

Теория статистики

Корреляционно-регрессионный анализ: статистическое моделирование зависимостей

Часть 1.

Задача изучения зависимостей

- Исследование объективно существующих связей между явлениями и их показателями – одна из важнейших задач анализа
- Различают классы статистических признаков:
 - независимые (факторные)
 - и зависимые (результативные)
- Причинность, корреляция, регрессия

Виды зависимости

- Зависимости бывают функциональными и нет, т.е. с элементом случайности
- При Функциональной зависимости каждому значению независимой переменной соответствует определенное значение зависимой

Балансовая зависимость

- Пример функциональной связи – балансовая:

$$O_n + \Pi = P + O_k$$

O_n – остаток средств на начало изучаемого периода;

Π – поступление средств в течении данного периода;

P – расход средств за период;

O_k – остаток средств на конец периода

Статистическая зависимость

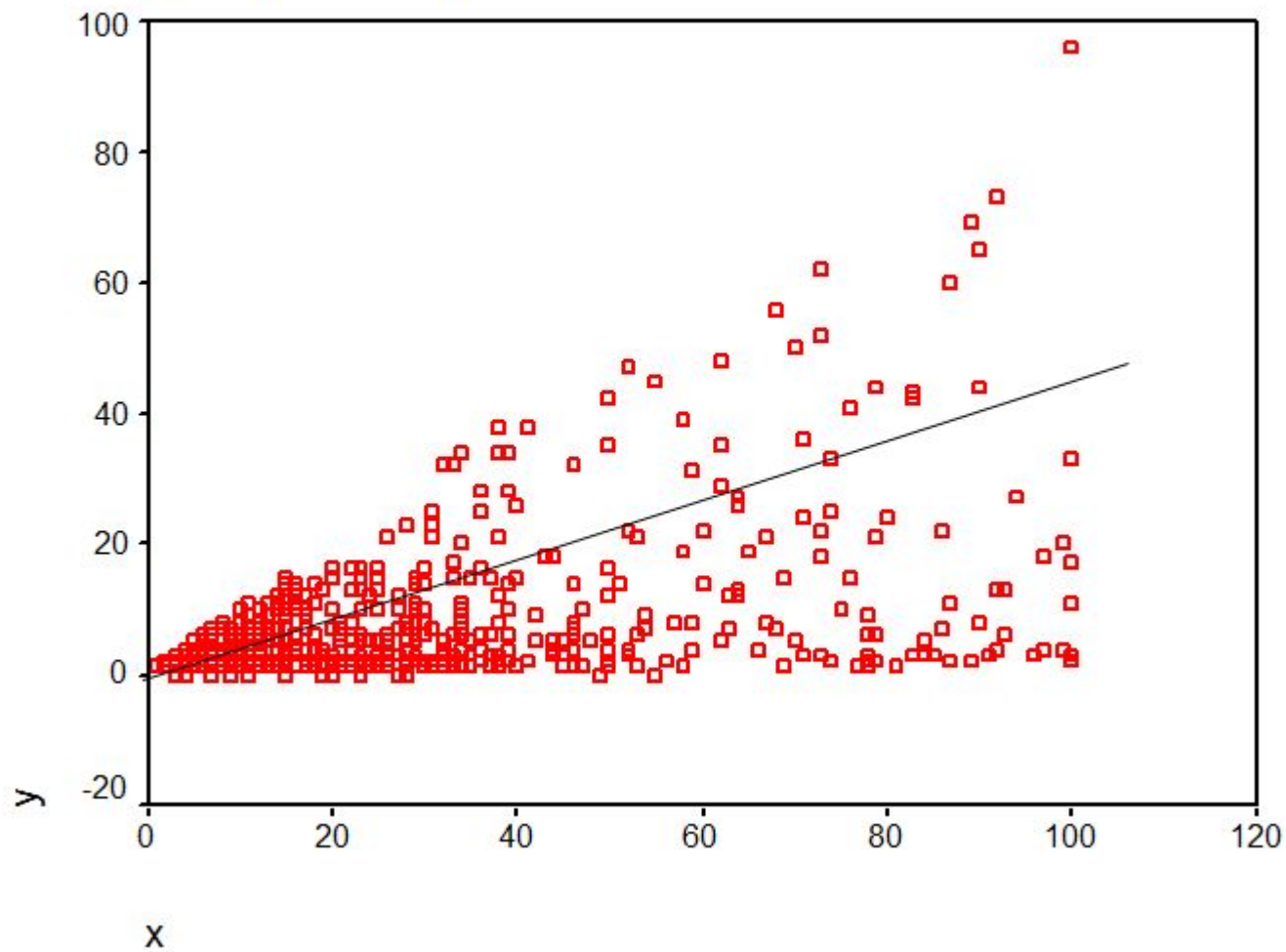
- В социально-экономических исследованиях в большинстве случаев наблюдается связь, при которой каждому значению одной переменной соответствует некоторое множество возможных значений другой переменной
- Такая зависимость называется статистической

