



Регрессионный анализ

- **Регрессионный анализ** – это статистический метод исследования зависимости величины Y от величин X_j ($j = \overline{1, k}$).

Задачи регрессионного анализа:

- установление формы зависимости между переменными (***спецификация***),
- определение параметров выбранного уравнения (***параметризация***),
- анализ качества уравнения (***верификация***) и проверка адекватности уравнения эмпирическим данным,
- определение неизвестных значений (***прогноз значений***).

Если каждому значению X соответствует свое значение $M(Y|X)$, то зависимость

$$M(Y | X) = f(X)$$

называется **функцией регрессии Y на X** .

При этом X называется **экзогенной**, Y – **эндогенной**.

При рассмотрении зависимости

- двух переменных говорят о **парной регрессии**:

$$M(Y | X) = f(X)$$

- нескольких переменных говорят о **множественной регрессии**

$$M(Y | X_1, X_2, \dots, X_k) = f(X_1, X_2, \dots, X_k).$$

Реальные значения Y не всегда совпадают с $M(Y | X)$. Поэтому фактическая зависимость дополняется случайной величиной ε .

Статистическую модель вида:

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

или

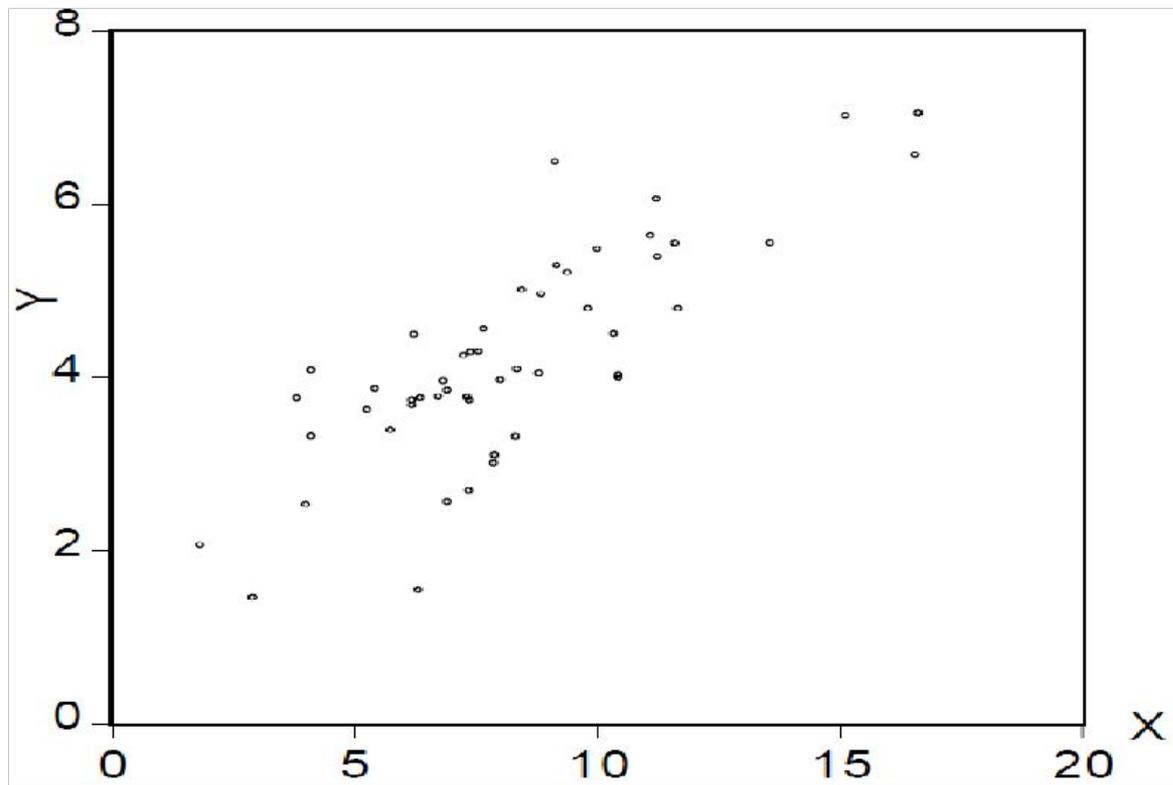
$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon$$

называют *регрессионными моделями (уравнениями)*.

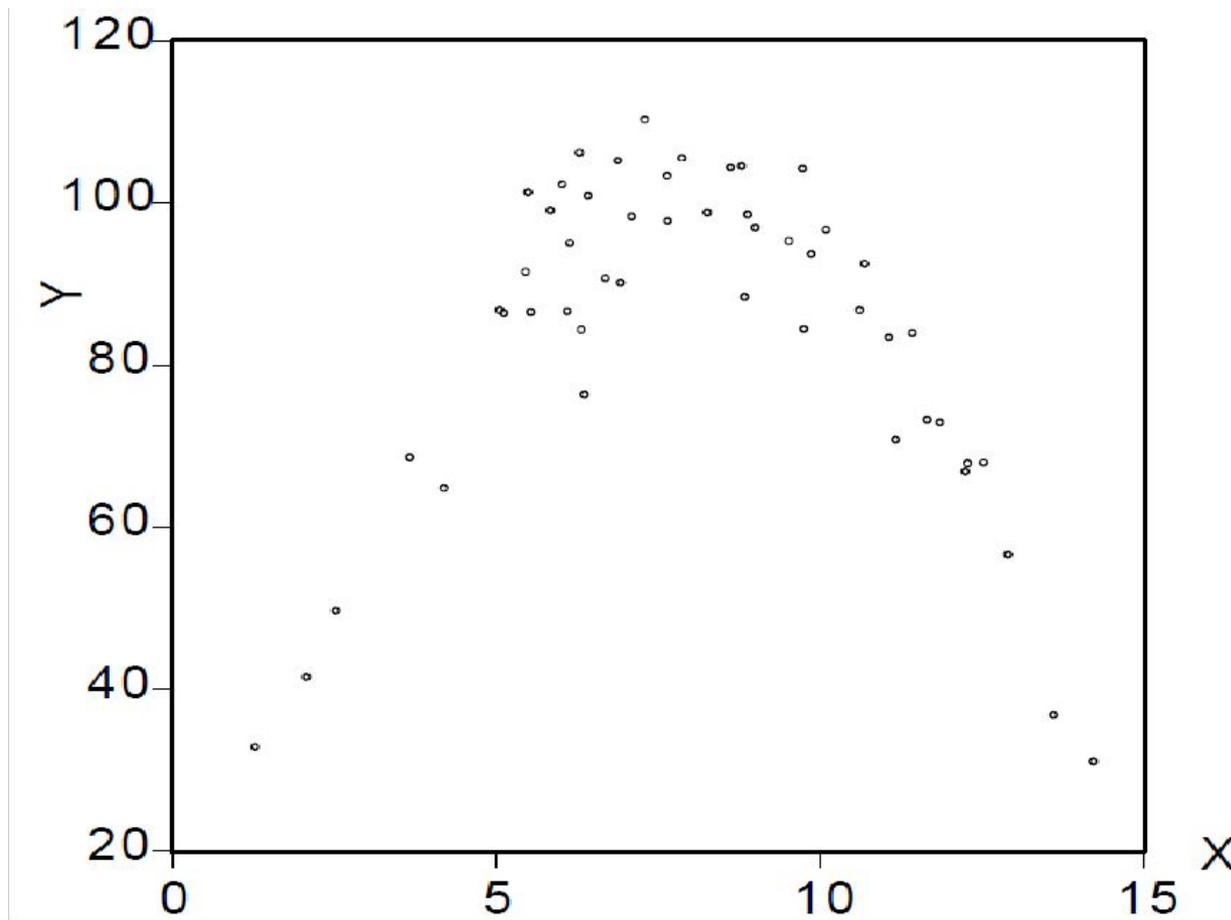
В зависимости от вида функции $f(X)$ модели делятся на *линейные* и *нелинейные*.

