ЛАБОРАТОРНАЯ РАБОТА №7. ПОНЯТИЕ «РЕГРЕССИЯ». ЛИНЕЙНАЯ И НЕЛИНЕЙНАЯ ВЗАИМОСВЯЗЬ.

Каждый программист должен хорошо разбираться не только в языках разработки, но и в некоторых точных науках. Примеры – статистика, информатика, математика. За счет этого получается не только логически рассуждать и делать грамотные выводы, но и решать большое количество разнообразных бизнес-задач.

В данной статье речь зайдет о моделях регрессии. Предстоит разобраться в определении соответствующего термина, целях применения регрессионного анализа, а также рассмотреть виды упомянутого компонента. Все это поможет разбираться лучше в наиболее вероятных характеристиках имеющихся факторов, случайных ошибках моделей.

Что это такое

Регрессия – метод, используемый для моделирования и анализа отношений между переменными. Позволяет просматривать влияние этих самых переменных на получение того или иного результата.

В теории вероятностей и математической статистике это:

- обратное движение (от латинского regressio);
- односторонняя стохастическая зависимость, устанавливающая соответствие между переменными (математическое выражение, отвечающее за связи зависимой переменной y и независимыми x при условии статистической значимости).

Существуют различные модели регрессии. Особое внимание уделяется линейной. Именно она встречается на практике чаще остальных. Далее акцент тоже будет сделан на линейных регрессиях, но и примеры остальных видов изучаемого компонента тоже изучим.

Линейный тип

Линейная регрессия – регрессионная модель одной переменной y от другой или нескольких переменных x. В процессе используется зависимость линейной функции. Отсюда и произошло соответствующее «название».

Пусть будут даны две непрерывные функции (это – переменные):

$x = (x_1, x_2, x_3,\ldots,x_n)$;

$y = (y_1, y_2, y_3,\ldots,y_n)$.

Нужно провести построение графика – разместить соответствующие точки на двумерном графике рассеяния. Данный прием позволяет получать линейные соотношения. Картина актуальна для ситуаций, при которых информация аппроксимируется прямой линией.

Если y зависит от x, а изменения в первой переменной вызваны корректировками во второй, можно определить линию регрессии. В данном случае целесообразно говорить о том, что имеет место регрессия y на x. Полученная линия лучшим образом опишет прямолинейное соотношение между указанными компонентами.

Простая классическая регрессия – способ выбора из заданного семейства функции той, что минимизирует функцию потерь. Последняя подчеркнет степень отклонения функции от заданных в точках значений. Это – основная задача линейной модели. Построить соответствующий ситуации график достаточно просто. Обычно он представляет собой линию.

Допущения

Простая классическая регрессия представляется зависимостью одной величины от другой. Ее элементарный вариант предусматривает такие требования (условия):

- значения зависимой переменной определяются без ошибок;
- модель имеет всего два параметра, которые будут задаваться заранее (предварительно);
- ошибки распределения стремятся к нулю и имеют постоянное отклонение;
- значения параметров неизвестны и не могут быть ясны заранее – их получают путем подбора.

В простой линейной ре грес сии параметры иногда выбирают самостоятельно, вручную. Но чаще всего для этого используют специальное программное обеспечение. Также существуют специальные формулы, которые позволяют провести необходимые вычисления и расчета собственноручно.

Нюансы расчетов

Простая регрессионная модель включает в себя функции. Если соответствующая информация имеет линейный вид, то и регрессия окажется линейной. Ее вычисление заключается в том, чтобы подобрать выборку по результатам анализа вычислений, информация в которых отвечает установленным правилам.

Данные в linearregressionmodel должны соответствовать следующим критериям:

- результаты являются адекватными;
- используются статические гипотезы в параметрах модели;
- оптимальные точечные и интервальные оценки.

