Курс: Статистические Методы Обработки Данных

Лекция 6. Регрессионный и корреляционный анализ

Специальность: 1-53 01 02 – Автоматизированные системы обработки информации

УО «ГГУ им. Ф. Скорины»

Преподаватель: Бабич К.С, ст. преподаватель, 2016

Раздел 3 – Регрессионный и корреляционный анализ.

(исследование связей между случайным физическими величинами)

11. Некоторые сведения о двумерных случайных величинах.

Пример:

1) исследовать возможную связь м-ду активностью Солнца и вспышкой вирусов.
2) Котировки акций от погоды

и др.

Раздел 3 – Регрессионный и корреляционный анализ.

11. Некоторые сведения о двумерных случайных величинах.

Пусть: $x(k)$ $y(k)$ **- две случайные величины**

$F(x)$ $F(y)$ **- Интегральные функции распределения**

Определение: Совместной функцией распределения $F(x,y)$ **называется вероятность, которая приписывается выборочному множеству** k **удовлетворяющему одновременно условиям:**

$$\begin{cases} x(k) \le x \\ y(k) \le y \end{cases}$$

т.е.

$$F(x,y) = \mathrm{Prob}\left[\, x(k) \le x \;\; u \;\; y(k) \le y \,\right]$$

Раздел 3 – Регрессионный и корреляционный анализ.

11. Некоторые сведения о двумерных случайных величинах.

$$F(x,y)=\Pr ob\left[\, x(k)\le x \;\; u \;\; y(k)\le y \,\right]$$

Свойства:

1) $F(-\infty,y)=0=F(x,-\infty)$

2) $F(+\infty,+\infty)=1$

Дифференциальная плотность вероятности:

$$p(x,y)=\frac{\partial}{\partial y}\left[\frac{\partial F(x,y)}{\partial x}\right]$$

Плотность вероятности для x, y:

$$p(x)=\int_{-\infty}^{\infty}p(x,y)dy$$

$$p(y)=\int_{-\infty}^{\infty}p(x,y)dx$$

Раздел 3 – Регрессионный и корреляционный анализ.

11. Некоторые сведения о двумерных случайных величинах.

Важным частным случаем 2-мерных случайных величин является случай статистической независимости:

Опр: **Если** $p(x)p(y)=p(x,y)$ **то** $x(k)$ **и** $y(k)$ **называются статистически независимыми**

Для таких величин:

$$F(x,y)=F(x)F(y)$$

Раздел 3 – Регрессионный и корреляционный анализ.

11. Некоторые сведения о двумерных случайных величинах.

Математическое ожидание (М.О.) произвольной функции $g(x,y)$ **от случайных величин** $x(k)$ **и** $y(k)$

$$M.o.[g(x,y)] = \int_{-\infty}^{+\infty}\int g(x,y)p(x,y)dx\,dy$$

Раздел 3 – Регрессионный и корреляционный анализ.

11. Некоторые сведения о двумерных случайных величинах.

Если $\bar{x}$ **-М.о.** $x(k)$ **, а** $\bar{y}$ **-М.о.** $y(k)$

то можно найти Коэффициент Ковариации (аналог Дисперсии)

$$C_{xy} = \int_{-\infty}^{+\infty}\int (x-\bar{x})(y-\bar{y})p(x,y)dx\,dy$$

Свойство: $|C_{xy}| \le \sigma_x \sigma_y$

Раздел 3 – Регрессионный и корреляционный анализ.

11. Некоторые сведения о двумерных случайных величинах.

Вводят коэффициент корреляции: $\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$ $\left|\rho_{xy}\right| \le 1$

Свойство:

(Необходимое условие, но не достаточное):

Если $\rho_{xy} = 0$ **, то случайные величины не коррелированны.**

т.е. статистически независимые.

Обратное неверно.

(хотя для нормального распределения величин, обратное утверждение верно)

Коэффициент корреляции может дать информацию о наличии взаимосвязи.

Существует линейная и нелинейная зависимость между x и y.

Раздел 3 – Регрессионный и корреляционный анализ.

12. Линейный корреляционный анализ.

Пусть имеется n значений пар случайных величин: (x_i, y_i)

На эксперименте может быть найдена точечная оценка коэффициента корреляции ρ_{xy} **:**

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}}$$

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x}\,\bar{y}}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^{1/2}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)^{1/2}}$$

Раздел 3 – Регрессионный и корреляционный анализ.

12. Линейный корреляционный анализ.

$r_{xy} = 0$ - Связи между x и y не существует

$r_{xy} = \pm 1$ - Существует 100% линейная взаимосвязь

Мы будем рассматривать только случаи: -1, 0, 1

В реальности чаще $r_{xy} = 0{,}70$

Раздел 3 – Регрессионный и корреляционный анализ.

12. Линейный корреляционный анализ.