Корреляционная связь – частный случай статистической зависимости

- Корреляционной зависимостью между двумя переменными величинами называется функциональная зависимость между значениями одной из них и средним значением другой
- Поле корреляции – графическое изображение взаимосвязи двух признаков

Поле корреляции

Диаграмма рассеяния



Классификация статистических связей

- Связи между явлениями и их признаками классифицируются:
 - По тесноте:
сильная, умеренная, слабая или отсутствует
 - По направлению:
прямая или обратная
 - По аналитическому выражению:
линейная или нелинейная

Виды корреляционной зависимости

- Парная корреляция – линейная зависимость между двумя переменными
- Частная корреляция – линейная зависимость между двумя переменными при исключении влияния других
- Множественная корреляция - линейная зависимость между набором переменных

Этапы статистического изучения СВЯЗИ

1. Качественный анализ на наличие объективной зависимости
2. Построение модели связи:
 - Метод приведения параллельных данных и построение поля корреляции
 - Корреляционный анализ
 - Регрессионный анализ
3. Содержательная интерпретация полученных результатов моделирования

Характеристика тесноты и направления связи

- Цель состоит в количественном описании тесноты и направления связи
- В качестве характеристики используется коэффициент корреляции (r):

$ r $	Связь
0,0 – 0,3	Отсутствует
0,3 – 0,5	Слабая
0,5 – 0,8	Умеренная
0,8 – 1,0	Сильная

Регрессионный анализ

- Регрессионный анализ заключается в аналитическом выражении связи:
 - Нахождение функциональной зависимости среднего (математического ожидания) признака (y) от значений независимой переменной (x):

$$\bar{y}_x = f(x) + \varepsilon,$$

ε – случайный остаток

Определение параметров регрессии

- Определение класса функций для выражения функциональной зависимости среднего признака (y) от значений переменной (x)
- Оценка параметров функции регрессии: метод наименьших квадратов

$$\sum_{k=1}^n (y_k - f(x_k))^2 \rightarrow \min$$

- Проверка случайности остатков и адекватности модели связи

Пример

- Пусть имеются данные по 9 студентам:
 - Признак (x) – количество пропущенных студентом занятий по дисциплине
 - Признак (y) – полученная студентом оценка на экзамене

Данные о числе пропущенных занятий и полученных оценках на экзамене студентами

№	A	B	C	D	E	F	G	H	I
(x)	3	4	1	2	10	6	2	5	3
(y)	4	3	5	4	2	3	4	5	4

Пример

- Исследуем зависимость среднего значения (y) от признака (x)
1. Ясно, что такая объективная зависимость может существовать (хотя и не функциональная)