Спецификация уравнения регрессии.

В случае парной регрессии – графический анализ реальных статистических данных (наблюдений).

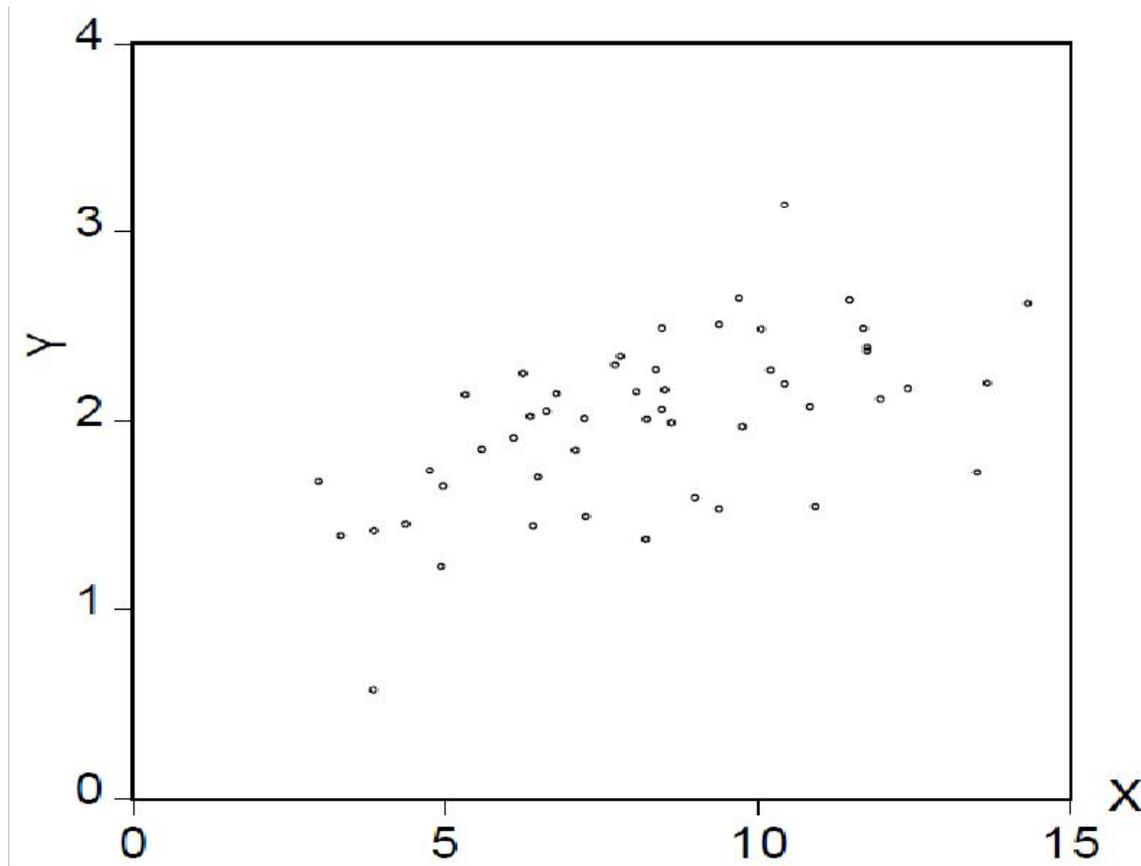


Линейная зависимость $\hat{Y} = \beta_0 + \beta_1 X$.



