Data Science.
Lectures. Weeks 3-4.
Nonstationary Time Series.
Нестационарные временные ряды.

Pogorelova Elena

6 октября 2022 г.

Contents

1 Причины нестационарности временных рядов.

Во многих наблюдаемых временных рядах (например, ценах акций или деривативах) мы видим нестационарность, и причины этой нестационарности выделяются следующие:
- Наличие временных трендов
- Наличие сезонности
- Наличие случайных блужданий

Есть и другие причины нестационарности. Например, трейдеры адаптируются к текущей ситуации на рынке, в связи с чем возникает нестационарность распределений. Но эту часть мы включаем в случайные блуждания.

2 Как бороться с нестационарностью

2.1 Временные тренды

Первый компонент нестационарности - это временные тренды.
Под трендом понимается долговременная тенденция изменения временного ряда. Тренд может быть линейным или нелинейным, детерминированным (неслучайным) или случайным.
Если удалось определить детерминированный тренд (т.е. мы знаем, что временная последовательность колеблется таким образом, что есть неслучайная часть, зависящая от времени, а также случайное колебание), то мы можем вычесть неслучайную часть и получить остатки, которые уже представляют собой стационарный в широком смысле процесс. Данное преобразование возможно, когда нестационарность определяется детерминированным трендом, который мы можем найти. Однако это простое преобразование не позволяет учесть стохастические тренды, т.е. тренды, возникающие случайным образом.

2.2 Сезонность

Второй компонент нестационарности - это сезонность.
Сезонность - это изменение цены, связанное с календарными датами (например, продажа пшеницы). В моделях сезонность может учитываться двумя способами. Первый вариант - с использованием фиктивных (вспомогательных) переменных. Допустим, если сезонность имеет квартальный период, то мы введем фиктивную переменную, которая будет отвечать за то, какой сейчас квартал. Второй вариант - включить в модель значение временного ряда с соответствующим лагом. Скажем, если сезонность имеет годовой вид, то мы будем рассматривать текущие значения одновременно со значениями год назад (этот подход используется в модели ARMA).

2.3 Стохастические тренды

Третий компонент нестационарности - это стохастические тренды.
Они возникают из случайных блужданий, явным образом их найти и вычесть не удается. В этом случае помогает взятие разности значений на текущий момент и значений с лагом k ($Y_t - Y_{t-k}$). Если правильно подобрать k, то мы уберем из наших значений сезонную компоненту. Также разность помогает избавиться от стахостических трендов и в некоторых случаях - от детерменированных трендов.

3 Детерминированный тренд

Линейный тренд: $Y_t = B_0 + B_1 t + \varepsilon_t$
t - это дискретная переменная, обозначающая шаг времени.
Как видно, линейный тренд - это линейная регрессия.

Полиномиальный тренд: $Y_t = B_0 + B_1 t + B_2 t^2 + ... + B_k t^k + \varepsilon_t$
Полиномиальный тренд - это нелинейная регрессия.
Вот что нужно про нее знать:
- Оценки коэффициентов могут быть получены с помощью МНК (метод наименьших квадратов)
- С помощью критерия Стьюдента можно проверять, что соответствующие коэффициенты равны нулю
- Важно выбрать сбалансированный k (если k слишком маленький, то предсказания модели могут быть неточными, но если k слишком большой, то модель будет включать случайные ошибки, которые она должна была, наоборот, отбросить)

Линейная модель не применима к ценам, так как они не обладают свойством постоянного роста. Также, если мы рассматриваем падение цены, то прогноз может дать отрицательные значения, что невозможно в действительности.
Поэтому, работая с ценами, используют **модель логарифма цены**: $ln(Y_t)$

4 Сезонность

- Сезонность, как уже было сказано, происходит в связи с календарными событиями: праздникам, бизнес-процессами, природными явлениями и т. д. Важно, что сезонность - это периодическое явление.

4.1 Вспомогательные переменные

Один из способов учесть сезонность в модели - использование вспомогательных (сезонных/фиктивных) переменных.

