Временной ряд и его основные свойства

Временной ряд — это измерения одной и той же случайной величины в разные моменты времени.

Несколько формальных свойств временного ряда:

1. Данные временного ряда структурированы, а атрибуты (так иногда называют признаки) зависимы от времени.
2. Данные временного ряда, в отличие от любых других данных, имеют определённую последовательность.
3. В отличие от анализа других данных, в анализе временных рядов важно, чтобы последовательные значения в данных наблюдались через равные промежутки времени, например каждый час, неделю, год, каждый понедельник и так далее.

Тренд, сезонность, шум

Анализ временных рядов следует стандартному плану:

1. Выявление тренда.
2. Анализ сезонности и цикличности.
3. Проверка на шум.

Выявление тренда

Тренд — это основная тенденция изменения величины со временем.

Большое преимущество тренда — его можно прогнозировать как функцию времени, не учитывая предыдущие значения временного ряда.

Анализ сезонности и цикличности

Сезонность задаёт периодические колебания ряда вокруг тренда. Сезонность есть не всегда, но очень часто.

Цикличность — это колебания временного ряда относительно тренда.

Отличием цикличности от сезонности является то, что сезонность возникает из периода в период (каждый декабрь, каждые выходные и т. д.), а цикличность проявляется на более длительных дистанциях и может слегка меняться от цикла к циклу.

Проверка на шум

Белый шум — это значения, которые являются независимыми друг от друга и одинаково распределены в районе нуля на протяжении всего временного интервала.

Как только мы получили белый шум в остатке ряда, дальше прогнозировать бессмысленно.

Давайте сведём основную информацию в небольшую таблицу:

Тренд	Описывает чистое влияние долговременных факторов, изменяется плавно. ***Пример:*** *рост численности населения.*
Цикличность	Состоит из циклов, меняющихся по длительности и амплитуде, описывает периоды подъёма и спада. ***Пример:*** *циклы в экономике, связанные с изменением спроса и предложения*

	или с переменами в финансовой и налоговой политике.
Сезонность	Представляет собой последовательность почти повторяющихся циклов. ***Пример:*** *объёмы продаж цветов накануне 8 марта или авиабилетов в сезон отпусков летом.*
Шум (случайная компонента)	Останется после вычитания всех вышеперечисленных компонентов. Не несёт никакого глубокого смысла.

Инструменты для декомпозиции временного ряда

Разделить ряд на компоненты в *Python* можно с помощью библиотеки statsmodels. Если вы не устанавливали её ранее, это можно сделать стандартным способом (pip install statsmodels) или следуя рекомендациям в официальной документации.

Декомпозиция выполняется методом seasonal_decompose(), который принимает на вход временной ряд с одним признаком. Индексом ряда должна быть дата или время.

Более подробно о необязательных параметрах можно узнать в документации.

Экспоненциальное сглаживание

> **Экспоненциальное сглаживание** — это метод прогнозирования временных рядов для одномерных данных с систематическим трендом или сезонным компонентом. Оно также известно как метод простого экспоненциального сглаживания, или метод Брауна.

Формула для получения экспоненциального сглаживания выглядит так:

$$S_1 = X_0$$

$$S_t = \alpha \cdot X_{t-1} + (1 - \alpha) \cdot S_{t-1}$$

где:

S_t — итоговый сглаженный ряд;

X_{t-1} — фактическое наблюдение в момент времени t-1;

α — коэффициент сглаживания, который выбирается априори $[0 < \alpha < 1]$.

Проще говоря, под экспоненциальным сглаживанием понимается **взвешенная линейная сумма наблюдений**, при этом веса для наблюдений экспоненциально уменьшаются для более старых наблюдений. Тем самым мы не обращаем особого внимания на поведение в прошлом, а недавнему поведению присваиваем больший вес. Если быть точнее, наблюдения взвешиваются с геометрически уменьшающимся коэффициентом.

Коэффициент экспоненциального сглаживания подбирается интуитивно. Чем выше коэффициент, тем меньше внимания мы обращаем на старые данные. Если коэффициент близок к 0, данным в далёком прошлом будет уделено больше внимания.

Простое экспоненциальное сглаживание используется в задачах сглаживания и краткосрочного прогнозирования временных рядов.

Стационарность

Стационарность означает, что сам временной ряд может меняться с течением времени, однако статистические свойства генерирующего его процесса не меняются.

Говоря простым языком, **стационарный процесс (стационарный временной ряд)** — это процесс, который не меняет свои основные характеристики со временем.

- В нашем же случае временной ряд будет стационарным, если у него отсутствуют тренд и сезонность, а математическое ожидание и дисперсия при этом остаются постоянными на протяжении всего периода времени.
- У нестационарного временного ряда статистики (математическое ожидание и дисперсия) будут изменяться со временем, а сам ряд будет иметь сезонность и тренд.

