Статистические методы изучения взаимосвязи социально-экономических явлений

Лекция 5

Вопросы лекции:

Корреляционная связь

- (частный случай стохастической) – связь, проявляющаяся при достаточно большом числе наблюдений в виде определенной зависимости между средним значением результативного признака и признаками-факторами.

  Задача корреляционного анализа – измерение тесноты связи между варьируемыми признаками и оценка факторов, оказывающих наибольшее влияние.

Задача регрессионного анализа

– выбор типа модели (формы связи), устанавливающих степени влияния независимых переменных.

Связь признаков проявляется в их согласованной вариации, при этом одни признаки выступают как факторные, а другие – как результативные. Причинно-следственная связь факторных и результативных признаков характеризуется по степени:

- тесноты;
- направлению;
- аналитическому выражению.

Регрессионный анализ

Для оценки параметров уравнений регрессии наиболее часто используется метод наименьших квадратов (МНК), суть которого заключается в следующем требовании: искомые теоретические значения результативного признака должны быть такими, при которых бы обеспечивалась минимальная сумма квадратов их отклонений от эмпирических (фактических) значений, т. е.

$$S = \sum (y - \overline{y}_x)^2 \rightarrow \min$$

При изучении связей показателей применяются различного вида уравнения прямолинейной и криволинейной связи. Так, при анализе прямолинейной зависимости применяется уравнение:

$$y = a_0 + a_1 x$$

Это наиболее часто используемая форма связи между коррелируемыми признаками, при парной корреляции она выражается уравнением,

где a_0 – среднее значение в точке $x=0$, поэтому экономической интерпретации коэффициента нет;

a_1 – коэффициент регрессии, показывает, на сколько изменяется в среднем значение результативного признака при увеличении факторного на единицу собственного измерения.

При криволинейной зависимости применяется ряд математических функций:

полулогарифмическая	$y = a_0 + a_1 \lg x$
показательная	$y = a_0 + a_1^{\,x}$
степенная	$y = a\ x^{a_1}$
параболическая	$y = a_0 + a_1 x + a_2 x^2$
гиперболическая	$y = a_0 + a_1 \frac{1}{x}$

Система нормальных уравнений

МНК для линейной парной регрессии имеет следующий вид:

$$\begin{cases} na_0 + a_1 \sum x = \sum y; \\ a_0 \sum x + a_1 \sum x^2 = \sum xy. \end{cases}$$

Отсюда можно выразить коэффициенты регрессии:

$$a_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2};$$

$$a_0 = \overline{y} - a_1 \overline{x}.$$

При численности объектов анализа до 30 единиц

возникает необходимость проверить, насколько вычисленные параметры типичны для отображаемого комплекса условий, не являются ли полученные значения параметров результатом действия случайных причин. Значимость коэффициентов регрессии применительно к совокупности $n<30$ определяется с помощью ***t-критерия Стьюдента***. При этом вычисляются фактические значения *t-критерия*:

для параметра a_0:

$$t_{a_0} = |a_0| \frac{\sqrt{n-2}}{\sigma_{\text{ост}}},$$

для параметра a_1:

$$t_{a_1} = |a_1| \frac{\sqrt{n-2}}{\sigma_{\text{ост}}} \sigma_x .$$

$\sigma_{ост} = \sqrt{\frac{\sum (y_i - \hat{y}_{x_i})^2}{n}}$ – среднее квадратическое отклонение результативного признака y_i от выровненных значений $\hat{y}_{x_i}$.

$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$ – среднее квадратическое отклонение факторного признака x_i от общей средней $\bar{x}$.

Полученные по формулам (1) и (2) фактические значения и сравниваются с критическим , который получают по таблице Стьюдента с учетом принятого уровня значимости и числа степеней свободы v (v=*n*-*k*-1), где *n* – число наблюдений, *k* – число факторов, включенных в уравнение регрессии). Рассчитанные параметры a_0 и a_1 уравнения регрессии признаются типичными, если ***t*** фактическое больше ***t*** критического.

