Лекция № 13

тема: **Парная линейная регрессия. Метод наименьших квадратов**

План лекции:

- 1. Парная регрессия
- 2. Парная линейная регрессия
- 3. Метод наименьших квадратов
- 4. Линейный коэффициент корреляции
- 5. Линейный коэффициент детерминации
- 6. Средняя ошибка аппроксимации

1. Парная регрессия

Парная регрессия представляет собой регрессию между двумя переменными – y и x, т. е. модель вида:

$$y = \hat{f}(x)$$

где y – зависимая переменная (результативный признак);
x – независимая, или объясняющая, переменная (признак-фактор).

Знак «^» означает, что между переменными x и y нет строгой функциональной зависимости, поэтому практически в каждом случае величина y складывается из двух слагаемых:

$$y = \hat{y}_x + \varepsilon,$$

где y – фактическое значение результативного признака;

$\hat{y}_x$ – теоретическое значение результативного признака, найденное из уравнения регрессии;

ε – случайная величина, характеризующая отклонения реального значения результативного признака от теоретического, найденного по уравнению регрессии.

Случайная величина **ε** называется также **возмущением**.

Она включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения.

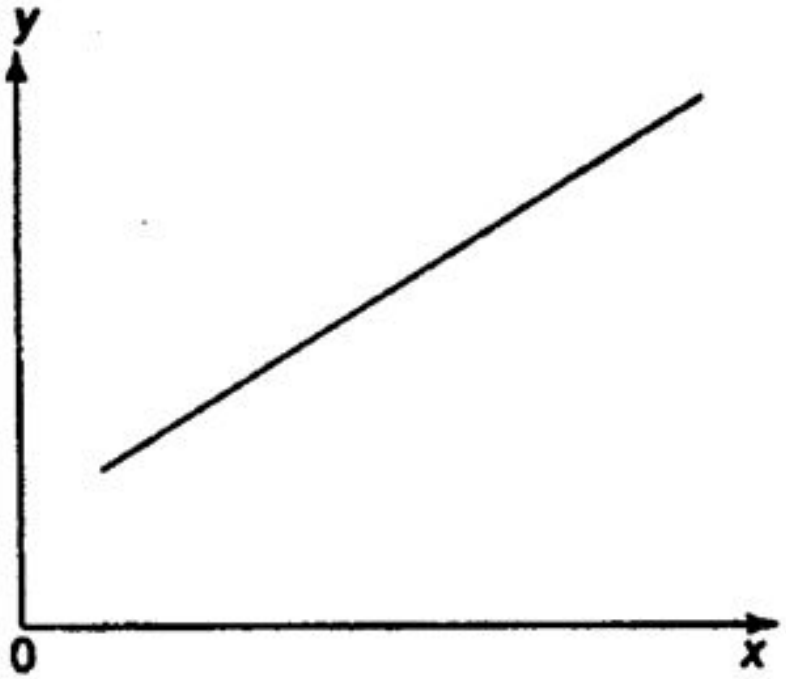
Ее присутствие в модели порождено тремя источниками: спецификацией модели, выборочным характером исходных данных, особенностями измерения переменных.

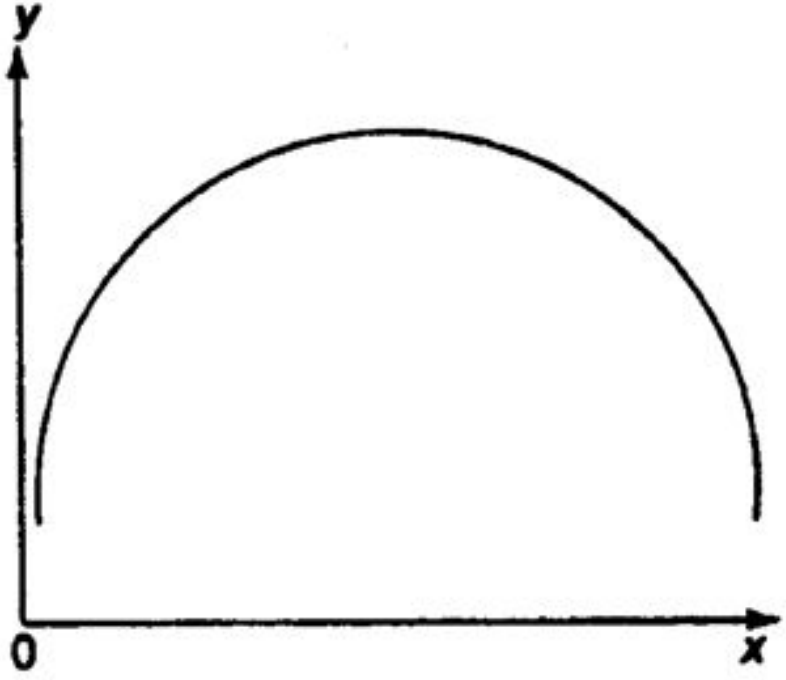
В парной регрессии выбор вида математической функции $\hat{y}_x = f(x)$ может быть осуществлен тремя методами:

1) графическим;
2) аналитическим, т.е. исходя из теории изучаемой взаимосвязи;
3) экспериментальным.

При изучении зависимости между двумя признаками графический метод подбора вида уравнения регрессии достаточно нагляден.

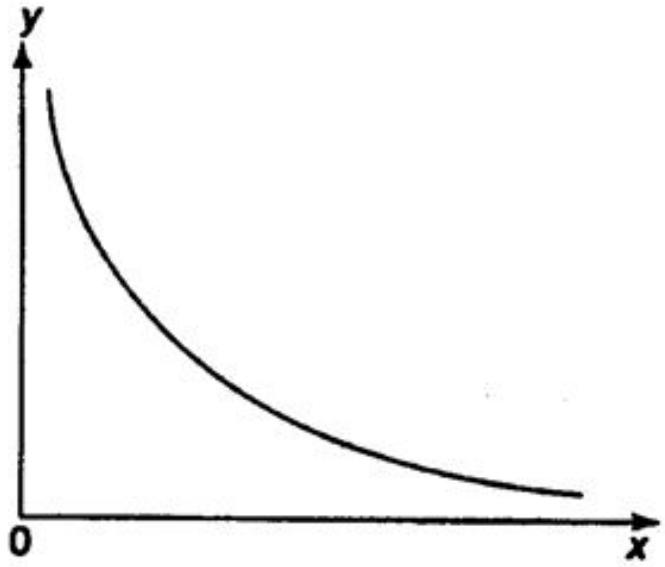
Он основан на поле корреляции. Основные типы кривых, используемые при количественной оценке связей, представлены на рисунках:

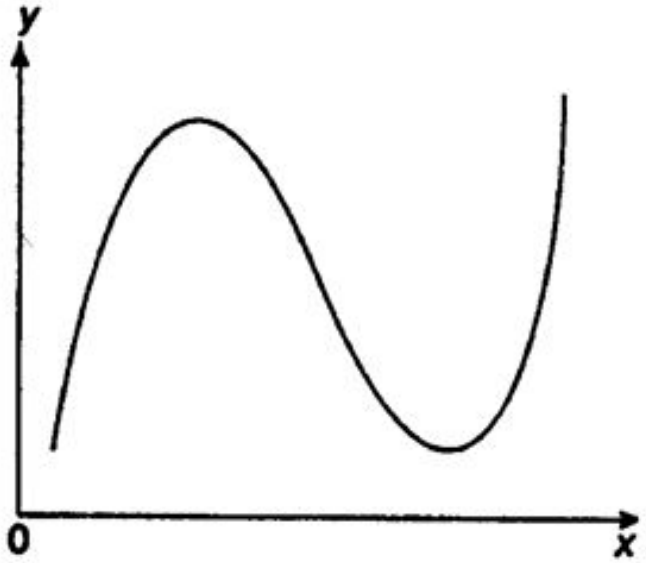




$$\hat{y}_x = a + b \cdot x$$

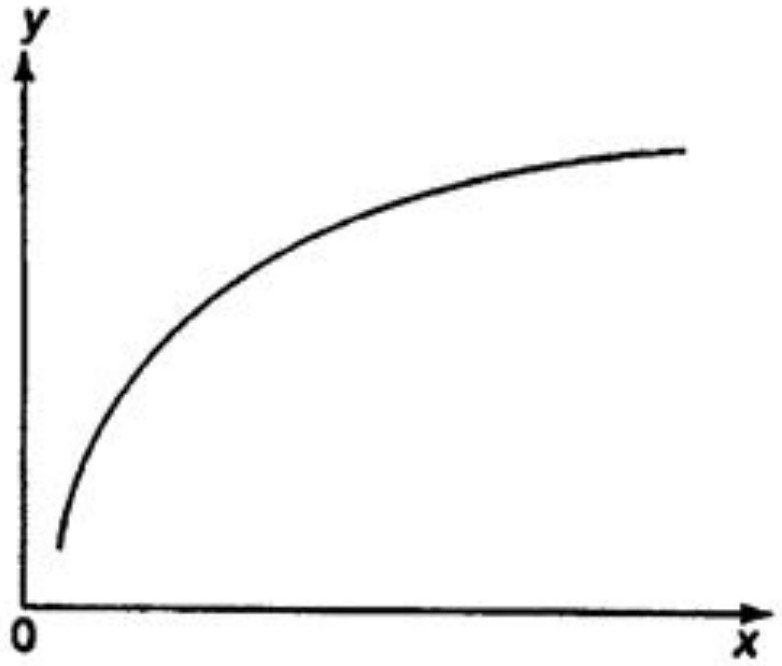
$$\hat{y}_x = a + b \cdot x + c \cdot x^2$$

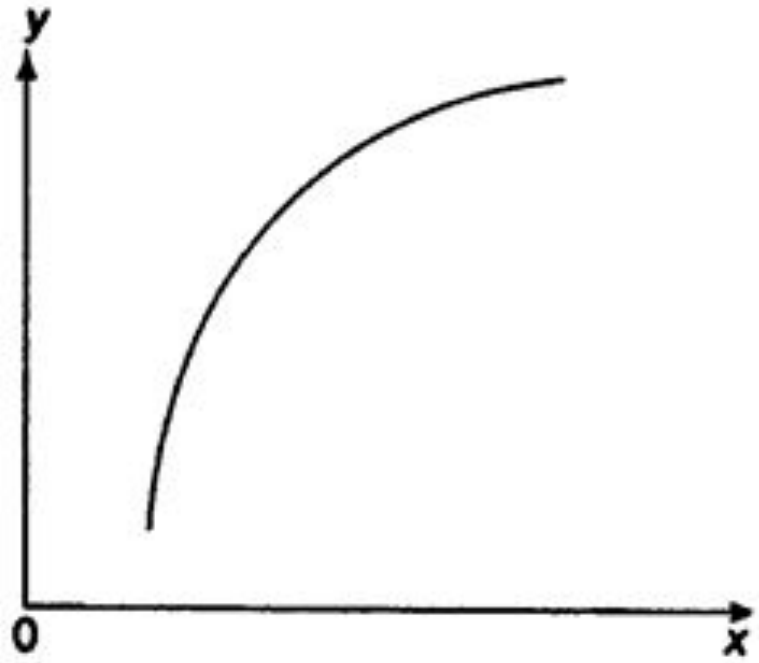




$$\hat{y}_x = a + b/x$$

$$\hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$





$$\hat{y}_x = a \cdot x^b$$

$$\hat{y}_x = a \cdot b^x$$

Считается, что число наблюдений должно в **7-8** раз превышать число рассчитываемых параметров при переменной .

Это означает, что искать линейную регрессию, имея менее 7 наблюдений, вообще не имеет смысла. Если вид функции усложняется, то требуется увеличение объема наблюдений, ибо каждый параметр при должен рассчитываться хотя бы по 7 наблюдениям.

2. Парная линейная регрессия

Рассмотрим простейшую модель парной регрессии – **линейную регрессию**. Линейная регрессия находит широкое применение в эконометрике ввиду четкой экономической интерпретации ее параметров.