Так как нестационарный ряд анализировать труднее, в анализе временных рядов принято приводить любой временной ряд к стационарности. Это можно сделать путём выявления и устранения тренда и сезонности.

Существует несколько методов проверки временного ряда на стационарность:

- Визуально оценить по графику данных, есть ли какие-либо очевидные тенденции или сезонность. Например, на графике ниже нет ни выраженного тренда, ни сезонности.
- Просмотреть сводную статистику для данных по сезонам, чтобы понять, есть ли очевидные и существенные различия.
- Использовать статистические тесты, чтобы проверить, выполняются ли ожидания стационарности.

Одним из наиболее распространённых тестов на проверку временного ряда на стационарность является расширенный **тест Дики — Фуллера**. В тесте формулируется две гипотезы:

- нулевая гипотеза *(H0)*: временной ряд нестационарный, то есть имеет некоторый тренд и сезонную компоненту;
- альтернативная гипотеза *(H1)*: временной ряд стационарный, то есть не имеет тренда и сезонной компоненты, и данные скорее случайны.

В результате проведения теста мы получим два значения: *p*-значение из теста и пороговое значение.

- Если значение p ниже порога, отвергаем гипотезу H0 и принимаем гипотезу *H1* (ряд стационарный).
- Если значение *p* выше порога, принимаем гипотезу *H0* (ряд нестационарный).

Что делать, если ряд нестационарный?

Если тест на стационарность показал, что ряд нестационарный и в нём присутствуют тренд и сезонность, необходимо избавиться от них.

Обычно для этого достаточно взять разность рядов. Разность выполняется путём дифференцирования первого порядка (periods=1). Если полученная первая разность ряда окажется стационарной, то этот ряд называется интегрированным рядом первого порядка.

Для определения порядка интегрированного ряда необходимо сделать следующее:

1. Получить новый ряд посредством взятия разности.
2. Провести для нового ряда тест на стационарность (например, тест Дики — Фуллера).

Автокорреляция

- **Лаг** — это предыдущее наблюдение (например, лаг в шесть дней относительно сегодняшнего дня указывает на значение чего-либо, полученное шесть дней назад).
- **Положительная корреляция** — это отношение, при котором увеличение одного значения предсказывает увеличение другого.
- **Отрицательная корреляция** — это отношение, при котором увеличение одного значения предсказывает уменьшение другого.
- **Доверительный интервал** — это рассчитанный диапазон значений, в котором, вероятно, будет содержаться неизвестное (предсказанное) значение для наших данных.

➔ **Уровень достоверности** — это вероятность того, что доверительный интервал будет содержать наблюдаемое значение (фактическое значение для предсказания).

Как мы уже сказали, автокорреляция — это корреляция ряда с самим собой, сдвинутым во времени, а значит, в формуле автокорреляции вместо X и Y будут X и значения сдвинутого X:

$$r_1 = \frac{\sum_{t=2}^{n}(x_t - \overline{x_1})\cdot(x_{t-1} - \overline{x_2})}{\sqrt{\sum_{t=2}^{n}(x_t - \overline{x_1})^2 \cdot \sum_{t=2}^{n}(x_{t-1} - \overline{x_2})^2}}$$

График автокорреляций разного порядка называется **коррелограмма**. Его довольно просто построить с помощью метода plot_acf из пакета statsmodels.graphics.tsaplots. Методу необходимо передать всё тот же временной ряд с индексом-датой.

Значения на коррелограмме будут близки к 0 в случае, если данные ряда не зависят от себя в прошлом. Если скрытая зависимость всё-таки имеется, то одно или несколько значений будут значительно отличаться.

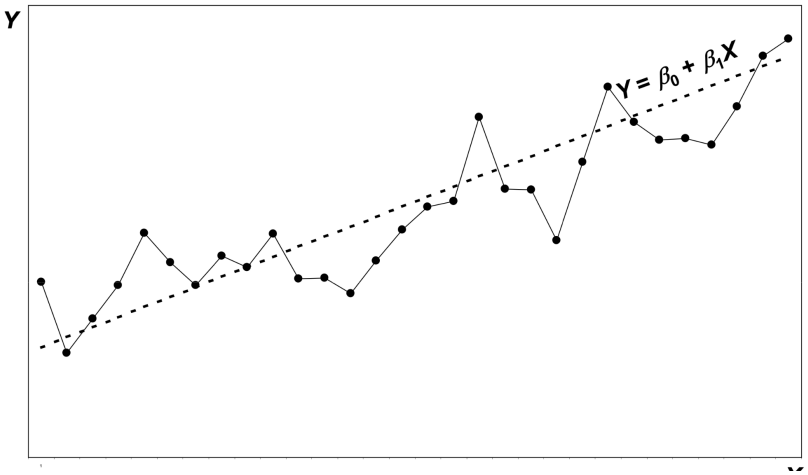
Частичная автокорреляция

Для определения сезонного периода используется **частичная автокорреляция**. Она похожа на классическую автокорреляцию, однако дополнительно избавляется от линейной зависимости между сдвинутыми рядами.

