1. Эконометрика: основные понятия и определения

Эконометрика – это наука, изучающая методами математической статистики количественные закономерности и связи в экономике, выражаемые в виде математических моделей.

Целевое назначение эконометрики – эмпирический вывод экономических закономерностей.

Основные задачи эконометрики состоят в построении моделей, выражающих выводимые закономерности, оценка их параметров и проверка гипотез о закономерностях изменения и связях экономических показателей.

Регрессионная модель – это уравнение, в котором объясняемая переменная представляется в виде функции от объясняющих переменных (например, модель спроса на некоторый товар в зависимости от его цены и дохода покупателей). По виду функции различают **линейные** и **нелинейные** регрессионные модели. Наиболее детально изучены и потому наиболее часто встречается в эконометрическом анализе методы оценки и анализа линейных регрессионных моделей.

2. Основные задачи эконометрических исследований

Эконометрическая модель, как правило, основана на теоретическом предположении о круге взаимосвязанных переменных и характере связи между ними. При всем стремлении к «наилучшему» описанию связей приоритет отдается качественному анализу. Поэтому в качестве этапов эконометрического исследования можно указать:

- постановку проблемы;
- получение данных, анализ их качества;
- спецификацию модели;
- оценку параметров;
- интерпретацию результатов.

На начальном этапе решения любой эконометрической задачи необходимо сформулировать эконометрическую модель, т. е. представить модель в виде уравнений, характеризующих связи между экономическими показателями. Например, уравнение связи между доходами семей (x) и сбережениями семей (y), которое необходимо получить путем обработки результатов опроса нескольких сотен случайно отобранных семей:

$$y = \alpha + \beta \cdot x + \varepsilon,$$

где:

x – объясняющая (независимая) переменная (доходы семей);

y – объясняемая (зависимая) переменная (сбережения семей);

ε – случайная составляющая (ошибка);

α и β – параметры уравнения, заранее не известные и подлежащие определению в результате эконометрического анализа задачи.

При решении любой задачи эконометрики необходима проверка соответствия полученной модели реальным экономическим данным. Если модель соответствует реальным данным, то возникает задача определения (оценки) параметров модели. Различают два уровня анализа: теоретический и эмпирический.

На *теоретическом* уровне предполагается, что известны все возможные реализации экономических показателей (т. е. имеется вся генеральная совокупность в целом). Теоретически параметры модели можно оценить, если известны (или предполагаются заданными) статистические свойства генеральной совокупности. Как правило, все возможные исходы (т. е. возможные значения показателей) заранее неизвестны; на практике можно наблюдать только выбранные значения интересующих показателей, т. е. выборочную совокупность.

На *эмпирическом* уровне на основе выборочной совокупности нельзя точно определить значения параметров модели, можно лишь получить их оценки, являющиеся случайными величинами. Таким об-

разом, цель оценивания параметров состоит в получении как можно более точных значений неизвестных параметров модели, которые характерны для всей генеральной совокупности.

Одной из основных задач экономических исследований является анализ зависимости между переменными (показателями), которая может быть функциональной (встречается очень редко) или статистической (в экономике, как правило, является преобладающей).

Функциональная зависимость (иначе ее называют детерминированной) задается в виде формулы, которая каждому значению одной переменной ставит в соответствие строго определенное значение другой переменной, при этом воздействием случайных факторов пренебрегают.

Статистическая зависимость – это связь переменных, на которую накладывается воздействие случайных факторов, при этом изменение одной переменной приводит к изменению математического ожидания другой переменной. Наиболее распространенной формулой статистической связи между переменными является уравнение регрессии. Если эта формула линейная (нелинейная), то регрессию называют линейной (нелинейной). Многие нелинейные модели можно преобразовать в линейные.

3. Парная регрессия и корреляция

Пусть имеется два ряда эмпирических данных **X** *(x_1, x_2, …, x_n)* и **Y** *(y_1, y_2, …, y_n)*, соответствующие им точки с координатами *(x_i, y_i)*, где *i=1,2,…,n,* Тогда, в общем виде *теоретическую линейную парную регрессионную модель* можно представить в виде:

$$Y = \beta_0 + \beta_1 X + \varepsilon \text{ или } y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i=1,2,\ldots,n;$$

где Y – объясняемая (результирующая, зависимая, эндогенная) переменная,

X – объясняющая (факторная, независимая, экзогенная) переменная или регрессор;

β_0 и β_1 – теоретические параметры (числовые коэффициенты) регрессии, подлежащие оцениванию;

ε_i – случайное отклонение (возмущение, ошибка).

Задача линейного регрессионного анализа состоит в том, чтобы по имеющимся статистическим данным *(x_i, y_i), i=1,2,…,n*, для переменных X и Y получить наилучшие оценки неизвестных параметров β_0 и β_1, т. е. построить так называемое *эмпирическое уравнение регрессии*

$$\hat{y}_i = b_0 + b_1 x_i,$$

где $\hat{y}_i$ – оценка объясняемой переменной *y*; b_0 и b_1 – оценки неизвестных параметров β_0 и β_1, называемые *эмпирическими коэффициентами регрессии.* В каждом конкретном случае можно записать

$$y_i = b_0 + b_1 x_i + e_i,\ i=1,2,\ldots,n,$$

где отклонения e_i – ошибки (остатки) модели, которые являются оценками теоретического случайного отклонения ε_i.

