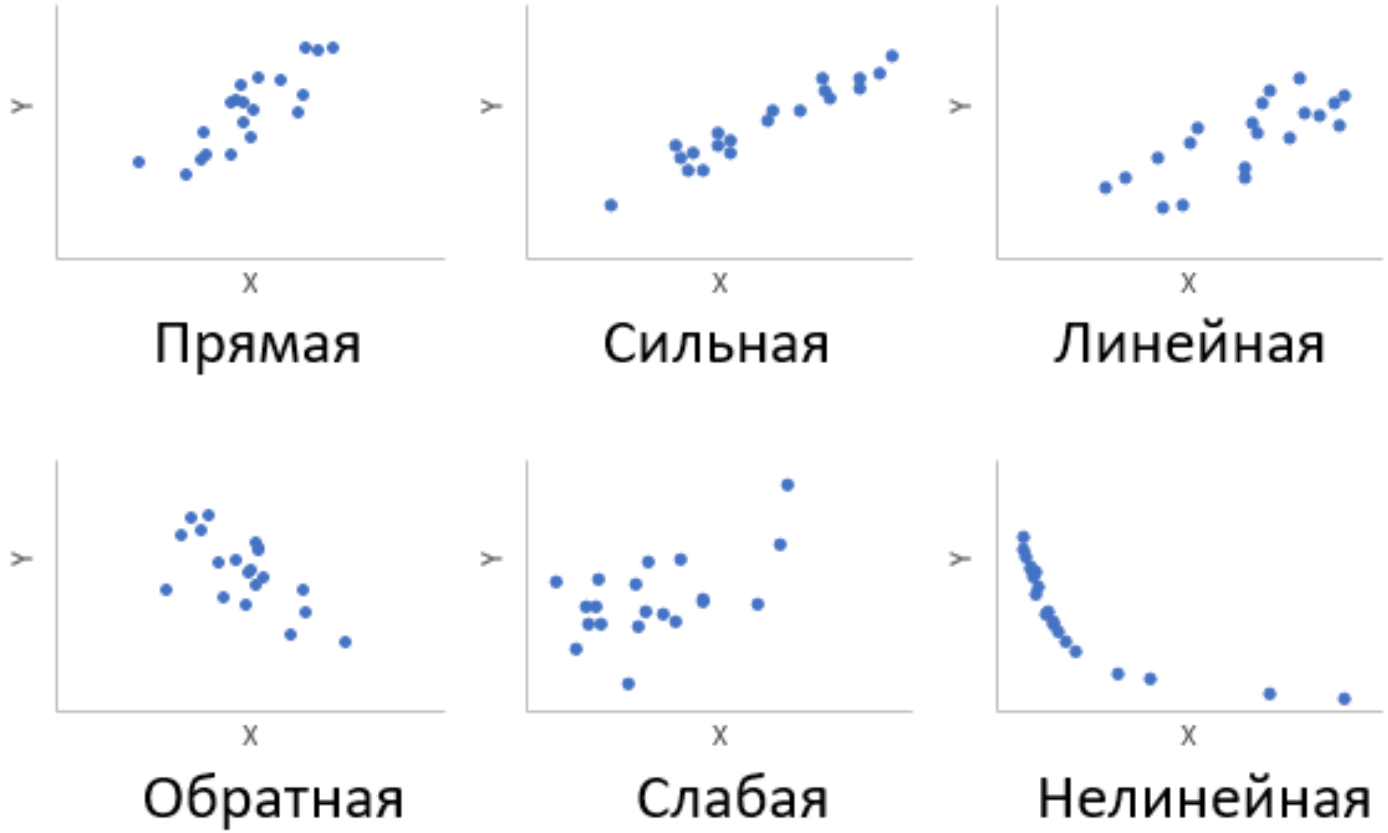
IT-Academy

Корреляция. Метод наименьших квадратов

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ

КОРРЕЛЯЦИЯ

Взаимосвязь двух переменных проявляется в **совместной вариации**: при изменении одного показателя имеет место тенденция изменения другого. Такая взаимосвязь называется **корреляцией.**



КОРРЕЛЯЦИЯ

IT-Academy

Линейность корреляции проявляется в том, что точки расположены вдоль прямой линии. Положительный или отрицательный наклон такой линии определяется направлением взаимосвязи.

$$VAR(X) = \frac{\sum(X_i - \bar{X})^2}{n-1} = \frac{\sum(X_i - \bar{X})(X_i - \bar{X})}{n-1}$$