Пример

2. Построение модели связи

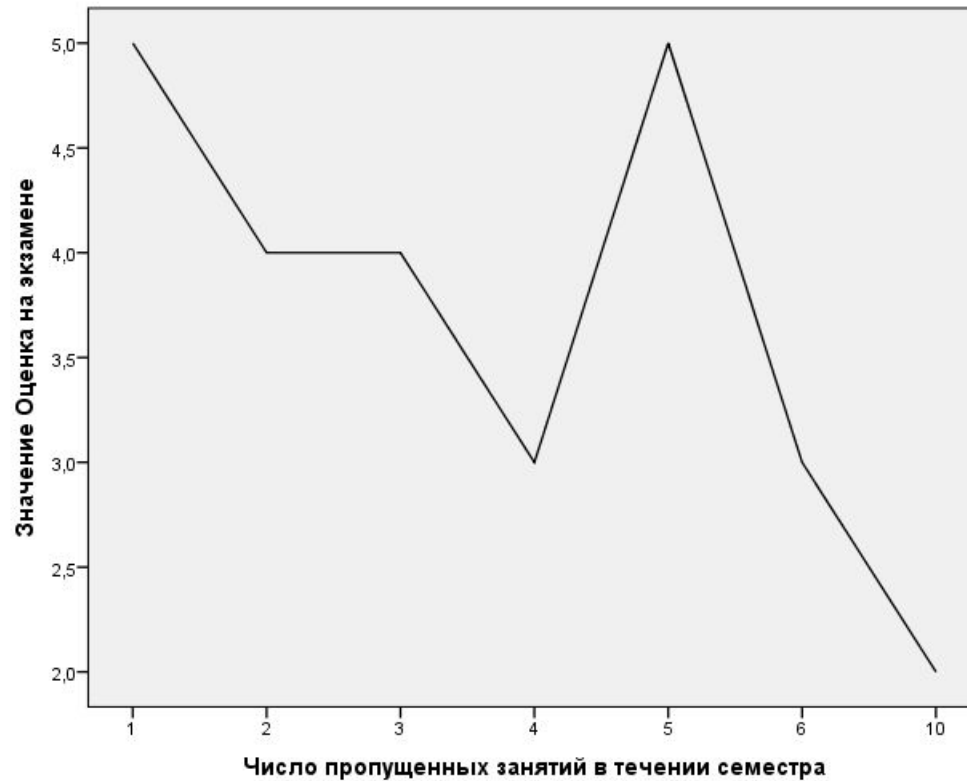
- Метод приведения параллельных данных

Данные о числе пропущенных занятий и полученных оценках на экзамене студентами, упорядоченные по переменной (x)

№	C	D	G	A	I	B	H	F	E
(x)	1	2	2	3	3	4	5	6	10
(y)	5	4	4	4	4	3	5	3	2

Пример

- Поле корреляции



Пример

- Теснота и направление связи между количественными переменными измеряются с помощью коэффициента корреляции Пирсона:

$$r = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}$$
$$= \frac{1/n \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}}{\sqrt{1/n \sum_{i=1}^n (x_i - \bar{x})^2 \cdot 1/n \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Пример

$$n = 9; \quad \sum_{i=1}^9 x_i = 36; \quad \sum_{i=1}^9 y_i = 34; \quad \bar{x} = 4; \quad \bar{y} = 3,78;$$

$$\sum_{i=1}^9 x_i y_i = 120 \quad 1/n \sum_{i=1}^9 x_i y_i = 13,33$$

$$\sigma_x^2 = 1/n \sum_{i=1}^9 (x_i - \bar{x})^2 = 6,67 \quad \sigma_y^2 = 1/n \sum_{i=1}^9 (y_i - \bar{y})^2 = 0,84$$

$$r = \frac{1/n \sum xy - \bar{x} \cdot \bar{y}}{\sqrt{1/n \sum (x_i - \bar{x})^2 \cdot 1/n \sum (y_i - \bar{y})^2}} = \frac{13,33 - 4 \cdot 3,78}{\sqrt{6,67 \cdot 0,84}} = \frac{-1,79}{2,37} = -0,76$$

Пример

- Делать выводы о тесноте и направлении связи пока преждевременно: нужно проверить значимость коэффициента корреляции (r)
- Гипотеза H_0 : истинное значение коэффициента корреляции (R) равно «0»
- Для проверки значимости коэффициента корреляции (r) применяется T -критерий Стьюдента

Пример

- По выборке рассчитываем значение

статистики:

$$t_r = |r| \cdot \sqrt{\frac{n-2}{1-r^2}} = |-0.76| \cdot \sqrt{\frac{9-2}{1-0.76^2}} = 3,09$$

Т-распределение Стьюдента (Фрагмент таблицы)			
К	Вероятность: $\alpha = St(t) = P(T > t_{tabl})$		
	0,05	0,02	0,01
7	2,365	2,998	3,499