Различают линейные и нелинейные регрессии. Нелинейные регрессии делят на два класса: регрессии, нелинейные относительно включенных объясняющих переменных, но линейных по оцениваемым параметрам, и, регрессии, нелинейные по оцениваемым параметрам.

Линейная: $y = a + bx + \varepsilon, \qquad a + bx = f(x).$

Нелинейные по объясняющим параметрам:

$$y = a + b_1 x + b_2 x^2 + \ldots + b_k x^k + \varepsilon,$$

$$y = a + \frac{b}{x}$$

Регрессии, нелинейные по оцениваемым параметрам:

Степенная: $y = ax^b + \varepsilon$

Показательная: $y = ab^x + \varepsilon$

Экспоненциальная: $y = e^{a+bx} + \varepsilon$

Логарифмическая: $\ln y = a + b \ln x + \varepsilon$

Полулогарифмическая: $y = a + b \ln x + \varepsilon$

$$y = a + bx^c + \varepsilon$$

Обратная: $y = \dfrac{1}{a + bx} + \varepsilon$

Если у нас есть набор значений двух переменных x_i и y_i, $i = \overline{1, n}$ то на плоскости XY эти значения можно отобразить точками, таким образом получаем поле корреляции, которое изображено на рис. 1.

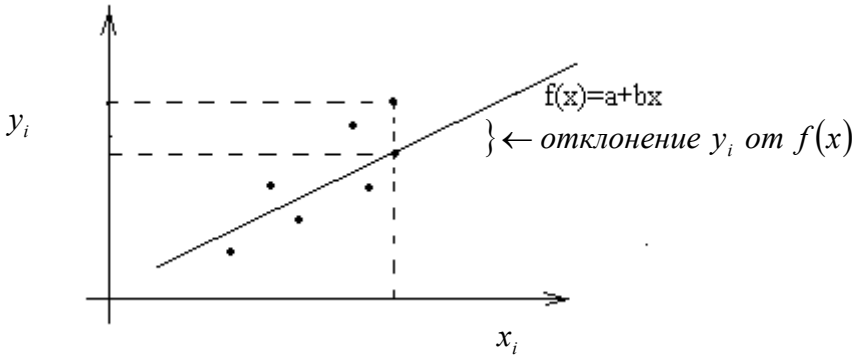

Рис.1. Поле корреляции

4. Метод наименьших квадратов

Построение уравнения регрессии сводится к оценке её параметров. Для оценки параметров регрессии, линейной по параметрам, будем использовать МНК. Согласно МНК поиск наилучшей аппроксимации набора наблюдений линейной функцией сводится к минимизации функционала

$$g = \sum_{i=1}^{n} \left(y_i - \left(a + bx_i\right)\right)^2 .$$

Необходимые условия экстремума:

$$\frac{\partial g}{\partial a} = -2\sum_{i=1}^{n} \left(y_i - \left(a + bx_i\right)\right) = 0 ,$$

$$\frac{\partial g}{\partial b} = -2\sum_{i=1}^{n} \left(y_i - \left(a + bx_i\right)\right)x_i = 0 ,$$

или

$$\sum_{i=1}^{n} \left(y_i - a - bx_i\right) = 0$$

$$\sum_{i=1}^{n} \left(y_i - a - bx_i\right)x_i = 0$$

Введем обозначения:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ , \ \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \ , \ \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \ .$$

Введем обозначения для: $\sigma_x^2 = \overline{x^2} - \bar{x}^2$ – выборочной дисперсии переменной x;

$\sigma_y^2 = \overline{y^2} - \bar{y}^2$ – выборочной дисперсии переменной y;

$\mathrm{cov}(x, y) = \overline{yx} - \bar{y} \cdot \bar{x}$ – выборочной ковариации.

В новых обозначениях система определения a и b принимает вид:

$$\left.\begin{array}{l} a + b\bar{x} = \bar{y} \\ a\bar{x} + b\overline{x^2} = \overline{xy} \end{array}\right\} \times \bar{x}$$

Тогда

$$b = \frac{\bar{x} \cdot \bar{y} - \overline{xy}}{\bar{x}^2 - \overline{x^2}} = \frac{\operatorname{cov}(x, y)}{\sigma_x^2},\ a = \bar{y} - b\bar{x},$$

Замечание 1. Из уравнения для определения параметра a: $\bar{y} = a + b\bar{x}$ следует, что уравнение прямой

$y = a + bx$ проходит через точку $(\bar{x}, \bar{y})$[1].

Замечание 2. Мы предполагаем здесь, что среди X_t t=1,..,n, не все числа одинаковые, т. е $\operatorname{var}(X) \neq 0$ и $b = \frac{\bar{x} \cdot \bar{y} - \overline{xy}}{\bar{x}^2 - \overline{x^2}} = \frac{\operatorname{cov}(x, y)}{\sigma_x^2}$ имеет смысл [1].

