ЛИНЕЙНАЯ МОДЕЛЬ ПАРНОЙ РЕГРЕССИИ

ВОПРОСЫ

1. Суть регрессионного анализа.
2. Парная линейная регрессия.
   Метод наименьших квадратов (МНК).
3. Основные предпосылки регрессионного анализа. Теорема Гаусса-Маркова.

1.СУТЬ РЕГРЕССИОННОГО АНАЛИЗА

В регрессионном анализе рассматривается односторонняя зависимость случайной переменной ***Y*** от одной (или нескольких) неслучайной независимой переменной ***X***.

Такая зависимость не является однозначной в том смысле, что каждому конкретному значению переменной ***X*** может соответствовать не какое-то определенное, а множество возможных значений переменной ***Y***.

Поэтому анализируют, как независимая переменная ***X*** влияет на зависимую переменную ***Y*** «в среднем».

Зависимость такого типа выражается соотношением:

$$\boldsymbol{M}\,(\,\boldsymbol{Y}\,/\,\boldsymbol{X} = \boldsymbol{x}\,) = \boldsymbol{f}\,(\boldsymbol{x}) \qquad (1)$$

где $\boldsymbol{M}\,(\,\boldsymbol{Y}\,/\,\boldsymbol{X} = \boldsymbol{x}\,)$ – математическое ожидание (среднее значение) случайной переменной ***Y***, вычисленное в предположении, что переменная ***X*** приняла значение ***x***.

Уравнение (1) называется

уравнением регрессии,

а функция $\boldsymbol{f}$ ($\boldsymbol{x}$) - ***функцией регрессии.***

Реальные (наблюдаемые) значения зависимой переменной $\boldsymbol{Y}$ будут в большей или меньшей мере отклонятся от функции регрессии $\boldsymbol{f(X)}$.

Поэтому уравнение взаимосвязи двух переменных должно быть дополнено некоторой случайной переменной $\boldsymbol{\varepsilon}$ **,** характеризующей отклонение реального значения зависимой переменной $\boldsymbol{Y}$ от теоретического значения, найденного по уравнению регрессии.

Переменную $\boldsymbol{\varepsilon}$ называют ***возмущением*** (или ***отклонением***).

Тогда взаимосвязь двух переменных будет представлена в виде **парной регрессионной модели**:

$$Y = f(X) + \varepsilon \qquad (2)$$

наблюдаемое значение зависимой переменной

объясненная часть, зависящая от объясняющей переменной и выраженная уравнением регрессии

случайная составляющая (отклонение, возмущение)

Причинами присутствия в регрессионной модели случайной величины ε могут быть:

- ✓ невключение в модель всех объясняющих переменных;
- ✓ неправильный выбор функциональной формы модели;
- ✓ ошибки измерений;
- ✓ ограниченность статистических данных;
- ✓ непредсказуемость человеческого фактора;
- ✓ и другие.

2. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (МНК)

Наиболее простым видом зависимости между экономическими переменными является *модель парной линейной регрессии*.

Теоретическая линейная регрессионная модель представляется в виде:

$$Y = \alpha + \beta X + \varepsilon, \quad (3)$$

где α, β - параметры (коэффициенты) уравнения регрессии.

Для определения значений параметров регрессии $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$ необходимо знать и использовать все значения переменных $\boldsymbol{X}$ и $\boldsymbol{Y}$ генеральной совокупности, что практически невозможно.

Обычно исследователь располагает лишь выборочными наблюдениями над величинами $\boldsymbol{X}$ и $\boldsymbol{Y}$ ограниченного объема $\boldsymbol{n}$:

$$x_1,\ x_2, \ldots\ldots, x_n;$$

$$y_1,\ y_2, \ldots\ldots, y_n$$

Таким образом, задача линейного регрессионного анализа состоит в том, чтобы

по имеющимся статистическим данным (x_i, y_i) (i =1,2,…n) для переменных $\boldsymbol{X}$ и $\boldsymbol{Y}$ получить наилучшие оценки неизвестных параметров $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$.

По выборке ограниченного объема ***n*** можно построить ***выборочное уравнение регрессии:***

$$\hat{y} = a + bx \ , \qquad (4)$$

где

$\hat{y}$ – условная средняя переменной Y при конкретном значении переменной X.

a и ***b*** - оценки параметров α и β.

Неизвестные значения ***a*** и ***b*** определяются **МНК**.

Согласно МНК,
неизвестные параметры a и b выбираются таким образом, чтобы сумма квадратов отклонений наблюдаемых значений y_i от значений $\hat{y}_i$, найденных по уравнению регрессии (4), была минимальной.

$$S = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i \right)^2 \to \min$$

где

($\boldsymbol{x}_i$, $\boldsymbol{y}_i$) - известные значения наблюдений,

a и ***b*** - неизвестные параметры;

n - число наблюдений.

