Статистические методы изучения взаимосвязи социально-экономических явлений

План лекции:

1. Виды связей между явлениями
2. Простейшие методы изучения стохастических связей
3. Статистическое моделирование связи методом корреляционного и регрессионного анализа
   а) однофакторные модели
   б) многофакторные модели
4. Непараметрические методы

1. Виды связей между явлениями

- Знание характера и силы связей позволяет управлять социально-экономическими явлениями и процессами и предсказать их развитие

- Среди многих форм связей важнейшей является причинная. Причинно-следственные отношения - это связь явлений и процессов, когда изменение одного из них (причины) ведет к изменению другого (следствия).
- Причина- это совокупность условий, обстоятельств, действие которых приводит к появлению следствия.

- Между различными явлениями и их признаками выделяют два типа связей:
- *функциональную* (жестко детерминированную)
- и *статистическую* (стохастически детерминированную)

Связь признака *y* с признаком *x* называется *функциональной*, если каждому возможному значению независимого признака *x* соответствует одно или несколько строго определенных значений зависимого признака *y*. Определение функциональной связи может быть легко обобщено для случая многих признаков *x1, x2, ..., xn*.

Функциональную связь можно представить уравнением:

$$yi=f(xi),$$

где *yi*- результативный признак *(i=1,...,n); f(xi)*- известная функция связи результативного и факторного признаков; *x1*- факторный признак.

Стохастическая связь- это связь между величинами, при которой одна из них, случайная величина *y*, реагирует на изменение другой величины *x* или других величин *x1, x2,..., xn* (случайных или неслучайных) изменением закона распределения. Это обусловливается тем, что зависимая переменная (результативный признак), кроме рассматриваемых независимых, подвержен влиянию ряда неучтенных или неконтролируемых (случайных) факторов, а также некоторых неизбежных ошибок измерения переменных.

Поскольку значения зависимой переменной подвержены случайному разбросу, они не могут быть предсказаны с достаточной точностью, а только указаны с определенной *вероятностью*.

Модель стохастической связи может быть представлена в общем виде уравнением:

$$\hat{y}_i = f\left(x_i\right) + \varepsilon_i,$$

где - расчетное значение результативного признака; ***f (xi)***- часть результативного признака, сформировавшаяся под воздействием учтенных известных факторных признаков (одного или множества), находящихся в стохастической связи с признаком;

εi- часть результативного признака, возникшая вследствие действия неконтролируемых или неучтенных факторов, а также измерения признаков неизбежно сопровождающегося некоторыми случайными ошибками.

Частные случаи стохастических связей- это корреляционная и регрессионная.

- *Корреляция*- это статистическая зависимость между случайными величинами, не имеющими строго функционального характера. Корреляционный анализ имеет своей задачей определение тесноты связи между двумя признаками (при парной корреляции) и между результативным и несколькими факторными.
- *Регрессионный анализ* - заключается в определении аналитического выражения связи.
- *Корреляционный анализ*: измерение тесноты связи, направления и установления аналитического выражения связи.

- *Прямые и обратные связи*.
- *Прямолинейные и криволинейные связи*.
- *Однофакторные и многофакторные связи*.

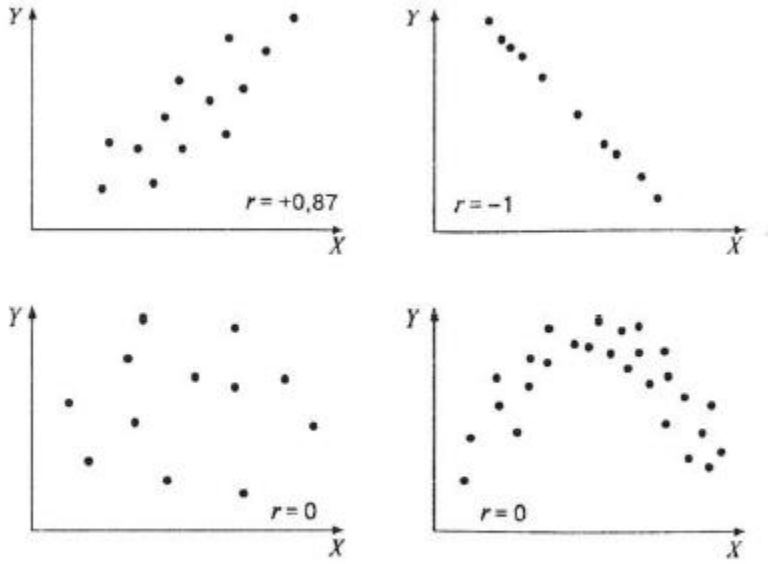
2. Простейшие методы изучения стохастических связей.