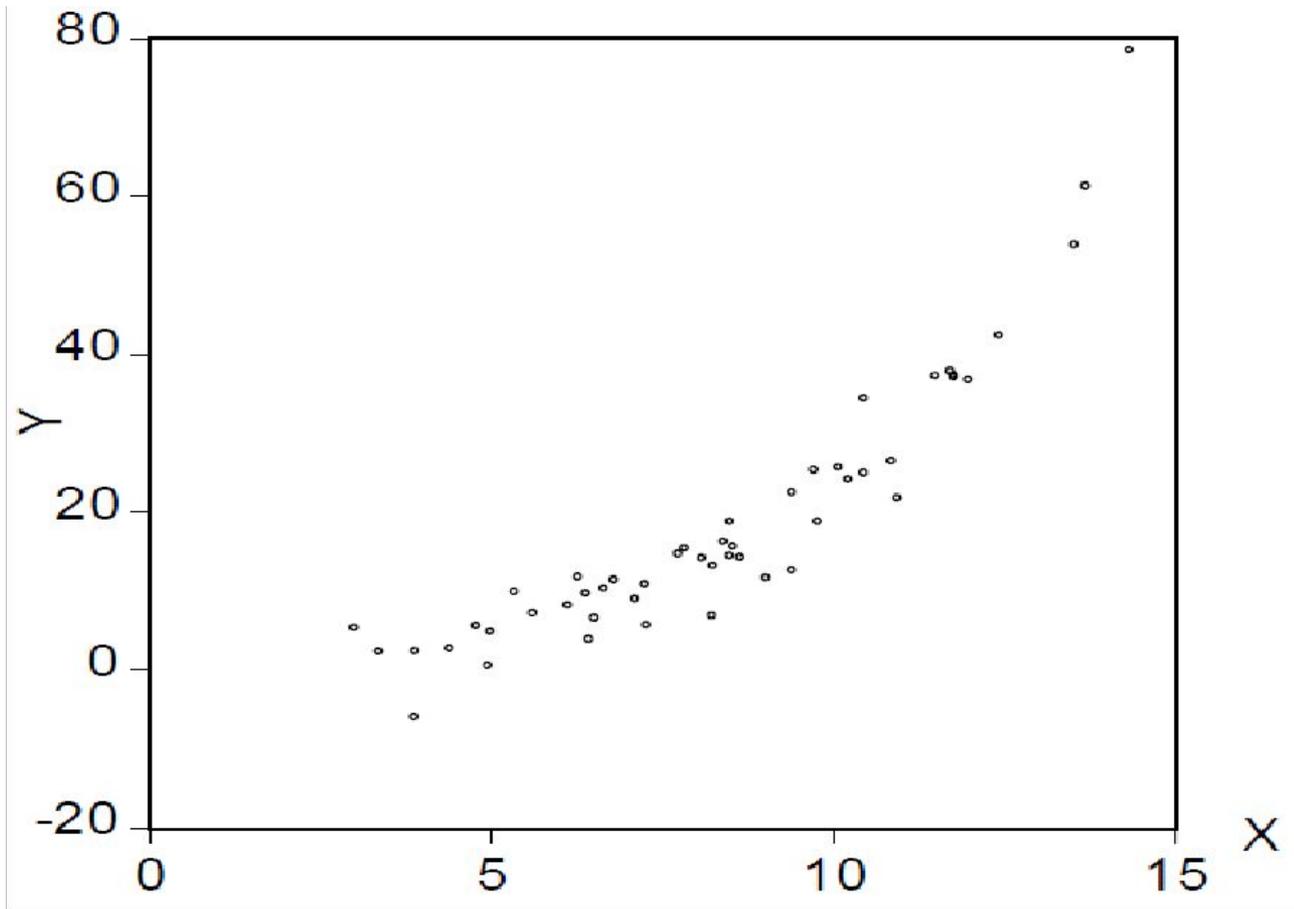
Квадратичная зависимость:

