КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

Модуль 3. Элементы математической статистики

Лекция 8. Понятие о регрессионном и корреляционном анализе

1. Понятие о регрессионном и корреляционном анализе. Линейная регрессия.
2. Вывод уравнения линейной регрессии методом наименьших квадратов (МНК).
3. Коэффициент корреляции Пирсона.
4. Коэффициенты ранговой корреляции (Спирмена, Кендалла)

ФАЙЛЫ:

- МС_ Карасев В.А., Лёвшина Г.Д. МИСИС. Статистика 2770
- МС_ Боголюбов_ ПР.ЗАН
- Л.8.2_ ПРИМЕРЫ ЛИНЕЙНОЙ РЕГРЕССИИ
- Л.8.1_ Коэффициент корреляции Спирмена, Кендалла

Корреляционно-регрессионный анализ

Корреляционный анализ – это статистический метод исследования связей случайных величин

$$Y \text{ и } X_j \quad (j = \overline{1,k}).$$

Задачи корреляционного анализа:

- существует ли зависимость между СВ,
- определение коэффициента корреляции для СВ,
- насколько сильно взаимосвязаны СВ

Инструмент исследований – коэффициент корреляции

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

Задачи регрессионного анализа:

- установление вида зависимости между СВ,
- определение параметров выбранной зависимости,
- анализ ***адекватности*** модели и ***значимости коэффициентов,*** т.е. соответствие эмпирическим данным,
- определение неизвестных значений (***прогноз***)

Инструмент исследований – регрессионная модель (линейная, квадратичная, полиномиальная, экспоненциальная и т.д.)

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

$$M(Y \mid X) = f(X)$$

называется ***функцией регрессии Y на X.***

При рассмотрении зависимости

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

$$M(Y \mid X) = f(X)$$

- нескольких переменных говорят о ***множественной регрессии***

$$M(Y \mid X_1, X_2, \ldots, X_k) = f(X_1, X_2, \ldots, X_k).$$

УРАВНЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Ковариация или **корреляционный момент случайных величин *X* и *Y*** - это математическое ожидание произведения отклонений случайных величин X и Y от их математических ожиданий: $\mathrm{Cov(X,Y)} = K_{xy} = M((X - m_x)\cdot(Y - m_y))$.

Ковариацию можно записать в виде: $\mathrm{Cov(X,Y)} = K_{xy} = M(X \cdot Y) - m_x \cdot m_y$

m_x, m_y - математическое ожидание X и Y.

Коэффициент корреляции (Пирсона) $r_{xy} = \dfrac{K_{xy}}{\sigma_x \cdot \sigma_y}$,

σ_x, σ_y - средние квадратические отклонения X и Y ;

m_x, m_y - математические ожидания X и Y.

- **Линейная регрессия *Y* на *X*** : $y - m_y = r_{xy}\dfrac{\sigma_y}{\sigma_x}(x - m_x)$
- **Линейная регрессия X на *Y*** : $x - m_x = r_{xy}\dfrac{\sigma_x}{\sigma_y}(y - m_y)$

Эти прямые пересекаются в точке (m_x , m_y) - в центре распределения

ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ ДВУМЕРНОЙ ВЫБОРКИ

Пусть (x_i, y_i), $i = 1,2,...,n$, — двумерная выборка объема n из значений случайного двумерного вектора (ξ,η). Изображение элементов выборки в виде точек на координатной плоскости называется *диаграммой рассеяния*.

Точечными оценками математических ожиданий $\mathbf{M}\xi$ и $\mathbf{M}\eta$ служат выборочные средние

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{и} \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i ,$$

дисперсий $\mathbf{D}\xi$ и $\mathbf{D}\eta$ — несмещенные оценки дисперсий

$$s^2(\xi) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2) \quad \text{и}$$

$$s^2(\eta) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2),$$

ковариации $\mathrm{cov}(\xi,\eta)$ — *несмещенная выборочная ковариация*

$$k(\xi,\eta) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1}(\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}),$$

коэффициента корреляции $\rho(\xi,\eta)$ — *выборочный коэффициент корреляции*

$$r(\xi,\eta) = \frac{k(\xi,\eta)}{s(\xi)s(\eta)}.$$

УРАВНЕНИЯ ВЫБОРОЧНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

s^2_x, s^2_y - выборочные исправленные оценки дисперсий X ,Y ;

$\bar{x}$, $\bar{y}$ - выборочные средние оценки математических ожиданий X, Y.