Определение. Коэффициент b называется выборочным коэффициентом регрессии (или просто коэффициентом регрессии) y по x.

Коэффициент регрессии y по x показывает, на сколько единиц в среднем изменяется переменная y при увеличении переменной x на одну единицу.

Предпосылки МНК:

1. Математическое ожидание случайного отклонения ε_i равно нулю для всех наблюдений: $M(\varepsilon_i) = 0, i = 1,2,\ldots,n$.

2. Постоянство дисперсии отклонений (гомоскедастичность): $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$ для любых наблюдений *i* и *j*.

3. Отсутствие автокорреляции: случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i \neq j$.

4. Случайное отклонение должно быть независимо от объясняющих переменных.

5. Модель является линейной относительно параметров.

6. Ошибки $\varepsilon_i, i = 1,2,\ldots,n$ имеют нормальное распределение. Выполнимость данной предпосылки важна для проверки статистических гипотез и построения интервальных оценок.

5. Коэффициент корреляции

Наряду с построением уравнения регрессии осуществляется оценка тесноты связи между переменными.

Тесноту связи в случае линейной зависимости характеризуют с помощью выборочного коэффициента корреляции r_{xy}.

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x^2}\cdot\frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y}.$$

Для практических расчетов наиболее удобна формула:

$$r = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x\sigma_y} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\cdot\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

т. к. по этой формуле r находится непосредственно из данных наблюдений, и на значении r не скажутся округление данных, связанные с расчетом средних и отклонений от них.

Коэффициент корреляции принимает значения от -1 до +1.

При значении коэффициента корреляции равном ± 1 связь представлена линейной функциональной зависимостью. При этом все наблюдаемые значения располагаются на линии регрессии.

При r_{xy}=0 корреляционная связь между признаками в линейной форме отсутствует. При этом линия регрессии параллельна оси Ох.

При $r_{xy} > 0$ – корреляционная связь между переменными называется прямой, а при $r_{xy} < 0$ – обратной.

Для характеристики силы связи можно использовать шкалу *Чеддока.*

Показатель тесноты связи	0,1–0,3	0,3–0,5	0,5–0,7	0,7–0,9	0,9–0,99
Характеристика силы связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

6. Оценка значимости уравнения регрессии

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

Проверка значимости уравнения регрессии производится на основе дисперсионного анализа.

Обозначим через $\hat{y} = a + bx$ - теоретически вычисляемые по формуле значения, тогда

$$y_i - \bar{y} = y_i - \bar{y} + \hat{y}_i - \hat{y}_i = \left(y_i - \hat{y}_i \right) + \left(\hat{y}_i - \bar{y} \right)$$

Преобразуем формулу вариации результирующего признака с учетом вышеуказанной суммы:

$$\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2 = \sum_{i=1}^{n} \left[\left(\hat{y}_i - \bar{y} \right) + \left(y_i - \hat{y}_i \right) \right]^2 = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2 + \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + 2 \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right) \left(y_i - \hat{y}_i \right)$$

Далее

$$\sum \left(\hat{y}_i - \bar{y} \right) \left(y_i - \hat{y}_i \right) = \sum \left(y_i - \hat{y}_i \right) \left(a + bx_i - \bar{y} \pm b\bar{x} \right) = \sum \left(y_i - \hat{y}_i \right) \left(a - \bar{y} + b\bar{x} \right) +$$

$$+ b \sum \left(y_i - \hat{y}_i \right) \left(x_i - \bar{x} \right) = b \sum \left(y_i - a - bx_i \right) x_i - b\bar{x} \sum \left(y_i - a - bx_i \right) = 0$$

Так как имеет место равенство $\left(a - \bar{y} + b\bar{x} \right) = 0$,

и из МНК следуют два соотношения $\begin{matrix} \sum \left(y_i - a - bx_i \right) x_i = 0 \\ \sum \left(y_i - a - bx_i \right) = 0 \end{matrix}$,

то

$$\underbrace{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}_{TSS} = \underbrace{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2}_{RSS} + \underbrace{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}_{ESS} \quad (*)$$

Введем обозначения:

TSS (total sum of sguares) – вся дисперсия: сумма квадратов отклонений от среднего.

RSS (regression sum of sguares) – объясненная часть всей дисперсии (обусловленная регрессией), факторная, объясненная дисперсия.

ESS (error sum of sguares) – остаточная сумма, дисперсия остаточная.

Замечание. Вектор остатков регрессии ортогонален константе, вообще говоря, только в том случае, когда константа включена в число объясняющих параметров регрессии. Поэтому (*) справедливо, вообще говоря, только в случае, когда константа включена в число объясняющих параметров регрессии.

Определение. Коэффициентом детерминации – это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными называется

$$R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS}.$$

В силу определения R^2 : $0 \le R^2 \le 1$.