В реальности же м.б. что $r_{xy} = 0{,}70$

Для таких ситуаций, следует вводить доверительный интервал для r_{xy}

Это можно сделать вводя функцию:

$$W = \frac{1}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right]$$

W **- это случайная величина, которая подчиняется нормальному распределению с М.о.**

$$\mu_W = \frac{1}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right]$$

и дисперсией:

$$\sigma_W^2 = \frac{1}{n-3}$$

Раздел 3 – Регрессионный и корреляционный анализ.

12. Линейный корреляционный анализ.

Если $r_{xy} = 0$ **то** $\mu_W = 0$ **(отсутствие корреляции)**

Можно построить доверительный интервал

$$-Z_{\alpha/2} \le \frac{\sqrt{n-3}}{2} \ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right] \le Z_{\alpha/2}$$

$Z_{\alpha/2}$ **- квантиль нормального распределения**

Если значение окажется вне интервала, то это будет признаком наличия корреляции

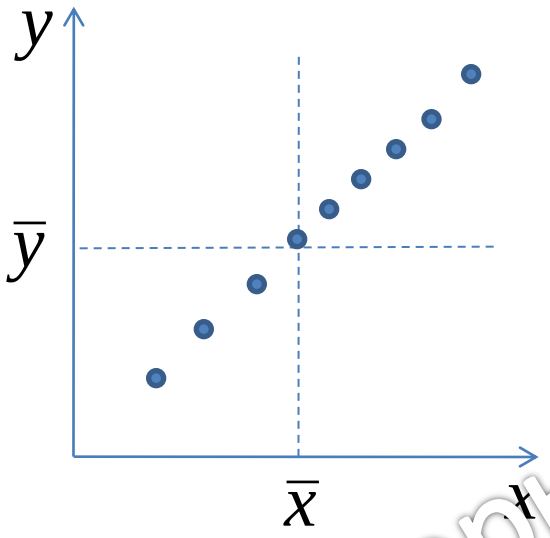
$$\Pr ob = 1-\alpha$$

Раздел 3 – Регрессионный и корреляционный анализ.

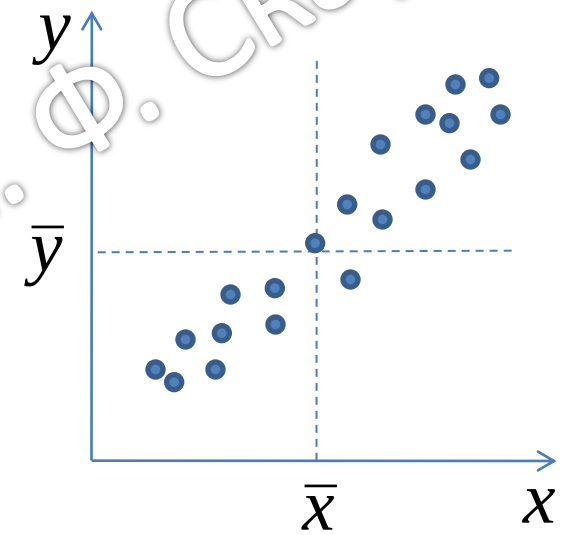
12. Линейный корреляционный анализ.

Нелинейная связь из-за разброса данных (или если неполная корреляция) приводит к уменьшению r_{xy}

а) полностью линейно коррелированны



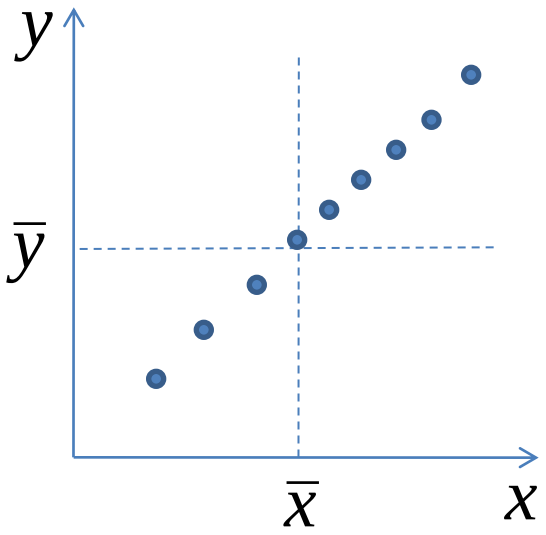
б) умеренно линейно коррелированны



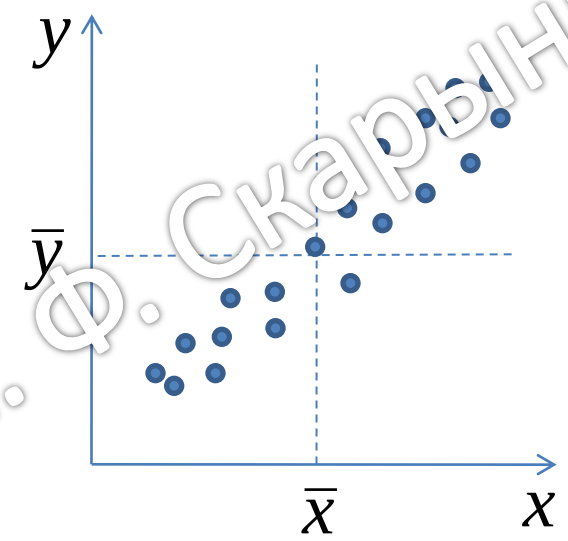
Раздел 3 – Регрессионный и корреляционный анализ.