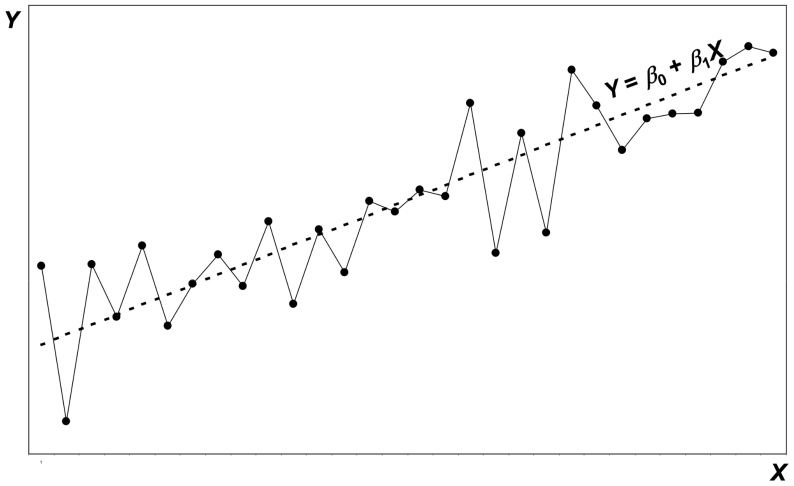
Частичная автокорреляция строится с помощью метода plot_pacf из пакета statsmodels.graphics.tsaplots.

Автокорреляция остатков

В хорошей модели остатки (ошибки) должны иметь случайный характер — это означает, что модель уловила все существующие зависимости.



На графике выше в большинстве случаев после положительных остатков следуют положительные, а после отрицательных — отрицательные. Это пример положительной автокорреляции.



А в этом случае после положительных остатков чаще всего следуют отрицательные и наоборот, на графике — отрицательная автокорреляция.

Причины автокорреляции остатков:

- Если в остатках имеется автокорреляция (наличие зависимости), это значит, что какая-то зависимость осталась незамеченной для вашей модели — возможно, какие-то важные признаки не были учтены.

- На появление автокорреляции в остатках может повлиять предварительное сглаживание данных, так как вы искусственно сглаживаете значения (накладывая соседние друг на друга, добавляя зависимость).

Остатки будут случайными, если автокорреляции нет. Статистически, а не только визуально проверить её наличие или отсутствие можно с помощью теста Дарбина — Уотсона.

Авторегрессия

Авторегрессионная модель — это модель временных рядов, которая описывает, как прошлые значения временного ряда влияют на его текущее значение. Как можно понять из значений частей слова, авторегрессия представляет собой линейную регрессию на себя.

В контексте прогнозирования временных рядов авторегрессионное моделирование будет означать создание модели, в которой переменная Y будет зависеть от предыдущих значений Y с заранее определённой постоянной задержкой во времени. Временной лаг может быть ежедневным (или два, три, четыре дня и т. д.), еженедельным, ежемесячным и т. п.

$$Y_t = \beta_0 + \beta_1 \times Y_{t-1} + error_t$$

В приведённой выше формуле берётся значение последнего временного лага (лаг = 1). Если выбрать лаг, равный неделе, то Y_{t-1} будет представлять значение Y за последнюю неделю, а Y_t — за текущую. Коэффициенты β_1, β_0 — это настраиваемые коэффициенты (как в линейной регрессии), которые мы получим после обучения модели, а $error$ — это ошибка, которую мы, скорее всего, не сможем предсказать, но будем иметь в виду, что итоговое значение включает в себя предсказание и некоторую ошибку.

Модель, в которой для расчёта следующего значения используется только предыдущее, называется моделью первого порядка, или *AR(1)*.

Как выбирать p?

Для определения значения *p* будем использовать график частичной автокорреляции — будем обращать внимание на последний лаг, сильно отличный от нуля, при условии, что ряд стационарный. Если ряд нестационарный, мы сначала приводим его к необходимому виду с использованием разностей, а затем определяем *p*.

AR-моделирование на Python

Для загрузки класса ar_model.AutoReg, который применяется для обучения одномерной авторегрессионной модели порядка *p*, используется пакет statsmodels.tsa.

Ниже приведены некоторые из ключевых шагов, которые необходимо выполнить для обучения *AR*-модели:

1. Отобразить временной ряд.
2. Проверить ряд на стационарность.
3. Выбрать параметр *p* (порядок модели *AR*).
4. Обучить модель.