Анализ данных.
1-2 неделя.
Стационарные временные ряды.

Кармацких Ирина

26 сентября 2022 г.

Содержание

1 Базовые понятия и свойства **2**
- 1.1 Составляющие временного ряда: тренд, сезонность, остатки . . . 2
- 1.2 Стационарность, авторегрессия, автоковариация, автокорреляция . . . 2

2 Белый шум **3**
- 2.1 Q - статистика . . . 4
- 2.2 Процесс скользящего среднего (MA) . . . 5
- 2.3 Процесс авторегресии (AR) . . . 6
- 2.4 Модель ARMA . . . 7

3 Оценки качества модели **7**
- 3.1 Информационные критерии Акаики и Шварца . . . 8
- 3.2 Состоятельность и асимптотическая эффективность . . . 9

1 Базовые понятия и свойства

В данной лекции рассматриваются стационарные временные ряды.

Def. *Временной ряд - это набор наблюдений за переменной в течение последовательных периодов времени. Обычно наблюдения, которые представляет собой эти случайные величины, производятся через фиксированные промежуток времени.*

Например: ежедневно наблюдаем как меняется цена акции и тогда переменная, которая имеет смысл времени - день, начиная с начала года и нумерация будет последовательная, без учета выходных дней.

Временные последовательности - частный случай случайных процессов, то есть можно рассматривать последовательности с непрерывным времени (в любой момент времени своя случайная величина), но в этой лекции рассматривается более простой вариант: у нас будут дискретные наблюдения.

Глобальная задача: прогнозирование временных рядов.

1.1 Составляющие временного ряда: тренд, сезонность, остатки

Изменение временной последовательности можно разложить на 3 составляющие:

- Тренд
- Сезонность
- Случайные изменения

Тренд - некоторый паттерн (не случайный), который легко увидеть на графике. Чаще всего мы его определяем проводя регрессию относительно временной переменной (если линейный - берем линейную регрессию, если нет - берем многочлен или экспоненту).

Далее рассматриваем остатки, и из остатков пытаемся выделить сезонную компоненту. При этом сезонность понимается как изменение величины, связанное с коллинеарными событиями. То есть это могут быть времена года, изменения связанные с финансовым годовым, изменения, связанные с мероприятиями (например олимпиада, которая проходит раз в 4 года). Кстати, сезонная компонента тоже может быть найдена с помощью регрессии относительно фиктивных переменных, выражающих соответствующий сезон.

После того как мы вычитаем сезонную компоненту из данных, у нас остается только случайные изменения, которые мы тоже хотим предсказать. Если эти изменения малы, то мы можем их просто проигнорировать, но если они вносят существенную долю изменения интересующий нас зависимой переменной, то мы должны прогнозировать и случайные изменения.

1.2 Стационарность, авторегрессия, автоковариация, автокорреляция

Def. *Авторегрессия - регрессия, проведенная переменной по самой себе (на прошлых значениях этой переменной).*

Например мы можем проводить регрессию продаж в текущем месяце по продажам в предыдущем месяце.

Def. *Стационарность (в узком смысле): если мы возъмем любой набор случайных величин из временного ряда, то их распределение не будет зависеть от того из какого отрезка временного ряда мы их взяли.*

Например, если мы взяли 10 случайных величин соответствующие прошлому году, прошлому месяцу, прошлой неделе, важным будет являться только то, какое время происходит внутри (между этими случайными величинами), а не временной период из которого мы их взяли (год/месяц/неделю назад).

