Корреляция и корреляционный анализ

Олег Булыгин

Python для анализа данных

Олег Булыгин

IT-аудитор, ментор и наставник

Аккаунты в соц.сетях

fb.com/obulygin91

linkedin.com/in/obulygin

vk.com/obulygin91

@obulygin91

obulygin91@ya.ru

Сегодня на лекции

1. Узнаем как искать и анализировать взаимосвязи в данных
2. Познакомимся с понятием корреляции
3. Научимся предсказывать значение одной переменной по другой

Зависимости в данных

И их виды

1

Олег Булыгин

Статистика

нетология

Вопросы

1. Существует ли зависимость между доходом семьи и ее расходами на питание?
2. Связан ли уровень безработицы в стране с ВВП?
3. Влияет ли количество часов, которые студент тратит на подготовку к экзамену на его итоговую оценку?
4. …

Изучение связи между переменными

Корреляционный и регрессионный анализ предназначены для изучения статистических связей между переменными.

Изучение связи между переменными

Корреляционный анализ

1. Существует ли связь (не причинно-следственная!) между явлениями?
2. Насколько сильная связь между явлениями?

Регрессионный анализ

1. Каков характер связи между явлениями?
2. Построение и исследование регрессионной модели.

Корреляционный анализ

2

Олег Булыгин

Статистика

нетология

Корреляция

Изменения значений одной из величин сопутствуют систематическому изменению значений другой или других величин.

Коэффициент корреляции (линейный коэффициент корреляции Пирсона) показывает:

1. силу линейной взаимосвязи между двумя переменными,
2. направление взаимосвязи (прямая или обратная)

Формула

$$r_{X,Y} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\cdot\left(n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}};$$

Это ковариация двух переменных поделить на их дисперсии.

Величина коэффициента корреляции заключена в пределах -1 <= r <= 1

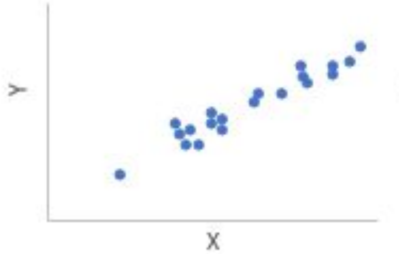
Свойства

1. Если 0 <= r <= 1, то при увеличении значений одной из величин значения другой имеют тенденцию к увеличению (прямая связь)

2. Если -1 <= r <= 0, то при увеличении значений одной из величин значения другой имеют тенденцию к уменьшению (обратная связь)

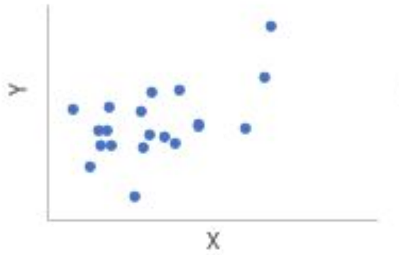
Свойства

1. Чем ближе |r| к единице, тем сильнее линейная связь между случайными величинами, т.е. тем меньше точки рассеяны вокруг прямой.



Сильная

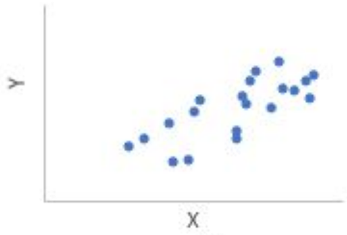
2. |r| = 1 тогда и только тогда, когда когда случайные величины X и Y линейно связаны, т.е. точки лежат на одной прямой.



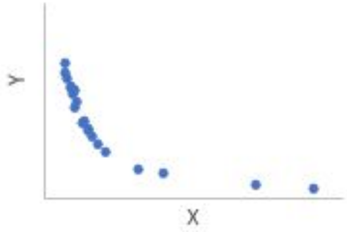
Слабая

Свойства

1. Если |r| = 0, то
    a. связь между случайными величинами либо отсутствует
    b. либо не носит линейного характера