Если $R^2 = 0$, то это означает, что регрессия ничего не дает, т. е. x_i не улучшает качество предсказания y_i, по сравнению с тривиальным $\hat{y}_i = \bar{y}$.

Если $R^2 = 1$, то (x_i, y_i) лежат на линии регрессии и между x и y существует линейная функциональная зависимость, т. е. абсолютно точное совпадение: $\hat{y}_i = y_i$.

Для линейной регрессии определяется коэффициент регрессии по формуле:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} \text{ или } r_{xy}^2 \cdot \sigma_y^2 = b^2 \cdot \sigma_x^2.$$

Тогда

$$b^2\sigma_x^2 = b^2\sum\left(x_i - \bar{x}\right)^2 = \sum\left(bx_i - b\bar{x}\right)^2 = \sum\left(\left(\hat{y}_i - a\right) - \left(\bar{y} - a\right)\right)^2 = \sum\left(\hat{y}_i - a - \bar{y} + a\right)^2$$

– получившаяся формула есть дисперсия объясненная, факторная, тогда $r_{xy}^2 = \frac{b^2\sigma_x^2}{\sigma_y^2} = \frac{\sigma^2_{объясн}}{\sigma^2_{y\ общ}} = \frac{RSS}{TSS} = R^2$;

7. Оценка качества модели

Оценку качества построенной модели можно определить через коэффициент (индекс) детерминации, а также с помощью средней ошибки аппроксимации.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических в процентах:

$$\overline{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}_x}{y}\right|\cdot 100\%.$$

Предел значений $A \le 0.08-0.1 \quad (8-10\%)$ считаем допустимым при построении модели.

В эконометрических исследованиях широкое применение находит такой показатель, как коэффициент эластичности. Если зависимость между переменными *x* и *y* имеет вид *y=f(x),* то коэффициент эластичности *Э* вычисляется по формуле

$$Э = f'(x)\frac{x}{y}$$

Коэффициент эластичности *Э* показывает, на сколько процентов в среднем изменится результативный признак *y* при изменении фактора *x* на 1% от своего номинального значения.

Средний коэффициент эластичности $\overline{Э}$ показывает, на сколько % в среднем по совокупности изменится результат y от своей средней величины при изменении фактора x на 1% от своего среднего значения

$$\overline{Э} = f'(x)\frac{\overline{x}}{\overline{y}} \quad \Rightarrow \quad \overline{Э}\cdot\overline{y} = f'(x)\cdot\overline{x}$$

f' - характеризует соотношение прироста результата и фактора для соответствующей формы связи.

Т. к., коэффициент Э не всегда const, то используем среднее значение – $\overline{Э}$.

В таблице представлены формулы эластичности для наиболее употребительных функций.

y	y'	Э
$y = a + bx$	b	$Э = \frac{bx}{a+bx}$
$y = a + bx + cx^2$	$b + 2cx$	$Э = \frac{(b+2cx)x}{a+bx+cx^2}$
$y = a + \frac{b}{x}$	$-\frac{b}{x^2}$	$Э = \frac{-b}{a+bx}$

$y = ab^x$	$\ln b \cdot a \cdot b^x$	$x \ln b$
$y = ax^b$	$a \cdot b \cdot x^{b-1}$	b
$y = a + b \ln x$	$\frac{b}{x}$	$\frac{b}{a + b \ln x}$
$y = \frac{1}{a + bx}$	$\frac{-b}{(a + bx)^2}$	$\frac{-bx}{a + bx}$

Иногда коэффициент Э экономического смысла не имеет. Это происходит тогда, когда для рассматриваемых признаков бессмысленно определение изменения значений в процентах. Например, изменение роста заработной платы с ростом стажа работы на 1%.

Использование F-критерия

С помощью F-критерия можно оценить качество построенного уравнения регрессии.

Поскольку при заданном объеме наблюдений (x, y) факторная сумма квадратов при линейной регрессии зависит только от одной константы – коэффициента регрессии b, то говорят, что данная сумма квадратов имеет одну степень свободы.

К этому же выводу мы придем формальным путем, а именно, $y_x = a + bx$. Но свободный член $a = \bar{y} - b\bar{x}$, тогда

$$y_x = \bar{y} - b\bar{x} + bx = \bar{y} - b(x - \bar{x})$$

при заданном наборе переменных x и y, расчетное значение y_x является в линейной регрессии функцией только одного параметра b. Соответственно факторная сумма квадратов отклонений имеет число степеней свободы равное 1.

Любая сумма квадратов отклонений связана с числом степеней свободы, т. е. с числом свободы независимого варьирования признака. Значит число степеней свободы связано с числом единиц совокупности n и с числом определяемых по ней констант. Применительно к исследуемой проблеме число степеней свободы должно показывать, сколько независимых отклонений из n возможных $y_1 - \bar{y}$, $y_2 - \bar{y}$, …, $y_n - \bar{y}$ требуется для образования данной суммы квадратов. Так для общей суммы квадратов $\sum (y_i - \bar{y})^2$ требуется $(n-1)$ независимое отклонение, ибо по совокупности из n единиц после расчета среднего уровня $\bar{y}$, свободно варьируют лишь $(n-1)$ числом отклонений.

