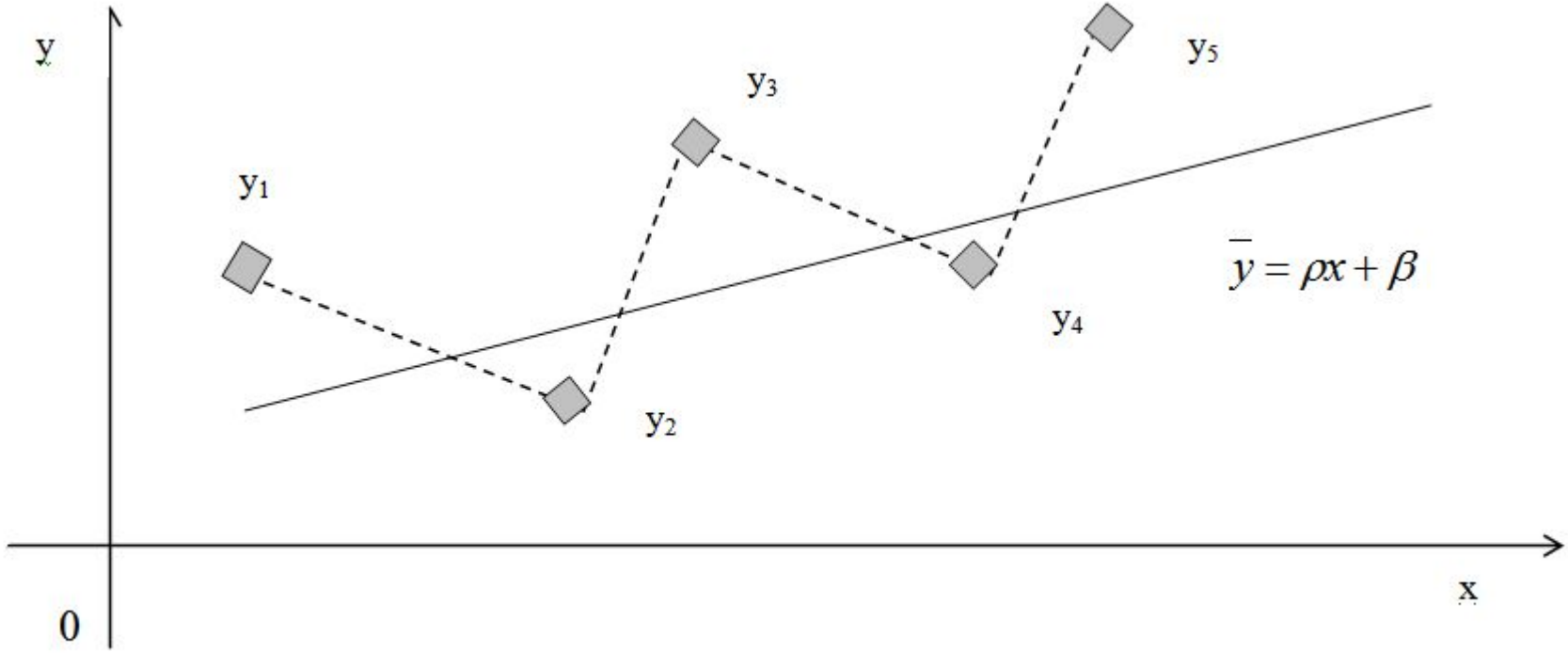


Рис. 5.1. Динамика изменения признака *Y*

Применение МНК (продолжение)

Уточним смысл этого требования. Для этого введем следующее понятие. Назовем отклонением разность вида:

$$Y_i - y_i (y = 1,2,...,n),$$

где Y_i — вычисляется по уравнению (5.10) и соответствует наблюдаемому значению x_i;

y_i — наблюдаемая ордината, соответствующая x_i.

