

Корреляционно- регрессионный анализ данных. Парная регрессия

Кагирова Мария Вячеславовна,
доцент, кандидат экономических наук,
кафедра статистики и кибернетики
РГАУ-МСХА им. К. А. Тимирязева





Вопросы:

- Показатели связи между признаками
- Парная линейная регрессия
- Оценка достоверности модели регрессии.
- Нелинейная регрессия

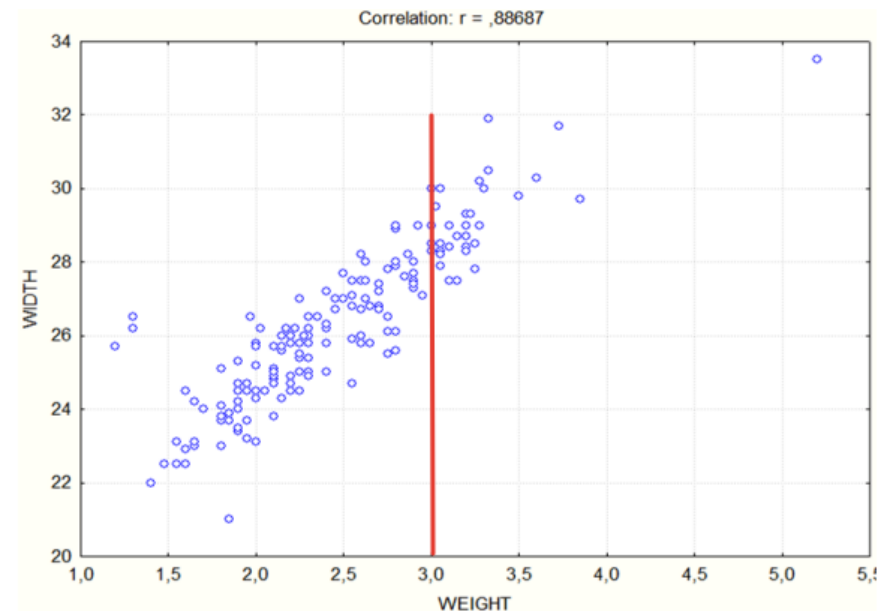




Показатели связи между признаками

Функциональные связи являются «жесткими», четко определяющими значение одной переменной (признака) при изменении другой переменной (признака).

Статистические связи формируются под влиянием случайных факторов. При этом конкретному значению одной переменной может соответствовать не одно значение другой переменной, а множество, т.е. некоторое распределение второй переменной





Показатели связи между признаками

Выборочная ковариация ($Cov(x, y)$) является мерой взаимосвязи между двумя переменными. Основой формирования данного показателя является оценка взаимной изменчивости переменных, выраженной в произведении отклонений каждой из них от среднего уровня.

Если $Cov(x, y) > 0$, то связь прямая, если $Cov(x, y) < 0$, то связь обратная, $Cov(x, y) = 0$ – связь отсутствует.

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$Cov(x, y) = \frac{\sum xy}{n} - \bar{x} \cdot \bar{y}$$



Показатели связи между признаками

Коэффициент корреляции

$0 < |r_{xy}| < 1$:

0 – 0,3 – связь слабая;

0,3 – 0,5 – умеренная;

0,5 – 0,7 – средняя;

0,7 – 0,99 – тесная.

$Cov(x, y)$ – выборочная ковариация между указанными переменными,
 $Var(x)$ и $Var(y)$ – выборочные дисперсии переменных x и y .

$$r_{xy} = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}}$$



Показатели связи между признаками

	y	x	z
y	1,000		
x	-0,470	1,000	
z	0,055	0,825	1,000

Матрица парных коэффициентов корреляции

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} = \frac{-0.47 - 0.825 \cdot 0.055}{\sqrt{(1-0.825^2)(1-0.055^2)}} = -0.9146$$