На практике часто приходится исследовать зависимость результативного признака от нескольких факторных признаков. Аналитическая форма связи результативного признака от ряда факторных признаков выражается и называется **многофакторным (множественным) уравнением регрессии**.

Линейное уравнение множественной регрессии

$$\overline{y}_{1,2,..k} = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_k x_k$$

Система нормальных линейных уравнений МНК для оценки коэффициентов двухфакторной регрессии $\bar{y}_{x_1x_2} = a_0 + a_1x_1 + a_2x_2$

имеет вид:

$$\begin{cases} na_0 + a_1\sum x_1 + a_2\sum x_2 = \sum y; \\ a_0\sum x_1 + a_1\sum x_1^2 + a_2\sum x_1x_2 = \sum x_1y; \\ a_0\sum x_2 + a_1\sum x_1x_2 + a_2\sum x_2^2 = \sum x_2y. \end{cases}$$

Корреляционный анализ

Различают:

- парную корреляцию – это зависимость между результативным и факторным признаком;
- частную корреляцию – это зависимость между результативным и одним факторным признаком при фиксированном значении других факторных признаков;
- множественную – многофакторное влияние в статической модели

$$y_x = f(x_1 x_2 \ldots x_k)$$

Теснота связи при линейной зависимости измеряется с помощью

линейного коэффициента корреляции, который рассчитывается по одной из формул:

$$r = a_1 \frac{\sigma_x}{\sigma_y}$$

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}.$$

Оценка линейного коэффициента корреляции

Значение r	Характер связи	Интерпретация связи
$r = 0$	Отсутствует	Изменение x не влияет на изменения y
$0 < r < 1$	Прямая	С увеличением x увеличивается y
$-1 > r > 0$	Обратная	С увеличением x уменьшается y и наоборот
$r = 1$	Функциональная	Каждому значению факторного признака строго соответствует одно значение результативного

Значимость линейного коэффициента корреляции проверяется на основе *t*-критерия Стьюдента. Для этого определяется фактическое значение критерия *t расч*

$$t_{\text{расч}} = \frac{|r|}{\sigma_r} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Вычисленное по формуле значение $t_{\text{расч}}$ сравнивается с критическим t_k, который получают по таблице Стьюдента с учетом принятого уровня значимости α и числа степеней свободы ν. Коэффициент корреляции считается статистически значимым, если $t_{\text{расч}}$ превышает t_k: $t_{\text{расч}} > t_k$.

Универсальным показателем тесноты связи является **теоретическое корреляционное отношение**:

$$\eta_{\text{теор}} = \sqrt{\frac{\delta_{\text{ф}}^2}{\sigma_y^2}} = \sqrt{\frac{\sigma_y^2 - \sigma_{\text{ост}}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sigma_{\text{ост}}^2}{\sigma_y^2}}$$

где σ_y^2 – *общая* дисперсия эмпирических значений *y*, характеризует вариацию результативного признака за счет всех факторов, включая *x*;

δ_ϕ^2 – *факторная* дисперсия теоретических значений результативного признака, отражает влияние фактора *x* на вариацию *y*;

$\sigma_{ост}^2$ – *остаточная* дисперсия эмпирических значений результативного признака, отражает влияние на вариацию *y* всех остальных факторов кроме *x*.

Оценка связи на основе теоретического корреляционного отношения (шкала Чеддока)

Множественный коэффициент корреляции в случае зависимости результативного признака от двух факторов вычисляется по формуле:

$$R_{y/x_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \times r_{yx_2} \times r_{x_1x_2}}{1 - r_{x_1x_2}^2}},$$

где $r_{yx_1}, r_{yx_2}, r_{x_1x_2}$ – парные коэффициенты корреляции между признаками.

Множественный коэффициент корреляции изменяется в пределах от 0 до 1 и по определению положителен: $0 \le R \le 1$.

Условие включения факторных признаков в регрессионную модель – наличие тесной связи между результативным и факторными признаками и как можно менее существенная связь между факторными признаками.

Значимость коэффициента множественной детерминации, а соответственно и адекватность всей модели и правильность выбора формы связи можно проверить с помощью критерия Фишера:

$$F_{\text{расч}} = \frac{R^2}{1-R^2}\frac{n-k-1}{k},$$

- где R^2 – коэффициент множественной детерминации (R^2);
- k – число факторных признаков, включенных в уравнение регрессии.