Например, имеем ряд 1, 2, 3, 4, 5. Среднее $\bar{y}$ =3, тогда *n* отклонений от среднего: -2, -1, 0, 1, 2. Т.к. $\sum (y - \bar{y}) = 0$, то свободно варьируют 4 отклонения, а пятое может быть определено, если 4 известны.

Число степеней свободы в левой и правой частях соотношения (*) должно совпадать, то число степеней свободы второго слагаемого должно быть равно $(n-2)$.

То есть $\underbrace{\sum(y-\bar{y})^2}_{n-1} = \underbrace{\sum(y_x-\bar{y})^2}_{1} + \underbrace{\sum(y-y_x)^2}_{n-2}$.

Разделив каждую сумму квадратов на соответствующее ей число степеней свободы, получим средний квадрат отклонений, или, что то же самое, дисперсию на одну степень свободы D

$$\frac{TSS}{n-1} = D_{общ}, \quad \frac{RSS}{1} = D_{факт}, \quad \frac{ESS}{n-2} = D_{остат}\cdot\cdot$$

Это приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточные дисперсии в расчете на одну степень свободы, получим величину F – отношения (F- критерия):

$F = \dfrac{D_{факт}}{D_{остат}}$, где F- критерий для проверки нулевой гипотезы H_0 о статистической незначимости уравнения регрессии.

Эта гипотеза отвергается при выполнении условия $F_{табл} < F_{факт}$

$F_{табл}$ – это максимальная величина отношения дисперсий, которая может иметь место при случайном их расхождении для данного уровня вероятности.

F-критерий – это оценивание качества уравнения регрессии, которое состоит в проверке гипотезы H_0 о статистической незначимости уравнения регрессии и показателя тесноты связи. Для этого производится сравнение фактического $F_{факт}$ и $F_{табл}$ значений F критерия Фишера-Снедекора. $F_{факт}$ определяется из соотношения значений факторной и остаточной дисперсий, рассчитанных на одну степень свободы

$$F_{факт} = \frac{\sum(y_x-\bar{y})^2/m}{\sum(y-y_x)^2/(n-m-1)} = \frac{r_{xy}^2}{1-r_{xy}^2}(n-2), \quad m=1\cdot$$

$F_{табл}$ - это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости α. Уровень значимости α - это вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно $\alpha = 0{,}05 \ \ (0{,}01)$.

Если $F_{табл} < F_{факт}$, то H_0 - гипотеза отклоняется и признается их статистическая значимость и надежность т. е. построенное уравнение регрессии признается значимым.

Если $F_{табл} > F_{факт}$, то H_0 - гипотеза не отклоняется и признается статистическая незначимость, ненадежность уравнения регрессии.

8. Интервальная оценка функции регрессии и её параметров.

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитывается t-критерий Стьюдента и доверительные интервалы каждого из показателей. Выдвигается гипотеза H_0 о случайной природе показателей, т. е. о незначимом их отличии от нуля. Оценка значимости коэффициентов регрессии и корреляции с помощью t-критерия Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки.

Рассчитываются: $t_b = \dfrac{b}{m_b}$; $t_a = \dfrac{a}{m_a}$; $t_r = \dfrac{r}{m_r}$.

Уравнение регрессии $\hat{y} = a + bx$ представимо в виде:

$$\hat{y} = \bar{y} + b(x - \bar{x}).$$

Стандартная ошибка $m^2_{\hat{y}} = m^2_{\bar{y}} + m^2_b (x - \bar{x})^2$, т. е. стандартная ошибка прогнозного значения зависит от ошибки $\bar{y}$ и ошибки коэффициента регрессии.

Для доказательства этого рассмотрим дисперсии: $\sigma^2_{\hat{y}}$, $\sigma^2_{\bar{y}}$, σ^2_b.

$\sigma^2_{\hat{y}} = \sigma^2_{\bar{y}} + \sigma^2_b (x - \bar{x})^2$ - здесь учтено, что $(x - \bar{x})$ неслучайная (детерминированная) величина, при вынесении которой за знак дисперсии её необходимо возвести в квадрат,

$$\sigma^2_{\bar{y}} = \sigma^2 \left(\frac{\sum_{i=1}^{n} y_i}{n} \right) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2_{y_i} = \frac{\sigma^2}{n^2} n = \frac{\sigma^2_y}{n}$$

$$\sigma^2_b : \; b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum x_i' y_i'}{\sum x_i'^2};$$

$$\sigma^2_b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma^2_y}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} = \frac{\sigma^2_y \sum_{i=1}^{n} (x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} = \frac{\sigma^2_y}{\sigma^2_x}$$

Т.о. $\sigma^2_{\hat{y}} = \frac{\sigma_y^2}{n} + \frac{\sigma_y^2 (x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sigma_y^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right)$.