- *Метод сопоставления двух параллельных рядов.*
- *Метод аналитических группировок*

3. Статистическое моделирование связи методом корреляционного и регрессионного анализа.

- **а)** *Двухмерная линейная модель корреляционного и регрессионного анализа (однофакторный линейный корреляционный и регрессионный анализ)*.
- Наиболее разработанной является методология так называемой *парной корреляции*, рассматривающая влияние вариации факторного признака *x* на результативный признак *y* и представляющая собой *однофакторный корреляционный и регрессионный анализ*.

- Выбор типа функции может опираться на теоретические знания об изучаемом явлении, опыт предыдущих аналогичных исследований, или осуществляться эмпирически — перебором и оценкой функций разных типов и т.п.



Примеры рассеивания и соответствующих коэффициентов корреляции

При изучении связи экономических показателей производства (деятельности) используют различного вида уравнения прямолинейной и криволинейной связи.

$y_x = a_0 + a_1 \cdot x$ – *прямолинейная зависимость*

$y_x = a_0 + a_1 \cdot Lq\, x$ – *полулогарифмическая функция*

$y_x = a_0 + a_1^x$ – *показательная*

$y_x = a_0 \cdot x^{a_1}$ – *степенная*

$y_x = a_0 + a_1 \cdot x + a_2 \cdot x^2$ – *параболическая*

$y_x = a_0 + a_1 \cdot \frac{1}{x}$ – *гиперболическая и др.*

Система уравнений:

для прямой $y_x = a_0 + a_1 \cdot x$

$$\left.\begin{aligned} n \cdot a_0 + a_1 \cdot \Sigma x &= \Sigma y \\ a_0 \cdot \Sigma x + a_1 \cdot \Sigma x^2 &= \Sigma yx \end{aligned}\right\}$$

для параболы $y_x = a_0 + a_1 \cdot x + a_2 \cdot x^2$

$$\left.\begin{aligned} a_0 \cdot n + a_1 \cdot \Sigma x + a_2 \cdot \Sigma x^2 &= \Sigma y \\ a_0 \cdot \Sigma x + a_1 \cdot \Sigma x^2 + a_2 \cdot \Sigma x^3 &= \Sigma yx \\ a_0 \cdot \Sigma x^2 + a_1 \cdot \Sigma x^3 + a_2 \cdot \Sigma x^4 &= \Sigma yx^2 \end{aligned}\right\}$$

для гиперболы $y_x = a_0 + a_1 \cdot \frac{1}{x}$

$$\left.\begin{aligned} a_0 \cdot n + a_1 \cdot \Sigma \frac{1}{x} &= \Sigma y \\ a_0 \cdot \Sigma \frac{1}{x} + a_1 \cdot \Sigma \frac{1}{x} &= \Sigma y \cdot \frac{1}{x} \end{aligned}\right\}$$

- *Уравнение однофакторной (парной) линейной корреляционной связи* имеет вид:
- $$\hat{y} = a_0 + a_1 x,$$
- где — теоретические значения результативного признака, полученные по уравнению регрессии; *a0, a1* – коэффициенты (параметры) уравнения регрессии.
- Поскольку *a0* является средним значением *y* в точке *x = 0*, экономическая интерпретация часто затруднена или вообще невозможна.

Интерпретация:

- Коэффициент парной линейной регрессии *a1* имеет смысл показателя *силы связи* между вариацией факторного признака *x* и вариацией результативного признака *y*. Уравнение показывает среднее значение изменения результативного признака *y* при изменении факторного признака *x* на одну единицу его измерения, т.е. вариацию *y* , приходящуюся на единицу вариации *x*. Знак *a1* указывает направление этого изменения.