Def. *Стационарность (в широком смысле) чуть более мягкое понятие, под стационарностью в широком смысле понимают стационарность числовых характеристик случайных величин: мат ожидания, дисперсии и ковариации.*

Def. *Если мы рассмотрим ковариацию временного ряда с самим собой, то есть ковариацию случайной величины в текущий момент времени и случайно величины соответствующей значению месячной давности, то такие ковариации образуют функцию автоковариации.*

Def. *Автокорреляция - это автоковариация, деленная на корень из произведения дисперсий случайных величин, рассматриваемых в моменты времени.*

Частная автокорреляция - некоторое линейное преобразование автокорреляции (подробно рассматриваться не будет). Смысл частной автокорреляции: из зависимости между рассматриваемыми случайными величинами, то есть допустим мы считаем частную корреляцию между случайными величинами в текущий момент времени и месяц назад, тогда в этой автокорреляции исключается вся зависимость, связанная со значениями в промежуточный момент времени. То есть мы исключаем часть зависимости, связанную с тем, каким путем мы пришли из точки месяц назад в текущую.

Временной ряд называется *стационарным в широком смысле*, если удовлетворяет следующим условиям:

- Математическое ожидание должно быть постоянным и конечным;
- Дисперсия должна быть постоянной и конечной;
- Автоковариация должна зависеть только от времени прошедшего между случайными величинами, но не от их положения во временном ряде. То есть если наш процесс обозначается x_t и мы рассмотрим автоковариацию, то это будет функция от 2 переменных (рассматриваемых в 2 временных точках t_1, t_2) и если эта функция на самом деле зависит только от разности $t_1 - t_2$, то эта функция соответствует стационарному процессу в широком смысле.

Заметим, что если в данных присутствует тренд или сезонные компоненты, то соответствующая временная последовательность не будет стационарной. Таким обраом сначала мы должны из наших данных исключить тренд и сезонные компоненты, а потому будем смотреть являются ли остатки стационарным процессом.

Почему важна стационарность? Она важна тем, что позволяет нам делать прогнозы. Если же последовательность не обладает свойством стационарности, то часто невозможно получить прогнозы или же их качество оставляет желать лучшего.

2 Белый шум

Def. *Рассмотрим частный случай стационарного процесса: процесс белого шума. Он характеризуется следующим:*

- *Математическое ожидание равно 0*
- *Дисперсия постоянна*
- *Функция автокорреляции отлична от 0 в одной точке. То есть если считается корреляция точки с собой.*

Заметим, что из некоррелированости не следует независимость. Если накладываются дополнительные требования, то надо это отразить в названии. Можно рассматривать независимый белый шум, в котором требуется независимость случайных величин, соответствующих разным моментам времени или же можно сделать предположение о распределении и сказать, что мы рассматриваем гауссовский белый шум, его свойства: некоррелированный, независимый (следует из некоррелированности) с нормальным (гауссовским) распределением случайных величин.

Важным свойством белого шума является непредсказуемость. Это лучше видно для независимого белого шума: если у нас все составляющие величины независимы, то есть будущие никаким образом не зависит от прошлого. Для некоррелированного белого шума: случайные величины все-таки могут быть зависимы, что позволяет говорить о некотором прогнозе, но его качество сложно оценить

Def. *При рассмотрении временных последовательностей часто удобно рассматривать оператор лага (L). Оператор L определяется следующим образом:*

$$Ly_t = y_{t-1}$$

.

Первый разностный оператор: $\delta y_t = y_t - Y_{t-1}$

Распределенное отставание (отставание - lag, лаг) определяется как взвешенная сумма текущих и прошлых значений:

$$\omega_0 y_t + \omega_1 y_{t-1} + \cdots + \omega_k y_{t-k}$$

Центральной теоремой при рассмотрении стационарных в широком смысле последовательностей является разложение Вольда.

Th (Вольд)**.** *Если у нас есть случайная временная последовательность, обладающая свойством стационарности в широком смысле, тогда ее можно представить в следующем виде:*

$$Y_t = \sum_0^\infty b_j \epsilon_{t-j}$$

где ϵ_t - процесс белого шума, b_j - коэффициенты; $t-j$ означает, что мы рассматриваем процесс j шагов назад.