$t_r = 3.09 > t_{0.05,7} = 2.365$: коэфф. корреляции значим

Вывод

- Корреляционная связь:
 - Обратная - коэффициент корреляции (r) отрицательный
 - Умеренная ($|r| = 0,76$) ~~и~~ близкая к сильной

Регрессионный анализ

- Наблюдается существенная линейная корреляционная зависимость, поэтому аналитическое выражение связи будем искать в линейной форме:

$$\bar{y}_x = a_0 + a_1 x \Rightarrow \begin{cases} a_0 = \bar{y} - a_1 \bar{x}; \\ a_1 = (\overline{xy} - \bar{x} \cdot \bar{y}) / \sigma_x^2; \end{cases}$$

$$a_1 = \frac{-1.78}{6.67} = -0.27; \quad a_0 = 3.78 - (-0.27) \cdot 4 = 4.86$$

$$\bar{y}_x = 4.86 - 0.27 \cdot x$$

Регрессионный анализ

- Необходима проверка значимости полученного уравнения регрессии
 - в целом
 - каждого коэффициента в отдельности
- Тем не менее, пользуясь полученным уравнением регрессии, находим, что, например, при $x = 3$, оценка ожидается 4:

$$\overline{y}_x(3) = 4.86 - 0.27 \cdot 3 = 4.05$$

Регрессионный анализ

- Значимость полученного уравнения регрессии (в целом) проверяется по F -критерию Фишера:
 - Гипотеза H_0 : все коэффициенты регрессии равны «0»

Регрессионный анализ

- Уравнение регрессии в целом значимо, если выполняется условие:

$$F = \frac{\sigma_R^2 (n - 2)}{\sigma_E^2} = \frac{Q_R (n - 2)}{Q_E} > F_{\alpha; 1; n-2}$$

$F_{\alpha; 1; n-2}$ табличное значение F -критерия на уровне значимости α (обычно $\alpha = 0.05$) при числе степеней свободы числителя $n - 2$ и знаменателя 1.

$Q_R = \sum_{k=1}^n (\hat{y}_k - \bar{y})^2$ - суммы квадратов отклонений, обусловленные регрессией;

$Q_E = \sum_{k=1}^n (y_k - \hat{y}_k)^2$ - суммы квадратов регрессионных остатков;

$\hat{y}_k = a_0 + a_1 x_k$ - значение, предсказанное регрессией.

Регрессионный анализ

- Так как $\hat{y}_k - \bar{y} = (a_0 + a_1 x_k) - (a_0 + a_1 \bar{x}) = a_1 (x_k - \bar{x})$ то объясненное регрессией отклонение от среднего уровня:

$$Q_R = \sum_{k=1}^n (\hat{y}_k - \bar{y})^2 = \sum_{k=1}^n a_1^2 (x_k - \bar{x})^2 = a_1^2 \cdot (n\sigma_x^2)$$

$$= 0,27^2 \cdot 9 \cdot 6,67 = 4,38$$

Полное отклонение от среднего уровня:

$$Q = n \cdot \sigma_y^2 = 9 \cdot 0,84 = 7,56$$

Отклонение, необъясненное регрессией:

$$Q_E = Q - Q_R = 7,56 - 4,38 = 3,18$$

Регрессионный анализ

- Значение F -статистики:

$$F = \frac{Q_R(n-2)}{Q_E} = \frac{4.38 \cdot (9-2)}{3.18} = 9.61$$

- Вывод: так как вычисленное значение F -критерия:

$$F = 9,61 > F_{0,05;1;7} = 5,59,$$

то уравнение регрессии значимо

Регрессионный анализ: коэффициент детерминации

$$R^2 = \frac{Q_R}{Q} = \frac{\sigma_R^2}{\sigma^2}$$

- В силу правила сложения дисперсий для R^2 имеем $0 \leq R^2 \leq 1$; $\sigma^2 = \sigma_R^2 + \sigma_E^2$

- В примере коэффициент детерминации:

$$R^2 = Q_R / Q = \frac{4,38}{7,56} = 0,58$$

- Вывод: предсказанные по регрессии значения объясняют вариацию результативного признака (y) на 58%