Математические основы искусственного интеллекта. Анализ статистических связей

Солодушкин Святослав Игоревич

Кафедра вычислительной математики и компьютерных наук,
УрФУ имени первого Президента России Б.Н. Ельцина

Март 2022

Постановка задачи

Пусть даны наблюдения за двумя случайными величинами X и Y, которые мы будем трактовать как цену на нефть в долларах за баррель и доход в бюджет в миллионах рублей. Требуется проверить, существует ли статистическая зависимость между этими величинами, и если да, то какая и насколько тесная.
Для ответа на этот вопрос используем корреляционный анализ.

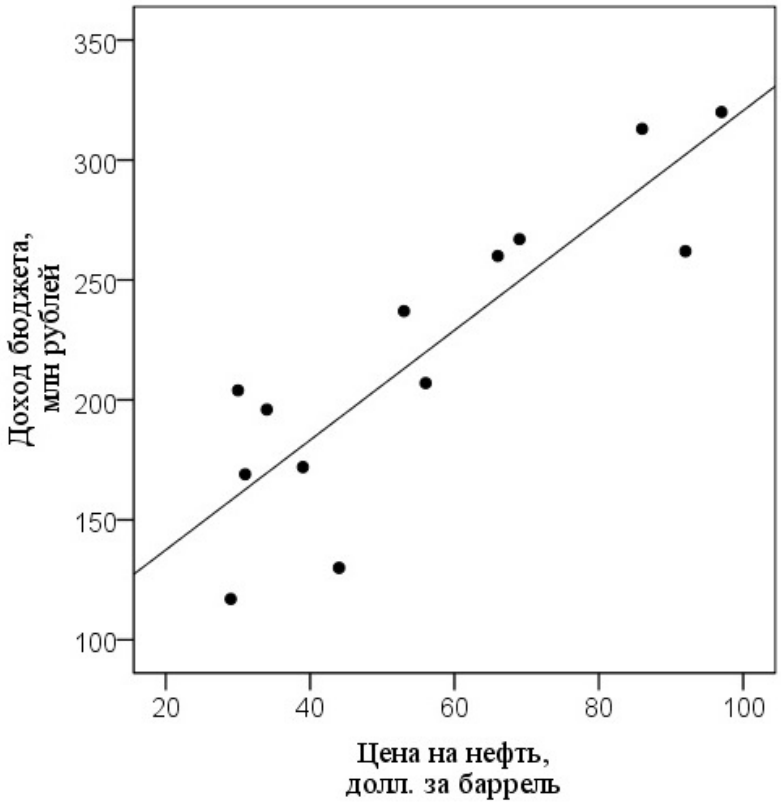
X	30	69	86	56	44	97	53
Y	204	267	313	207	130	320	237