$$\hat{Y} = \beta_0 + \beta_1 X + \beta_2 X^2$$



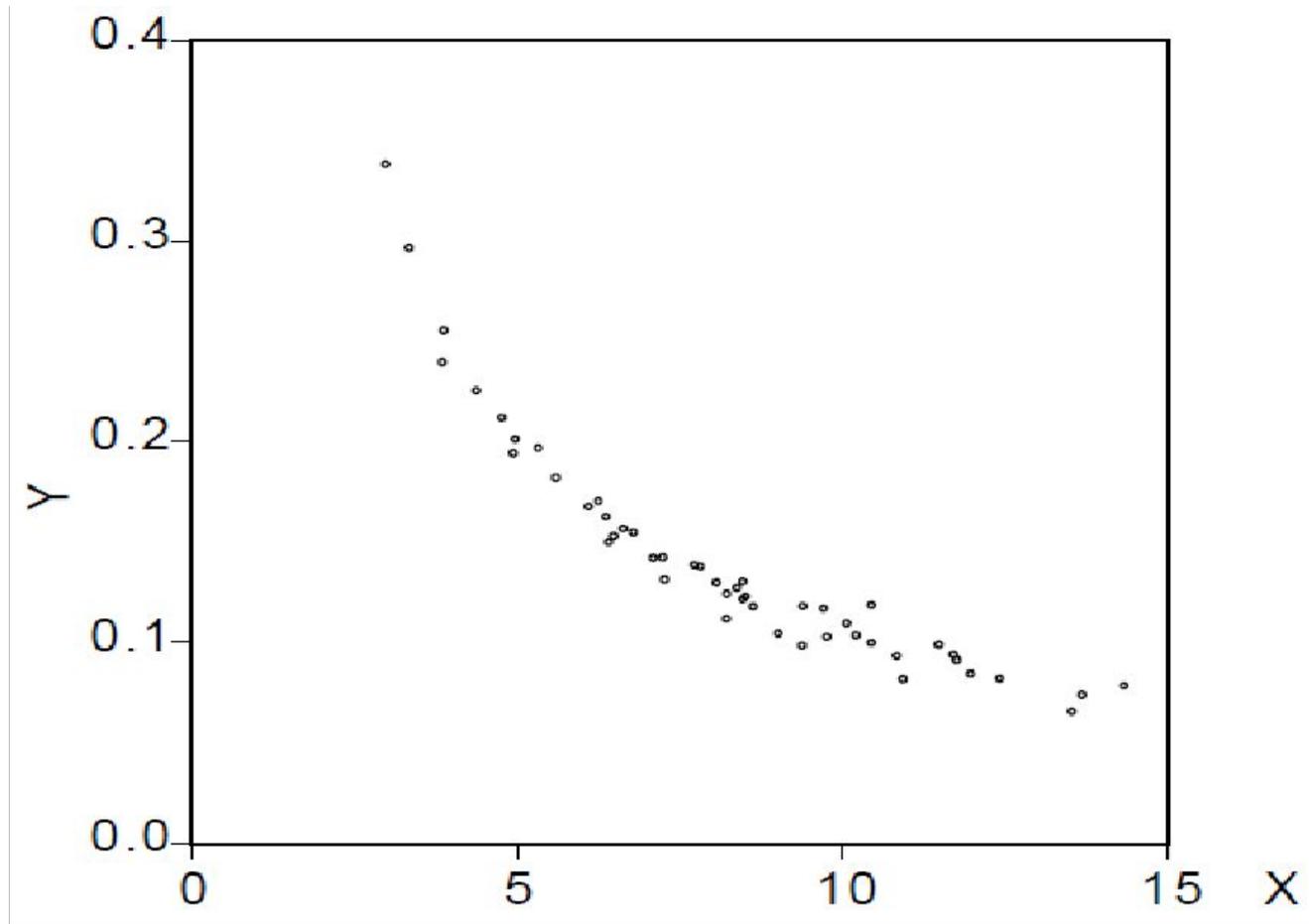
Степенная зависимость

$$\hat{Y} = \beta_0 X^{\beta_1}$$

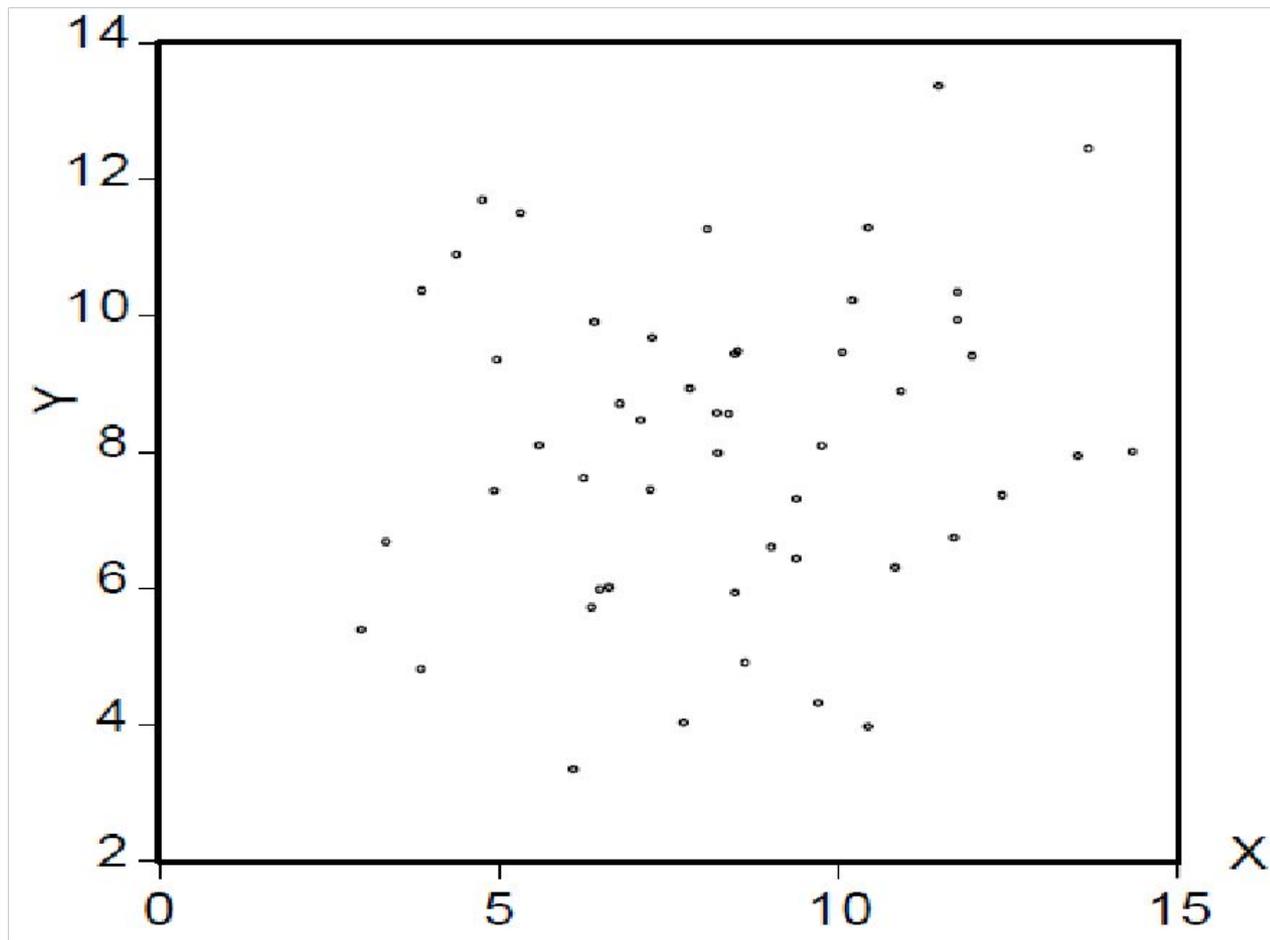


Показательная зависимость

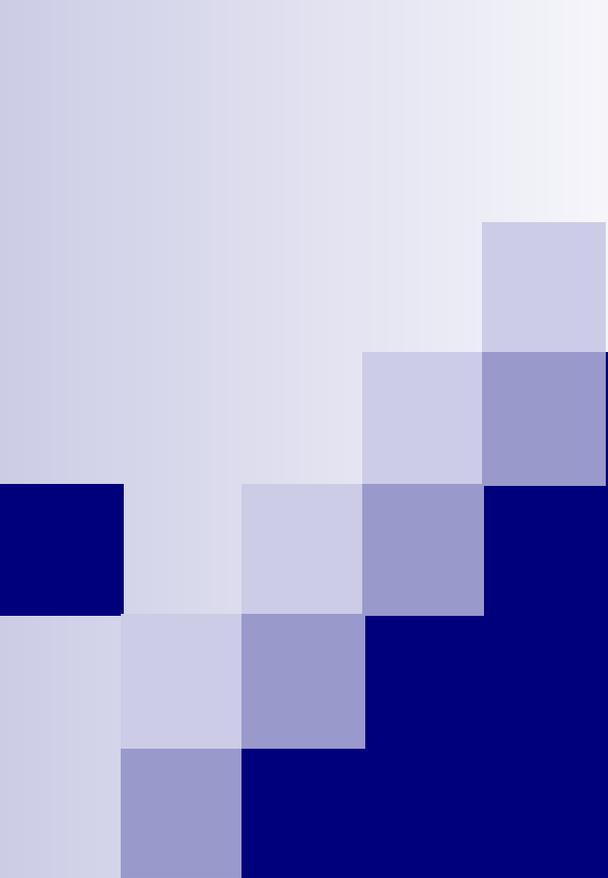
$$\hat{Y} = \beta_0 e^{\beta_1 X}$$



Гиперболическая зависимость $\hat{Y} = \beta_0 + \frac{\beta_1}{X}$



X и Y независимы



*Классическая модель
парной линейной регрессии.*

Общий вид модели *парной линейной регрессии*:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \text{ где}$$