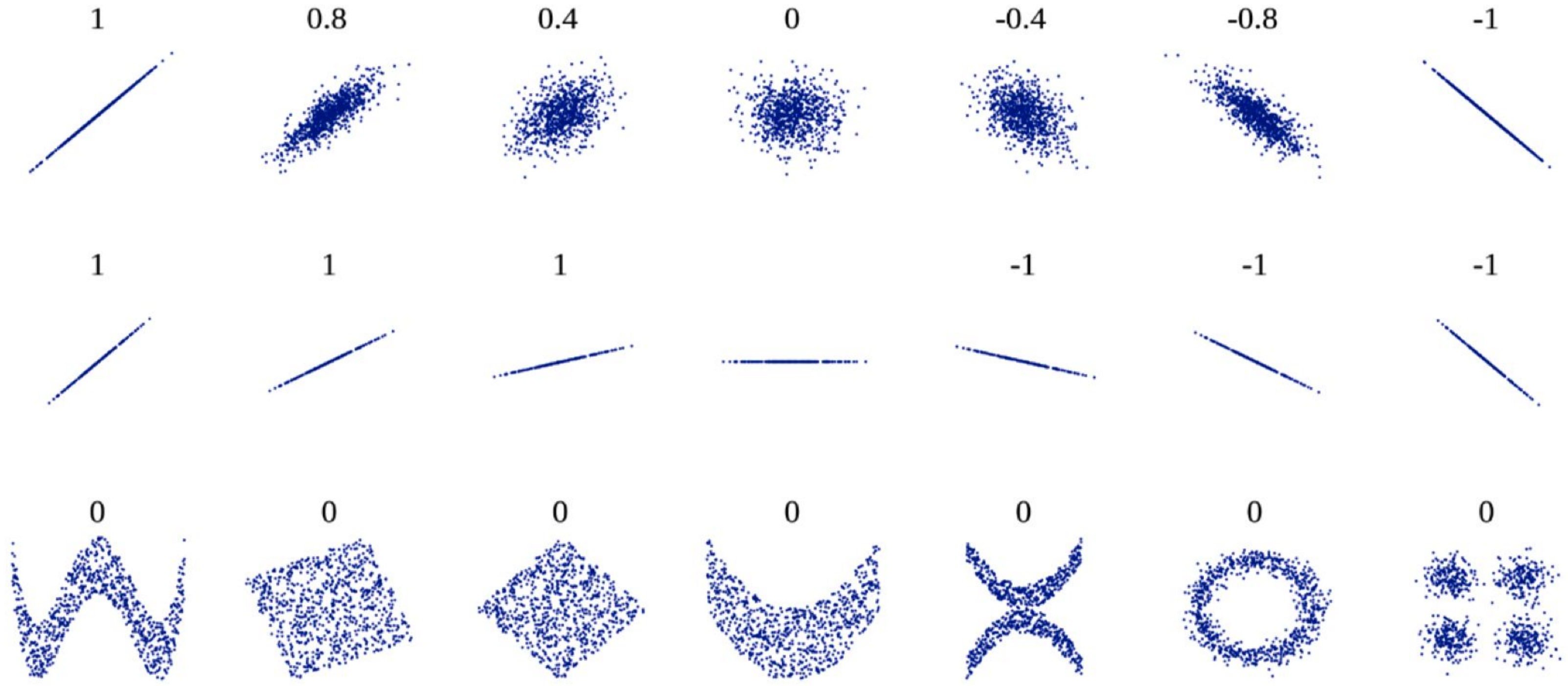


Линейная

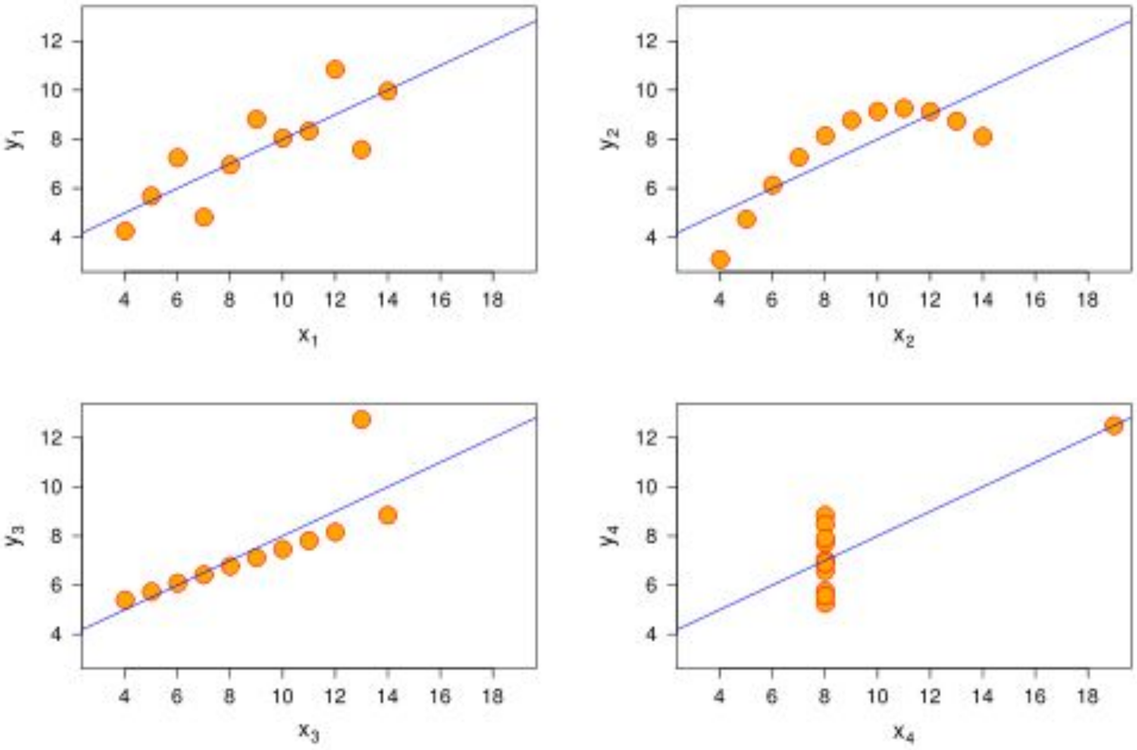


Нелинейная

Примеры



Корреляция - просто число



Квартет Энскомба

Одинаковые:

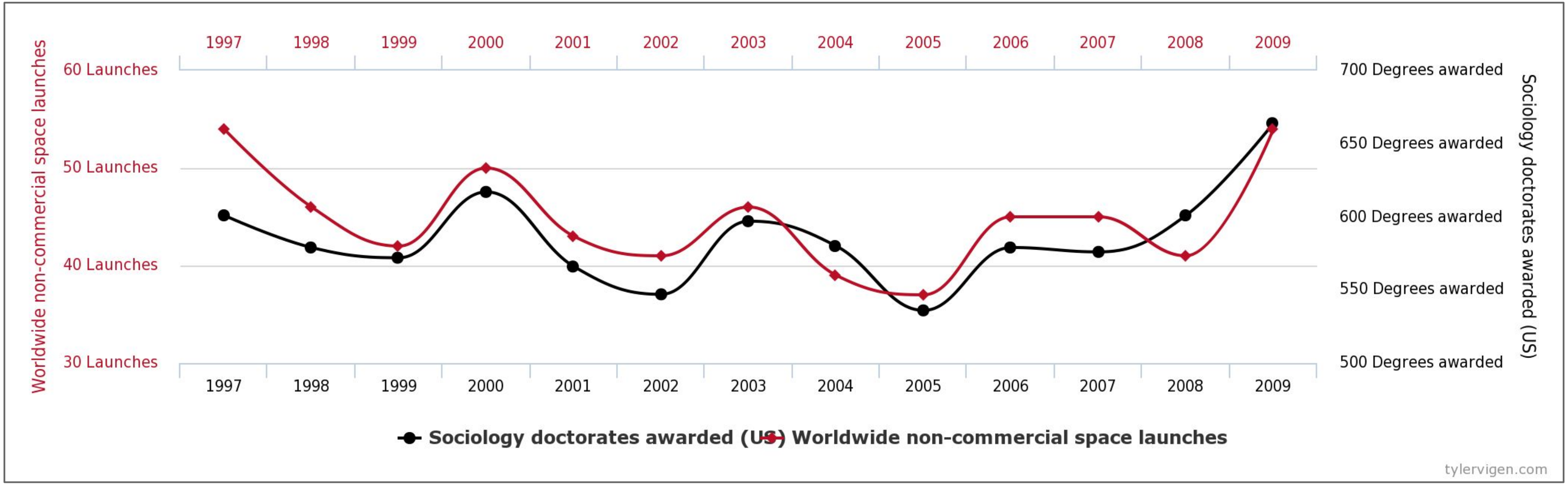
- среднее ***x***,
- среднее ***y***,
- дисперсия ***x***,
- дисперсия ***y***,
- уравнение прямой ***y=ax+b***,
- коэффициент корреляции ***ρ***

«Корреляция не подразумевает причинно-следственной связи»

Пример: Уровень определенного типа холестерина обратно пропорционален риску развития сердечно сосудистых заболеваний. Т.е. чем больше «хорошего» холестерина, тем лучше. Однако, если давать пациентам препараты с таким веществом – это никак не повлияет на болезни сердца.

Ошибка вывода

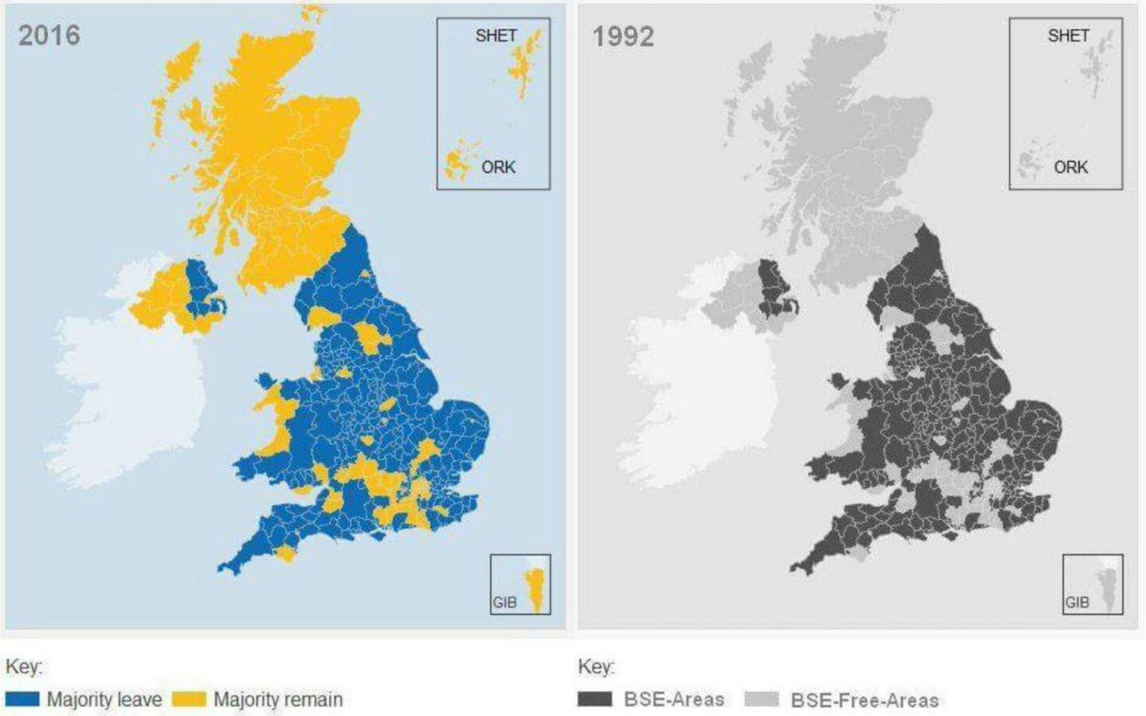
Корреляция не подразумевает причинно-следственных связей!