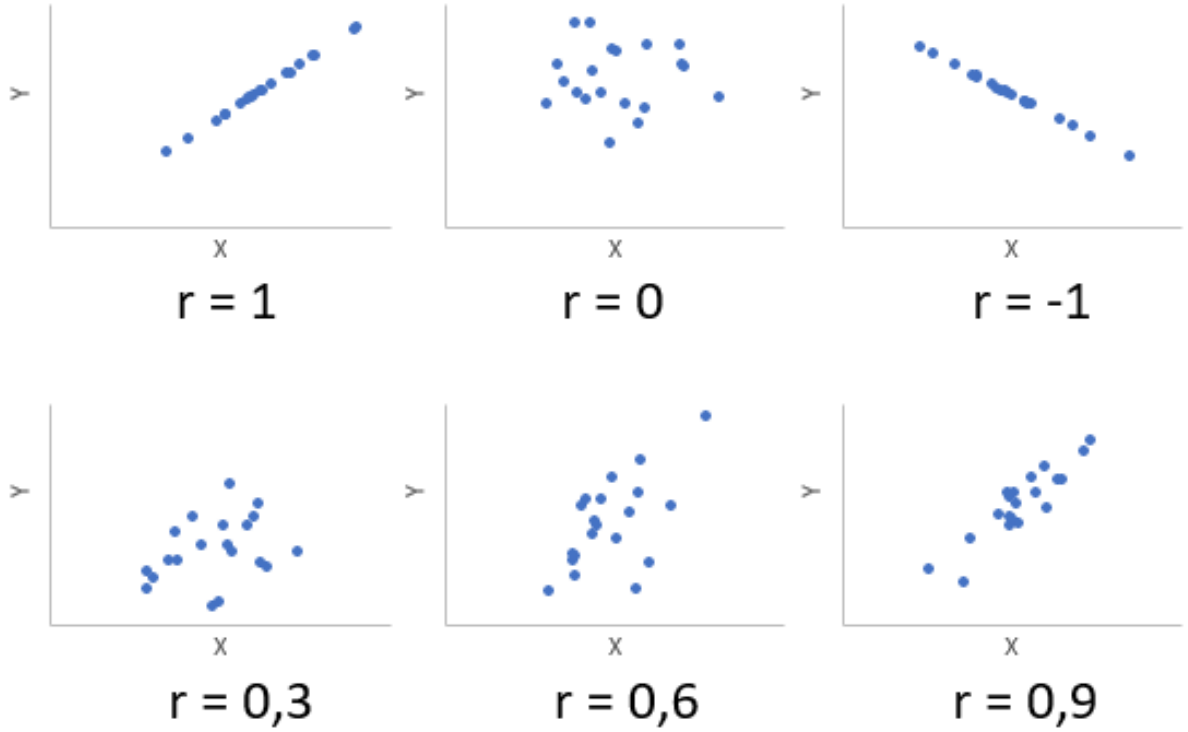
Квадрат отклонения от средней измеряет вариацию показателя как бы относительно самого себя. Если второй множитель в числителе заменить на отклонение от средней второго показателя, то получится совместная вариация двух переменных, которая называется **ковариацией**.

$$COV(X, Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

КОРРЕЛЯЦИЯ

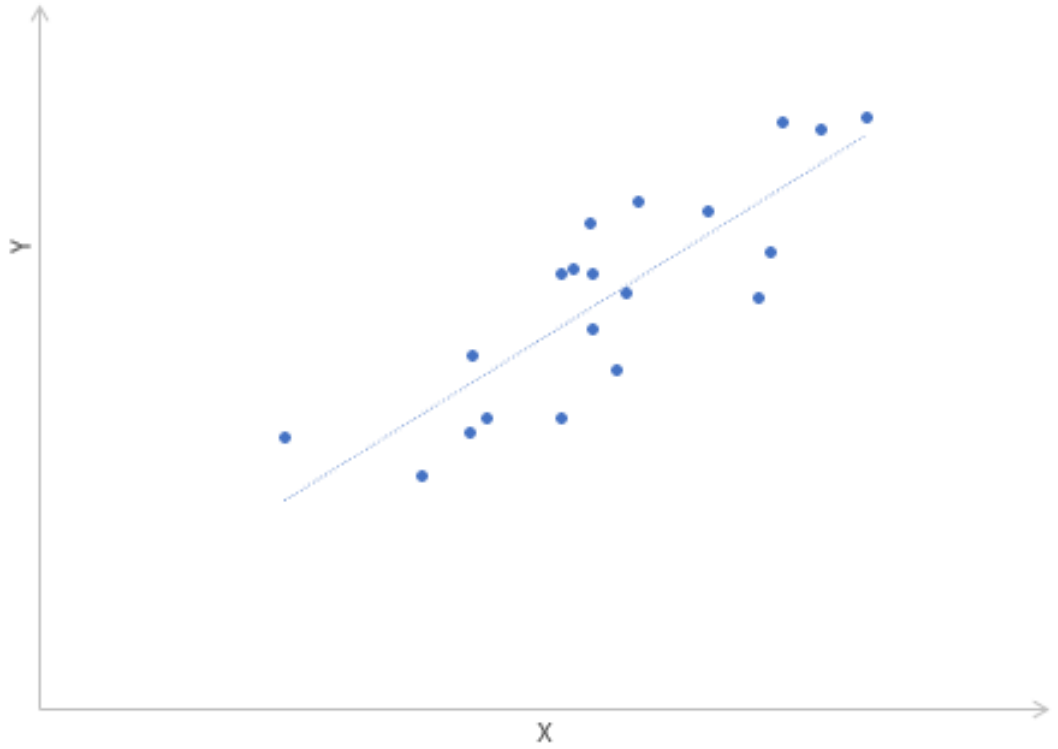
Для получения стандартизованной величины тесноты взаимосвязи нужно избавиться от единиц измерения путем деления ковариации на произведение стандартных отклонений обеих переменных. В итоге получится формула коэффициента корреляции Пирсона.