X	66	39	29	34	31	92
Y	260	172	117	196	169	262

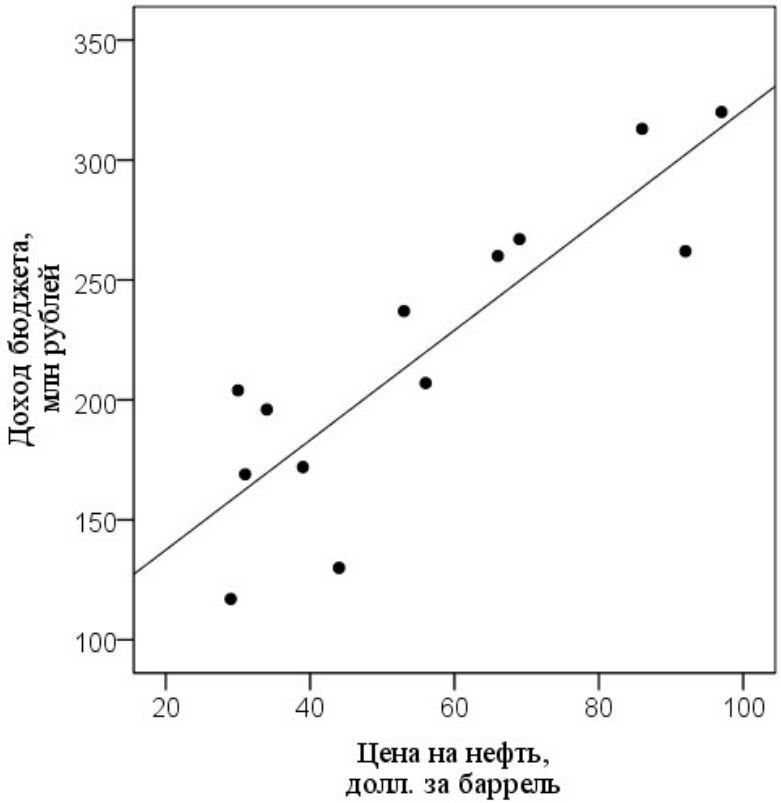
Корреляционный анализ

Определение

Корреляционный анализ — метод обработки статистических данных, с помощью которого измеряется теснота связи между двумя или более переменными.



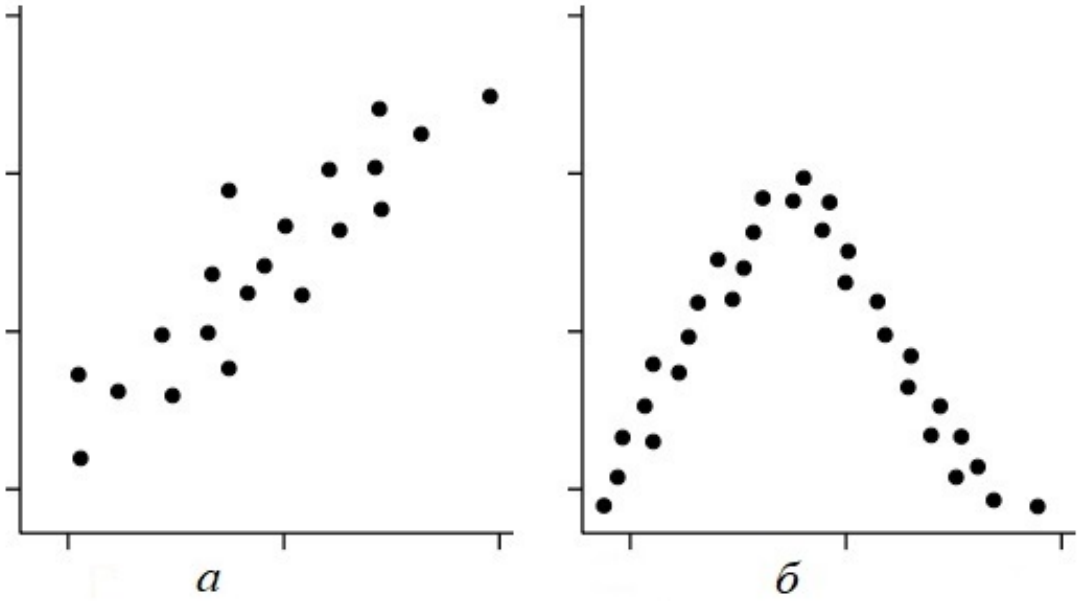
Точки образовали некоторое облако и выстроились вдоль воображаемой наклонной прямой, т. е. подчинены *линейной статистической связи*.



Уравнения прямой: $y = kx + b$. Здесь y — зависимая величина; x — независимая величина; k — угловой коэффициент; b — свободный член.

Если $x = 0$, то $y = b$, т. е. b — это начальное смещение; при увеличении x на единицу y увеличивается на k единиц.

Линейные и нелинейные связи



Корреляционное отношение определяется через отношение межгрупповой дисперсии к общей:

$$\eta^2_{Y|X} = 1 - M\left[\frac{D(Y|X)}{D(Y)}\right],$$

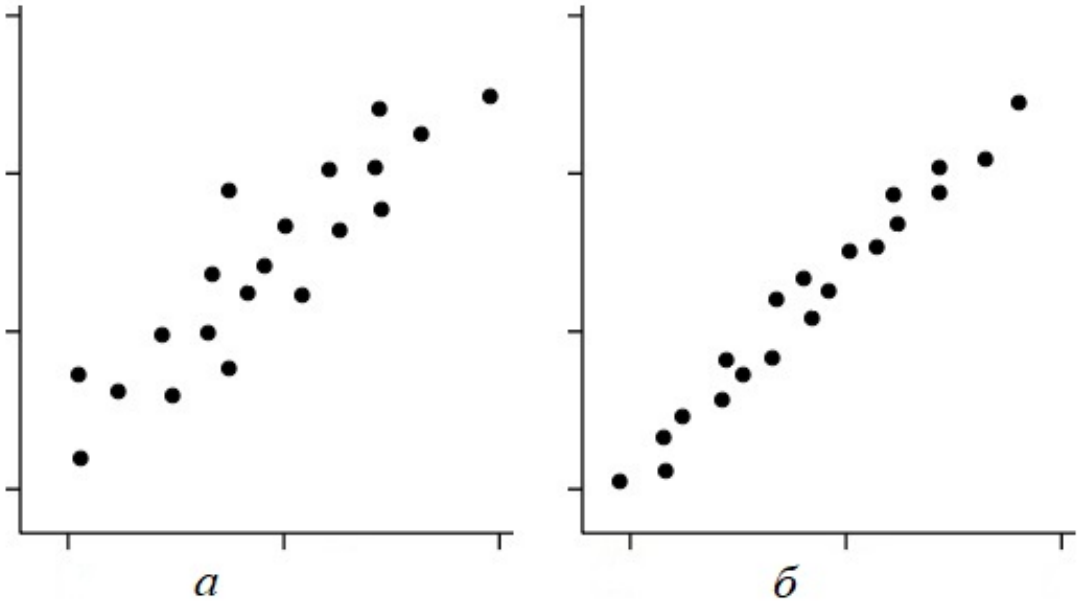
где $D(Y)$ — дисперсия Y; $D(Y|X)$ — условная дисперсия Y при данном X, характеризующая рассеяние Y около условного математического ожидания $M(Y|X)$ при данном значении X.

Корреляция

Определение

Корреляция — это линейная связь между парой случайных величин.

Две случайные величины могут быть связаны более тесной линейной связью, тогда соответствующее корреляционное облако будет узкое, или менее тесной линейной связью, тогда соответствующее корреляционное облако будет широкое.



Коэффициент корреляции

Для оценки тесноты линейной связи между парой случайных величин введем понятие коэффициента корреляции (Пирсона).

Definition

Коэффициент корреляции — это числовая величина, характеризующая тесноту линейной связи между парой случайных величин.

Коэффициент корреляции случайных величин X и Y

$$r_{XY} = \frac{M(XY) - M(X)M(Y)}{\sigma_X \sigma_Y},$$

где $M(X)$ — математическое ожидание X;
σ_X — среднее квадратическое отклонение X;
$M(XY)$ — математическое ожидание произведения X и Y.

Свойства коэффициента корреляции

1. Коэффициент корреляции не превосходит единицы по модулю, $-1 \leq r_{XY} \leq 1$.
2. Если величины связаны строгой линейной связью и с увеличением X величина Y также возрастает, то $r_{XY} = 1$. В этом случае точки корреляционного облака идеально ложатся на прямую, направленную вверх.
3. Если величины связаны строгой линейной связью и с увеличением X величина Y убывает, то $r_{XY} = -1$. В этом случае точки корреляционного облака идеально ложатся на прямую, направленную вниз.
4. Если $r_{XY} = 0$, то величины некоррелированы. В этом случае точки корреляционного облака вдоль прямой не ложатся и обычно хаотично разбросаны.