Определим стандартную ошибку $\bar{y}$ через остаточную дисперсию на одну степень свободы:

$$\sigma_{ост} = \sqrt{\frac{\sum \left(y - \hat{y} \right)^2}{n-2}} \text{ или } \sigma^2_{ост} = \frac{\sum \left(y - \hat{y} \right)^2}{n-2}$$

$$m^2_{\bar{y}} = \frac{\sigma^2_{ост}}{n}, \; m^2_b = \frac{\sigma^2_{ост}}{\sigma^2_x}, \; m^2_{\hat{y}} = \sigma^2_{ост} \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$$

$$m^2_a = \sigma^2_{ост} \frac{\sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}, \; m^2_r = \frac{1 - r^2_{xy}}{n-2}$$

Сравнивая фактические и табличные значения t - статистики $t_{факт}$ и $t_{табл}$ принимаем или отвергаем гипотезу H_0. Установим связь между F-критерием Фишера и t - статистикой Стьюдента:

$$\sigma^2_{факт} = r^2 \sigma^2_y \; ; \; F_{крит} = \frac{r^2 \sigma^2_y}{\left(1 - r^2\right)\sigma^2_y} (n-2) = \frac{r^2}{1-r^2}(n-2),$$

но $t_r = \frac{r}{m_r} = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}$, очевидно $t_r^2 = F$.

$$t^2_b = \frac{b^2}{m^2_b} = \frac{b^2}{\sum \left(y - \hat{y} \right)^2 / n-2} \cdot \sum (x - \bar{x})^2 = \frac{\sum \left(\hat{y} - \bar{y} \right)^2}{\frac{\sum \left(y - \hat{y} \right)^2}{n-2}} = \frac{D_{факт}}{D_{ост}} = F$$

Следовательно: $t_r^2 = t_b^2$.

Т. о. проверка гипотез о значимости коэффициентов регрессии и корреляции равносильна проверке гипотезы о существенности линейного уравнения регрессии.

Если $t_{табл} < t_{факт}$, но H_0 отклоняется, т. е. a, b, r не случайно отличаются от нуля и сформировались под влиянием систематически

действующего фактора x. Если $t_{табл} > t_{факт}$, то H_0 не отклоняется и признается случайная природа формирования a, b или r_{xy}.

Для расчета доверительного интервала определяем предельную ошибку Δ для каждого показателя:

$$\Delta_a = t_{табл} \cdot m_a \ , \ \Delta_b = t_{табл} \cdot m_b ,$$

тогда формула для расчета доверительных интервалов имеют следующий вид:

$$\gamma_{a\min} = a - \Delta_a , \ \gamma_{a\max} = a + \Delta_a \ ;$$

$$\gamma_{b\min} = b - \Delta_b , \ \gamma_{b\max} = b + \Delta_b \ ;$$

Если в границы доверительного интервала попадает ноль, т. е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, т. к. он не может одновременно принимать и положительное и отрицательное значения.

Прогнозное значение y_p определяется путем подстановки в уравнение регрессии $y_x = a + bx$ соответствующего прогнозного значения x_p. Вычисляется средняя стандартная ошибка прогноза m_{yp}

$$m_{yp} = \sigma_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\sum \left(x - \bar{x}\right)^2}} \ , \ \sigma_{ост} = \sqrt{\frac{\sum \left(y - y_x\right)^2}{n - m - 1}}$$

и строится доверительный интервал прогноза

$$\gamma_{yp} = y_p \pm t_{табл} \cdot m_{yp} .$$

9. Решение задач регрессии без помощи специальных средств

Пример 1. В качестве примера модели линейной парной регрессии рассмотрим зависимость между сменной добычей угля на одного рабочего Y (т) и мощностью пласта X (м) по следующим (условным) данным, характеризующим процесс добычи угля в $n = 10$ шахтах и представленных в следующей таблице:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	11	12	9	8	8	9	9	8	12
y_i	5	10	10	7	5	6	6	5	6	8

По данным исходной таблицы требуется:

1) найти уравнение регрессии Y по X,

2) вычислить коэффициент корреляции между переменными X и Y,

3) оценить сменную среднюю добычу угля на одного рабочего для шахт с мощностью пласта 8 м,

4) найти 95%-ные доверительные интервалы для индивидуального и среднего значений сменной добычи угля на 1 рабочего для таких же шахт,

5) найти с надежностью 0,95 интервальные оценки коэффициента регрессии β_1 и дисперсии σ^2,

6) оценить на уровне $\alpha = 0{,}05$ значимость уравнения Y по X,

7) найти коэффициент детерминации и пояснить его смысл.