- **Выборочная линейная регрессия *Y* на *X*** :

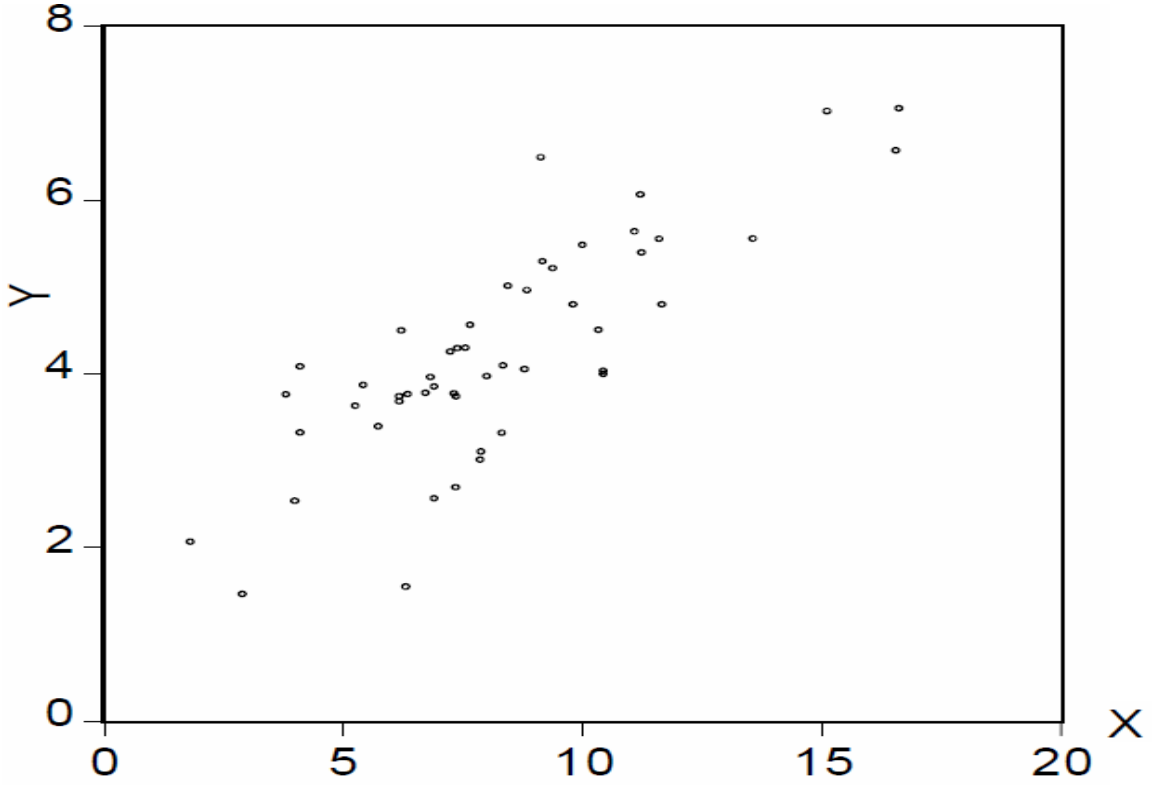
$$y - \bar{y} = r_{xy(в)} \frac{s_y}{s_x}(x - \bar{x})$$

- **Выборочная линейная регрессия X на *Y*** :