Линейная регрессия сводится к нахождению уравнения вида $y_x = a + b \cdot x$

$$y = a + b \cdot x + \varepsilon$$

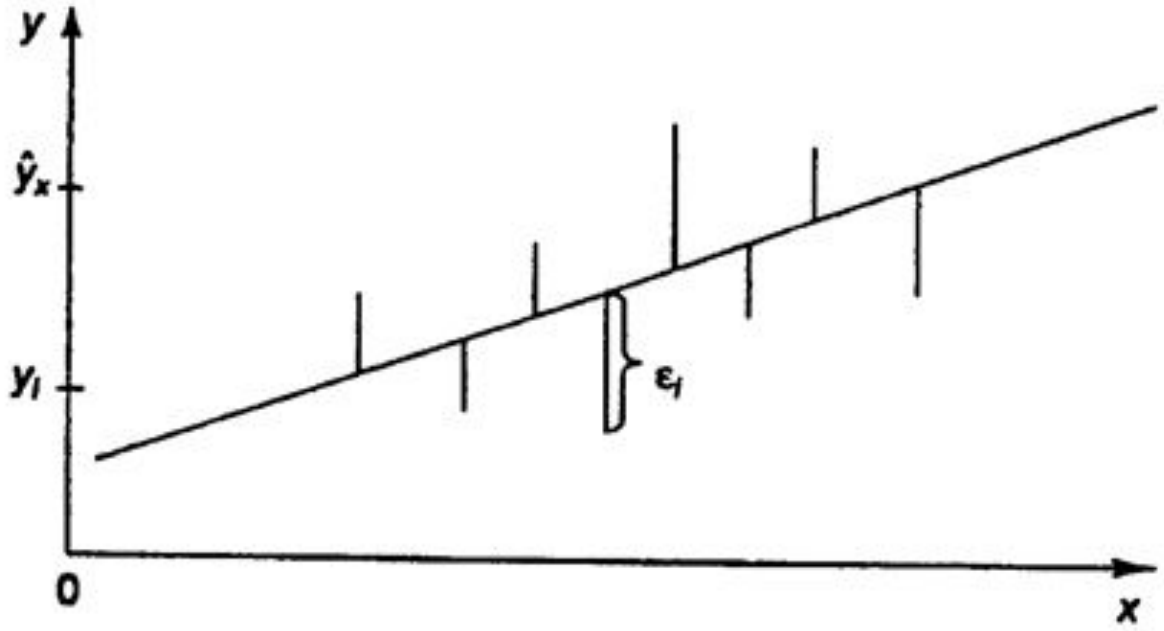
Уравнение вида $\hat{y}_x = a + b \cdot x$ позволяет по заданным значениям фактора x находить теоретические значения результативного признака, подставляя в него фактические значения фактора x.

3. Метод наименьших квадратов

Построение линейной регрессии сводится к оценке ее параметров – a и b. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических $\hat{y}_x$ минимальна:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_{x_i}\right)^2 = \sum_{i=1}^{n}\varepsilon_i^2 \rightarrow \min .$$

Т.е. из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной:



После несложных преобразований, получим следующую систему линейных уравнений для оценки параметров a и b:

$$\begin{cases} a \cdot n + b \cdot \sum x = \sum y; \\ a \cdot \sum x + b \cdot \sum x^2 = \sum x \cdot y. \end{cases}$$

Можно воспользоваться следующими готовыми формулами, которые следуют непосредственно из решения системы:

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

4. Линейный коэффициент корреляции

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает ***линейный коэффициент корреляции,*** который можно рассчитать по следующим формулам:

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

- Линейный коэффициент корреляции находится в пределах: $-1 \le r_{xy} \le 1$
- Чем ближе абсолютное значение к единице, тем сильнее линейная связь между факторами (при $r_{xy} = \pm 1$ имеем строгую функциональную зависимость). Но следует иметь в виду, что близость абсолютной величины линейного коэффициента корреляции к нулю еще не означает отсутствия связи между признаками. При другой (нелинейной) спецификации модели связь между признаками может оказаться достаточно тесной.

Значение r_{xy}	Оценка связи
$r_{xy} < 0$	Обратная
$0 < r_{xy} < 0,1$	Отсутствует
$0,1 <= r_{xy} < 0,3$	Слабая
$0,3 <= r_{xy} < 0,5$	Умеренная
$0,5 <= r_{xy} < 0,7$	Заметная
$0,7 <= r_{xy} < 0,9$	Сильная
$0,9 <= r_{xy} <= 0,99$	Весьма сильная
$0,99 < r_{xy} <= 1$	Полная функциональная

5. Линейный коэффициент детерминации

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции , называемый коэффициентом детерминации r_{xy}^2

Коэффициент детерминации характеризует долю дисперсии результативного признака у, объясняемую регрессией, в общей дисперсии результативного признака

$$r_{xy}^2 = 1 - \frac{\sigma_{ост}^2}{\sigma_y^2},$$

где $\sigma_{ост}^2 = \frac{1}{n}\sum\left(y - \hat{y}_x\right)^2$, $\sigma_y^2 = \frac{1}{n}\sum\left(y - \bar{y}\right)^2 = \overline{y^2} - \bar{y}^2$.

6. Средняя ошибка аппроксимации

После того как найдено уравнение линейной регрессии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров

Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации.

Средняя ошибка аппроксимации не должна превышать 8–10%.

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