Решение

1) Уравнения для определения величин b_0 и b_1 удобно предварительно преобразовать к так называемой системе нормальных уравнений:

$$\begin{cases} b_0 + b_1 \cdot \bar{x} = \bar{y}; \\ b_0 \cdot \bar{x} + b_1 \cdot \overline{x^2} = \overline{x \cdot y}, \end{cases}$$

где соответствующие средние определяются по формулам:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}; \quad \overline{x \cdot y} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{n}; \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}; \quad \overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}.$$

Подставляя значение

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

из первого уравнения последней системы в уравнение регрессии, получим:

$$\hat{y} = \bar{y} - b_1 \cdot \bar{x} + b_1 \cdot x$$

или

$$\hat{y} - \bar{y} = b_1 \cdot (x - \bar{x}).$$

Для нахождения уравнения регрессии Y по X вычислим все необходимые суммы:

$$\sum_{i=1}^{10} x_i = 8 + 11 + 12 + 9 + 8 + 8 + 9 + 9 + 8 + 12 = 94;$$

$$\sum_{i=1}^{10} x_i^2 = 8^2 + 11^2 + 12^2 + 9^2 + 8^2 + 8^2 + 9^2 + 9^2 + 8^2 + 12^2 = 908;$$

$$\sum_{i=1}^{10} y_1 = 5 + 10 + 10 + 7 + 5 + 6 + 6 + 5 + 6 + 8 = 68;$$

$$\sum_{i=1}^{10} x_i \cdot y_i = 8 \cdot 5 + 11 \cdot 10 + 12 \cdot 10 + 9 \cdot 7 + 8 \cdot 5 + 8 \cdot 6 + 9 \cdot 6 + 9 \cdot 5 + 8 \cdot 6 + 12 \cdot 8 = 664$$

Теперь находим выборочные характеристики и параметры уравнений регрессии:

$$\bar{x} = \frac{94}{10} = 9{,}4;\ \bar{y} = \frac{68}{10} = 6{,}8;\ s_x^2 = \frac{908}{10} - 9{,}4^2 = 2{,}44;$$

$$\hat{Cov}(X, Y) = \frac{664}{10} - 9{,}4 \cdot 6{,}8 = 2{,}48;\ b_1 = \frac{2{,}48}{2{,}44} = 1{,}016.$$

Итак, уравнение регрессии Y по X:

$\hat{y} - 6{,}8 = 1{,}016 \cdot (x - 9{,}4)$ или $\hat{y} = -2{,}75 + 1{,}016 \cdot x$.

Из полученного уравнение регрессии следует, что при увеличении мощности пласта X на 1 м добыча угля на одного рабочего Y увеличивается в среднем на 1,016 т (в усл. ед.) (отметим, что свободный член в данном уравнении не имеет экономического смысла).

2) Для практических расчетов коэффициента корреляции r_{xy} между переменными X и Y целесообразно формулу

$$r_{xy} = \frac{cov(x, y)}{\sqrt{var(x) \cdot var(y)}} = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y}$$

преобразовать к виду:

$$r_{xy} = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}},$$

так как по ней r_{xy} определяется непосредственно из данных наблюдений, и на значении r_{xy} не скажутся округления данных, связанные с расчетом средних и отклонений от них.

Используя ранее подсчитанные суммы

$$\sum_{i=1}^{10} x_i = 94;\ \sum_{i=1}^{10} x_i^2 = 908;\ \sum_{i=1}^{10} y_i = 68;\ \sum_{i=1}^{10} x_i \cdot y_i = 664$$

и вычислив сумму

$$\sum_{i=1}^{10} y_i^2 = 5^2 + 10^2 + 10^2 + 7^2 + 5^2 + 6^2 + 6^2 + 5^2 + 6^2 + 8^2 = 496,$$

определим искомый коэффициент корреляции:

$$r_{xy} = \frac{10 \cdot 664 - 94 \cdot 68}{\sqrt{10 \cdot 908 - 94^2} \cdot \sqrt{10 \cdot 496 - 68^2}} = 0{,}866,$$

величина которого показывает достаточно тесную связь между переменными X и Y.

3) Для построения *доверительного интервала* для функции регрессии (накрывающего с доверительной вероятностью $\gamma = 1 - \alpha$ неизвестное значение $M_X(Y)$) определим дисперсию групповой средней $\hat{y}$, представляющей выборочную оценку $M_X(Y)$. С этой целью уравнение регрессии $\hat{y} - \bar{y} = b_1 \cdot (x - \bar{x})$ представим в виде:

$$\hat{y} = \bar{y} + b_1 \cdot (x - \bar{x})$$

Так как дисперсия групповой средней равна сумме дисперсий двух независимых слагаемых:

$$\sigma_{\hat{y}}^2 = \sigma_{\bar{y}}^2 + \sigma_{b_1}^2 \cdot (x - \bar{x})^2,$$

то остается рассчитать каждую из них в отдельности.