Правая часть в разложении Вольда называется обобщенным линейным процессом. Если b_j быстро убывают очень быстро убывают к 0 (с ростом j), то часто ряд обрывают для какого-то m, и вместо бесконечной суммы рассматривают конечную. В этом случае мы получаем модель скользящего среднего. Альтернативный вариант: эту сумму можно переписать в виде оператора лага, то есть $\epsilon_{t-j} = \underbrace{L \dots L}_{j} \epsilon_t \Rightarrow \sum_0^\infty b_j \epsilon_{t-j} = \epsilon_t \sum_0^\infty b_j L^j$ - бесконечная сумма представляет ряд для какой-то функции от оператора лага.

Чтобы работать со стационарными временными рядами, нам нужно научиться оценивать математическое ожидание, дисперсию, автоковариации, автокорреляции.

Для оценки мат. ожидания берется среднее арифметическое значений процесса в некотором окне ширины T :

$$\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t$$

Для стационарных временных рядов эта оценка состоятельна. Для нестационарных временных рядов эта оценка существенным образом будет зависеть от выбранного окна и таким образом задача оценить сводится к более сложной.

Дисперсии, автоковариация, автокорреляция вычисляются по аналогичным подходам. Например, мы можем оценить автокорреляцию следующим образом: рассмотрим автокорреляцию точки с шагом τ и вычисляется она как корреляция нашего случайного процесса в момент времени t и момент времени $t-\tau$.

$$\hat{\rho}(\tau) = \frac{\sum_{t=\tau+1}^{T} (y_t - \bar{y})(y_{t-\tau} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}$$

В знаменателе стоит оценка дисперсии. Поскольку речь идет о стационарных временных рядах, то дисперсия не зависит от того момента времени, в который мы рассматриваем

2.1 Q - статистика

Нужно понять, являются ли остатки, то есть наша наблюдаемая временная последовательность, процессом белого шума. Если является, то в этом случае прогноз может быть невозможен. Либо если он не является процессом белого шума, мы можем попытаться построить какой-нибудь прогноз.

Для определения является ли стационарная последовательность процессом белого шума используется критерий Q-статистика. В частности используются статистики Бокса - Пирса или Льюнга - Бокса. Q-статистика может быть использована для измерения степени отклонения автокорреляции от нуля, и наличия белого шума в наборе данных. Эта статистика распределена приблизительно как хи-квадрату с m степенями свободы в больших выборках при выполнении нулевой гипотезы.

Нулевая гипотеза: автокорреляции равны 0 с первого по шаг m. Заметим, что с нулевым шагом автокорреляция отлична от 0 (автокорреляция процесса самого с собой = 1). Формально мы не проверяем гипотезу о том, что процесс является процессом белого шума, а проверяем только автокорреляции до какого-то шага m.

- Статистика Бокса-Пирсона:

$$Q = T \sum_{k=1}^{m} \hat{\rho_k^2}$$

  где $\hat{\rho}$ - оценки автокорреляции, T - размер окна, по которому оценивается автокорреляция.

- Статистика Льюнга - Бокса:

$$Q = T \sum_{k=1}^{m} \frac{T+2}{T-k} \hat{\rho_k^2}$$

  где $\hat{\rho}$ - оценки автокорреляции, T - размер окна, по которому оценивается автокорреляция

Результат: если нулевая гипотеза верна, то статистики Бокса-Пирсона и Льюнга - Бокса стремятся к распределению хи - квадрат с числом степеней свободы m. Таким образом если значения статистики окажется больше критического значения распределения хи - квадрат, то нулевая гипотеза отклоняется.

Q-статистика Бокса-Пирса отражает абсолютные величины корреляций, поскольку она суммирует квадраты автокорреляций. Таким образом, знаки не компенсируют друг друга, и большие положительные или отрицательные коэффициенты автокорреляции приведут к большой Q-статистике Q-статистика Бокса - Льюнга аналогична Q-статистике Бокса-Пирса, но она импольует взвешенную сумму квадратов автокорреляций.

Если рассматривается большой объем выборки, большой период, большое окно, то статистики Бокса - Пирсона и Льюнга - Бокса эквивалентны. В этом случае можно использовать любой из этих критериев.

Однако, если размер окна небольшой, то статистика Льюнга - Бокса быстрее сходится к распределению хи - квадрат, таким образом этот критерий оказывается точнее, а значит, предпочтительнее.