Связь считается существенной

если $F_{расч} > F_{табл}$ – табличного значения

F-критерия для заданного уровня значимости α и числе степеней свободы

$v_1 = k$, $v_2 = n - k - 1$.

Частные коэффициенты корреляции

характеризуют *степень тесноты связи результативного признака и фактора,* при элиминировании его взаимосвязи с остальными факторами, включенными в анализ. Расчет частных коэффициентов корреляции в случае двухфакторной регрессии (в первом случае исключено влияние факторного признака x_2, во втором – x_1):

$$r_{yx_1/x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)\cdot(1-r_{x_1x_2}^2)}}; \qquad r_{yx_2/x_1} = \frac{r_{yx_2} - r_{x_1y} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)\cdot(1-r_{x_1x_2}^2)}},$$

- где *r* – парные коэффициенты корреляции между указанными в индексе переменными.

Для оценки сравнительной силы влияния факторов, по каждому фактору рассчитывают **частные коэффициенты эластичности**:

$$Э_{x_i} = a_i \frac{\overline{x_i}}{\overline{y}} \quad ,$$

где $\overline{x_i}$ – среднее значение соответствующего факторного приз

$\overline{y}$ – среднее значение результативного признака;

a_i – коэффициент регрессии при i-м факторном признаке.

Данный коэффициент показывает, на сколько процентов следует ожидать изменения результативного показателя при изменении фактора на 1% и неизменном значении других факторов.

Частный коэффициент детерминации показывает,

на сколько процентов вариация результативного признака объясняется вариацией *i*-го признака, входящего в множественное уравнение регрессии, рассчитывается по формуле:

$$d_{x_i} = r_{yx_i} \times \beta_{x_i} \quad ,$$

где r_{yx_i} – парный коэффициент корреляции между результативным и *i*-

факторным признаком;

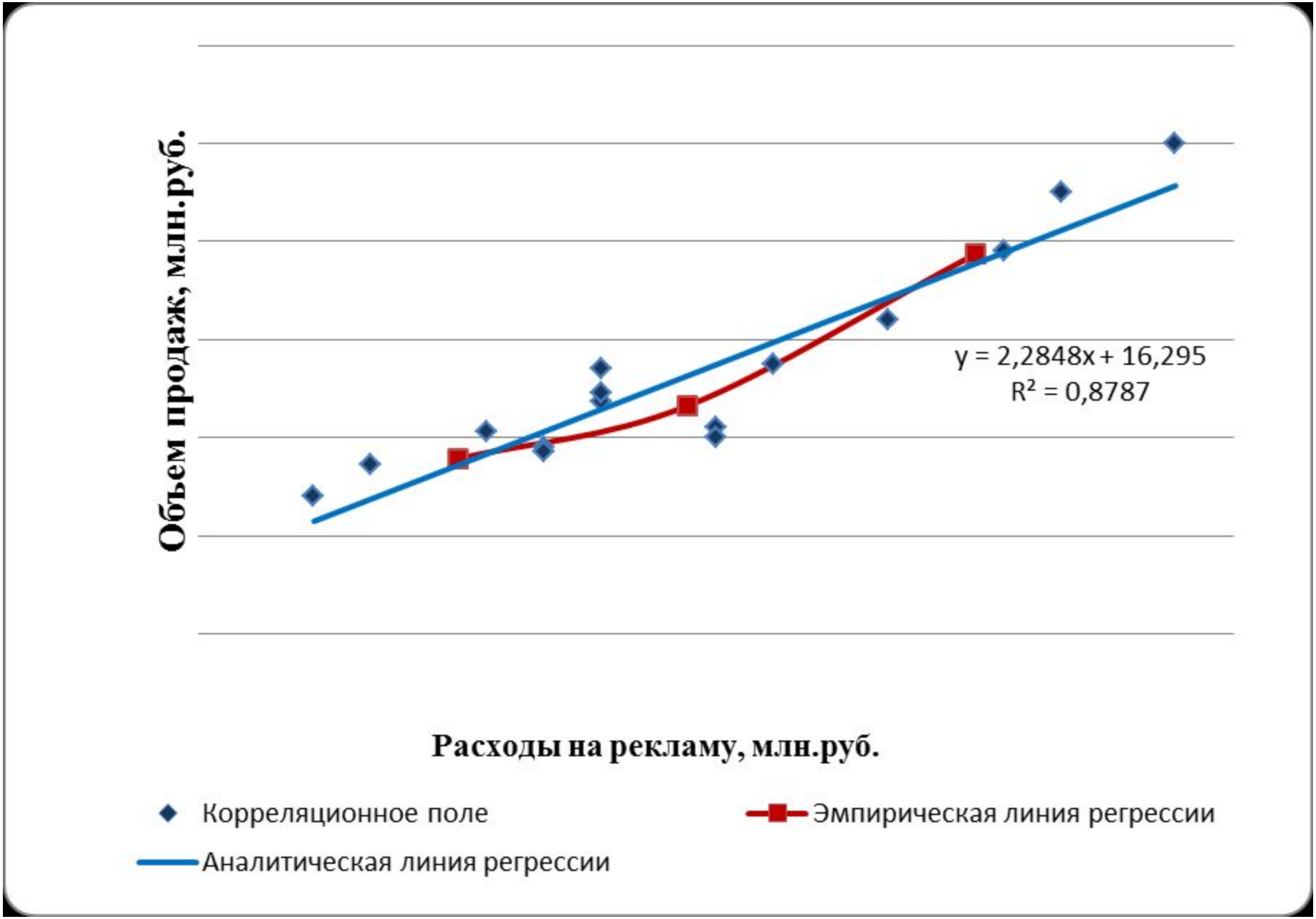
β_{x_i} – соответствующий стандартизованный коэффициент