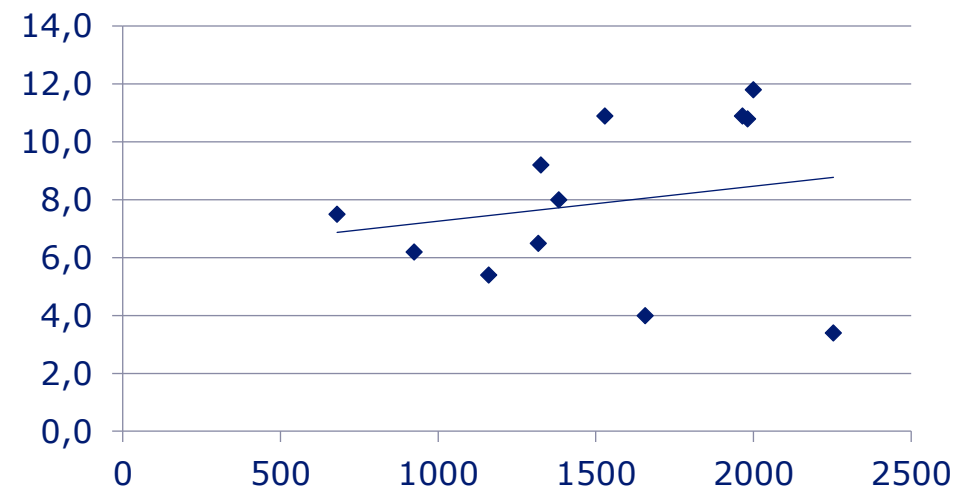
Парная линейная регрессия

$$y = \alpha + \beta x + u$$

Величина y , рассматриваемая как зависимая переменная, состоит из двух составляющих:

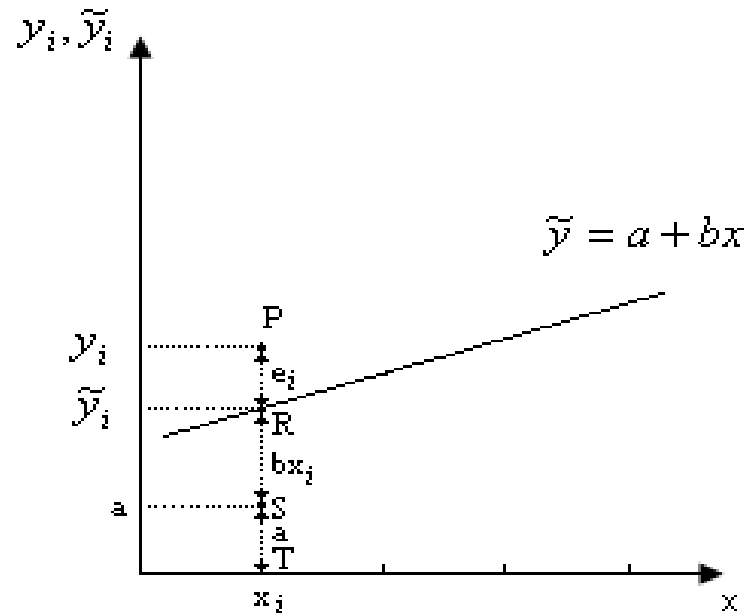
- неслучайной составляющей $\alpha + \beta x$, где x выступает как объясняющая (или независимая) переменная, а постоянные величины α и β — как параметры уравнения;
- случайного члена u .

Зависимость ВРП от инвестиций





Парная линейная регрессия



a – условное начало, характеризует значение зависимой переменной y при равенстве независимой переменной x нулю

b – коэффициент полной регрессии показывает, на сколько в среднем изменится значение зависимой переменной y (в своих единицах измерения) при изменении независимой переменной x на единицу (в своих единицах измерения)



Парная линейная регрессия

Коэффициент детерминации

характеризует долю вариации зависимой переменной y , объясненную влиянием фактора x , включенного в уравнение регрессии.

Позволяет оценить качество модели регрессии. Чем ближе значение R^2 к 1, тем выше качество модели.

$$R^2 = \frac{Var(\tilde{y})}{Var(y)}$$

$$R^2 = 1 - \frac{Var(e)}{Var(y)}$$



Оценка достоверности модели регрессии

- Выдвигаются гипотезы:

$$H_0: \sigma^2_{\text{регр}} = \sigma^2_{\text{ост}}$$

$$H_a: \sigma^2_{\text{регр}} \neq \sigma^2_{\text{ост}}$$

- Выбирается уровень значимости критерия.
- Производится разложение общего объема вариации:
- Определяется число степеней свободы
- Рассчитываются выборочные несмещенные оценки дисперсий:

$$S^2_{\text{регр.}} = \frac{W_{\text{регр.}}}{1}$$

$$S^2_e = \frac{W_e}{n-2}$$

- Определяется фактическое значение F-критерия Фишера:
- Определяется критическое (табличное) значение критерия:
- Делается статистический вывод:

$$F = \frac{S^2_{\text{регр.}}}{S^2_e}$$

$$F_{\alpha; V_{\text{регр.}}; V_{\text{ост.}}}$$

$$F_{\text{факт.}} \leq F_{\text{табл.}} \Rightarrow H_0 (\sigma^2_{\text{регр}} = \sigma^2_{\text{ост}})$$

$$F_{\text{факт.}} > F_{\text{табл.}} \Rightarrow H_a (\sigma^2_{\text{регр}} \neq \sigma^2_{\text{ост}})$$



Оценка достоверности модели регрессии

t-Тест

- Формулируются рабочая и альтернативная гипотезы:

$$H_0 : \alpha = 0; \beta = 0; \rho = 0 \quad H_a : \alpha \neq 0; \beta \neq 0; \rho \neq 0$$