Ошибка вывода

Корреляция не подразумевает причинно-следственных связей!

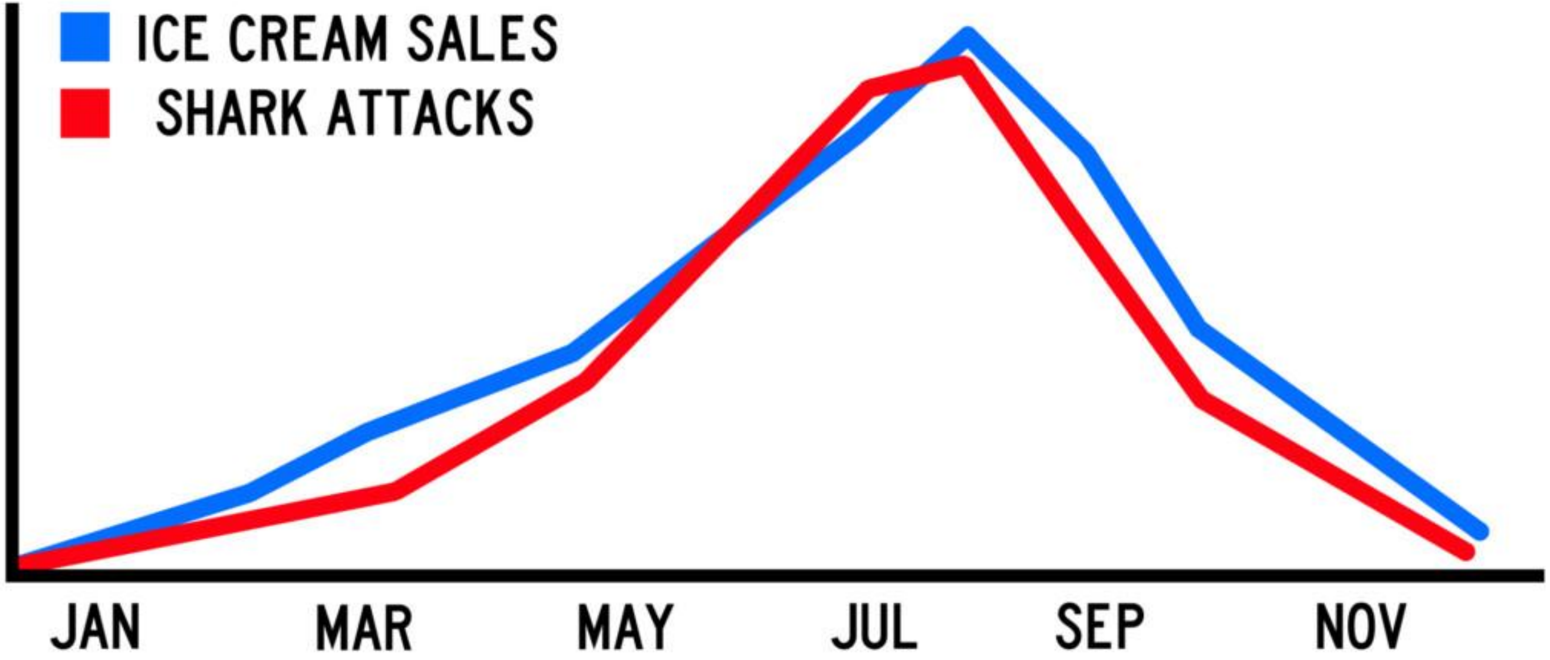
За и против брекзита (2016)



Коровье бешенство (1992)

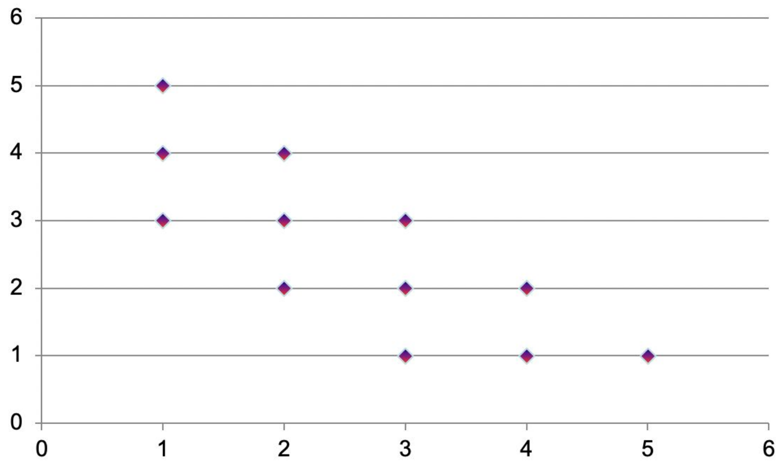
Ошибка вывода

Корреляция не подразумевает причинно-следственных связей!