12. Линейный корреляционный анализ.

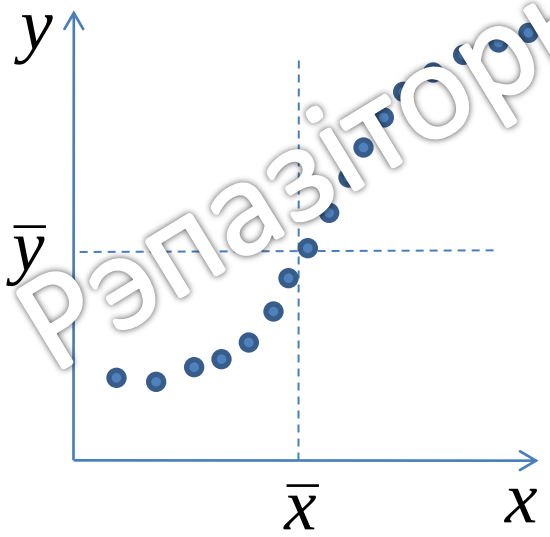
а) полностью линейно коррелированны



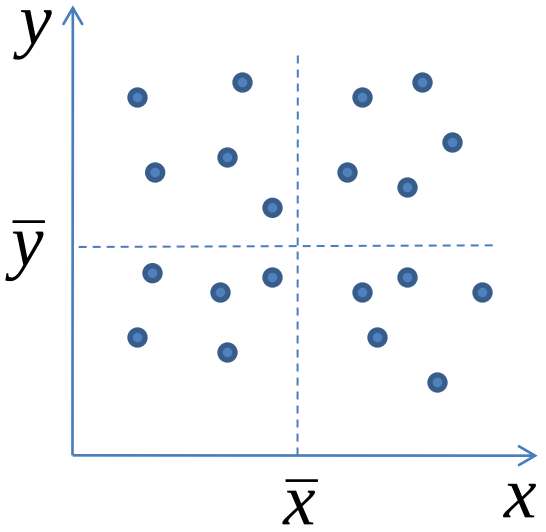
б) умеренно линейно коррелированны



в) нелинейно коррелированны

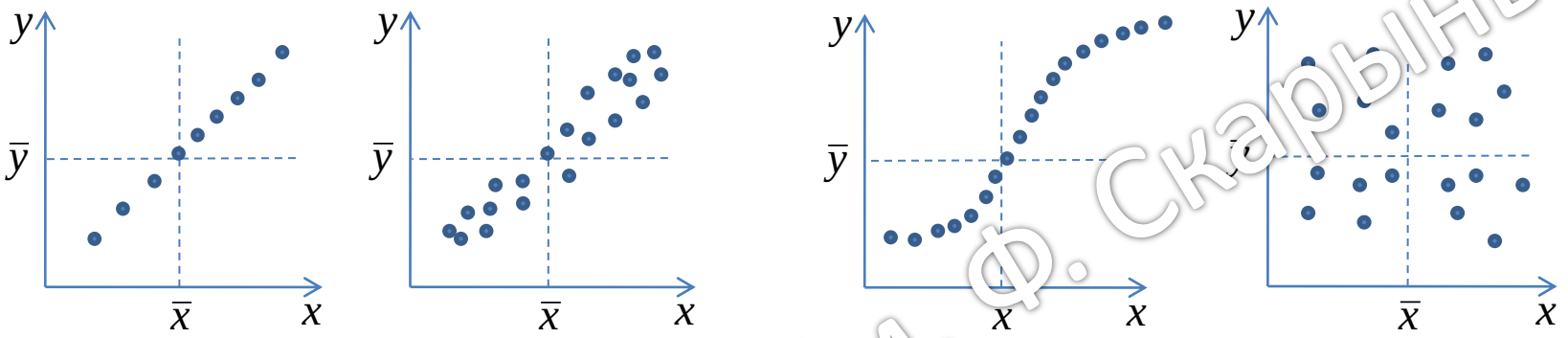


г) отсутствие корреляции



Раздел 3 – Регрессионный и корреляционный анализ.

12. Линейный корреляционный анализ.