$$r = \frac{COV_{xy}}{s_x s_y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_x s_y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$



КОРРЕЛЯЦИЯ

Роль формальной интерпретации выполняет квадрат коэффициента корреляции *r2*, который называется **коэффициентом детерминации**, и обычно применяется при оценке качества **регрессионных моделей**.



КОРРЕЛЯЦИЯ

Важные замечания

1. Коэффициент корреляции Пирсона чувствителен к выбросам. Одно аномальное значение может существенно исказить коэффициент. Поэтому перед проведением анализа следует проверить и при необходимости удалить выбросы. Другой вариант – перейти к ранговому коэффициенту корреляции Спирмена. Рассчитывается также, только не по исходным значениям, а по их рангам.

2. Синоним корреляции – это взаимосвязь или совместная вариация. Поэтому наличие корреляции ($r \neq 0$) еще не означает причинно-следственную связь между переменными. Вполне возможно, что совместная вариация обусловлена влиянием третьей переменной. Совместное изменение переменных без причинно-следственной связи называется **ложная корреляция**.

3. Отсутствие линейной корреляции ($r = 0$) не означает отсутствие взаимосвязи. Она может быть нелинейной. Частично эту проблему решает ранговая корреляция Спирмена, которая показывает совместный рост или снижение рангов, независимо от формы взаимосвязи

КОРРЕЛЯЦИЯ. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Уравнение множественной регрессии

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon$$

Цель множественной регрессии:

- Построить модель с большим числом признаков, определив влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый фактор.

Спецификация модели включает в себя два круга вопросов:

- отбор признаков;

- выбор вида уравнения регрессии.

КОРРЕЛЯЦИЯ

IT-Academy

Два этапа отбора признаков:

- исходя из сущности проблемы;
- на основе корреляционной матрицы и t - статистики параметров регрессии

1) Проверка парной корреляции.

Принцип исключения факторов:

- Если две переменные явно коллинеарны ($r_{x_i x_j} \geq 0,7$), то одну из них исключаем.
- Включаем фактор, имеющий наименьшую тесноту связи с другими факторами

2) Оценка мультиколлинеарности факторов (когда более, чем два фактора связаны между собой линейной зависимостью):

- Проверка гипотезы H_0:

$$|R| = |(r_{x_i x_j})| = 1,$$

R – матрица коэффициентов корреляции

Чем ближе к 1 определитель матрицы межфакторной корреляции, тем меньше мультиколлинеарность факторов

КОРРЕЛЯЦИЯ

IT-Academy

Пути преодоления сильной межфакторной корреляции

Исключение одного или нескольких факторов

Преобразование факторов для уменьшения корреляции между ними

- Переход к первым разностям
- Переход к линейным комбинациям (метод главных компонент)

Переход к совмещенным уравнениям регрессии

Переход к уравнениям приведенной формы

IT-Academy

Пример

Дана матрица парных коэффициентов корреляции зависимости :