β_0 – свободный член уравнения (*среднее значение Y при условии, что $X=0$*),

β_1 – коэффициент регрессии, характеризует изменение среднего значения переменной Y , при изменении значения X на единицу своего измерения:

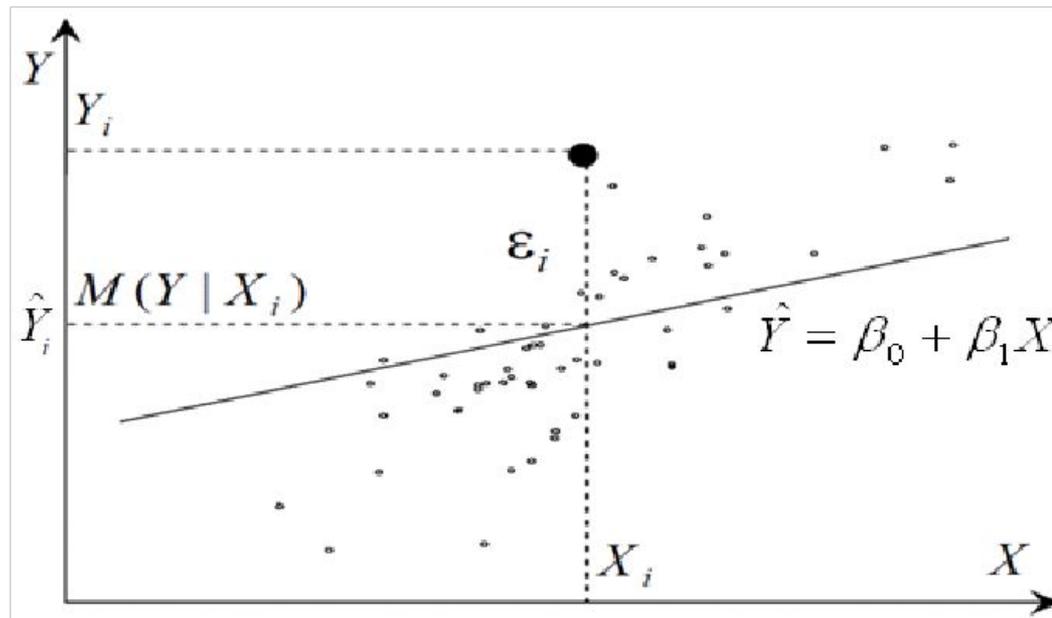
- если $\beta_1 > 0$ – переменные X и Y положительно коррелированные,
- если $\beta_1 < 0$ – отрицательно коррелированы.

ε_i – случайная составляющая.

Выборка: (x_i, y_i) – результат i -го наблюдения.

Для каждого наблюдения *модель парной линейной регрессии*:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i .$$



Выборочная линия регрессии

$$\hat{y} = b_0 + b_1 x , \text{ где}$$

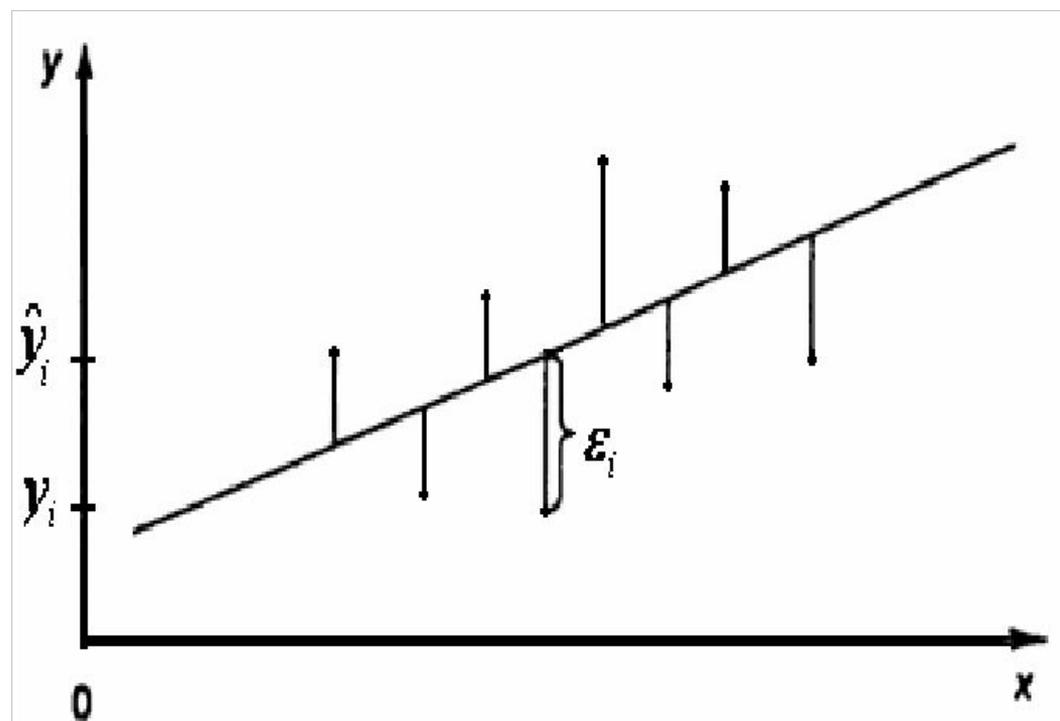
b_0 и b_1 – оценки параметров β_0 и β_1 .