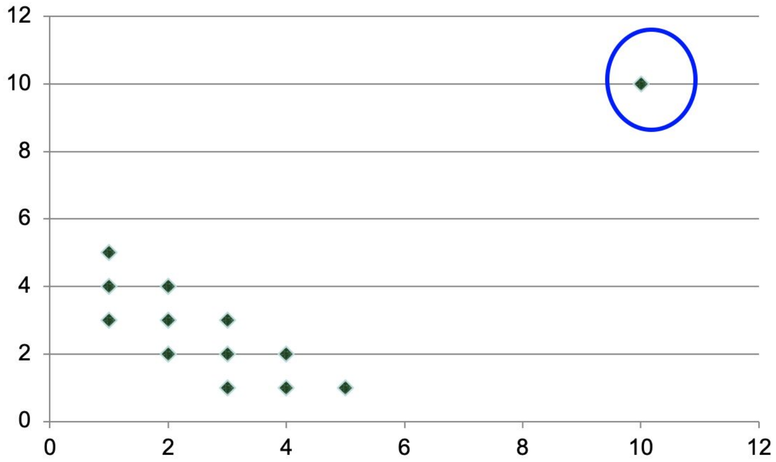


Выбросы

Коэффициент корреляции очень чувствителен к выбросам!



Обратная связь
r=-0,80



Прямая связь
r=0,51

Проблемы коэффициента Пирсона

1. Выбросы
2. Работает только с непрерывными данными (а как же порядковые?)
3. Может испытывать проблемы при не нормальном распределении данных

Ранговый коэффициент корреляции Спирмена

$$\rho = 1 - \frac{6}{n(n^2-1)} \cdot \sum_{k=1}^{n} \left(A_k - B_k\right)^2$$

Ak - ранг k-го наблюдения в первой выборке

Bk - ранг k-го наблюдения во второй выборке

n - число пар наблюдений

Ранговый коэффициент корреляции Кенделла

$$\tau = \frac{2S}{n(n-1)}$$

S – сумма баллов

Баллом +1 оценивается пара рангов, имеющих по обоим показателям одинаковый порядок

Баллом -1 – пара с разным порядком.

Пример

X	Y	Nx	Ny	D=Nx-Ny	d2	+	-
46	45	1	1	0	0	7	0
60	69	2	6	-4	16	2	4
66	59	3	5	-2	4	2	3
68	49	4	2	2	4	4	0
71	54	5	3	2	4	3	0
78	70	6	7	-1	1	1	1
82	58	7	4	3	9	1	0
90	75	8	8	0	0	-	-
					38	20	8

Расчеты

Спирмен:

$$\rho = 1 - \frac{6*38}{8(64-1)} = 1 - 0.453 = 0.547$$

Кенделл:

$$\tau = \frac{2(20-8)}{8(8-1)} = \frac{24}{56} = 0.429$$

Другие меры взаимосвязи

- Коэффициент ассоциации
- Коэффициент контингенции
- Коэффициенты сопряженности Пирсона
- Коэффициент сопряженности Чупрова
- Коэффициент корреляции знаков Фехнера
- ...

Практика

1. Возьмем датасет с boston'ом
2. Посмотрим на имеющиеся в нем корреляции

Регрессионный анализ

3

Олег Булыгин

Статистика

нетология

Регрессионный анализ

Если суточное потребление калорий и вес связаны, то можем ли мы предсказать конкретный вес человека?

Регрессионный анализ – инструмент для количественного предсказания значения одной переменной на основании другой.

Регрессия vs Корреляция

РЕГРЕССИЯ – предсказание одной переменной на основании другой. Одна переменная – независимая, а другая – зависимая.

Пример: чем больше студент занимается перед экзаменом, тем выше его оценка

КОРРЕЛЯЦИЯ показывает, в какой степени две переменные СОВМЕСТНО ИЗМЕНЯЮТСЯ. Нет зависимой и независимой переменных, они эквивалентны.

Пример: рост человека положительно связан с его массой

Регрессионный анализ

По количеству независимых переменных:

- простой (регрессия между двумя переменными);
- множественной (регрессия между зависимой переменной Y и несколькими независимыми переменными (X_1, X_2, ..., X_n)).