Вывод: желательно строить графики, т.к. м.б. нелинейная связь или др. случаи.

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Корреляционный анализ помогает «уловить» взаимосвязь двух и более величин.

Но! Хочется иметь модель этой взаимосвязи, которая позволит предсказывать значение одной случайной величины по значению другой.

Пусть есть 2 случайные величины x и y.

Линейная связь между x и y означает, что прогноз значения $\tilde{y}$ по данному x имеет вид:

$$\tilde{y} = A + Bx$$

Если $r_{xy} = 1$, то y и $\tilde{y}$ совпадают (идеальная связь)

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Линейная связь между x и y означает, что прогноз значения $\tilde{y}$ по данному x имеет вид:

$$\tilde{y} = A + Bx$$

Коэффициенты A и B нужно подобрать, чтобы предсказать ожидаемое y_i для любого x_i .

Т.е. не обязательно, что y_i и $\tilde{y}_i$ совпадают, но они будут равны среднему значению всех таких наблюдений.

Теперь нужно найти A и B , чтобы удовлетворить этому требованию.

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Наиболее популярный метод - Метод наименьших квадратов (МНК)

(предложен Лежандром и Гауссом в 1795-1805 гг.)

МНК работает не только для линейно регрессии, но и для любых ситуаций

Сущность МНК:

1) **Находят** $\Delta_i = y_i - \tilde{y}_i$ $\qquad \tilde{y} = A + Bx$

2) **Находят коэффициенты** A, B **так, чтобы** $Q = \sum_{i=1}^{n} \Delta_i^2 \to \min$

$$Q = Q(A, B)$$

т.е.

$$\begin{cases} \dfrac{\partial Q}{\partial A} = 0 \\ \dfrac{\partial Q}{\partial B} = 0 \end{cases}$$

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Метод наименьших квадратов (МНК)

$$\frac{\partial Q}{\partial A} = -2\sum_{i=1}^{n}\left(y_i - A - Bx_i\right) = 0$$

или

$$A \cdot n + B \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\frac{\partial Q}{\partial B} = -2\sum_{i=1}^{n}\left(y_i - A - Bx_i\right)x_i = 0$$

или

$$A \cdot \sum_{i=1}^{n} x_i + B \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Метод наименьших квадратов (МНК)

A и B найдем решая систему уравнений

$$\begin{cases} A \cdot n + B \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ A \cdot \sum_{i=1}^{n} x_i + B \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$A = \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$B = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

На сленге название процедуры - *фитирование*

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Метод наименьших квадратов (МНК)

$$A=\frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i^2\right)-\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right)-\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$B=\frac{n\left(\sum_{i=1}^{n} x_i y_i\right)-\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right)-\left(\sum_{i=1}^{n} x_i\right)^2}$$

На сленге название процедуры *фитирование*

Далее коэффициенты используются в формуле $\tilde{y}=A+Bx$

это прямая регрессия y на x.

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Свойства (МНК):

1) Из уравнения $\frac{\partial Q}{\partial A} = 0$ **имеем** $A + B \cdot \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} y_i$

или $A + B \cdot \bar{x} = \bar{y}$

т.е. кривая регрессии проходит через точку $(\bar{x}, \bar{y})$ **- центр тяжести экспериментальных точек.**

Если $\bar{x} = \bar{y} = 0$ **, то уравнение превратится в уравнение, где на один параметр меньше** *(что удобно когда много параметров)*

Используя это свойство:

$$\tilde{y} = A + B \cdot x = \bar{y} + B(x - \bar{x})$$

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Свойства (МНК):

2) Уравнение МНК необратимо, т.е нельзя получить обратную регрессию x на y:

$$\tilde{x} = A' + B'y$$ **обращая уравнение**

$$\tilde{y} = A + B \cdot x$$

$$x = \frac{1}{B} y - A$$ **- это неверно**

Проделав все с начала для $\tilde{x} = A' + B'y$ **получим**

$$A' = -A \qquad B' = r_{xy}^2 \frac{1}{B}$$

r_{xy}^2 **- коэффициент корреляции**

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Свойства (МНК):

3) Этот метод чувствителен к неоднородности выборок.

Необходимо провести цензурирование выборок

Раздел 3 – Регрессионный и корреляционный анализ.

13. Линейный регрессионный анализ.

Ряд задач м.б. сведен к линейному регрессионному анализу

Модель

Пр. 1 $$\frac{1}{y} = a_1 + b_1 x \quad \Rightarrow \quad y' = \frac{1}{y} \quad \Rightarrow \quad y' = a_1 + b_1 x$$

аналогично $$y = a_1 + b_1 \frac{1}{x}$$

Пр. 2 $$y = a\,x^b \qquad \lg y = \lg a + b \lg x$$

$$y' \Rightarrow \lg y \qquad x' \Rightarrow \lg x$$