**КЛАССИЧЕСКИЙ (ОБЫЧНЫЙ)
МЕТОД НАИМЕНЬШИХ КВАДРАТОВ
(МНК)**

Суть метода состоит в минимизации суммы квадратов отклонений фактических значений результатного признака от его расчетных значений, т.е. y_i от \hat{y}_i :

$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \rightarrow \min$$



Найдем частные производные Q и приравняем их к нулю:

$$\begin{cases} \frac{\partial Q}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0; \\ \frac{\partial Q}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i = 0. \end{cases}$$

Получим *систему нормальных уравнений*:

$$\begin{cases} nb_0 + b_1 \sum x_i = \sum y_i \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i \end{cases} \quad \text{или} \quad \begin{cases} b_0 + b_1 \bar{x} = \bar{y} \\ b_0 \bar{x} + b_1 \overline{x^2} = \overline{xy} \end{cases}$$

Решая систему, получаем:

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

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\text{cov}(X, Y)}{s_x^2},$$

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

$$\bar{y} = \frac{\sum y_i}{n}; \quad \overline{xy} = \frac{\sum x_i y_i}{n}.$$

По полученному уравнению регрессии

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

получают *расчетные (прогнозные)* значения переменной y для каждого i наблюдения, т.е. $\hat{y}_i(x_i)$.

Величина b_1 – *выборочный коэффициент регрессии* Y по X , который показывает, на сколько единиц в среднем изменяется переменная Y при увеличении переменной X на одну единицу.



МАТРИЧНАЯ ФОРМА ЗАПИСИ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Матричная форма модели:

$$Y = X\beta + \varepsilon, \text{ где}$$

$$Y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} \text{ – вектор значений зависимой;}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{ – вектор неизвестных параметров;}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{bmatrix} \text{ – вектор случайных ошибок.}$$

$X = \begin{bmatrix} 1 & x_1 \\ \dots & \dots \\ 1 & x_n \end{bmatrix}$ - матрица значений независимых переменных размерности.

Оценка модели по выборке

$$\hat{Y} = Xb, \text{ где}$$

$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$ - вектор оценок неизвестных параметров.

Решение в матричной форме:

$$b = (X^T X)^{-1} X^T Y.$$



ОСНОВНЫЕ ПРЕДПОСЫЛКИ МНК

Условия Гаусса – Маркова.

1. ε_i ($i = \overline{1, n}$) (или y_i) есть величина случайная, а объясняющая переменная x_i – величина неслучайная: $\text{cov}(\varepsilon_i, X_i) = 0$.
2. $M(\varepsilon_i) = 0$ для всех наблюдений Y .
3. $D(\varepsilon_i) = \sigma^2 = \text{const}$ для всех наблюдений Y .

Это условие называется условием *гомоскедастичности*.

В матричной форме:

$$D(\varepsilon_i) = \sigma^2 E_n, \text{ где}$$

E_n — единичная матрица n -го порядка.

4. ε_i и ε_j независимы в любых двух наблюдениях:
 $\text{cov}(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$, т.е. отклонения регрессии
не коррелируют:

$$\text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j. \end{cases}$$

Матричная форма записи предпосылки:

$$\text{cov}(\varepsilon\varepsilon^T) = \sigma^2 E_n, \text{ где}$$

E_n – единичная матрица n -го порядка, а $\text{cov}(\varepsilon\varepsilon^T)$ –
ковариационная матрица возмущений

$$\text{cov}(\varepsilon\varepsilon^T) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}.$$

Модель парной линейной регрессии, построенная с учетом условий Гаусса–Маркова называется *классической регрессионной моделью*.

Если с условиями Гаусса – Маркова также предполагается нормальность распределения случайного члена:

$$\varepsilon_i \sim N(0; \sigma^2)$$

(если ε – вектор возмущений, то $\varepsilon \sim N(0; \sigma^2 E_n)$),

то модель называется *классической нормальной регрессионной моделью*.

СВОЙСТВА ОЦЕНОК МНК

Несмещенность оценки означает, что $M(\varepsilon) = 0$.

Вектор b – несмещенная оценка вектора β : $M(b) = \beta$.

Оценки считаются **эффективными**, если они характеризуются наименьшей дисперсией.

Вектор b – наиболее эффективная оценка вектора β , т.е. обладает наименьшей дисперсией:

$$\sigma^2(b) \rightarrow \min.$$

Состоятельность оценок характеризует увеличение их точности с увеличением объема выборки.

Вектор b – состоятельная оценка вектора β :

$$\lim_{n \rightarrow \infty} b_j = \beta_j.$$



ОЦЕНКА КАЧЕСТВА (ВЕРИФИКАЦИЯ) МОДЕЛИ

Качество модели регрессии связывают с адекватностью (или соответствия) модели эмпирическим данным.

Проверка адекватности модели регрессии – на основе анализа остатков - e_i .

Качество модели регрессии оценивается по следующим направлениям:

- 1) проверка общего качества уравнения регрессии;
- 2) проверка значимости уравнения регрессии;
- 3) проверка статистической значимости коэффициентов уравнения регрессии;
- 4) проверка выполнения предпосылок МНК.



ПРОВЕРКА ОБЩЕГО КАЧЕСТВА УРАВНЕНИЯ РЕГРЕССИИ

Вычисляют коэффициенты, по которым делаются выводы об ее адекватности и точности.

1. Качество парной линейной регрессии определяется с помощью **выборочного коэффициента парной линейной корреляции** – показателя близости наблюдений к линейной регрессии:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - (\bar{x})^2} \cdot \sqrt{\overline{y^2} - (\bar{y})^2}} = \frac{\text{cov}(X, Y)}{s_x \cdot s_y} .$$

2. Коэффициент детерминации – наиболее эффективная оценка адекватности регрессионной модели:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{Q_{ост}}{Q_{общ}}, \text{ где}$$

$Q_{ост} = \sum (y_i - \hat{y}_i)^2$ – сумма квадратов остатков.

$Q_{общ} = \sum (y_i - \bar{y})^2$ – общая сумма квадратов.

R^2 показывает на сколько процентов вариация результативного признака Y учтена в модели и обусловлена влиянием на него фактора X в общем объеме вариации.

Свойства коэффициента детерминации.

1. $0 \leq R^2 \leq 1$.
2. $R^2 = 0$ – вывод о независимости Y и X .
3. $R^2 = 1$ – вывод о наличии функциональной линейной зависимости между переменными Y и X .
4. $0 < R^2 < 1$ – чем ближе R^2 к 1, тем лучше качество подгонки кривой к нашим данным, тем точнее Y .

3. Для оценки точности прогноза используются характеристики: несмещенная оценка остаточной дисперсии, стандартная ошибка остатков и средняя относительная ошибка аппроксимации.