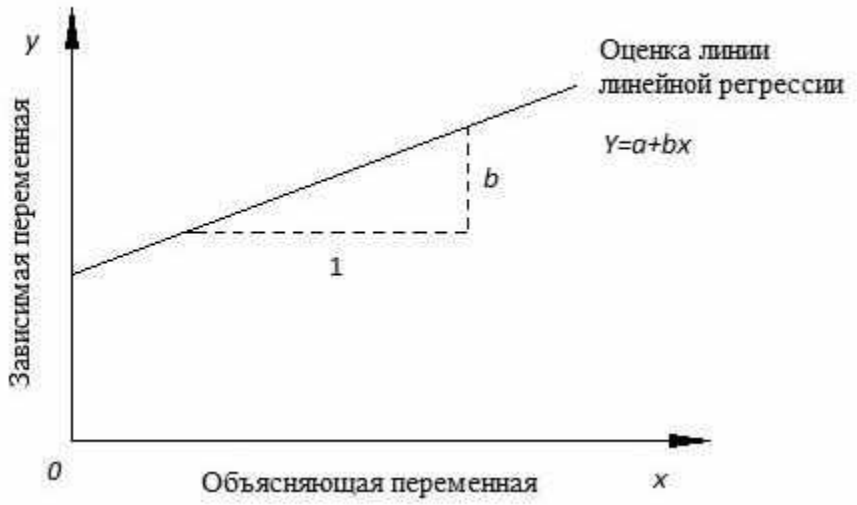
Все это требуется принимать во внимание при прогнозировании и анализе, а также расчете линейной регрессии.

Регрессионная линия

Регрессионный анализ – это набор статистических методов исследования влияния одной или нескольких независимых переменных на зависимую. Это – ключевая задача соответствующего процесса.

Если нужно построить график линейной регрессии, придется иметь дело с одноименной линией. Простая модель описывается формулой: $Y = a+bx$. Здесь:

- Y – переменная отклика (зависимая);
- b – градиент оцененной линии (угловой коэффициент);
- a – свободный член линии оценки (пересечение, значение Y в x = 0);
- x – предиктор (независимое значение).



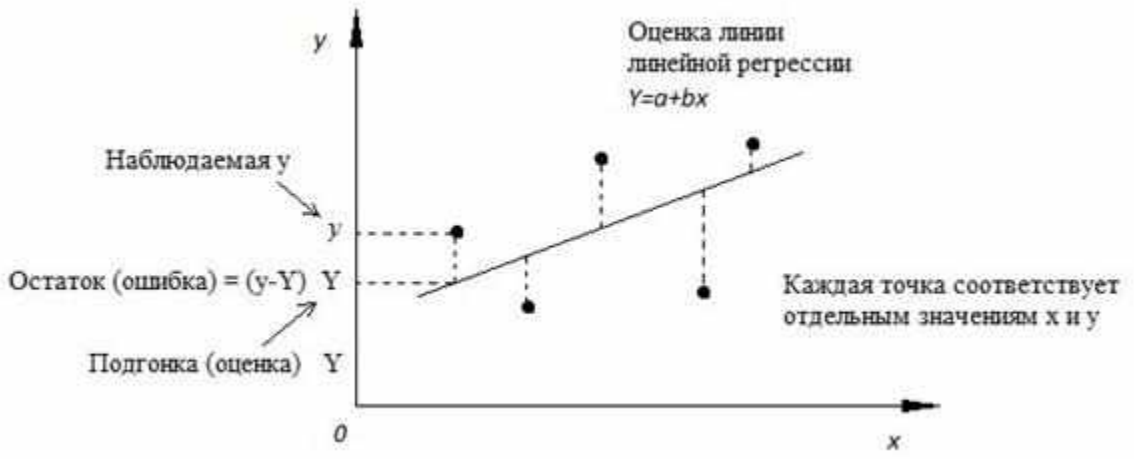
Выше – примеры того, как можно построить график регрессионной линии. Расширить соответствующую интерпретацию можно путем включения в функции новых независимых переменных. В подобной ситуации целесообразно говорить не о простой вариации. Она будет множественной линейной регрессией. Оба варианта активно используются при прогнозировании в статистике и программировании.

Метод наименьших квадратов

Для того, чтобы определить в рассмотренной линейной формуле a и b, нужно использовать различные программы и приложения. Это позволит быстрее организовать не только расчеты, но и сопутствующее построение графиков.