Зависимость и коррелированность

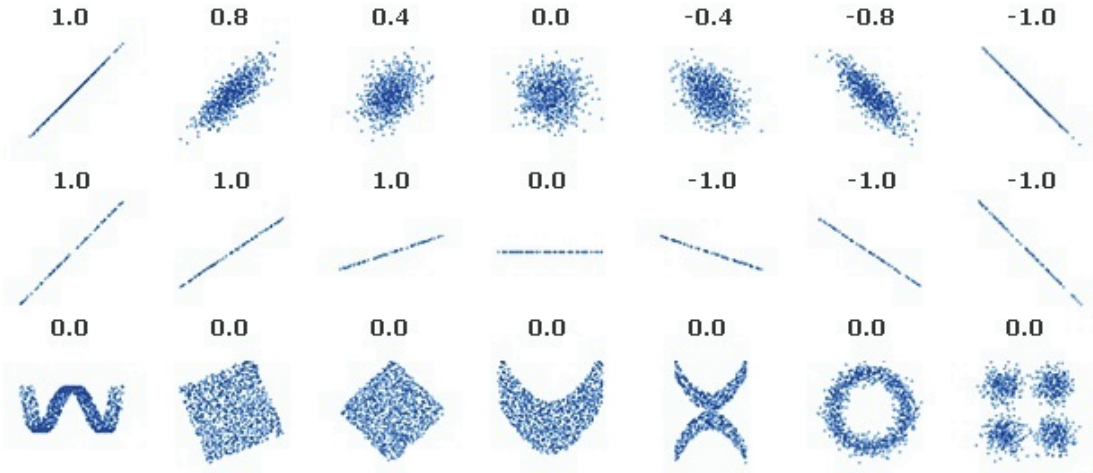
Поскольку коррелированность — это частный случай зависимости (линейная зависимость), то из коррелированности следует зависимость. Обратное неверно: из зависимости не следует коррелированность.

Часто, желая сделать свою речь более наукообразной и показать свою «образованность», люди допускают ошибку, говоря «это некоррелированные величины», имея в виду то, что величины несвязаны.

Зависимость и коррелированность — это не синонимы.

Теснота и сила связи

Если величины связаны линейной связью, то коэффициент корреляции показывает только тесноту корреляционного облака, отсутствие «зашумленности», но не показывает наклон прямой. Коэффициент корреляции не показывает, на сколько изменяется одна величина при изменении другой на единицу. Если такая оценка требуется, то необходимо провести регрессионный анализ.



Оцнека коэффициента корреляции

Мы работаем с выборкой, которая (как мы надеемся) отражает пропорции генеральной совокупности. В данном случае генеральная совокупность является двумерной случайной величиной, первая компонента которой — это цена на нефть, вторая — доход бюджета.

Для оценки коэффициента корреляции находят выборочный коэффициент корреляции r^*_{XY}.

Пусть, например, $r^*_{XY} \approx 0.874$. Означает ли это, что величины коррелированы? Вообще говоря, нет.

X	30	69	86	56	44	97	53
Y	204	267	313	207	130	320	237

X	66	39	29	34	31	92
Y	260	172	117	196	169	262

Формулировка основной и альтернативных гипотез.

Гипотеза H_0 состоит в том, что $r_{XY} = 0$, а полученное отличие расчетного значения r^*_{XY} от нуля — результат случайности, гипотеза H_1 состоит в том, что величины коррелированы, т. е. $r_{XY} \neq 0$.

Этап 2

Назначение уровня значимости α. В наших задачах мы, следуя общепринятым в социологии, психологии и иных гуманитарных науках рекомендациям, положим $\alpha = 0.05$.

Этап 3

Нахождение величины[1] p-value. Если p-value $< \alpha$, то нулевую гипотезу отвергают как противоречащую экспериментальным данным, иначе нет оснований отвергнуть нулевую гипотезу.

В рассматриваемом примере получаем p-value $= 9.35 \cdot 10^{-5}$. Таким образом, нулевую гипотезу отвергаем. Содержательный смысл величины p-value следующий: p-value — это вероятность получить такую (или более тесно связанную, с более узким корреляционным облаком) выборку, если бы величины действительно были бы некоррелированы.

[1] В некоторых статистических пакетах эта величина называется Sig.

Этап 4

Нахождение коэффициента корреляции, если нулевая гипотеза была отвергнута. Только если на этапе 3 было показано, что коэффициент корреляции статистически значимо отличается от нуля, находят его числовое значение. В данном примере получено $r_{XY} = 0.874$. Можно сказать, что доходы бюджета тесно связаны с ценами на нефть.