Несмещенная оценка остаточной дисперсии:

$$\hat{S}^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2 = \frac{Q_{ост}}{n-2},$$

$Q_{ост} = \sum (y_i - \hat{y}_i)^2$ – сумма квадратов остатков.

Величину $S = \sqrt{\hat{S}^2}$ называют *стандартной ошибкой* остатков.

Чем меньше значения этих характеристик, тем выше точность модели.

Средняя относительная ошибка аппроксимации – среднее относительное отклонение расчетных значений зависимой переменной \hat{y}_i от фактических значений y_i :

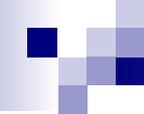
$$\bar{\Delta} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%$$

Если средняя ошибка аппроксимации составляет менее 6–7%, то качество модели считается хорошим.

Максимально допустимым значением данного показателя считается 12-15%.



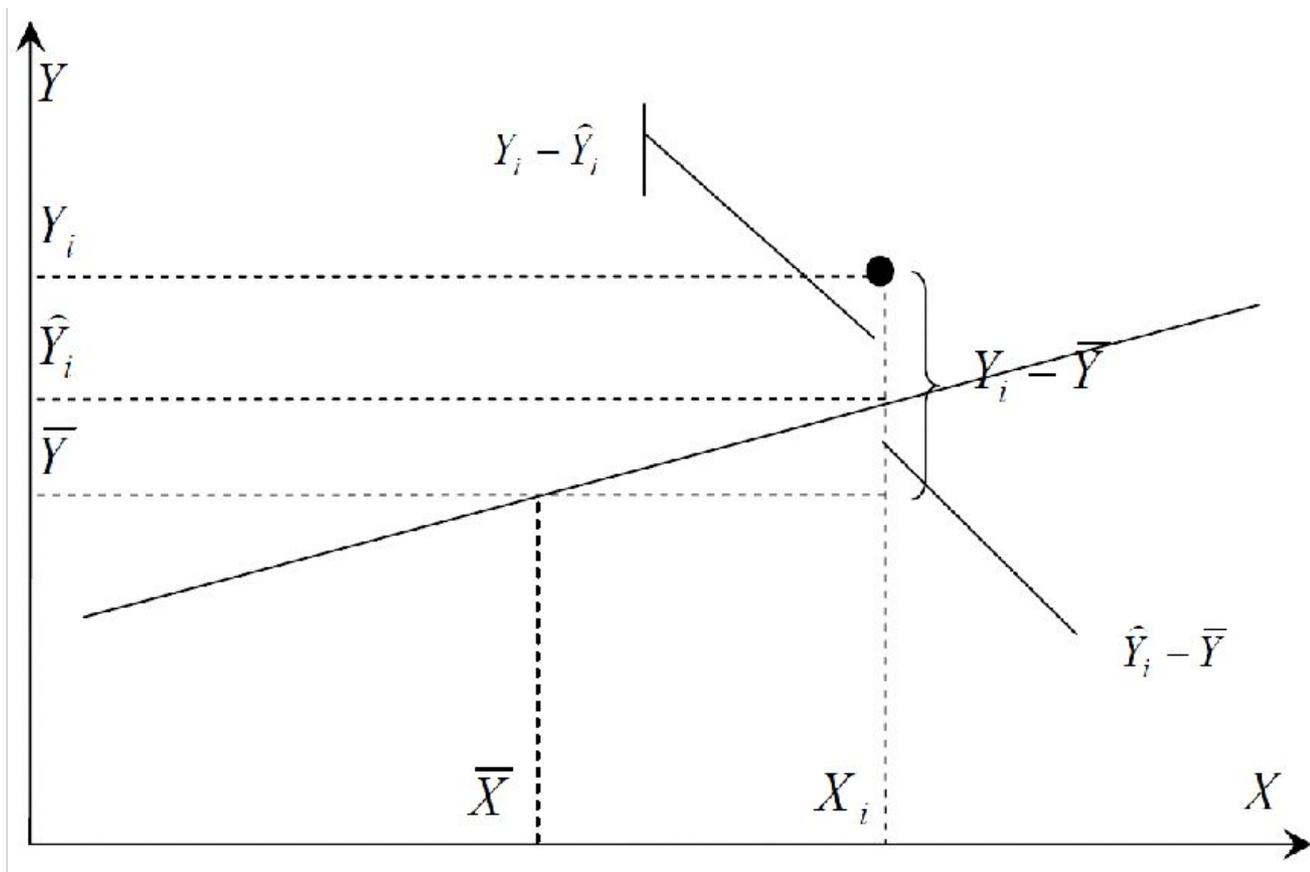
ПРОВЕРКА ЗНАЧИМОСТИ УРАВНЕНИЯ РЕГРЕССИИ



Проверить значимость уравнения регрессии –
установить:

- соответствует ли модель исходным данным и
- достаточно ли включенных в уравнение объясняющих переменных.

Проверка значимости уравнения регрессии происходит на основе дисперсионного анализа.



Основное положение дисперсионного анализа

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2, \text{ или}$$

$$Q_{\text{общ}} = Q_{\text{факт}} + Q_{\text{ост}}$$

СХЕМА ДИСПЕРСИОННОГО АНАЛИЗА

(n – число наблюдений, k – число объясняющих переменных).

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$Q_{общ} = \sum (y_i - \bar{y})^2$	$n - 1$	$S_{общ}^2 = \frac{Q_{общ}}{n - 1}$
Факторная (объясненная регрессией)	$Q_{факт} = \sum (\hat{y}_i - \bar{y})^2$	k	$S_{факт}^2 = \frac{Q_{факт}}{k}$
Остаточная	$Q_{ост} = \sum (y_i - \hat{y}_i)^2$	$n - k - 1$	$S_{ост}^2 = \frac{Q_{ост}}{n - k - 1}$

Выдвигают гипотезу о не значимости уравнения в целом, которая формально сводится к гипотезе о равенстве нулю параметров регрессии:

$$H_0 : \beta_1 = 0.$$

Альтернативная ей гипотеза о значимости уравнения – гипотеза о неравенстве нулю параметров регрессии:

$$H_1 : \beta_1 \neq 0.$$

Значимость уравнения проверяют с помощью F -критерия Фишера:

$$F_{набл} = \frac{S_{регр}^2}{S_{ост}^2} = \frac{Q_{регр}/k}{Q_{ост}/(n-k-1)} = \frac{Q_{регр} \cdot (n-2)}{Q_{ост}}, \text{ где}$$

n – число выборочных наблюдений, k – число объясняющих переменных.