Параметры уравнения a_0, a_1 находят *методом наименьших* квадратов (метод решения систем уравнений, при котором в качестве решения принимается точка минимума суммы квадратов отклонений), т.е. в основу этого метода положено требование минимальности сумм квадратов отклонений эмпирических данных y_i, от выровненных $\hat{y}$:

$$\Sigma(y_i - \hat{y})^2 = \Sigma(y_i - a_0 - a_1 x_i)^2 \to \min$$

Для нахождения минимума данной функции приравняем к нулю ее частные производные и получим систему двух линейных уравнений, которая называется *системой нормальных уравнений*:

$$\left.\begin{aligned} na_0 + a_1\Sigma x &= \Sigma y; \\ a_0\Sigma x + a_1\Sigma x^2 &= \Sigma xy. \end{aligned}\right\}$$

Решим эту систему в общем виде:

$$a_0 = \frac{\Sigma y \Sigma x^2 - \Sigma xy \Sigma x}{n\Sigma x^2 - \Sigma x \Sigma y}; \quad a_1 = \frac{n\Sigma xy - \Sigma y \Sigma x}{n\Sigma x^2 - \Sigma x \Sigma y}.$$

Параметры уравнения парной линейной регрессии иногда удобно исчислять по следующим формулам, дающим тот же результат:

$$a_1 = \frac{\Sigma(y - \bar{y})(x - \bar{x})}{\Sigma(x - \bar{x})^2}, \quad \text{или } a_1 = \frac{\overline{xy} - \overline{xy}}{\overline{x^2} - \bar{x}^2};$$

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

Определив значения a_0, a_1 и подставив их в уравнение связи $\hat{y} = a_0 + a_1 x$, находим значения $\hat{y}$, зависящие только от заданного значения x.

- *Проверка адекватности регрессионной модели.* Для практического использования моделей регрессии большое значение имеет их адекватность, т.е. соответствие фактическим статистическим данным.
- При численности объектов анализа до 30 единиц возникает необходимость проверки значимости (существенности) каждого коэффициента регрессии. При этом выясняют насколько вычисленные параметры характерны для отображения комплекса условий: не являются ли полученные значения параметров результатами действия случайных причин.

Значимость коэффициентов простой линейной регрессии (применительно к совокупностям, у которых $n < 30$) осуществляют с помощью *t-критерия Стьюдента.* При этом вычисляют расчетные (фактические) значения t –критерия

для параметра a_0:

$$t_{a_0} = |a_0| \frac{\sqrt{n-2}}{\sigma_{ост}}; \qquad (1.3)$$

для параметра a_1:

$$t_{a_1} = |a_1| \frac{\sqrt{n-2}}{\sigma_{ост}} \sigma_x, \qquad (1.4)$$

где n- объём выборки; $\sigma_{ост} = \sqrt{\Sigma(y-\hat{y})^2 / n}$ - среднее квадратическое отклонение результативного признака y от выровненных значений $\hat{y}$;

$\sigma_x = \sqrt{\Sigma(x-\bar{x})^2 / n}$ *или* $\sigma_x = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$ – среднее квадратическое отклонение факторного признака x от общей средней $\bar{x}$.

Вычисленные по формулам (1.3) и (1.4) значения, сравнивают с критическими, *t,* которые определяют по таблице Стьюдента с учетом принятого уровня значимости α и числом степеней свободы вариации $\nu = n-2$. В социально-экономических исследованиях уровень значимости α обычно принимают равным 0,05. Параметр признается значимым (существенным) при условии, если $t_{расч} > t_{табл}$. В таком случае практически невероятно, что найденные значения параметров обусловлены только случайными совпадениями.

- *Проверка адекватности регрессионной модели* может быть дополнена корреляционным анализом. Для этого необходимо определить *тесноту* корреляционной связи между переменными *x* и *y* .
- Теснота корреляционной связи, как и любой другой, может быть измерена *эмпирическим корреляционным отношением* ηэ, когда δ2 (межгрупповая дисперсия) характеризует отклонения групповых средних результативного признака от общей средней: .
- Говоря о корреляционном отношении как о показателе измерения тесноты зависимости, следует отличать от эмпирического корреляционного отношения — *теоретическое*.

Теоретическое корреляционное отношение η представляет собой относительную величину, получающуюся в результате сравнения среднего квадратического отклонения выровненных значений результативного признака δ, т.е. рассчитанных по уравнению регрессии, со средним квадратическим отношением эмпирических (фактических) значений результативности признака σ:

$$\eta = \sqrt{\delta^2 / \sigma^2},$$

$$\text{где } \delta^2 = \sqrt{\frac{\Sigma(\hat{y} - \bar{y})^2}{n}} = \sqrt{\delta^2_{\hat{y}_x}}; \quad \sigma = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n}} = \sqrt{\sigma^2_y}. \qquad (1.6)$$

$$\text{Тогда } \eta = \sqrt{\frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}}.$$

Изменение значения η объясняется влиянием факторного признака.

