ЛАБОРАТОРНАЯ РАБОТА № 8. УРАВНЕНИЕ РЕГРЕСИИ, ПОДГОНКА ЛИНИИ РЕГРЕСИИ. ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ РЕГРЕССИИ.

Уравнение регрессии: Что это такое и как его использовать

Статистические определения > Что такое уравнение регрессии?

Уравнение регрессии: Обзор

Уравнение регрессии используется в статистике для того, чтобы выяснить, какая связь, если таковая существует, существует между наборами данных. Например, если каждый год измерять рост ребенка, то можно обнаружить, что он растет примерно на 3 дюйма в год. Эта тенденция (которая растет на 3 дюйма в год) может быть смоделирована с помощью уравнения регрессии. Фактически, большинство вещей в реальном мире (от цен на газ до ураганов) можно смоделировать с помощью некоего уравнения, что позволяет нам предсказывать будущие события.

Линия регрессии – это “самая подходящая” линия для ваших данных. По сути, вы рисуете линию, которая наилучшим образом представляет точки данных. Она представляет собой среднее арифметическое того, где выравниваются все точки. В линейной регрессии линия регрессии является абсолютно прямой линией:

линия регрессии

Линия линейной регрессии.

Линия регрессии представлена уравнением. В данном случае уравнение равно -2.2923x + 4624.4. Это означает, что если построить график уравнения -2.2923x + 4624.4, то линия будет представлять собой грубую аппроксимацию для Ваших данных.

Не очень распространено, чтобы все точки данных действительно попадали на линию регрессии. На рисунке выше точки немного рассеяны вокруг линии. На следующем изображении точки падают на линию. Изогнутая форма этой линии является результатом полиномиальной регрессии, которая укладывает точки в уравнение полинома.

В результате полиномиальной регрессии получается кривая линия.

Результатом полиномиальной регрессии является кривая линия.

Регрессия и линии прогнозирования

Регрессия полезна, так как позволяет делать прогнозы о данных. Первый график выше – с 1995 по 2015 год. Если вы хотите предсказать, что произойдет в 2020 году, вы можете поместить его в уравнение:

-2.2923(2020)+4626.4 = -4.046.

Отрицательное выпадение осадков не имеет особого смысла, но можно сказать, что до 2020 года осадки выпадут на 0 дюймов. Согласно этой конкретной линии регрессии, рано или поздно это произойдет в 2018 году:

-2.2923(2018)+4626.4 = 0.5386

-2.2923(2019)+4626.4 = -1.7537

Для чего нужно уравнение регрессии?

Уравнения регрессии могут помочь вам понять, подходят ли ваши данные для уравнения. Это чрезвычайно полезно, если вы хотите сделать прогноз на основе своих данных – как будущих прогнозов, так и указаний на прошлое поведение. Например, вы можете захотеть узнать, сколько ваших сбережений будет стоить в будущем. Или, возможно, вы захотите предсказать, сколько времени понадобится на выздоровление от болезни.

Существуют различные типы уравнений регрессии. К наиболее распространенным относятся экспоненциальная линейная регрессия и простая линейная регрессия (для адаптации данных к экспоненциальному уравнению или линейному уравнению). В элементарной статистике уравнение регрессии, с которым вы, скорее всего, столкнетесь, является линейной формой.

Расчет линейной регрессии

Есть несколько способов найти линию регрессии, даже вручную и с помощью технологий, таких как Excel (см. ниже). Поиск линии регрессии очень скучен вручную. Следующее видео иллюстрирует шаги:

Линию регрессии также можно найти в калькуляторах TI:

TI 83 Регрессия.

Как выполнять регрессию TI-89.

Уравнение линейной регрессии показано ниже.

Обратная сторона регрессионного анализа

Для того, чтобы данные вписались в уравнение, необходимо сначала понять, какая общая схема подходит для данных. Общие шаги для выполнения регрессии включают в себя составление дисперсионной диаграммы, а затем гипотезу о том, какой тип уравнения может быть наиболее подходящим. Затем можно выбрать наилучшее уравнение регрессии для задания.