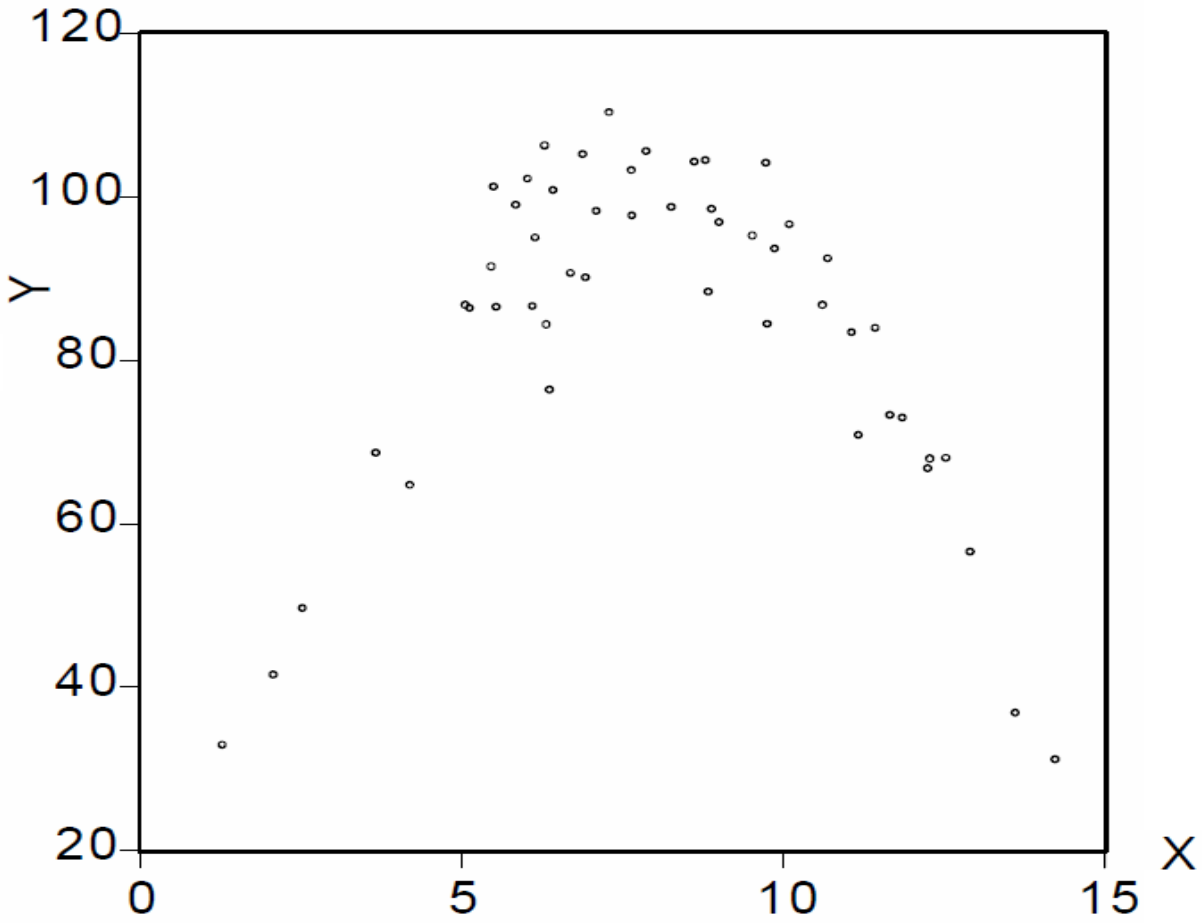
$$x - \bar{x} = r_{xy(в)} \frac{s_x}{s_y}(y - \bar{y})$$

Спецификация уравнения регрессии.

В случае парной регрессии – графический анализ реальных статистическихданных (поле рассеивания).