В основе расчета корреляционного отношения лежит правило сложения дисперсий, т.е. $\sigma^2 = \delta^2 + \overline{\sigma_i^2}$, где $\overline{\sigma_i^2}$ - отражает вариацию y за счет всех остальных факторов, кроме x, т.е. является *остаточной дисперсией*:

$$\overline{\sigma_i^2} = \sigma^2_{ост} = \frac{\Sigma(y - \hat{y})^2}{n}.$$

Тогда формула теоретического корреляционного отношения примет вид :

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}} = \sqrt{\frac{\sigma^2 - \sigma^2_{ост}}{\sigma^2}} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2}}, \qquad (1.6)$$

$$\text{или } \eta = \sqrt{1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}}. \qquad (1.7)$$

Подкоренное выражение корреляционного отношения представляет собой *коэффициент детерминации* (меры определенности, причинности).

- Теоретическое корреляционное отношение применяется для измерения тесноты связи при линейной и криволинейной зависимостях между результативным и факторным признаком. При криволинейных связях теоретическое корреляционное отношение, исчисляемое по формулам часто называют *индексом корреляции R*. При значительной корреляции расчет по формулам (1.6) и (1.7) значительно проще, так как отклонение (), как правило, по значению меньше, чем отклонение ().
- Корреляционное отношение может находиться в пределах от 0 до 1, т.е. (). Чем ближе корреляционное отношение к 1, тем связь между признаками теснее.
- Кроме того, при линейной форме уравнения применяется другой показатель тесноты связи — линейный коэффициент корреляции.

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n \cdot \sigma_x \sigma_y}, \qquad (1.8)$$

где n — число наблюдений.

Для практических вычислений при малом числе наблюдений $(n \le 20 \div 30)$ линейный коэффициент корреляции удобнее исчислять по следующей формуле:

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\left[\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right]}}. \qquad (1.9)$$

Значение линейного коэффициента корреляции важно для исследования социально-экономических явлений и процессов распределение которых близко к нормальному. Он принимает значения в интервале : $-1 \le r \le 1$.

Отрицательные значения указывают на обратную связь, положительные — на прямую. При r =0 линейная связь отсутствует. Чем ближе коэффициент корреляции по абсолютной величине к единице, тем теснее связь между признаками. И, наконец, при $r = \pm 1$ связь — функциональная.

Квадрат линейного коэффициента корреляции r^2 называется *линейным коэффициентом детерминации*.

- Факт совпадений и несовпадений значений теоретического корреляционного отношения η и линейного коэффициента корреляции *r* используется для оценки формы связи..
- Посредством теоретического корреляционного отношения измеряется теснота связи любой формы, а с помощью линейного коэффициента корреляции — только прямолинейной. Следовательно, значения η и *r* совпадают только при наличии прямолинейной связи. Несовпадение этих величин свидетельствует, что связь между изучаемыми признаками не прямолинейная, а криволинейная. Установлено, что если разность квадратов *η2* и *r2* не превышает 0,1, то гипотезу о прямолинейной форме связи можно считать подтвержденной.

- Показатели тесноты связи, исчисленные по данным сравнительно небольшой статистической совокупности, могут искажаться действием случайных причин. Это вызывает необходимость проверки их *существенности*, дающей возможность распространять выводы по результатам выборки на генеральную совокупность.
- Для оценки *значимости коэффициента корреляции r* используют *t*-критерий Стьюдента, который применяется при *t*-распределении, отличном от нормального.

- При линейной однофакторной связи *t*-критерий можно рассчитать по формуле:
- $$t_{\text{рас÷}} = r\sqrt{\frac{n-2}{1-r^2}}\ , \qquad (1.10)$$
- где (*n*—2) — число степеней свободы при заданном уровне значимости α и объеме выборки *n*.
- Полученное значение *tрасч* сравнивают с табличным значением *t*-критерия (для α = 0,05 и 0,01). Если рассчитанное значение *tрасч* превосходит табличное значение критерия *tтабл*, то практически невероятно, что найденное значение обусловлено только случайными колебаниями (т.е. отклоняется гипотеза о его случайности).

- **б)** *Многофакторный корреляционный и регрессионный анализ.*

- Между факторами существуют сложные взаимосвязи, поэтому их влияние комплексное и его нельзя рассматривать как простую сумму изолированных влияний.

- Многофакторный корреляционный и регрессионный анализ позволяет оценить меру влияния на исследуемый результативный показатель каждого из включенных в модель (уравнение) факторов при фиксированном положении (на среднем уровне) остальных факторов, а также при любых возможных сочетаниях факторов с определенной степенью точности найти теоретическое значение этого показателя (важным условием является отсутствие между факторами функциональной связи).