Однако, как видно на следующем рисунке, не всегда легко выбрать подходящее уравнение регрессии, особенно при работе с реальными данными. Иногда получаются “шумные” данные, которые, кажется, не подходят ни под одно уравнение. Если большинство данных, кажется, следуют шаблону, вы можете пропустить пропуски. На самом деле, если игнорировать промахи, данные, кажется, моделируются экспоненциальным уравнением.

Как вес и смещение влияют на положение линии

Дадим термину “вес” новое название — “наклон” (на это есть причина).

Теперь вспомним, что вышеприведенное уравнение состоит из двух компонентов:

- наклон;
- y-пересечение (или смещение).

Если мы попробуем нарисовать линию, используя эти две метрики, то получим нечто подобное:

Пример того, как будет выглядеть линия y = 0.5x + 2

Наклон показывает, насколько крута линия (отсюда и название “наклон”), а y-пересечение — место, где находится линия. Первый определяется как подъем, поделенный на спуск, а второе — как точка пересечения линии с осью y.

Итак, что означает уравнение y = 0,5x + 2?

Если наклон равен 0,5, это означает следующее: когда мы движемся вдоль этой линии, смещаясь вправо на каждую единицу, мы смещаемся на 0,5 единицы вверх. Наклон может быть нулевым, если мы не двигаемся вверх, или отрицательным, если мы двигаемся вниз.

Если провести любую параллельную линию к линии на рисунке выше, то эта линия также будет подниматься на 0,5 единицы на каждую единицу движения вправо.

Именно здесь и вступает в дело y-пересечение. Оно показывает, где линия пересекает ось y. Эта конкретная линия пересекает ось x на высоте 2 — это и есть y-пересечение.

Другими словами, наклон линии говорит о **направлении**, в котором она указывает, а y-пересечение — о **местоположении** линии.

Вот что произойдет, если мы изменим наклон и смещение

Освежим в памяти нашу задачу по прогнозированию цен на жилье, а также возьмем вышеприведенные сдвиги линий для изменения наклона (цена за комнату) и базовой цены (базовая цена дома).

Если мы добавим еще несколько деталей к сдвигам на рисунке выше, то получится нечто подобное:

- Если **увеличить наклон линии**, она будет **вращаться против часовой стрелки**.
- Если **уменьшить наклон линии,** то линия будет **вращаться по часовой стрелке**.

Эти вращения происходят в точке пересечения линии и оси y.

- Если мы **увеличим y-пересечение линии**, то линия будет **перемещена вверх**.
- Если мы **уменьшим y-пересечение линии**, линия будет **перемещена вниз**.

Теперь, когда у нас есть все “ингредиенты” — наклон, y-пересечение и уравнение линии, — можно переходить к практической части.

Простой способ перемещения линии ближе к множеству точек (к одной из точек за раз)

Способ действительно простой, и это можно понять, посмотрев на рисунок ниже:

Различные случаи реакции алгоритма на одну точку

Вспомните наше уравнение для прогнозирования цен на жилье. В нем y — это цена дома, а x — количество комнат. Следовательно, каждая точка данных будет определенной координатой (r,p).

Представим, что нужно написать псевдокод для простого способа перемещения. Вот что у нас есть.

Входные данные:

- Линия с наклоном m, y-пересечение b и уравнение $\hat{p}$=mr+b.
- Точка с координатами (r,p).

Выходные данные:

- Линия с уравнением $\hat{p}$=m'r+b, которая находится ближе к точке (здесь знак над "m" — это "хеш").

Как реализовать этот простой способ?

Выберите два очень маленьких случайных числа. Назовите их η1 и η2 (буква "эта" из греческого алфавита).