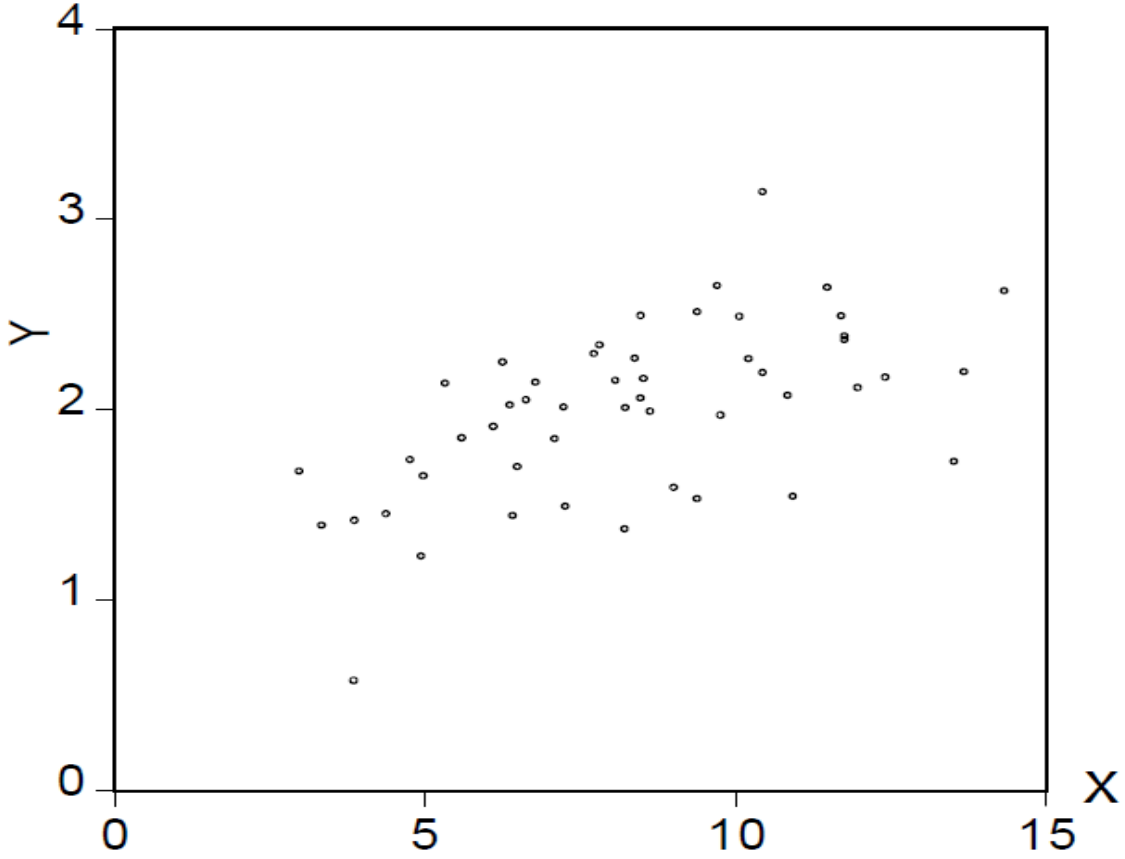


Линейная зависимость $\hat{Y} = \beta_0 + \beta_1 X$.



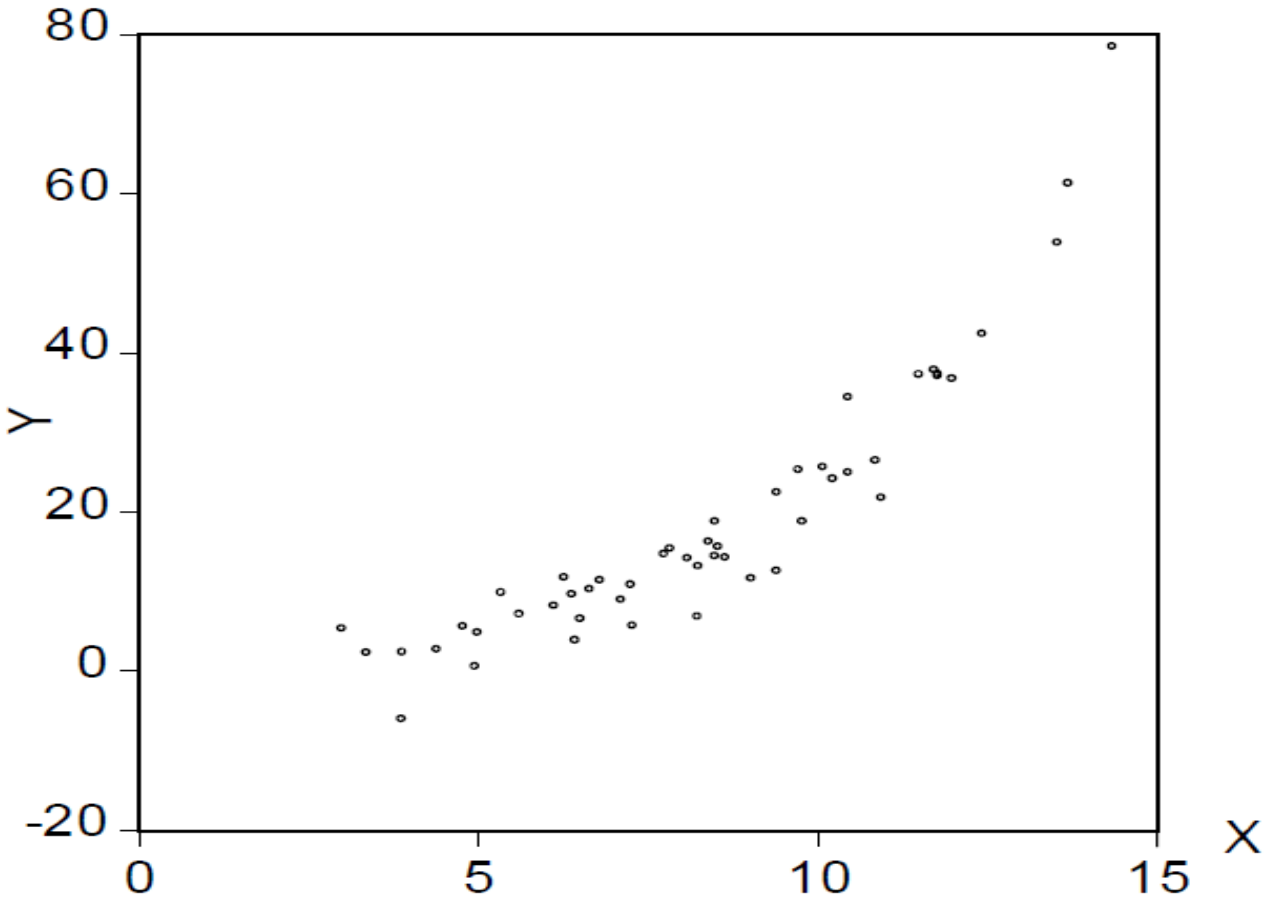
Квадратичная зависимость:

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



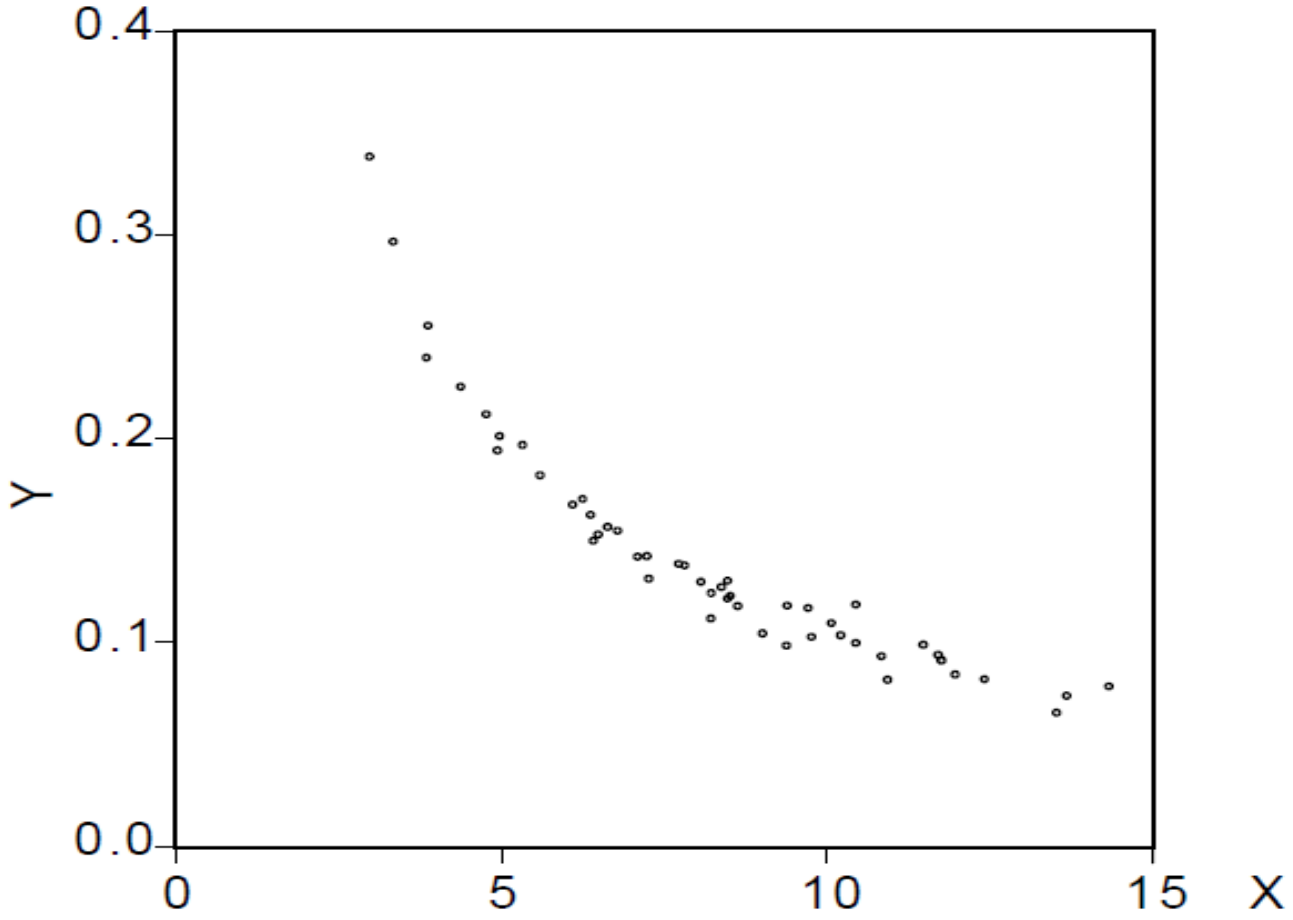
Степенная зависимость

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

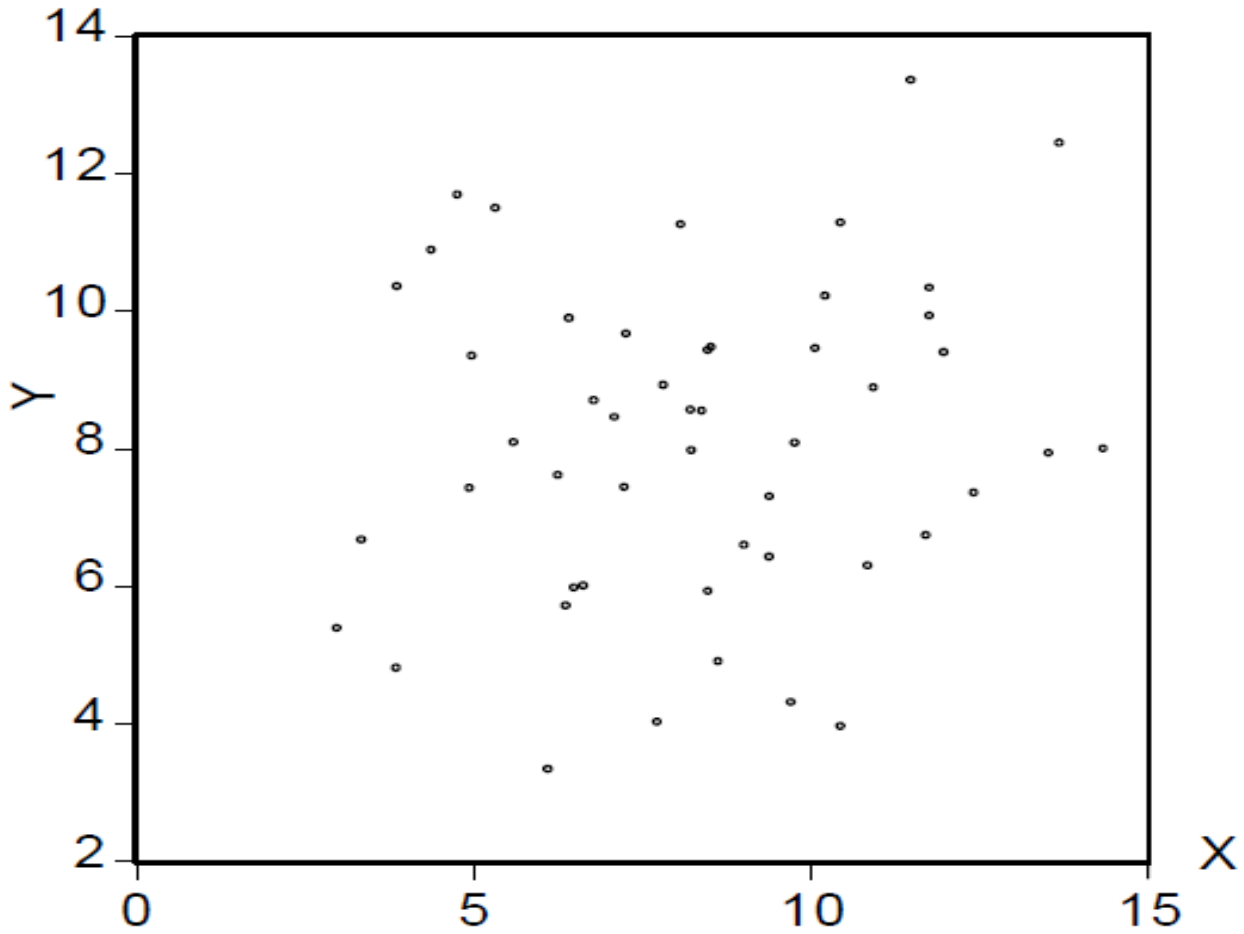


Показательная зависимость