- Математически задача формулируется следующим образом. Требуется найти аналитическое выражение, наилучшим образом отражающее установленную теоретическим анализом связь независимых признаков с результативным, т.е. функцию

$$y_i = f(x_1, x_2, \ldots, x_n) + \varepsilon_i$$

Построение и статистический анализ двухфакторной линейной модели (трехмерной регрессии). Для расчета параметров простейшего уравнения множественной линейной двухфакторной регрессии

$$\hat{y}_{x_1x_2} = a_0 + a_1x_1 + a_2x_2 ,$$

где $\hat{y}_{x_1x_2}$ - расчетные значения зависимой переменной (результативного признака); x_1, x_2 — независимые переменные (факторные признаки); a_0, a_1, a_2 — параметры уравнения.

Построим следующую *систему нормальных уравнений*:

$$\left.\begin{aligned} a_0 n + a_1\Sigma x_1 + a_2\Sigma x_2 &= \Sigma y; \\ a_0\Sigma x_1 + a_1\Sigma x_1^2 + a_2\Sigma x_1x_2 &= \Sigma yx_1; \\ a_0\Sigma x_2 + a_1\Sigma x_1x_2 + a_2\Sigma x_2^2 &= \Sigma yx_2. \end{aligned}\right\}$$

Параметры этой системы могут быть найдены, например, методом К.Гаусса.

После построения регрессионной модели необходимо исчислить различного рода характеристики тесноты связи между зависимой и независимой переменными: парные, частные и множественные коэффициенты корреляции, множественный коэффициент детерминации, а затем проверить адекватность данной модели.

Парные коэффициенты корреляции. Для измерения тесноты связи между двумя из рассматриваемых переменных (без учета их взаимодействия с другими переменными) применяются парные коэффициенты корреляции. Методика расчета таких коэффициентов и их интерпретация аналогичны методике расчета линейного коэффициента корреляции в случае однофакторной связи. Если известны средние квадратические отклонения анализируемых величин, то *парные коэффициенты корреляции* можно рассчитать проще, по следующим формулам:

$$r_{yx_1} = \frac{\overline{x_1 y} - \bar{x}_1 \bar{y}}{\sigma_{x_1} \sigma_y}; \qquad (1.12)$$

$$r_{yx_2} = \frac{\overline{x_2 y} - \bar{x}_2 \bar{y}}{\sigma_{x_2} \sigma_y}; \qquad (1.13)$$

$$r_{x_1 x_2} = \frac{\overline{x_1 x_2} - \bar{x}_1 \bar{x}_2}{\sigma_{x_1} \sigma_{x_2}}. \qquad (1.14)$$

Частные коэффициенты корреляции. Однако в реальных условиях все переменные, как правило, взаимосвязаны. Теснота этой связи определяется частными коэффициентами корреляции, которые характеризуют степень и влияние одного из аргументов на функцию при условии, что остальные независимые переменные закреплены на постоянном уровне. В зависимости от количества переменных, влияние которых исключается, частные коэффициенты корреляции могут быть различного порядка: при исключении влияния одной переменной получаем частный коэффициент корреляции первого порядка; при исключении влияния двух переменных - второго порядка и т.д. Парный коэффициент корреляции между функцией и аргументом обычно не равен соответствующему частному коэффициенту.

Частный коэффициент корреляции первого порядка между признаками x_1 и y при исключении влияния признака x_2 вычисляют по формуле:

$$r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}} \qquad (1.15)$$

то же - зависимость y от x_2 при исключении влияния x_1:

$$r_{yx_2(x_1)} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{x_1x_2}^2\right)}} \qquad (1.16)$$

Можно рассчитать взаимосвязь факторных признаков при устранении влияния результативного признака:

$$r_{x_1x_2(y)} = \frac{r_{x_1x_2} - r_{yx_1} r_{y\,x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{y\,x_2}^2\right)}} \qquad (1.17)$$

где r — парные коэффициенты корреляции между соответствующими признаками.