уравнения множественной регрессии:

$$\beta_{x_i} = a_i \frac{\sigma_{x_i}}{\sigma_y}$$.

Корреляционно-регрессионный анализ

- X – фактор (выпуск продукции)
- Y – результат (расход топлива)
- Поле корреляции


Объем продаж, млн.руб.
y = 2,2848x + 16,295
R² = 0,8787
Расходы на рекламу, млн.руб.
Корреляционное поле
Эмпирическая линия регрессии
Аналитическая линия регрессии

- В зависимости от формы связи уравнение регрессии может быть: линейным, гиперболическим, параболическим и т.д.
- **Уравнение линейной регрессии** имеет вид:

$$Y_x = a_0 + a_1 \cdot X \qquad (3)$$

где X - факторный признак;

Yx - результативный показатель;

a_0 – свободный параметр уравнения, который характеризует уровень результативного признака (при X=0);

a_1 - коэффициент регрессии. Он показывает, на сколько изменится результативный признак, если факторный увеличится на единицу.

$$a_0 = \frac{\sum y * \sum x^2 - \sum yx * \sum x}{n\sum x^2 - \sum x * \sum x}$$

$$a_1 = \frac{n\sum yx - \sum y * \sum x}{n\sum x^2 - \sum x * \sum x}$$

N	x	y	x*y	x2	y2
1	5	4			
2	6	4			
3	8	6			
4	8	5			
5	10	7			
6	10	8			
7	14	8			
8	20	10			
9	20	12			
10	24	16			
Итого	125	80			

N	x	y	x*y	x2	y2
1	5	4	20	25	16
2	6	4	24	36	16
3	8	6	48	64	36
4	8	5	40	64	25
5	10	7	70	100	49
6	10	8	80	100	64
7	14	8	112	196	64
8	20	10	200	400	100
9	20	12	240	400	144
10	24	16	324	576	256
Итого	125	80	1218	1961	

Расчет по формулам:

- $a_o = \frac{80*1961-1218*125}{10*1961-15625}$

  $= 4630/3985=1,16$

$$\frac{10*1218-125*80}{10*1961-15625} =$$

$$a_1 = 2191,75/3985 = 0,55$$

- $\hat{y} = 1,16+0,55\ y$