Выбор параметров для применения этих критериев играет существенную роль, то есть сколько мы должны проверить автокорреляций, чтобы сделать достоверное предположение о том, что наш процесс является процессом белого шума. И второй вопрос, каким нужно выбрать размер окна. Конкретных рекомендаций по этому поводу нет, можно сказать, что количество автокорреляций, которое участвует в гипотезе не должно быть меньше, чем количество переменных, используемых в предполагаемой модели.

2.2 Процесс скользящего среднего (MA)

Итак, если мы с вами в разложении Вольда вместо бесконечной суммы оставляем только конечную, то мы получаем процесс скользящего среднего (MA). Количество членов в сумме называется порядком процесса.

Например, процесс скользящего среднего первого порядка определяется как

$$y_t = \epsilon_t + \theta \epsilon_{t-1}$$

где ϵ_t - процесс белого шума, ϵ_{t-1} - процесс белого шума в предыдущий момент времени $(t-1)$, θ - коэффициент, а y_t - зависимая переменная.

Процесс скользящего среднего первого порядка называется процессом с короткой памятью.

Такое название этот процесс получил, так как для формирования y_t мы используем значения только в текущий момент времени и в предыдущий момент времени. Более дальние значения здесь никаким образом не участвуют.

Автокорреляция может быть рассчитана следующим образом:

$$\rho_0 = 1,\ \rho_1 = \frac{\theta}{1+\theta^2},\ \rho_k = 0,\ k > 1$$

.

Таким образом, мы наблюдаем эффект обрыва автокорреляции после шага 1 (после этого они все равны 0).

Процесс скользящего среднего первого порядка является стационарным в широком смысле.

Обратите внимание, что в этой форме модель трудно применить, поскольку мы обычно не наблюдаем непосредственно ϵ_t, а наблюдаем y_t. Поэтому процесс скользящего среднего можно переписать в авторегрессионной форме:

$$\epsilon_t = y_t - \theta\epsilon_{t-1}$$

Здесь y_t мы наблюдаем, а ϵ_{t-1} рассчитываем в предыдущий момент времени.

Здесь параметр θ может принимать любые значения. Но если взять и вместо θ подставить $\frac{1}{\theta}$, то автокорреляция не изменится. Поэтому чтобы θ однозначно определялась по автокорреляции мы должны наложить условие $|\theta| < 1$.

Def. *Процесс скользящего среднего порядка q получается очевидным образом:*

$$y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}$$

Здесь y_t - временной ряд, ϵ_{t-i} - процесс белого шума, θ_i - коэффициенты модели.

Эта модель так же называется моделью с короткой памятью, только память здесь уже q шагов.

Автокоррелция обладает похожим свойством, она отлична от 0 для значений лагов от 1 до q, а после этого в точности равна 0.

Данная модель напрямую неприменима, так как ϵ_t мы не наблюдаем, на парктике мы должны переписать в авторегрессионной форме.

2.3 Процесс авторегресии (AR)

Рассмотрим процесс, который называется процессом авторегресии 1-ого порядка (AR(1)):

$$y_t = \phi y_{t-1} + \epsilon_t$$

где ϕ - коэффициент модели, ϵ_t - значение белого шума.

То есть мы с вами рассматриваем 2-ой случай, когда в разложении Вольда бесконечная сумма может быть аппроксимирована какой-то простой функцией. Получается, что мы аппроксимируем оператором лага, примененным к переменной в предыдущий момент времени.

Здесь уже не всегда будет соблюдаться условие стационарности в широком смысле. Эта модель будет стационарной, если коэффициент $|\phi| < 1$.

Уравнения Юла-Уокера определяют автокорреляцию такой модели:

$$\rho_k = \phi^k,\ k = 0, 1, 2, \ldots$$

В отличие от процессов скользящего среднего здесь автокорреляция не равна 0 ни для какого лага k, то есть здесь нет отсечки автокорреляции. Так как $|\phi| < 1$, то она убывает к 0 с ростом k. При этом, если коэффициент $\phi > 0$, то она монотонно убывает к 0, а если $\phi < 0$, то мы получаем знакопеременный ряд.