Применение МНК (продолжение)

$$\begin{cases} \dfrac{dF}{d\rho} = 2\sum\limits_{i=1}^{n}(\rho x_i + b - y_i)x_i = 0, \\ \dfrac{dF}{d\rho} = 2\sum\limits_{i=1}^{n}(\rho x_i + b - y_i) = 0. \end{cases}$$

Далее запишем систему:

$$\begin{cases} (\sum\limits_{i=1}^{n} x_i^{\,2})\rho + (\sum\limits_{i=1}^{n} x_i)b - \sum\limits_{i=1}^{n} y_i x_i = 0, \\ (\sum\limits_{i=1}^{n} x_i)\rho + nb - \sum\limits_{i=1}^{n} y_i = 0. \end{cases}$$

Применение МНК (продолжение)

Для простоты вместо $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} y_i$

будем писать $\sum x, \sum x^2, \sum xy, \sum y$

(индекс / опускаем), тогда:

$$\begin{cases} (\sum x^2)\rho + (\sum x)b = \sum yx \ , \\ (\sum x)\rho + nb = \sum y \ . \end{cases}$$

Получили систему двух линейных уравнений относительно ρ и b.

Решая эту систему, получим:

$$\rho = \frac{n\sum xy - \sum x \sum y}{x^2 - (\sum x)^2}, \qquad (5.11)$$

$$b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}. \qquad (5.12)$$

Основные понятия корреляционно-регрессионного анализа

1. Среднее значение переменной определяется по следующей формуле

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad (5.13)$$

где x_i - эмпирическое значение переменной x;

n — число наблюдений.

4. Коэффициент корреляции

$$r_{xy} = \frac{\sum_{i=1}^{n}\left[\left(x_i - \bar{x}\right)\cdot\left(y_i - \bar{y}\right)\right]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \cdot \qquad (5.16)$$

- Коэффициент корреляции характеризует тесноту, или силу связи между переменными *y и x.*
- Значения - 1 =<r_{xy} =<+ 1.
- При изучении экономического явления, зависящего от многих факторов, строится множественная регрессионная зависимость.

В этом случае для характеристики тесноты связи используется коэффициент множественной корреляции:

$$R = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_{общ}}}, \qquad (5.17)$$

где $\sigma^2_{ост}$ - остаточная дисперсия зависимой переменной;

$\sigma^2_{общ}$ -общая дисперсия зависимой переменной.

5. Общая дисперсия определяется по формуле

$$\sigma^2_{общ} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}. \qquad (5.18)$$

Величина $\sigma^2_{общ}$ характеризует разброс наблюдений фактических значений от среднего значения $\bar{y}$.

6. Остаточная дисперсия определяется по следующей формуле

$$\sigma^2_{ост} = \frac{\sum_{i=1}^{n}(y_i - y_{iT})^2}{n-1}, \qquad (5.19)$$

где y_{iT} — теоретические значения переменной y, полученные по уравнению регрессии (5.1) при подстановке в него наблюдаемых фактических значений x_i.

Остаточная дисперсия характеризует ту часть рассеяния переменной y, которая возникает из-за всякого рода случайностей и влияния неучтенных факторов.

Вопрос 1. Общие сведения

7. Коэффициент детерминации служит для оценки точности регрессии, т. е. соответствия полученного уравнения регрессии имеющимся эмпирическим данным, и вычисляется по формуле

$$Д = 1 - \frac{\sigma^2_{ост}}{\sigma^2_{общ}}. \qquad (5.20)$$

8. Корреляционное отношение используется для оценки тесноты связи между двумя явлениями, в частности для определения тесноты связи исходного ряда y_i с теоретическим рядом y_{iT}.

Корреляционное отношение определяют по данным, сгруппированным по объясняющей переменной по следующей формуле

$$\eta = \frac{\sum_{i-1}^{n} (y_{iT} - \bar{y})^2}{\sum_{i-1}^{n} (y_i - \bar{y})^2}. \qquad \textbf{(5.2).}$$

Вопрос 2. Исходные предпосылки регрессионного анализа и свойства оценок

Применение метода наименьших квадратов для определения параметров регрессии предполагает выполнение некоторых предпосылок. Отметим наиболее существенные из них.

СВОЙСТВА ОЦЕНОК ПАРАМЕТРОВ РЕГРЕССИИ

СВОЙСТВА ОЦЕНОК ПАРАМЕТРОВ РЕГРЕССИИ (продолжение)

4. Достаточность оценки.