Случай 1. Если точка находится над линией и справа от оси y, поворачиваем линию против часовой стрелки и перемещаем ее вверх.

- Добавьте η1 к наклону m. Получается m'+η1.
- Добавьте η2 к y-пересечению b. Получается b'+η2.

Случай 2. Если точка находится выше линии и слева от оси y, поворачиваем линию по часовой стрелке и перемещаем ее вверх.

- Вычтите η1 из наклона m. Получается m'-η1.
- Прибавьте η2 к y-пересечению b. Получается b'+η2.

Случай 3. Если точка находится ниже линии и справа от оси y, поворачиваем линию по часовой стрелке и перемещаем ее вниз.

- Вычтите η1 из наклона m. Получается m'-η1.
- Вычтите η2 из y-пересечения b. Получается b'-η2.

Случай 4. Если точка находится ниже линии и слева от оси y, поворачиваем линию против часовой стрелки и перемещаем ее вниз.

- Прибавьте η1 к наклону m. Получается m'+η1.
- Вычтите η2 из y-пересечения b. Получается b'-η2.

Результат: линия с уравнением $\hat{p}$=m'r+b'.

Итак, подведем промежуточные итоги.

- *Если модель выдает цену дома, которая намного ниже фактической, добавьте небольшую случайную сумму к цене за комнату и к базовой цене дома.*
- *Если модель выдает цену дома, которая выше фактической, вычтите небольшое случайное количество из цены за комнату и базовой цены дома.*

Но в отношении этого способа возникает несколько вопросов.

- Можно ли выбирать лучшие значения для η1 и η2?
- Можно ли сократить четыре случая до двух или даже одного?

Вот тут-то и пригодится следующий способ!

Квадратичный способ перемещения линии ближе к одной из точек

С помощью квадратичного способа можно свести четыре вышеупомянутых случая к одному, найдя значения с правильными знаками (+ или -), которые нужно добавить к наклону и у-пересечению, чтобы линия всегда двигалась ближе к точке.

Используя простой способ, мы совершаем следующие действия.

- Когда точка находится выше линии, мы прибавляем небольшую величину к у-пересечению. Когда точка находится ниже линии, вычитаем небольшую величину.
- Если точка находится выше линии, значение p-$\hat{p}$ (разница между ценой и прогнозируемой ценой) положительное. Если точка находится ниже линии, это значение отрицательное.

Сложив эти две точки в одну, мы приходим к выводу, что если к у-пересечению добавить разницу p-$\hat{p}$, то линия всегда будет двигаться к точке. Дело в том, что это значение положительно, когда точка находится над линией, и отрицательно, когда точка находится под ней.

Но в МО следует быть осторожным при внесении корректировок и всегда двигаться не спеша. Введем еще один термин — скорость обучения.

Скорость обучения

Перед обучением модели мы выбираем очень маленькое число. Так мы будем знать, что модель в процессе обучения изменится незначительно.

Обозначим скорость обучения греческой буквой η *(“эта”)*.

Поскольку скорость обучения мала, то и значение η(p-$\hat{p}$) будет небольшим. Это значение мы прибавляем к у-пересечению, чтобы переместить линию в направлении точки.

Возвращаемся к наклону. Действия будут аналогичны тем, которые мы совершали в отношении у-пересечения, но только немного сложнее.

- Используя простой способ, когда точка находится в положении, описанном в случаях 1 и 4 (над линией и справа от вертикальной оси или под линией и слева от вертикальной оси), мы вращаем линию против часовой стрелки. В других случаях (случаи 2 и 3), мы вращаем ее по часовой стрелке.
- Если точка (r,p) находится справа от вертикальной оси, то значение r положительно. Если точка находится слева от вертикальной оси, оно будет отрицательно. Обратите внимание: в данном примере r никогда не будет отрицательным, так как оно указывает на количество комнат. Однако чисто теоретически признак может быть отрицательным.