Def. *Процесс авторегрессии порядка p (AR(p)) выглядит следующим образом:*

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

Здесь где ϕ_i - коэффициенты модели, ϵ_t - текущее значение белого шума, y_t - зависимая переменная, y_{t-i} - значения в предыдущие моменты времени.

Процесс авторегрессии порядка p будет стационарным в широком смысле, если $|\phi_i| < 1$. Но на самом деле чуть более правильно здесь составить так называемый характеристический многочлен, найти его корни и все эти корни должны быть < 1, в этом случае процесс будет стационарным в широком смысле.

Автокорреляция для такого процесса находится из системы уравнений, которая называется уравнениями Юла-Уокера. В случаю, когда у нас процесс первого порядка они легко выписываются и решаются, а в случае процесса порядка p уже нужно решать систему уравнений. Явное выражение не приводится, но важно отметить, что по-прежнему автокорреляция сходится к 0 и не происходит отсечка (не равна 0 ни для каких лагов).

2.4 Модель ARMA

Def. *Можно скомбинировать модель авторегрессии (AR) со скользящим среднем (MA), в этом случае модель будет называться ARMA (p, q). Здесь p - отвечает за то, сколько членов беертся авторегрессии, q - сколько скользящего среднего.*

Например, ARMA (1, 1) модель будет выглядеть следующим образом:

$$y_t = \phi y_{t-1} + \epsilon_t + \theta \epsilon_{t-1}$$

где y_t, y_{t-1} - зависимая переменная и зависимая переменная в предыдущий момент времени, $\epsilon_t, \epsilon_{t-1}$ - значение белого шума в момент t и момент $t-1$, ϕ, θ - коэффициенты модели.

И здесь мы накладываем условия для стационарности в широком смысле так же как и раньше можно строить характеристический многочлен и смотреть его корни), но в упрощенном варианте можно потребовать $|\phi| < 1$; $|\theta| < 1$.

Давайте разберемся, в каких случаях какой моделью пользоваться: обычно для начала вычисляются оценки автокорреляции и рисуют график, который называется коррелограммой. На этом графике отображаются по оси x лаги от 1 до k, а по оси y отображаются соответствующие оценки автокорреляций.

- Если мы на этом графике видем, что в какой-то момент все автокорреляции становятся очень близкими к 0, это означает, что скорее всего мы имеем дело с моделью скользящего среднего;
- Если же все автокорреляции постепенно стремятся к 0, но нельзя сказать, что в какой-то момент они все становятся очень близки к 0, в этом случае скорее всего будет модель авторегрессии, либо ARMA;
- Если мы видим периодические всплески на коррелограмме, это означает, что в наших данных присутствует сезонная компонента, в этом случае нам может помочь комбинированная модель ARMA.

Какое количество переменных выбрать для модели, обсудим позже, но общее правило такого, что нужно пытаться взять переменных как можно меньше, чтобы при этом сохранялось хорошее приближение данных моделей и чтобы количество переменных соответствовало наблюдаемой коррелограмме.

3 Оценки качества модели

Здесь применяются все те же характеристики, которые мы использовали в модели простой многомерной линейной регрессии. Для оценки параметров так же можно использовать метод наименьших квадратов , единственное, что оценки полученные по методу наименьших квадратов не всегда будут состоятельными.

Итак, какие характеристики показывают, насколько хорошо наша модель описывает данные

1. Среднеквадратичная ошибка (MSE) - это статистический показатель, вычисляемый как сумм квадратов невязок, деленная на общее количество наблюдений в выборке:

$$MSE = \frac{\sum e_i^2}{n}$$

где n - общий размер выборки, $e_i = y_t - \hat{y}_t, \hat{y}_t$ - прогнозное значение временного ряда.