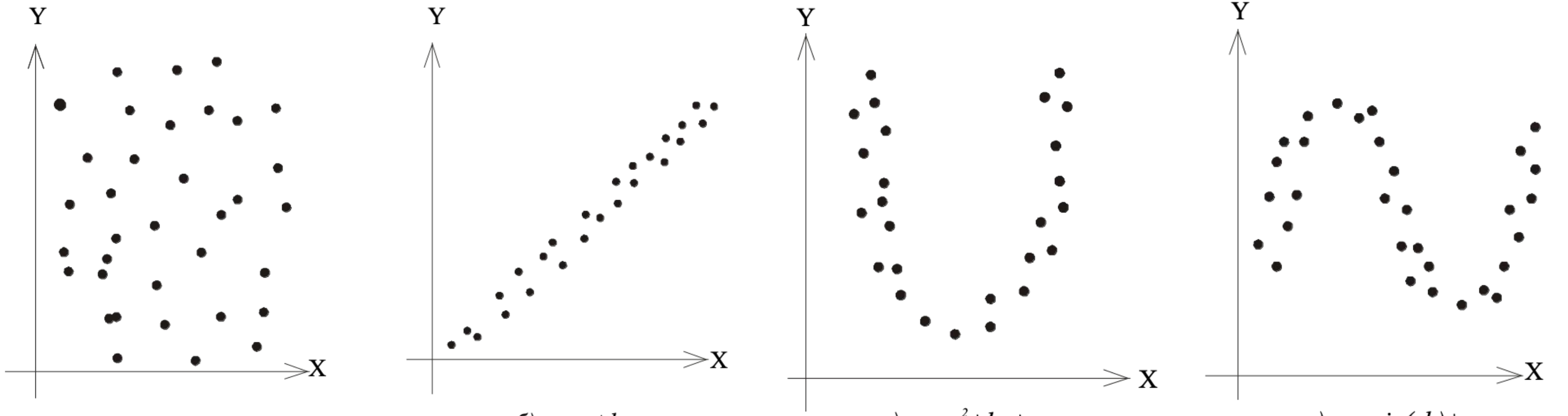
По типу зависимости:

- линейный
- нелинейный

Общий подход к решению

1. Определение формы зависимости
2. Построение модели регрессии
3. Оценка неизвестных значений зависимой переменной

Определение формы зависимости



а) связь отсутствует б) $y=ax+b$ в) $y=ax^2+bx+c$ г) $y=a\sin(xb)+c$

Построение модели регрессии

1. Мы выбрали форму регрессии. Предположим, это прямая линия $y=ax+b$
2. У нас есть выборка точек с измеренными значениями y и x
3. Нужно подобрать наилучшие параметры a, b, которые максимально точно описывают наши данные.
4. Как это сделать?

Построение модели регрессии

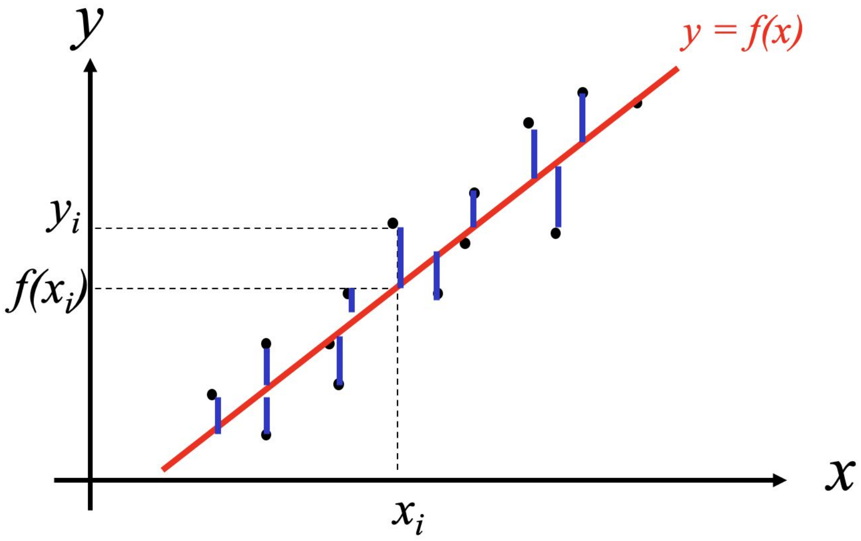
Рассмотрим одну точку - **(y, x)**.

Предсказание нашей модели для этой точки - **y_pred = ax + b**

Ошибка предсказания - **y - y_pred**

Нас интересует именно размер ошибки, а не его знак (+ или -).
Возведем ошибку в квадрат:

(y - y_pred) ^2



Построение модели регрессии

Тогда суммарная ошибка предсказания для всей выборки - сумма квадратов ошибок для каждой точки

S = (y_pred_1 - y1)^2 + (y_pred_2 - y2)^2 + ... + (y_pred_N - yN)^2

Очевидно - модель лучше, если такая суммарная ошибка меньше.