Если $F_{набл} > F_{кр}(\alpha; \nu_1=k=1, \nu_2=n-2)$, то гипотеза отвергается и уравнение считается значимым.

R^2 также применяется для проверки значимости уравнения регрессии.

$$H_0: R^2 = 0.$$

$$H_1: R^2 \neq 0.$$

Для этого рассчитывают статистику:

$$F = \frac{R^2}{1 - R^2} (n - 2).$$

Если $F_{набл} > F_{кр}$, то гипотеза отвергается и уравнение считается значимым.



**ПРОВЕРКА ЗНАЧИМОСТИ
КОЭФФИЦИЕНТОВ
УРАВНЕНИЯ РЕГРЕССИИ**

Коэффициент называется **значимым**, если есть достаточно высокая вероятность того, что его истинное значение отлично от нуля.

$$H_0: \beta_j = 0.$$

Для проверки гипотезы рассчитывают:

$$t_{\text{набл}_j} = \frac{b_j}{\hat{S}_{b_j}}, \text{ где}$$

$$\hat{S}_{b_0} = \sqrt{\frac{\hat{S}^2 \sum x_i^2}{n \sum_{i=1}^n x_i^2 - (\sum x_i)^2}} = S \frac{\sqrt{\sum x_i^2}}{n \cdot S_x}$$

$$\hat{S}_{b_1} = \sqrt{\frac{\hat{S}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = S \frac{\sqrt{n}}{n \cdot S_x}.$$

Если использовать матричную форму записи, то:

$\hat{S}_{b_j}^2 = \hat{S}^2 [(X^T X)^{-1}]_{jj}$ – дисперсия коэффициента регрессии b_j ;

\hat{S}^2 – несмещенная оценка остаточной дисперсии;

$[(X^T X)^{-1}]_{jj}$ – элементы обратной матрицы, стоящие на главной диагонали;

\hat{S}_{b_j} – стандартная ошибка коэффициента b_j .

Если $|t_{набл}| > t_{кр}(\alpha; \nu = n - 2)$, то гипотеза H_0 отвергается и коэффициент считается значимым.

Если $|t_{набл}| \leq t_{кр}$, то гипотеза H_0 не отвергается.



ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ КОЭФФИЦИЕНТОВ РЕГРЕССИИ

Доверительным интервалом называется интервал, относительно которого можно с заранее выбранной вероятностью утверждать, что он содержит значения прогнозируемого показателя.

Интервальная оценка для параметра β_0 :

$$\beta_0 \in \left(b_0 \pm t_{\alpha, n-2} \cdot \hat{S}_{b_0} \right), \text{ где}$$

$t_{кр}(\alpha; \nu=n-2)$ определяется из *таблицы распределения Стьюдента для двусторонней критической области для уровня значимости α* и числа степеней свободы $\nu=n-2$.

Аналогично определяется интервальная оценка для коэффициента β_1 :

$$\beta_1 \in \left(b_1 \pm t_{\alpha, n-2} \cdot \hat{S}_{b_1} \right)_1.$$



**ПРОГНОЗИРОВАНИЕ С
ПРИМЕНЕНИЕМ УРАВНЕНИЯ
РЕГРЕССИИ**

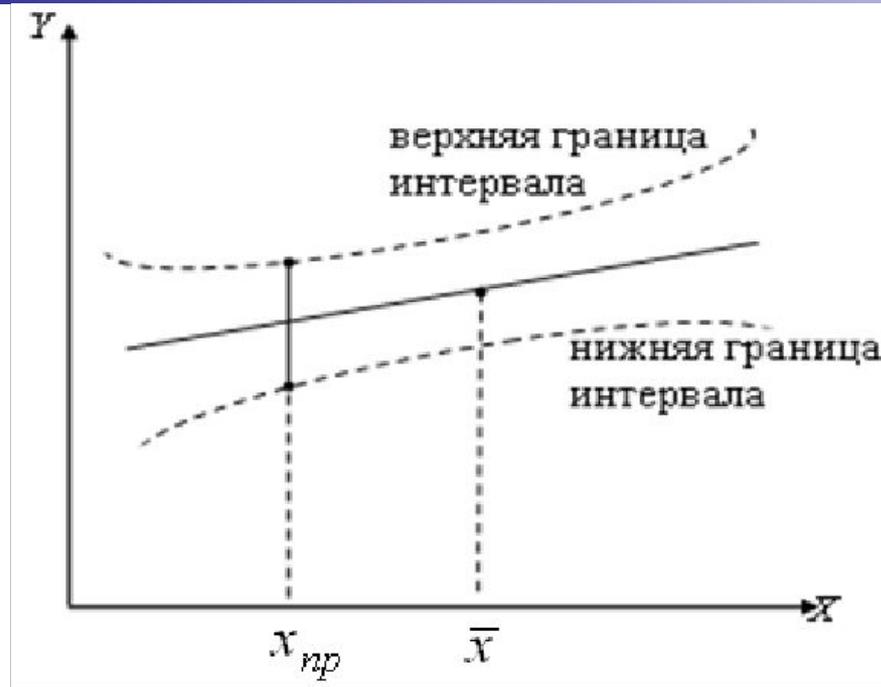
Регрессионные модели могут быть использованы для прогнозирования результативной переменной Y :

$$\hat{y}_{np} = b_0 + b_1 x_{np}.$$

Данный прогноз называется ***точечным***.

Интервальная оценка для уравнения регрессии \hat{y} в точке, определяемой начальным условием $X=x_{np}$ находится следующим образом:

$$y_{np} \in \left[\hat{y}_{np} \pm t_{\alpha, n-2} \cdot \hat{S} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{np} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right].$$



Доверительный интервал имеет наименьшую величину, когда $x_{np} = \bar{x}$, а по мере удаления x_0 от \bar{x} ширина доверительного интервала увеличивается, и точность оценки \hat{y} снижается.