$$y = f(x, z, u)$$

	y	x	z	u
y	1	0,8	0,7	0,6
x	0,8	1	0,8	0,5
z	0,7	0,8	1	0,2
u	0,6	0,5	0,2	1

КОРРЕЛЯЦИЯ

IT-Academy

Файл Корреляция.ipynb

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

IT-Academy

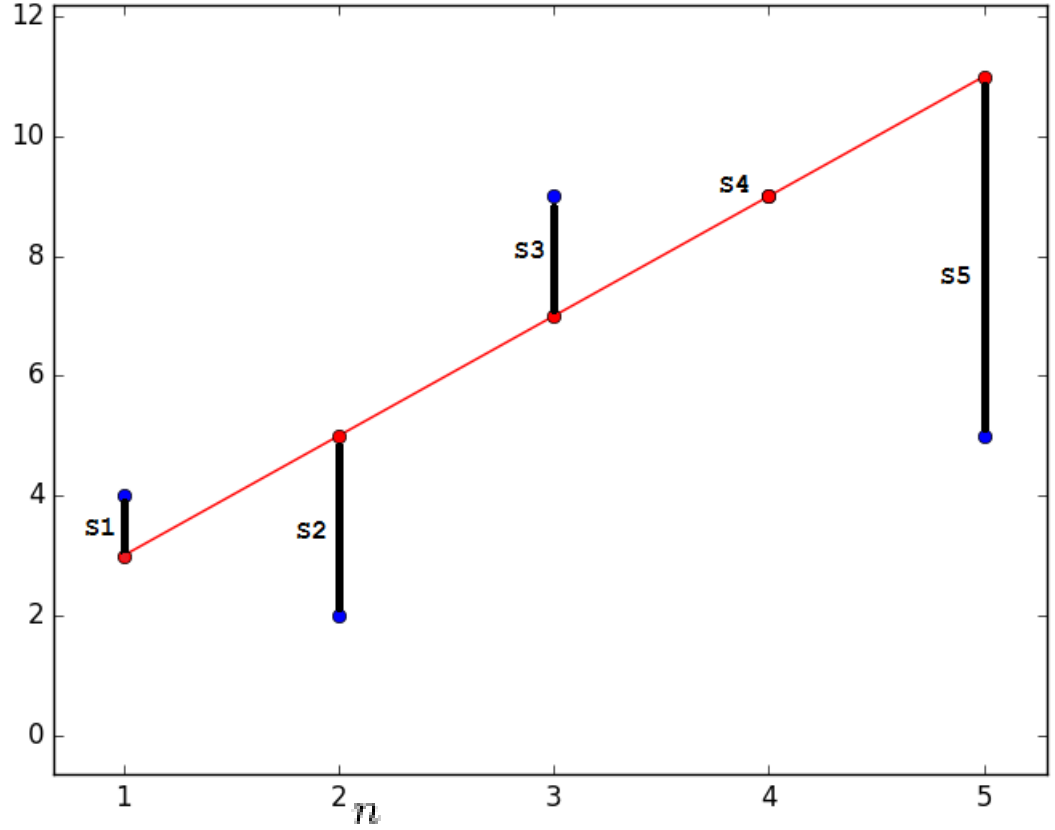
Линейная регрессия — метод восстановления зависимости между двумя переменными. Линейная означает, что мы предполагаем, что переменные выражаются через уравнение вида: $y = ax + b + \varepsilon$

Модель (гипотеза) регрессии запишется следующим образом: $\hat{y} = \theta_1 + \theta_2 x$

θ — неизвестные параметры — основная задача эти параметры отыскать, а x — свободная переменная, ее значения нам известны.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Суть МНК заключается в том, чтобы отыскать такие параметры θ, чтобы предсказанное значение было наиболее близким к реальному.



Математически это выглядит так: $\sum_{i=0}(y_i - \hat{y}_i)^2 \rightarrow \min_{\theta}$

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

$$\Theta = (A^T A)^{-1} A^T Y$$

Часть $(A^T A)^{-1} A^T$ называют псевдообратной матрицей.

Проблема заключается в том, что система может не иметь решений — иначе, у матрицы A может не существовать обратной матрицы. Простой пример системы без решения — любые три\четыре\n точки не на одной прямой\плоскости\гиперплоскости — это приводит к тому, что матрица A становится неквадратной, а значит по определению нет обратной матрицы

Невозможно построить линию через эти три точки — можно лишь построить примерно верное решение.

Такое отступление — это объяснение того, зачем вообще понадобился МНК. Минимизации функции стоимости (функции потерь) и невозможность (ненужность, вредность) найти абсолютно точное решение — одни из самых базовых идей, что лежат в основе нейронных сетей.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

IT-Academy

Файл MHK.ipynb

Последовательность действий:
1) Сгенерировать набор экспериментальных данных.
2) Создать матрицу A.
3) Найти псевдообратную матрицу .
4) Найти θ .

После этого задача будет решена — у нас в распоряжении будут параметры прямой линии, наилучшим образом обобщающей экспериментальные данные. Иначе, у нас окажутся параметры для прямой, наилучшим образом выражающей линейную зависимость одной переменной от другой — именно это и требовалось.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

IT-Academy

Файл MHK.ipynb

Последовательность действий:
1) Сгенерировать набор экспериментальных данных.
2) Создать матрицу A.
3) Найти псевдообратную матрицу .
4) Найти θ .

После этого задача будет решена — у нас в распоряжении будут параметры прямой линии, наилучшим образом обобщающей экспериментальные данные. Иначе, у нас окажутся параметры для прямой, наилучшим образом выражающей линейную зависимость одной переменной от другой — именно это и требовалось.

IT-Academy

Корреляция. Метод наименьших квадратов

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