Рассмотрим значение r(p-$\hat{p}$). Оно положительно, когда r и p-$\hat{p}$ оба положительны или оба отрицательны. Именно так обстоит дело в сценариях 1 и 4. В случаях 2 и 3 r(p-$\hat{p}$) отрицательно.

Поскольку это значение должно быть небольшим, то умножаем его на скорость обучения и делаем вывод, что добавление ηr(p-$\hat{p}$) к наклону всегда будет двигать линию в направлении точки.

Представим, что нужно написать псевдокод для квадратичного способа.

Входные данные:

- Линия с наклоном m, y-пересечение b и уравнение $\hat{p}=mr + b$.
- Точка с координатами (r,p).
- Небольшое положительное значение η (скорость обучения).

Выходные данные:

- Линия с уравнением $\hat{p} = m'r + b'$, которая ближе к точке.

Как реализовать квадратичный способ?

- Добавьте $\eta(p-(p))^2$ к y-пересечению b. Получается $b' = b+\eta(p-(p))^2$ (это переместит линию).
- Добавьте $\eta r(p-(p))^2$ к наклону m. Получается $m' = m+\eta r(p-(p))^2$ (это повернет линию).

Результат: линия с уравнением $p^= m'r + b'$.

И последний способ — абсолютный

Квадратичный способ очень эффективен, но есть еще один полезный трюк — *абсолютный метод*, который является промежуточным решением между простым и квадратичным способами.

Используя квадратичный способ, мы применяли две величины — $p-\hat{p}$ (цена — прогнозируемая цена) и r (количество комнат), чтобы свести четыре случая к одному.

Оперируя абсолютным значением, мы используем только r, чтобы свести четыре случая к двум.

Иными словами, если точка находится выше линии (т. е. если $p>\hat{p}$)), можно выполнить следующие действия.

- Добавьте η к y-пересечению b. Получается $b' = b+\eta$ (это переместит линию вверх).
- Добавьте ηr к наклону m. Получается $m' = m+\eta r$ (это повернет линию против часовой стрелки, если точка находится справа от оси y, и по часовой, если она расположена слева от оси y).

Если точка находится ниже линии (т. е. если $p < \hat{p}$), можно сделать следующее.

- Вычтите η из y-пересечения b. Получается $b' = b-\eta$ (это переместит линию вниз).
- Вычтите ηr из наклона m. Получается $m' = m-\eta r$ (это повернет линию по часовой стрелке, если точка находится справа от оси y, и против часовой, если она расположена слева от оси y).

CFA - Интерпретация коэффициентов регрессии

Рассмотрим расчет и интерпретацию коэффициентов простой линейной регрессии (точка пересечения, наклон), а также сравнение перекрестной регрессии и регрессии временных рядов, - в рамках изучения количественных методов по программе CFA (Уровень II).

См. начало:

- *CFA - Простая линейная регрессия*
- *CFA - Расчет параметров простой линейной регрессии*

Что означают коэффициенты регрессии?

Точка пересечения или константа (intercept) - это значение зависимой переменной Y, при котором независимая переменная X равна нулю. То есть, это точка, в которой линия регрессии пересекается с осью Y.

Важно отметить, что в некоторых случаях это не имеет смысла, особенно когда в реальности невозможно, чтобы независимая переменная была равна нулю.

Например, если у нас есть модель, в которой объем денежной массы объясняет рост ВВП, точка пересечения не имеет значения, потому что на практике нулевая денежная масса невозможна.

Однако, если бы независимая переменная была ростом денежной массы, то точка пересечения имела бы смысл.

Наклон (slope) - это изменение зависимой переменной Y при изменении на одну единицу независимой переменной X.

- Если наклон является *положительным*, то изменение независимой переменной и изменение зависимой переменной будут в одном и том же направлении.
- Если наклон *отрицательный*, то изменение независимой переменной и изменение зависимой переменной будут в противоположных направлениях.