Ранговая корреляция

Метод вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся переменные. Для переменных, представленных в интервальной шкале, необходимо использовать коэффициент корреляции Пирсона. Если по меньшей мере одна из двух переменных имеет порядковую шкалу либо не является нормально распределенной, то необходимо использовать ранговую корреляцию Спирмена или Кендалла.

Так, например, для выявления связи между степенью артериальной гипертензии и функциональным классом хронической сердечной недостаточности следует вычислять ранговые коэффициенты корреляции Спирмена или Кендалла.

Ложная корреляция

Существуют примеры весьма значимых и высоких корреляций между совершенно не связанными друг с другом величинами.

Определение

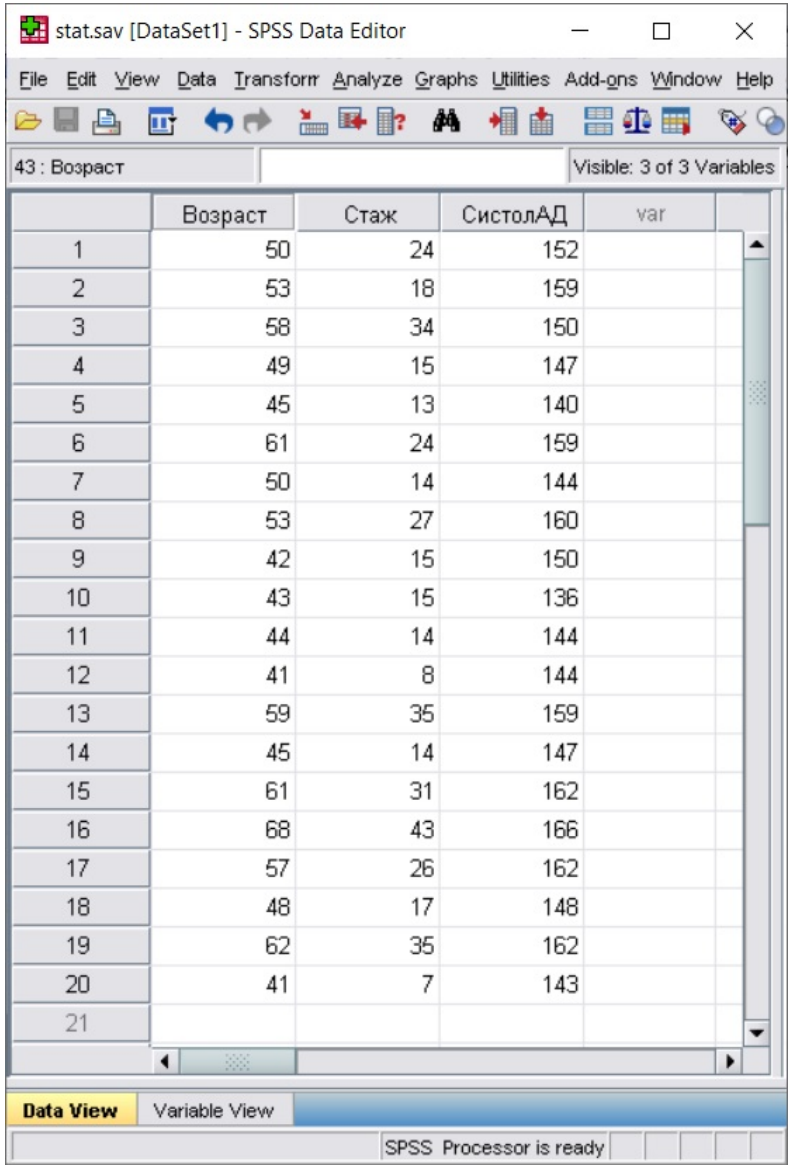
Ложная корреляция — корреляция, которая возникла не в результате прямого соотношения между оцениваемыми переменными, а в результате их связей с третьей переменной (или четвертой, или более); при этом нет никакой связи, объединяющей эти переменные.

Для исключения влияния третьей переменной вычисляют частный коэффициент корреляции при исключенном влиянии третьей.

Ложная корреляция: пример

В моногороде проведено поперечное исследование работников предприятия с целью выявления связи между наличием (а если есть, то между стадией) артериальной гипертензии и вредным стажем рабочих. Предполагается, что вредный стаж на данном производстве является причиной развития артериальной гипертензии.