На основе парных коэффициентов корреляции и средних квадратических отклонений можно легко рассчитать параметры уравнения линейной двухфакторной связи $\hat{y}_{x_1x_2} = a_0 + a_1x_1 + a_2x_2$ по следующим формулам:

$$a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2;\ a_1 = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{1 - r^2_{x_1x_2}} \cdot \frac{\sigma_y}{\sigma_{x_1}};\ a_2 = \frac{r_{yx_2} - r_{yx_1}r_{x_1x_2}}{1 - r^2_{x_1x_2}} \cdot \frac{\sigma_y}{\sigma_{x_2}}.$$

Совокупный коэффициент множественной корреляции. Показателем тесноты связи, устанавливаемой между результативными и двумя или более факторными признаками, является *совокупный коэффициент множественной корреляции* $R_{yx_1,x_2,\ldots,x_n}$. В случае линейной двухфакторной связи совокупный коэффициент множественной корреляции может быть рассчитан по формуле:

Совокупный коэффициент множественной корреляции. Показателем тесноты связи, устанавливаемой между результативными и двумя или более факторными признаками, является *совокупный коэффициент множественной корреляции* $R_{yx_1,x_2,\ldots,x_n}$. В случае линейной двухфакторной связи совокупный коэффициент множественной корреляции может быть рассчитан по формуле:

$$R_{yx_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r_{x_1x_2}^2}}, \qquad (1.18)$$

где *r*- линейные коэффициенты корреляции (парные); подстрочные индексы показывают, между какими признаками они исчисляются.

Совокупный коэффициент множественной корреляции измеряет одновременное влияние факторных признаков на результативный. Его значения находятся в пределах —1 до +1. Чем меньше наблюдаемые значения изучаемого показателя отклоняются от линии множественной регрессии, тем корреляционная связь является более интенсивной, а следовательно, значение *R* ближе к единице.

Совокупный коэффициент множественной детерминации.

Величина R^2 называется *совокупным коэффициентом множественной детерминации*. Она показывает, какая доля вариации изучаемого показателя объясняется влиянием факторов, включенных в уравнение множественной регрессии. Значение совокупного коэффициента множественной детерминации находится в пределах от 0 до 1. Поэтому, чем ближе R^2 к единице, тем вариация изучаемого показателя в большей мере характеризуется влиянием отобранных факторов.

Однако показатели множественной регрессии и корреляции могут оказаться подверженными действию случайных факторов. Поэтому только после проверки адекватности уравнения оно может быть пригодно, например, для выявления резервов повышения производительности труда.

Общая оценка адекватности уравнения может быть получена с помощью дисперсионного *F*-критерия Фишера. Применение же в этих целях множественного коэффициента корреляции недопустимо ввиду того, что многофакторный регрессионный анализ оперирует случайными наблюдениями, но не обязательно распределенными по многомерному нормальному закону (этому закону должны подчиняться отклонения фактических значений функции от расчетных). Совокупный коэффициент множественной детерминации определяет только качество выравнивания по уравнению регрессии.

Проверку значимости уравнения регрессии производят на основе вычисления F-критерия Фишера:

$$F = \frac{\sigma_y^2}{\sigma_{ост}^2} \cdot \frac{n-m}{m-1},$$

где m — число параметров в уравнении регрессии.

Полученное значение — критерия $F_{расч}$ сравнивают с критическим (табличным) для принятого уровня значимости 0,05 или 0,01 и чисел степеней свободы $\nu_1=m\text{-}1$ и $\nu_2=n\text{-}m$. Если оно окажется больше соответствующего табличного значения, то данное уравнение регрессии статистически значимо, т.е. доля вариации, обусловленная регрессией, намного превышает случайную ошибку.

Принято считать, что уравнение регрессии пригодно для практического использования в том случае, если $F_{расч} > F_{табл}$ не менее чем в 4 раза.

Для оценки значимости коэффициентов регрессии при линейной зависимости y от x_1 и x_2 — (двух факторов) используют t-критерий Стьюдента при n-m-1 степенях свободы:

$$t_{a_1} = \frac{a_1 \sigma_{x_1} \sqrt{1 - r^2_{x_1 x_2}} \cdot \sqrt{n - m - 1}}{\sigma_y \sqrt{1 - R^2_{y x_1 x_2}}}; \qquad (1.19, \text{а})$$

$$t_{a_2} = \frac{a_2 \sigma_{x_2} \sqrt{1 - r^2_{x_1 x_2}} \cdot \sqrt{n - m - 1}}{\sigma_y \sqrt{1 - R^2_{y x_1 x_2}}}; \qquad (1.19, \text{б})$$

Существенность совокупного коэффициента корреляции определяют по формуле :

$$t_{R_{yx_1x_2}} = \frac{R_{yx_1x_2} \sqrt{n - m - 1}}{1 - R^2_{yx_1x_2}}. \qquad (1.20)$$

Значения оцениваемых a_1, a_2, $R_{yx_1x_2}$, берутся по модулю.