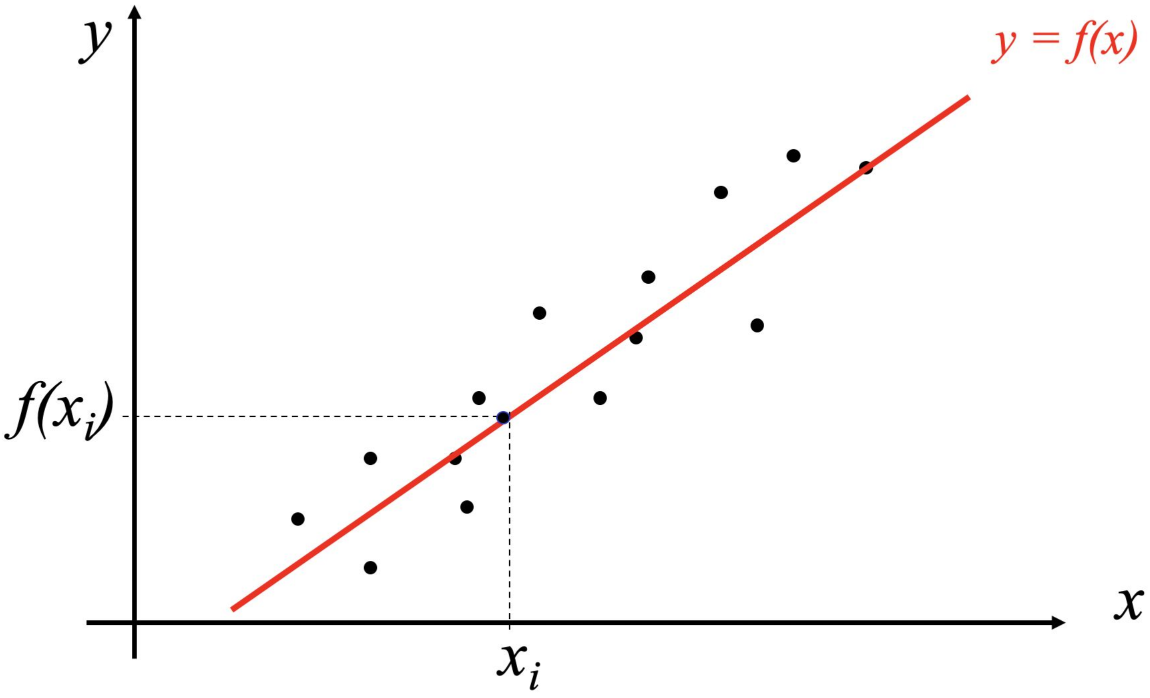
Оптимальные значения a, b для модели регрессии - те, при которых ошибка S достигает своего минимума.

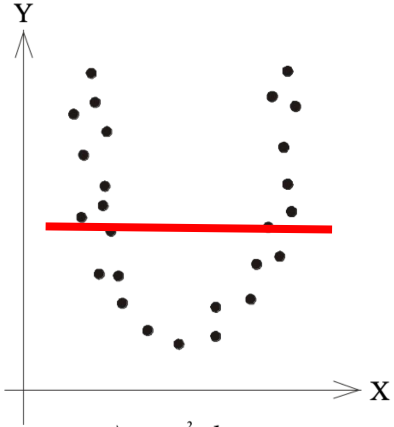
Линейная регрессия

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

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

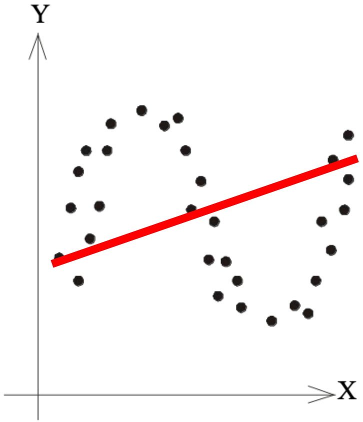
Оценка неизвестных значений





в) y=ax²+bx+c

А хорошо ли получилось?



г) y=asin(xb)+c

Оцениваем адекватность модели

1. Коэффициент детерминации
2. Анализ остатков

Немного порассуждаем

Какая может быть самая простая модель для регрессии?

Немного порассуждаем

Какая может быть самая простая модель для регрессии?

Оценка через среднее

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

Сравним нашу модель с такой наивной

Для этого рассчитаем сумму квадратов ошибок нашей модели:

$$SS_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

И наивной:

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Сравним их

Во сколько раз наши остатки “лучше”, чем остатки наивной модели?

SSres / SStot

Сравним их

Во сколько раз наши остатки “лучше”, чем остатки наивной модели?

SSres / SStot

Коэффициент Детерминации (R2):