Математики и статистики стараются пользоваться специальными формулами для реализации поставленной задачи. Лучше всего использовать выбор наблюдений. Здесь a и b – выборочные оценки генеральный параметров α и β. Они будут определять линию регрессионного компонента в совокупности. Такой подход носит название «метод наименьших квадратов» — МНК.

Здесь подборка оценивается путем рассмотрения остатков. Под последними подразумеваются вертикальные расстояния каждой точки от линии. Лучшая подгонка – такая, в которой сумма квадратов остатков является минимальной.



Вот – пример линейных соответствующих расчетов. При помощи такого графического представления удается лучше понять метод наименьших квадратов в действии.

Нелинейная регрессия - это способ нахождения нелинейной модели взаимосвязи между зависимой переменной и набором независимых переменных. В отличие от традиционной линейной регрессии, которая ограничена оценкой линейных моделей, нелинейная регрессия может оценивать модели с произвольными взаимосвязями между независимыми и зависимыми переменными. Это достигается при помощи итерационных алгоритмов оценки. Обратите внимание на то, что такая процедура не обязательна для простых полиномиальных моделей вида Y = A + BX**2. Определив новую переменную W = X**2, мы получаем простую линейную модель Y = A + BW, которую можно оценивать с использованием традиционных методов, таких как процедура линейной регрессии.

Пример. Можно ли предсказать размер совокупности в зависимости времени? На диаграмме рассеяния видно, что существует сильная взаимосвязь между размером совокупности и временем, но эта взаимосвязь нелинейна, поэтому требуются специальные методы оценки процедуры нелинейной регрессии. Задав соответствующее уравнение, например, логистическую модель роста совокупности, можно получить хорошую оценку модели, позволяющую предсказывать изменение размера совокупности от времени, когда реальные измерения не проводились.

Статистика. Для каждой итерации: оценки параметров и сумма квадратов остатков. Для каждой модели: сумма квадратов для регрессии, остаток, скорректированный и нескорректированный итог, оценки параметров, асимптотические среднеквадратичные ошибки и асимптотическая корреляционная матрица оценок параметров.

Данные для нелинейной регрессии

Данные. Зависимая и независимые переменные должны быть количественными. Категориальные переменные, такие как религия, основная об-

ласть исследования, регион проживания, должны быть перекодированы в бинарные (фиктивные) переменные или в другие типы переменных контрастов.

Допущения. Результаты допустимы только в том случае, если вы задали функцию, точно описывающую взаимосвязь между зависимыми и независимыми переменными. Кроме этого, очень важен выбор хороших начальных условий. Даже если вы задали правильную функциональную зависимость для модели, при использовании неудачных начальных условий модель может не сходиться или будет получено локальное оптимальное решение, а не нужное глобальное.

Родственные процедуры. Многие модели, которые на первый взгляд могут показаться нелинейными, можно преобразовать в линейные модели и использовать для них процедуру линейной регрессии. Если нет определенности с выбором подходящей модели, процедура Подгонка кривых может помочь в обнаружении полезных функциональных отношений для ваших данных.

Как выполнить нелинейный регрессионный анализ

Для этой возможности требуется модуль Настраиваемые таблицы и расширенная статистика.

1. Выберите в меню:
   Анализ > **Регрессия** > **Нелинейная...**
2. Из списка переменных в вашем активном наборе данных выберите одну численную зависимую переменную.
3. Для построения выражения модели введите выражение в поле **Выражение, задающее модель** или вставьте в это поле компоненты (переменные, параметры, функции).
4. Идентифицируйте параметры в вашей модели, нажав кнопку **Параметры**.

Сегментированная модель (у которой различные формы в разных частях ее домена) должна быть задана с использованием логических условий в одном операторе модели.

Эта процедура вставит синтаксис команды NLR.