Если в уравнении все коэффициенты регрессии значимы, то данное уравнение признают окончательным и применяют в качестве модели изучаемого показателя для последующего анализа.

- Оценку значимости коэффициентов регрессии с помощью *t*-критерия используют для завершения отбора существенных факторов в процессе многошагового регрессионного анализа. Он заключается в том, что после оценки значимости всех коэффициентов регрессии из модели исключают тот фактор, коэффициент при котором незначим и имеет наименьшее значение критерия. Затем уравнение регрессии строится без исключенного фактора, и снова проводится оценка адекватности уравнения и значимости коэффициентов регрессии.
- Такой процесс длится до тех пор, пока все коэффициенты регрессии не окажутся значимыми, что свидетельствует о наличии в регрессионной модели только существенных факторов. В некоторых случаях расчетное значение *tрасч* находится вблизи *tтабл* поэтому с точки зрения содержательности модели такой фактор можно оставить для последующей проверки его значимости в сочетании с другим набором факторов.

- Однако на основе коэффициентов регрессии нельзя сказать какой из факторных признаков оказывает наибольшее влияние на результативный признак, так как коэффициенты регрессии между собой не сопоставимы, поскольку они измерены разными единицами. На их основе нельзя также установить в развитии каких факторных признаков заложены наиболее крупные резервы изменения результативного показателя, потому что в коэффициентах регрессии не учтена вариация факторных признаков.
- Чтобы иметь возможность судить о сравнительной силе влияния отдельных факторов и о тех резервах, которые в них заложены, должны быть вычислены *частные коэффициенты эластичности Эi,,* а также *бета-коэффициенты* βi.

Различия в единицах измерения факторов устраняют с помощью *частных коэффициентов эластичности*, которые рассчитывают по формуле:

$$Э_i = a_i \frac{\bar{x}_i}{\bar{y}_i},$$

где a_i - коэффициент регрессии при i -м факторе; $\bar{x}_i$ — среднее значение i-го фактора; $\bar{y}_i$ - среднее значение изучаемого показателя.

Частные коэффициенты эластичности показывают на сколько процентов в среднем изменяется анализируемый показатель с изменением на 1 % каждого фактора при фиксированном положении других факторов.

Для определения факторов, в развитии которых заложены наиболее крупные резервы улучшения изучаемого показателя, необходимо учесть различия в степени варьирования вошедших в уравнение факторов. Это можно сделать с помощью β -*коэффициентов*, которые вычисляют по формуле :

$$\beta_i = a_i \frac{\sigma_{x_i}}{\sigma_y},$$

где σ_{x_i} — среднее квадратическое отклонение i-го фактора; σ_y — среднее квадратическое отклонение показателя; β - *коэффициент показывает* на какую часть среднего квадратического отклонения изменяется результативный признак с изменением соответствующего факторного признака на величину его среднего квадратического отклонения.

Исходя из соотношения $\sum_{i=1}^{n} \beta_i r_i = R^2$ и принимая во внимание, что коэффициент множественной детерминации R^2 есть доля изучаемых факторов в наличном приращении результативного показателя в анализируемой совокупности, можно сделать вывод, что произведение $\beta_i r_i (1 \le i \le n)$ является показателем силы влияния соответствующего фактора на данный показатель.

Поделив произведение $\beta_i r_i$, на коэффициент множественной детерминации R^2 , получим коэффициент, который показывает какова доля вклада анализируемого фактора в суммарное влияние всех отобранных факторов. Обозначив этот коэффициент Δ_i, получим

$$\Delta_i = \frac{\beta_i r_i}{R^2}.$$

Увеличение числа существенных факторов, включаемых в модель исследуемого показателя, позволяет выявить дополнительные резервы производства. Для этого могут быть использованы трех-, четырех- (и т.д.), n -факторные регрессии.

- Многофакторный корреляционный и регрессионный анализ может быть использован в экономико-статистических исследованиях:
- •для приближенной оценки фактического и заданного уровней;
- •в качестве укрупненного норматива (для этого достаточно в уравнение регрессии подставить вместо фактических значений факторов их средние значения);
- •для выявления резервов производства;
- •для проведения межзаводского сравнительного анализа и выявления на его основе скрытых возможностей предприятий;
- •для краткосрочного прогнозирования развития производства и др.