Этапы построения многофакторной корреляционно-регрессионной модели

1. априорное исследование экономической проблемы;
2. формирование перечня факторов и их логический анализ;
3. сбор исходных данных и их первичная обработка;
4. спецификация функции регрессии;
5. оценка функции регрессии;
6. отбор главных факторов;
7. проверка адекватности модели;
8. экономическая интерпретация;
9. прогнозирование неизвестных значений зависимой переменной.

Этапы разработки моделей и исследования экономических процессов

Вопрос 3. Этапы построения многофакторной корреляционно-регрессионной модели

Вопрос 3. Этапы построения многофакторной корреляционно-регрессионной модели

Ж. Проверка адекватности модели.

Данный этап анализа включает следующие процедуры:

- *оценку значимости коэффициента детерминации.*
- *проверку качества подбора теоретического уравнения*
- *вычисление специальных показателей*

Вопрос 3. Этапы построения многофакторной корреляционно-регрессионной модели

Вопрос 3. Этапы построения многофакторной корреляционно-регрессионной модели

Вопрос 3. Этапы построения многофакторной корреляционно-регрессионной модели

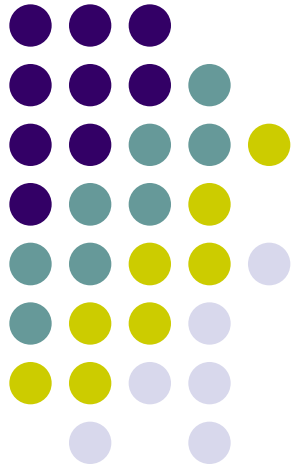
Вопрос 3. Этапы построения многофакторной корреляционно-регрессионной модели

ЛЕКЦИЯ 3
КОРРЕЛЯЦИЯ, ВЫЧИСЛЕНИЕ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ

Вопрос 1. Оценка тесноты линейной связи

- Коэффициент парной корреляции
- Матрица коэффициентов парной корреляции
- Множественный коэффициент корреляции
- Частный коэффициент корреляции
- Пример решения задачи по определению линейной связи

Вопрос 2. Оценка тесноты нелинейной связи



Типы связей

Функциональная	Стохастическая (вероятностная, статистическая).
• Характеризуется полным соответствием между изменением факторного признака и изменением результативной величины	• Между изменением двух признаков нет полного соответствием
• Зная величину факторного признака, можно определить величину результативного признака	• Устанавливается лишь тенденция изменения результативного признака при изменении факторного признака

Основная задача корреляционного анализа

- выявление связей между случайными переменными путем:

- точечной и интервальной оценки парных (частных) коэффициентов корреляции;
- вычисления и проверки значимости множественных коэффициентов корреляции и детерминации.

Другие задачи:

- Отбор факторов, оказывающих наиболее существенное влияние на результативный признак, на основании измерения тесноты связи между ними;
- Обнаружение ранее неизвестных причин связей.

Ковариация

- это статистическая мера взаимодействия двух переменных

$$Cov_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left[\left(x_i - \bar{x} \right) \cdot \left(y_i - \bar{y} \right) \right]$$

Ковариация зависит от единиц, в которых измеряются переменные X, Y, она является ненормированной величиной.

Коэффициент парной корреляции

Для двух переменных X и Y коэффициент парной корреляции определяется следующим образом:

$$r_{y,x} = \frac{Cov_{xy}}{S_x^2, S_y^2} = \frac{\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left[(x_i - \bar{x}) \cdot (y_i - \bar{y}) \right]}{S_x^2, S_v^2}$$

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

S_x^2, S_y^2 - оценки дисперсий величин X, Y;

- характеризуют степень разброса значений $x_1, x_2, \ldots, x_n$ ($y_1, y_2, \ldots, y_n$) вокруг своего среднего, или вариабельность (изменчивость) этих переменных на множестве наблюдений.