R2 = 1 - SSres/SStot

Коэффициент детерминации

доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными

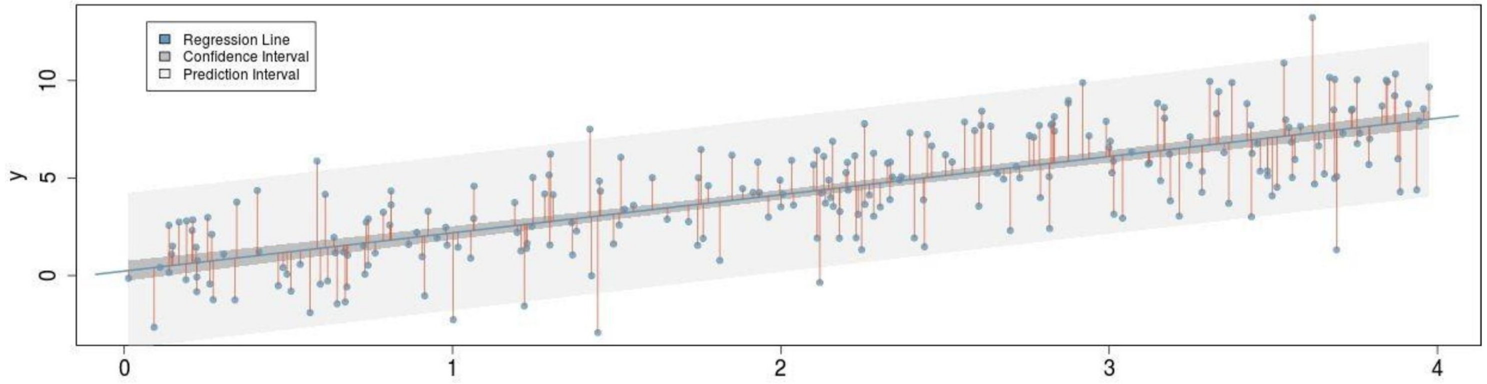
1. 0<=R2<=1;
2. Чем ближе коэффициент детерминации к 1, тем лучше регрессия «объясняет» зависимость данных;

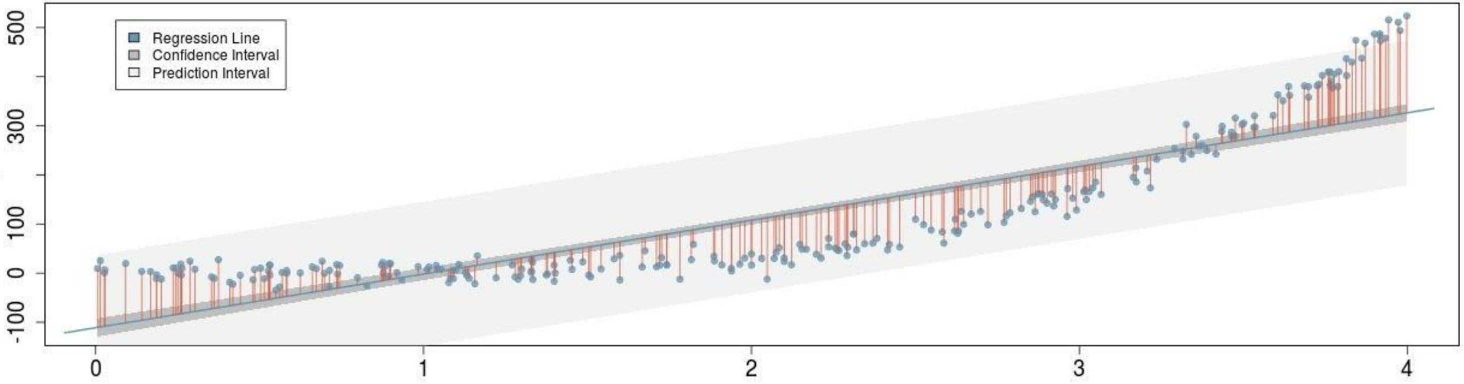
Анализ остатков

Если модель подобрана правильно, то

- остатки будут вести себя достаточно хаотично,
- в остатках не будет систематической составляющей, резких выбросов,
- в чередовании знаков не будет никаких закономерностей.

Анализ остатков





Практика

1. Попробуем построить регрессию, предсказывающую цену MEDV на основе среднего количества комнат в доме RM
2. В качестве инструментов попробуем использовать:
   a. LinearRegression из sklearn
   b. OLS из statsmodels

Итоги

4

Олег Булыгин

Статистика

Что мы узнали сегодня

- Познакомились с понятием корреляции и рассмотрели несколько способов ее расчета
- Узнали, что корреляция не всегда означает наличие причинно-следственной связи в данных
- Научились прогнозировать значение зависимого признака на основе независимых и строить модель линейной регрессии

Домашнее задание

5

Олег Булыгин

Статистика

нетология

Домашнее задание

1. Возьмите датасет Mortality and Water Hardness
   https://www.kaggle.com/ukveteran/mortality-and-water-hardness

   Дополнительно будет выложен в ЛК

   В этом датасете содержатся данные по средней годовой смертности на 100000 населения и концентрации кальция в питьевой воде для 61 большого города в Англии и Уэльсе. Города дополнительно поделены на северные и южные.

Домашнее задание

1. Задача - ответить на вопрос есть ли связь между жёсткостью воды и средней годовой смертностью?
    a. Построить точечный график
    b. Рассчитать коэффициенты корреляции Пирсона и Спирмена
    c. Построить модель линейной регрессии
    d. Рассчитать коэффициент детерминации
    e. Вывести график остатков
2. Сохраняется ли аналогичная зависимость для северных и южных городов по отдельности?
    a. Разделить данные на 2 группы
    b. Повторить аналогичные шаги из пункта 1 для каждой группы по отдельности

Корреляция и корреляционный анализ

Вопросы?

Олег Булыгин