Чтобы найти минимум функции $\boldsymbol{S = S\,(a, b)}$,

надо вычислить частные производные по каждому из параметров $\boldsymbol{a}$ и $\boldsymbol{b}$ и приравнять их к нулю.

$$\begin{cases} \dfrac{\partial S}{\partial a} = \sum 2\left(y_i - a - bx_i\right)\cdot\left(-1\right) = 0, \\ \dfrac{\partial S}{\partial b} = \sum 2\left(y_i - a - bx_i\right)\cdot\left(-x_i\right) = 0 \end{cases}$$

$$\begin{cases} 2\sum\left(-y_i + a + bx_i\right) = 0, \\ 2\sum\left(-x_i y_i + ax_i + bx_i^2\right) = 0 \end{cases}$$

$$\begin{cases} \sum -y_i + \sum a + \sum bx_i = 0, \\ \sum -x_i y_i + \sum ax_i + \sum bx_i^2 = 0 \end{cases}$$

$$\begin{cases} na + b\sum x_i = \sum y_i \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i \end{cases}$$

Разделим обе части уравнения на ***n***:

$$a + \frac{b\sum x_i}{n} = \frac{\sum y_i}{n}$$

$$\frac{a\sum x_i}{n} + \frac{b\sum x_i^2}{n} = \frac{\sum x_i y_i}{n}$$

Учитывая, что $\frac{\sum x_i}{n} = \overline{x}; \quad \frac{\sum y_i}{n} = \overline{y};$

$$\frac{\sum x_i^2}{n} = \overline{x^2}; \quad \frac{\sum x_i y_i}{n} = \overline{xy}$$

Получаем систему нормальных уравнений:

$$\begin{cases} a + b\overline{x} = \overline{y} \\ a\overline{x} + b\overline{x^2} = \overline{xy} \end{cases}$$

Решение системы

$$\begin{cases} a = \overline{y} - b\overline{x}, \\ b = \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \dfrac{\operatorname{cov}(x,y)}{\sigma_x^2} \end{cases}$$

где

$\operatorname{cov}(x, y)$ - выборочная ковариация переменных X и Y,

σ_x^2 - выборочная дисперсия переменной X.

Параметр ***b*** называется ***коэффициентом регрессии***. Его величина показывает, на сколько единиц в среднем изменяется переменная ***Y*** при увеличении переменной ***X*** на одну единицу.

Формально ***a*** - значение переменной ***Y*** при ***x***=0.

Если переменная ***X*** не имеет и не может иметь нулевого значения, то такая трактовка коэффициента ***a*** не имеет смысла.

3. ОСНОВНЫЕ ПРЕДПОСЫЛКИ РЕГРЕССИОННОГО АНАЛИЗА. ТЕОРЕМА ГАУССА- МАРКОВА

Пусть для оценки параметров линейной функции регрессии взята выборка, содержащая ***n*** пар значений переменных (x_i; y_i), где i =1, 2,…n.

В этом случае линейная регрессионная модель с двумя переменными имеет вид:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad (5)$$

$$(i = 1,2, \ldots,n)$$

Для того, чтобы регрессионный анализ, основанный на МНК, давал наилучшие из всех возможных результаты, должны выполняться определенные условия относительно возмущения ε -

основные предпосылки регрессионного анализа.

Основные предпосылки регрессионного анализа:
(условия Гаусса- Маркова)

1. *Математическое ожидание возмущения* $\boldsymbol{\varepsilon_i}$ *в любом наблюдении должно быть равно нулю, т.е.*

$$M(\varepsilon_i)=0,$$

$$i = 1, 2, \ldots, n$$

2. *Дисперсия возмущения* $\boldsymbol{\varepsilon_i}$ *должна быть постоянной для всех наблюдений:*

$$D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$$

3. *Возмущения* $\boldsymbol{\varepsilon_i}$ *и* $\boldsymbol{\varepsilon_j}$ *должны быть некоррелированы между собой.*

4. *В модели (5) возмущение* $\boldsymbol{\varepsilon_i}$ *есть величина случайная,*

а объясняющая переменная $\mathbf{x_i}$ **-** *величина неслучайная.*

При выполнении основных предпосылок регрессионного анализа модель (5) называется **классической нормальной линейной регрессионной моделью.**

Наряду с условиями 1 - 4 обычно предполагается, что возмущение ε_i имеет нормальное распределение.

Оценкой модели (5) по выборке является уравнение регрессии:

$$\hat{y} = a + bx$$

Параметры этого уравнения определяются

по МНК.

Выборочной оценкой возмущения $\boldsymbol{\varepsilon_i}$

является отклонение $\boldsymbol{e_i}$

значения $\boldsymbol{y_i}$ переменной Y от значения $\boldsymbol{\hat{y}_i}$,

найденного по уравнению регрессии (4), т. е.

$$e_i = y_i - \hat{y}_i$$

Являются ли оценки ***a*** и ***b*** параметров α и β «наилучшими»?

ТЕОРЕМА ГАУССА- МАРКОВА:

Если основные предпосылки регрессионного анализа выполняются (условия 1 - 4), то оценки (a, b), сделанные с помощью МНК, являются наилучшими линейными несмещенными оценками, т.е. обладают свойствами несмещенности, эффективности и состоятельности.

1) **несмещенность**: $M(\boldsymbol{a})=\boldsymbol{\alpha}$; $M(\boldsymbol{b})=\boldsymbol{\beta}$

2) **эффективность**: имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

3) **состоятельность**: при достаточно большом $\boldsymbol{n}$ оценки $\boldsymbol{a}$ и $\boldsymbol{b}$ близки к ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$);

(при увеличении объема выборки надежность оценок увеличивается).