Ложная корреляция: пример



	Возраст	Стаж	СистолАД	var
1	50	24	152	
2	53	18	159	
3	58	34	150	
4	49	15	147	
5	45	13	140	
6	61	24	159	
7	50	14	144	
8	53	27	160	
9	42	15	150	
10	43	15	136	
11	44	14	144	
12	41	8	144	
13	59	35	159	
14	45	14	147	
15	61	31	162	
16	68	43	166	
17	57	26	162	
18	48	17	148	
19	62	35	162	
20	41	7	143	
21				

Если просто найти коэффициент корреляции Пирсона для стажа и САД, то окажется, что между величинами существует тесная связь: $r_{\text{стаж, САД}} = 0.805$, $p < 0.001$, которая может быть неправильно интерпретирована как производственная обусловленность артериальной гипертензии у данных рабочих.

В моногородах связь возраста и стажа особенно сильная: $r_{\text{возраст, стаж}} = 0.924$, $p < 0.001$, а потому можно предположить, что у людей, проработавших на производстве более 30 лет, артериальная гипертензия возникла не в связи с вредным производством, а в силу их возраста.

Найдем частный коэффициент корреляции стажа и САД при исключенном влиянии возраста: $r_{\text{стаж, САД / возраст}} = 0.047$, $p = 0.849$, он недостоверно отличается от нуля. Это показывает, что имеет место ложная корреляция стажа и САД.

Задания для обсуждения

Пусть даны две случайные величины X и Y. Корреляционное отношение $\eta^2_{Y|X}$ — это число или функция? Если функция, то от чего? От случайной величины X от числа x?

Рассматривая пожары в конкретном городе, можно выявить весьма высокую корреляцию между ущербом, который нанес пожар, и количеством пожарных, участвовавших в ликвидации пожара, причем эта корреляция будет положительной.

Следует ли отсюда вывод, что увеличение количества пожарных приводит к увеличению причиненного ущерба? Можно ли для минимизации ущерба от пожаров ликвидировать пожарные бригады?

Задания для обсуждения

Профессор Р. А. Шамойлова учит студентов проводить корреляционный анализ. На лекции она дала следующее определение: «Корреляция — это статистическая зависимость между случайными величинами, не имеющая строгого функционального характера, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой». Согласны ли вы с профессором?

Также профессор утверждает, что для расчета коэффициента корреляции Пирсона необходимо, чтобы совокупность значений всех факторных и результативных признаков подчинялась многомерному нормальному распределению. Не ошибается ли она?

Задания для обсуждения

Далее Р. А. Шамойлова утверждает, что при проведении корреляционного анализа первым делом выборку нужно проверить на нормальность, и в случае если объем выборки недостаточен для проведения формального тестирования, то закон распределения определяется визуально на основе корреляционного поля. Если в расположении точек на этом поле наблюдается линейная тенденция, то исходные данные подчиняются нормальному закону распределения. Права ли профессор на этот раз?

Задания для обсуждения

В учебном пособии «Основы теории статистики» на стр. 88 находим «определение».
Корреляционная связь — это связь, выявленная при большом числе наблюдений между одним и тем же значение и разными значениями в виде определенной зависимости, которая предполагает следующее соотношение — каждому значению соответствует среднее значение результативного признака/ов .

https://elar.urfu.ru/bitstream/10995/34746/1/978-5-7996-1520-8.pdf

Что не так с этим «определением»?

Задания для обсуждения

В учебном пособии по спортивной метрологии В. В. Афанасьева[2] находим следующее определение: «Коэффициент корреляции — это статистический показатель зависимости двух случайных величин. Коэффициент корреляции может принимать значения от -1 до $+1$. При этом значение -1 будет говорить об отсутствии корреляции между величинами, 0 — о нулевой корреляции, а $+1$ — о полной корреляции величин. То есть чем ближе значение коэффициента корреляции к $+1$, тем сильнее связь между двумя случайными величинами».

Объясните, почему это определение неверное. Какие еще здесь ошибки?

[2] Спортивная метрология : учебник для среднего проф. образования.