Недостатки: во-первых размерная величина, то есть абсолютное значение этой величины нам ни о чем не говорит, во-вторых с ростом числа переменных в нашей модели она будет уменьшаться, значит, ее нельзя использовать для сравнения моделей с разным числом переменных.

2. Показатель качества регрессии - коэффициент детерминации R^2 :

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Он лучше MSE тем, что это безразмерная величина и значение лежит в диапазоне от 0 до 1, однако, если мы говорим о моделях с разным числом переменных, эта характеристика так же плоха, так как с ростом числа переменных она имеет склонность к увеличению. Таким образом все случай

3. MSE при большом количестве переменных уменьшается, так что мы можем рассмотреть MSE с некоторым штрафом k, такая оценка называется несмещенная MSE или s^2:

$$s^2 = \frac{\sum e_i^2}{n - k}$$

Таким образом мы учитываем количество переменных в модели, если k будет очень большим (близким к n), то эта оценка получится значительно хуже, чем если мы возьмем небольшое значение, таким образом мы накладываем штраф на использование большого числа переменных. Штраф за степени свободы увеличивается с увеличением количества параметров, но MSE все равно может упасть. Таким образом, наилучшая модель выбирается на основе s^2.

Заметим, что модифицированный R^2 может быть выражен через s^2 тогда мы получим :

$$R_a^2 = 1 - \frac{s^2}{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$

Таким образом, поведение величины s^2 соотносится с поведением модифицированного R^2, если эта величина мала, то R^2 близок к 1, и наоборот, чем больше s^2, тем меньше будет величина модифицированного R^2. Эти две оценки являются в некотором смысле эквивалентными.

3.1 Информационные критерии Акаики и Шварца

Какие еще способы оценки качества существуют? Перепишем s^2 в другом виде:

$$s^2 = \left(\frac{n}{n-k}\right)\frac{\sum e_i^2}{n}$$

Здесь $\frac{n}{n-k}$ - штрафной фактор (за использование большого числа переменных).

Если использовать другие штрафные факторы, то можно получить следующие информационные критерии:

1. Информационный критерий Акаике (AIC):

$$AIC = e^{\frac{2k}{n}}\left(\frac{\sum e_i^2}{n}\right)$$

2. Информационный критерий Шварца (SIC):

$$SIC = n^{\frac{k}{n}}\left(\frac{\sum e_i^2}{n}\right)$$

Чем меньше значение критерия, тем лучше модель. Также можно заметить, что SIC с ростом k/n растет быстрее и имеет больший штрафной коэффициент по сравнению с другими критериями.

Теперь разберемся, почему мы выделяем эти два критерия (Акаики и Шварца)?

3.2 Состоятельность и асимптотическая эффективность

3.2.1 Свойство состоятельности.

Состоятельность: если истинная модель зависимостью совпадает с выбранной регрессионной моделью, то состоятельность означает, что вероятность выбора правильной (истинной) модели стремится к 1 с увеличением количества рассматриваемых данных. То есть наша оценочная модель будет сходится к истинной модели.

Если же мы не угадали и истинная модель отличается от выбранной, то с ростом объема данных, мы сходимся к модели, которая наилучшим образом аппроксимирует истинную модель (из предложенных нами).

- Простая оценка среднеквадратического (MSE) не является состоятельной;
- Оценка s^2 не является состоятельной;
- Информационный критерий Акаики не является состоятельным;
- Информационный критерий Шварца является состоятельным.

3.2.2 Свойство асимптотической эффективности.

Асимптотическая эффективность определяется как наилучший прогноз, который может дать модель. Диспресия прогноза не может быть $<$ прогноза дисперсии в истинной модели. Если при этом с ростом объема выборки дисперсия прогноза в нашей оценочной модели сходится к диспресии в истинной модели, то такой критерий называется асимптотически эффективным.

- Информационный критерий Акаики является асимптотически эффективным;
- Информационный критерий Шварца не является асимптотически эффективным.

Информационный критерий Акаики используется для получения наиболее точного прогноза, в то же время информационный критерий Шварца используется для нахождения наиболее точной модели.