- **Логические условия (нелинейная регрессия)**
- **Параметры нелинейной регрессии**
- **Общие модели нелинейной регрессии**
- **Функция потерь нелинейной регрессии**
- **Ограничения на параметры нелинейной регрессии**
- **Сохранить новые переменные нелинейной регрессии**
- **Нелинейная регрессия: Параметры**
- **Интерпретация результатов нелинейной регрессии**
- **Команда NLR: дополнительные возможности**

Данные состоят из свободных от ошибок независимых переменных x и связанных наблюдаемых зависимых переменных (откликов) y. Каждая переменная y моделируется как случайная величина со средним значением, задаваемым нелинейной функцией $f(x,\beta)$. Методическая погрешность может при-

сутствовать, но её обработка выходит за границы регрессионного анализа. Если независимые переменные не свободны от ошибок, модель становится моделью с ошибками в переменных[англ.] и также выходит за рамки рассмотрения.

Например, модель Михаэлиса — Ментен для ферментативной кинетики

Другими примерами нелинейных функций служат показательные функции, логарифмические функции, тригонометрические функции, степенные функции, гауссова функция и кривые Лоренца. Регрессионный анализ с такими функциями, как показательная или логарифмическая, иногда может быть сведён к линейному случаю и может быть применена стандартная линейная регрессия, но применять её следует осторожно. Подробнее в разделе «Линеаризация» ниже.

В общем случае представления в замкнутом виде (как в случае линейной регрессии) может и не быть. Обычно для определения наилучших оценок параметров используются оптимизационные алгоритмы. В отличие от линейной регрессии может оказаться несколько локальных минимумов оптимизируемой функции и глобальный минимум даже может дать смещённую оценку. На практике используются оценочные значения[англ.] параметров совместно с оптимизационным алгоритмом в попытке найти глобальный минимум суммы квадратов.

Подробнее о нелинейном моделировании см. «Метод наименьших квадратов» и «Нелинейный метод наименьших квадратов[англ.]».

Предположение, лежащее в основе этой процедуры, заключается в возможности аппроксимации модели линейной функцией.

Статистика нелинейной регрессии вычисляется и используется как статистика линейной регрессии, но вместо **X** в формулах используется **J**. Линейная аппроксимация вносит смещение в статистику, поэтому следует более осторожно интерпретировать статистики, полученные из нелинейной модели.

Лучшей аппроксимирующей кривой часто предполагается та, что минимизирует сумму квадратов невязок[англ.]. Это подход (обычного) метода наименьших квадратов (МНК). Однако, в случае, когда зависимая переменная не имеет постоянной дисперсии, можно минимизировать сумму взвешенных квадратов. Каждый вес, в идеальном случае, должен быть равен обратной величине от дисперсии наблюдений, однако веса могут пересчитываться в итеративном алгоритме взвешенных наименьших квадратов на каждой итерации.

Некоторые задачи нелинейной регрессии могут быть сведены к линейным путём подходящего преобразования формулировки модели.

Тем не менее, из-за сильной чувствительности к ошибкам данных, а также вследствие сильного смещения, это не рекомендуется.

Для распределений ошибок, принадлежащих семейству экспоненциальных распределений, может быть использована связывающая функция для преобразования параметров под обобщённую линейную модель.

Независимая переменная (скажем, X) может быть разбита на классы или сегменты и может быть осуществлена линейная регрессия посегментно. Сегментированная регрессия с анализом достоверности может дать результат, в котором *зависимая переменная* или *отклик* (скажем, Y) ведёт себя различно в различных сегментах[1] .

График справа показывает, что засолённость почвы[англ.] (X) начально не оказывает никакого влияния на урожайность (Y) горчицы, пока не будет достигнуто *критического* или *порогового* значения, после которого сказывается отрицательное влияние на урожайность[2]

Правило Тициуса — Боде в виде математической формулы представляет собой одномерное уравнение нелинейной регресии, связывающее порядковые номера планет солнечной системы, считая от Солнца, с приближёнными значениями больших полуосей их орбит. Точность вполне удовлетворительная не для астрономических целей.

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 представлена уравнением. В данном случае уравнение равно -2.2923x + 4624.4. Это означает, что если бы вы строили график уравнения -2.2923x + 4624.4, то линия была бы грубой аппроксимацией 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.