Дисперсия выборочной средней $\bar{y}$ рассчитывается по формуле:

$$\sigma_{\bar{y}}^2 = \sigma^2 \cdot \left(\frac{\sum_{i=1}^{n} y_i}{n} \right) = \frac{\sum_{i=1}^{n} \sigma_{y_i}^2}{n^2} = \frac{\sum_{i=1}^{n} \sigma^2}{n^2} = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Для расчета дисперсии коэффициента b_1 удобно начало координат переместить в точку $(\bar{x}, \bar{y})$, тогда $x_i' = x_i - \bar{x}, y_i' = y_i - \bar{y}$, при этом $\overline{x'} = 0, \overline{y'} = 0$, а уравнение регрессии $\hat{y} - \bar{y} = b_1 \cdot (x - \bar{x})$ упрощается

$$\hat{y}' = b_1 \cdot x',$$

и коэффициент регрессии b_1 можно рассчитать по формуле:

$$b_1 = \frac{\overline{x' \cdot y'}}{\overline{x'^2}} = \frac{\sum_{i=1}^{n} x' \cdot y'}{\sum_{i=1}^{n} x'^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Тогда дисперсия коэффициента b_1 равна:

$$\sigma^2_{b_1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot \sigma^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2} = \frac{\sigma_2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Дисперсия групповых средних вычисляется с использованием соотношений для дисперсий выборочной средней $\bar{y}$ и коэффициента b_1 с заменой σ^2 ее оценкой s^2:

$$s^2_{\hat{y}} = s^2 \cdot \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right).$$

Доверительный интервал для условного математического ожидания $M_X(Y)$ можно построить, используя *статистику* $t = \frac{\hat{y} - M_X(Y)}{s_{\hat{y}}}$, имеющую $t-$ распределение Стьюдента с $k = n - 2$ *степенями свободы*:

$$\hat{y} - t_{1-\alpha:k} \cdot s_{\hat{y}} \le M_X(Y) \le \hat{y} + t_{1-\alpha:k} \cdot s_{\hat{y}},$$

где $s_{\hat{y}} = \sqrt{s^2_{\hat{y}}}$ – стандартная ошибка групповой средней $\hat{y}$.

Выборочной оценкой условного математического ожидания $M_{x=8}(Y)$ является групповая средняя $\hat{y}_{x=8}$, которая определяется по построенному уравнению регрессии:

$$\hat{y}_{x=8} = -2{,}75 + 1{,}016 \cdot 8 = 5{,}38 \text{ (т)}.$$

Построение доверительного интервала для $M_{x=8}(Y)$ предполагает знание дисперсию его оценки, т. е. $s^2_{\hat{y}_{x=8}}$. Результаты промежуточных расчетов (с учетом того, что $\bar{x} = 9{,}4$) удобно свести в таблицу:

x_i	8	11	12	9	8	8	9	9	8	12	∑
$(x_i - \bar{x})^2$	1,96	2,56	6,76	0,16	1,96	1,96	0,16	0,16	1,96	6,76	24,40

$\hat{y}_i = -2{,}75 + 1{,}016 \cdot x_i$	5,38	8,43	9,44	6,39	5,38	5,38	6,39	6,39	5,38	9,44	–
$e_i^2 = (\hat{y}_i - y_i)^2$	0,14	2,48	0,31	0,37	0,14	0,39	0,15	1,94	0,39	2,08	8,39

Несмещенной оценкой остаточной дисперсии σ^2 является выборочная остаточная дисперсия

$$s^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{8{,}39}{10-2} = 1{,}049,$$

а в дисперсии коэффициента b_1 заменой σ^2 ее оценкой s^2 получим:

$$s^2_{\hat{y}_{x=8}} = s^2 \cdot \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) = 1{,}049 \cdot \left[\frac{1}{10} + \frac{(8-9{,}4)^2}{24{,}4} \right] = 0{,}189$$

и $s_{\hat{y}_{x=8}} = \sqrt{0{,}189} = 0{,}435$ (т). Взяв из таблицы t – распределения Стьюдента $t_{0{,}95;8} = 2{,}31$, можно определить доверительный интервал для условного математического ожидания $M_X(Y)$ с помощью соотношения:

$$\hat{y} - t_{1-\alpha:k} \cdot s_{\hat{y}} \le M_X(Y) \le \hat{y} + t_{1-\alpha:k} \cdot s_{\hat{y}}, \text{ откуда}$$

$$5{,}38 - 2{,}31 \cdot 0{,}435 \le M_{x=8}(Y) \le 5{,}38 + 2{,}31 \cdot 0{,}435 \text{ или}$$

$$4{,}38 \le M_{x=8}(Y) \le 6{,}38 \text{ (т)}.$$

Таким образом, средняя сменная добыча угля на одного рабочего для шахт с мощностью пласта 8 м с надежностью 0,95 находится в пределах от 4,38 до 6,38 т.

4) Для построения доверительного интервала для индивидуального значения $y^*_{x_0=8}$ сначала определяется дисперсия его оценки:

$$s^2_{\hat{y}_{x=8}} = s^2 \cdot \left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) = 1{,}049 \cdot \left[1 + \frac{1}{10} + \frac{(8-9{,}4)^2}{24{,}4} \right] = 1{,}238;$$

откуда $s_{y_{x=8}} = \sqrt{1{,}238} = 1{,}113$ (т), а затем искомый доверительный интервал:

$$\hat{y}_0 - t_{1-\alpha:n-2} \cdot s_{\hat{y}_0} \le y_0^* \le \hat{y}_0 + t_{1-\alpha:n-2} \cdot s_{\hat{y}_0},$$

откуда