4. Непараметрические методы

- *непараметрические методы*, с помощью которых устанавливается связь между *качественными (атрибутивными) признаками.* Сфера их применения шире, чем параметрических, поскольку не требуется соблюдения условия нормальности распределения зависимой переменной, однако при этом снижается глубина исследования связей. При изучении зависимости между качественными признаками не ставится задача представления ее уравнением. Здесь речь идет только об установлении наличия связи и измерении ее тесноты.
- В практике экономических исследований приходится иногда анализировать связи между *альтернативными признаками*, представленными только группами с противоположными (взаимоисключающими) характеристиками. Тесноту связи в этом случае можно оценить, вычислив *коэффициент ассоциации*.

Для расчета *коэффициента ассоциации* строится четырехклеточная корреляционная таблица, которая носит название таблицы "четырех полей" и имеет следующий вид:

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d

Применительно к таблице "четырех полей" с частотами *a*, *Ь*, *c* и *d* коэффициент ассоциации выражается формулой:

$$k_a = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}.$$

Коэффициент ассоциации изменяется от —1 до +1; чем ближе к +1 или —1, тем сильнее связаны между собой изучаемые признаки.

Если k_a не менее 0,3, то это свидетельствует о наличии связи между качественными признаками.

Пример. Имеющиеся данные о росте отцов и сыновей представлены в табл. 1.1.

Таблица 1.1

Распределение отцов и сыновей по росту, чел.

Рост сына	Рост отца		Всего
	Ниже среднего	Выше среднего	
Ниже среднего	70	30	90
Выше среднего	20	80	110
Итого	**100**	**100**	**200**

Подсчитаем коэффициент ассоциации по данным табл. 1.1:

$$k_a = \frac{70 \cdot 80 - 30 \cdot 20}{\sqrt{90 \cdot 110 \cdot 100 \cdot 100}} \approx 0{,}51..$$

Поскольку $k_a > 0{,}3$, между ростом отцов и сыновей существует корреляционная связь.

Если по каждому из взаимосвязанных признаков выделяется число групп более двух, то для подобного рода таблиц теснота связи между качественными признаками может быть измерена с помощью *показателя взаимной сопряженности А.А. Чупрова:*

$$k_{\varphi} = \frac{\varphi^2}{\sqrt{(k_1 - 1)(k_2 - 1)}},$$

где k_1 — число возможных значений первой статистической величины (число групп по столбцам); k_2 - число возможных значений второй статистической величины (число групп по строкам); φ^2— показатель взаимной сопряженности (определяется как сумма отношений квадратов частот клетки таблицы распределения к произведению итоговых частот соответствующего столбца и строки).

Вычтя из этой суммы единицу, получим φ^2.

Коэффициент взаимной сопряженности А.А-Чупрова изменяется от 0 до 1, но уже при значении 0,3 можно говорить о тесной связи между вариацией изучаемых признаков.

Пример. Данные об уровне образования членов 100 семей приведены в табл. 1.2

Таблица 1.2

Распределение семей по уровню образования мужа и жены

Образование мужа	Образование жены			Итого	
	неполное среднее	среднее и среднее специальное	высшее	A	B
Неполное среднее	15 (225) 9,375	11 (121) 2,373	2 (4) 0,160	28 - 19,08	0,425
Среднее и среднее специальное	8 (64) 2,666	32 (1024) 20,078	8 (64) 2,560	48 - 25,304	0,527
Высшее	1 (1) 0,042	8 (64) 1,255	15 (225) 9,00	24 - 10,297	0,429
Итого	**24**	**51**	**25**	**100**	**1,381**

Примечание: частоты — верхние строки; их квадраты (в скобках) - средние строки; квадраты частот, деленные на суммы частот по столбцу — нижние строки; в итоговых столбцах — сумма частот, сумма результатов деления (*A*), а также результат деления нижнего числа на верхнее — последний столбец (*B*).

Тогда $\phi^2_{\vee} = 1{,}381\text{-}1 = 0{,}381$; $k_1 = k_2 = 3$.

Коэффициент взаимной сопряженности А.А.Чупрова

$$k_{\varphi} = \frac{0{,}381}{\sqrt{(3-1)(3-1)}} \approx 0{,}19.$$

Его значение показывает заметную связь между уровнями образования мужа и жены при формировании семьи.


КОРРЕЛЯЦИЯ
НЕ ПОДРАЗУМЕВАЕТ
ПРИЧИННУЮ
ОБУСЛОВЛЕННОСТЬ