Дисперсия

- Оценка дисперсии определяется по формуле

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- Более естественно измерять степень разброса значений переменой в тех же единицах, в которых измеряется и сама переменная.
- Эту задачу решает показатель, называемый среднеквадратическим отклонением (стандартным отклонением) или стандартной ошибкой переменной X (переменной Y) и определяемый соотношением

$$S_x = \sqrt{S_x^2}$$

- Корреляция и ковариация представляют, по сути, одну и ту же информацию, однако корреляция представляет эту информацию в более удобной форме.
- Для качественной оценки коэффициентов корреляции применяются различные шкалы.

Шкала Чеддока:

0,1-0,3 – слабая;

0,3-0,5- заметная;

0,5-0,7 – высокая;

0,9-1,0 – весьма высокая.

! Величина коэффициента корреляции не является доказательством того, что между исследуемыми признаками существует причинно-следственная связь, а представляет собой оценку степеней взаимной согласованности в изменениях признаков.

Оценка существенности линейного коэффициента корреляции

- - при малых объемах выборки используется t-критерий Стьюдента:

$$t_{\text{набл}} = \sqrt{\frac{r_{y,x}^2}{1 - r_{y,x}^2}(n-2)}.$$

- - t-наблюдаемое сравнивается с табличным значением с учетом заданного уровня значимости а и числа степеней свободы (n-2).
- - если t-набл. >ттабл. , то :
- полученное значение коэффициента парной корреляции признается значимым.
- Между исследуемыми переменными есть тесная статистическая взаимосвязь.

Матрица коэффициентов парной корреляции

$$R = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_m} \\ r_{yx_1} & 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{yx_2} & r_{x_1x_2} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_m} & r_{x_1x_m} & r_{x_2x_m} & \dots & 1 \end{pmatrix}.$$

- Анализ матрицы коэффициентов парной корреляции используют при построении моделей множественной регрессии

Задачи многомерного корреляционного анализа

1. Определение тесноты связи одной случайной величины с совокупностью остальных величин, включенных в анализ.
2. Определение тесноты связи между двумя величинами при фиксировании или исключении влияния остальных величин.

Множественный коэффициент корреляции

Задача: Определение тесноты связи одной случайной величины с совокупностью остальных величин, включенных в анализ.

Решение: определение выборочного коэффициента множественной корреляции:

$$R_{j,1,2,\ldots,j-1,j+1,\ldots,m} = \sqrt{1 - \frac{|R|}{R_{jj}}},$$

где |R| - определитель корреляционной матрицы;

R_{jj} – алгебраическое дополнение элемента r_{jj} той же матрицы R.

- R2 – выборочный множественный коэффициент детерминации;
- показывает, какую долю вариации (случайного разброса) исследуемой величины Xj объясняет вариация остальных случайных величин X1, X2, … Xm
- $0 \leq RR2 \leq 1$.

Проверка значимости коэффициента детерминации

- Используется сравнение расчетного значения F-критерия Фишера с табличным F.

$$F_{\text{расч}} = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}$$

$F_{\text{табл.}}$ определяется заданным уровнем значимости и степенями свободы $v_1=p-1$ и $v_2=n-p$,

где p - количество параметров модели.

R^2 значимо отличается от нуля, если выполняется неравенство:

$$F_{\text{расч}} > F_{\text{табл.}}$$

Частный коэффициент корреляции

- Возникает необходимость исследования частной корреляции между величинами при исключении влияния других случайных величин (одной или нескольких).
- **Выборочный частный коэффициент корреляции:**

$$r_{jk(1,2,\ldots,m)} = -\frac{R_{jk}}{\sqrt{R_{jj}R_{kk}}},$$

где R_{jk}, R_{jj}, R_{kk} – алгебраические дополнения к соответствующим элементам матрицы R.

- $-1 \leq r_{jk} \leq 1$

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{yx_1} & r_{yx_2} & \ldots & r_{yx_m} \\ r_{yx_1} & 1 & r_{x_1x_2} & \ldots & r_{x_1x_m} \\ r_{yx_2} & r_{x_1x_2} & 1 & \ldots & r_{x_2x_m} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{yx_m} & r_{x_1x_m} & r_{x_2x_m} & \ldots & 1 \end{pmatrix}.$$

Вопрос 2. Оценка тесноты нелинейной связи