- Выбирается уровень значимости критерия .
- Рассчитываются средние ошибки выборочных характеристик:

$$m_a = \sqrt{S_e^2 \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}} = \sqrt{S_e^2 \frac{\bar{x}^2}{n \sigma_x^2}} \quad m_b = \sqrt{\frac{S_e^2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{S_e^2}{n \sigma_x^2}} \quad m_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

- Определяются фактические значения t-критерия:
- Определяется критическое значение:

$$t_{\alpha; \nu_{OCT}}$$

$$t_a = \frac{a}{m_a};$$

$$t_b = \frac{b}{m_b};$$

$$t_r = \frac{r}{m_r}.$$

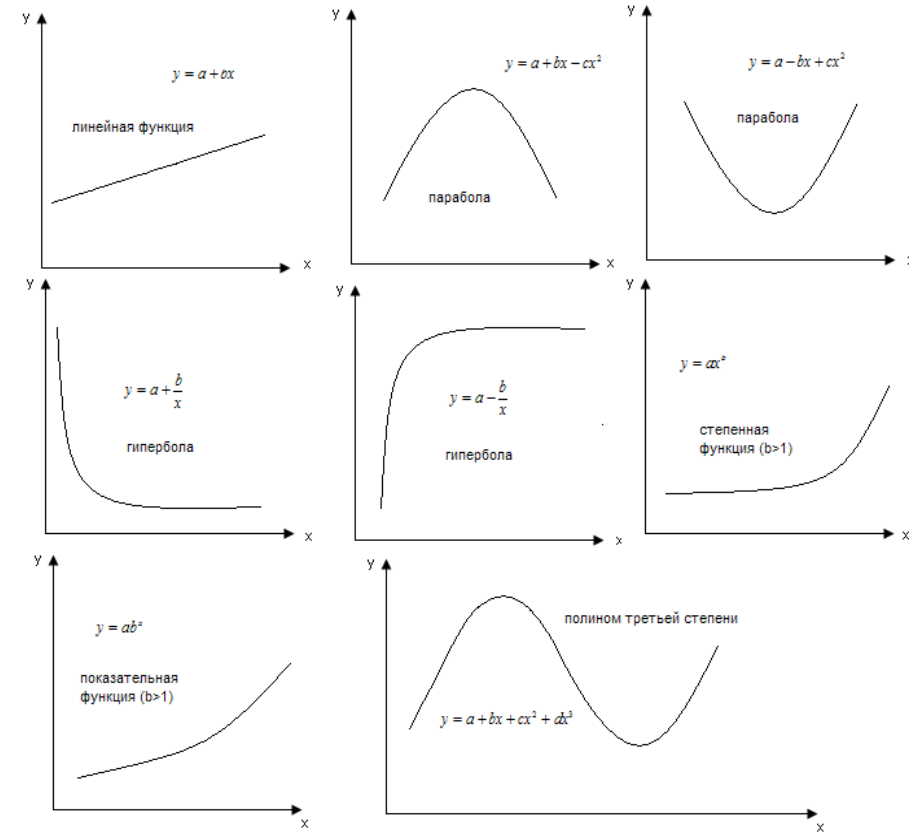
- Фактические значения сравниваются с критическими. Тестируемые параметры будут значимыми, если фактическое значение критерия превысит табличное

Нелинейная регрессия



Способы определения формы уравнения регрессии:

- аналитический;
- графический;
- экспериментальный





Нелинейная регрессия

Два класса нелинейных регрессий

- регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам. полиномы разных степеней

$$y = a + vx + cx^2;$$

$$y = a + vx + cx^2 + dx^3,$$

равносторонняя гиперболола $y = v + a/x$.

- нелинейные регрессии по оцениваемым параметрам:

степенная $y = a x^b$

показательная $y = a v^x$

экспоненциальная $y = e^{a+vx}$



Нелинейная регрессия

Подходы к линеаризации нелинейных уравнений

В параболе $y = a + vx + cx^2$, заменяя переменные $x_1 = x$, а $x_2 = x^2$, получаем двухфакторное уравнение линейной регрессии $y = a + vx_1 + cx_2$.

Для равносторонней гиперболы мы можем заменить $1/x$ на z и получим линейное уравнение регрессии, оценка параметров которого может быть дана МНК.

Для степенной функции $y = a x^b$ применяется логарифмирование левой и правой части уравнения

$\log y = \log \alpha + \log x^\beta$, а затем замена переменных

$y_1 = \log y$; $z = \log x$ и $\alpha_1 = \log \alpha$, получаем уравнение в линейном виде:

$$y_1 = \alpha_1 + \beta z$$



Спасибо за внимание!