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



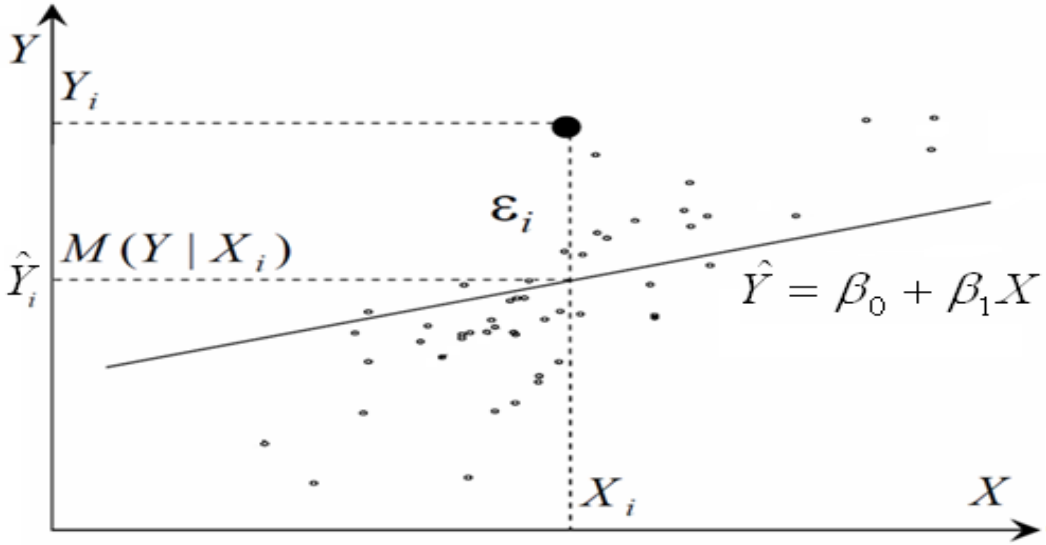
Гиперболическая зависимость $\hat{Y} = \beta_0 + \frac{\beta_1}{X}$



X и *Y* независимы

Вывод уравнения линейной выборочной регрессии (МНК).

Выборка: (x_i,y_i) – результат *i*-го наблюдения, i=1,…n .

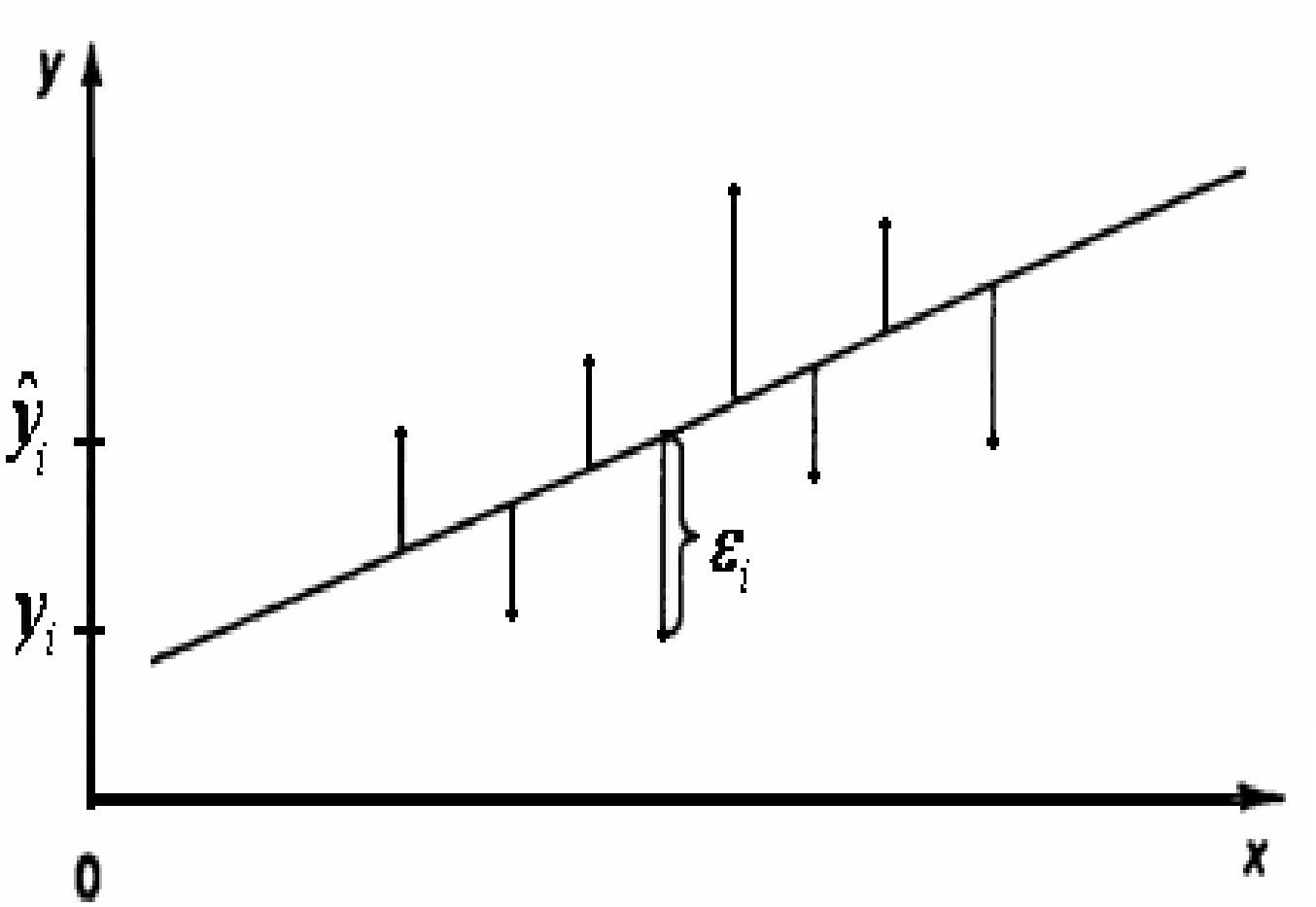


Выборочная линия регрессии

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

где b_0 и b_1 – неизвестные параметры, которые подбираются по методу наименьших квадратов.

МНК: минимизация суммы квадратов отклонений выборочных значений y_i от $\hat{y}_i$



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

Найдем частные производные Q и приравняем их к нулю:

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

Получим ***систему линейных 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:

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

$$\boxed{b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2} = \frac{K_{xy(в)}}{s_x^2} = r_{xy(в)} \frac{s_y}{s_x}},$$

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

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

$$\hat{y} = b_0 + b_1 x = \bar{y} + r_{xy(в)} \frac{s_y}{s_x}(x - \bar{x})$$

АДЕКВАТНОСТЬ

ЛИНЕЙНОЙ МОДЕЛИ РЕГРЕССИИ

Коэффициент детерминации – наиболее эффективная оценка адекватности линейной регрессионной модели:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = r^2_{xy(в)},$$

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

y_i – выборочные значения,

$\bar{y}$ – выборочное среднее,

$\hat{y}_i$ – теоретические значения,

Свойства коэффициента детерминации.

1. $0 \le R^2 \le 1$.
2. $R^2 = 0$ – Y и X независимы.
3. $R^2 = 1$ – функциональная линейная зависимость между Y и X.
4. $0 < R^2 < 1$ – чем ближе R^2 к 1, тем точнее подгонка кривой к опытным данным.

R^2 применяется для проверки значимости уравнения регрессии. Гипотеза H_0:модель неадекватна;

гипотеза H_1:модель адекватна.

$$H_0: R^2 = 0; \qquad H_1: R^2 \neq 0$$

.

Рассчитывают статистику Фишера:

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

Если $F_{набл} > F_{кр} = F_{1-a,1,n-2}$ то гипотеза отвергается и уравнение считается значимым.

Интервальное оценивание коэффициентов линейной 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:

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

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

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

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