Пусть s - это количество сезонов в году (s = 4 для кварталов, s = 12 для месяцев, s = 52 для недель и т.д.) Например, если s = 4, у нас получится 4 вспомогательные переменные: $d_1 = \{1, 0, 0, 0, 1, 0, 0, 0, ...\}$, $d_2 = \{0, 1, 0, 0, 0, 1, 0, 0...\}$, $d_3 = \{0, 0, 1, 0, 0, 0, 1, 0, ...\}$, $d_4 = \{0, 0, 0, 1, 0, 0, 0, 1, ...\}$.
В определенный момент времени **только одна из них равняется единице** (она и определяет сезон), остальные - нулю.

Далее мы строим модель множественной линеной регрессии: $Y_i = \Sigma B_i d_i + \varepsilon_i$
Заметим, что в этой модели отстутствует коэффициент свободного члена B_0. Это связано с тем, что, если мы включим этот коэффициент в модель, то получим мультиколлинеарность, а значит оценки коэффициентов B_i мы получить не сможем.
С мультколлинеарностью можно бороться разными способами. Первый вариант - исключить одну из переменных, допустим d_1, и включить B_0. Второй способ - сделать так, как мы уже поступили - убрать коэффициент B_0 и оставить сумму по сезонным переменным.

Также мы можем расширить нашу модель, например, добавив линейный тренд:
$Y_i = S_{s+1} T_i + \Sigma_1^S B_k d_{k,i} + \varepsilon_i$, где T_i - это время, соответствующее шагу i, B_k - это коэффициент регрессии, $d_{k,i}$ - это сезонная переменная d_k, значение, соответствующее наблюдению i, ε_i - значение случайной ошибки
Конечно, в эту модель можно добавлять и другие сезонные факторы. Например, сезонность, связанную с праздниками. При это коэффициенты регрессии оцениваются с помощью МНК.

4.2 Модель прогноза на h шагов

Рассмотрим, как строится модель для прогноза на h шагов.
В модель регрессии подставляется время $t+hT$, где T - это шаг во времени, h - количество шагов, t - время последнего наблюдения.
Чтобы построить прогноз, нужно взять математическое ожидания уравнения регрессии в этой точке: $\widehat{EY_{t+hT}}$.

Уравнение регрессии в этой точке выглядит следующим образом:
$\widehat{EY_{t+hT}} = S_{s+1}(t+hT) + \Sigma_1^S B_k d_{k,t+hT} + \varepsilon_{t+hT}$
Далее вместо B мы ставим их оценки и вычисляем математическое ожидание.
Детерменированная компонента остается как есть.
Математическое ожидание всех ε равно нулю (условие регрессионной модели).
Из сезонных переменных d_k только одна равна 1 в данный момент времени, остальные - ноль.
Поэтому получаем $\widehat{EY_{t+hT}} = S_{s+1}(t+hT) + B_l$, где l - это ненулевая вспомогательная переменная.

Чтобы построить доверительный интервал, требуется предположение о том, что ε имеет нормальное распределение (в этом случае Y тоже имеет нормальное распределение).
Доверительный интервал строится следующим образом: $\widehat{EY_{t+hT}} \pm \sigma Z_{\alpha/2}$
σ чаще всего не известна, поэтому берут ее оценку, построенную по остаткам.
Т.е. фактически вместо σ здесь берется среднеквадратическая ошибка.

5 Случайные блуждания

Третий компонент нестационарности возникает из случайных блужданий.

5.1 Unit root

Рассмотрим независимое случайное блуждание, также называемое **винеровсим процессом.**
Известно, что с ростом t траектория винеровского процесса удовлетворяет закону повторного логарифма. Это значит, что траектория всегда ходит от верхней границы до нижней. Верхняя при этом растет от t как повторный логарифм от t, нижняя - как минус повторный логарфим от t.
Таким образом, **с ростом t колебания становятся все больше и больше. Между колебаниями мы наблюдаем тренд**, который как раз порожден случайными блужданиями.

Простое случайное блуждание можно записать формулой $Y_t = Y_{t-1} + \varepsilon_t$, т.е. к значению предыдущего времени добавляется значение белого шума в текущий момент времени.
Если развернуть это формулу, то получится $Y_t = Y_0 + \sigma_{k=1}^{t}\varepsilon_i$.
Дисперсия процесса $VarY_t = t^2 VarY_0$ - растет, поэтому мы наблюдаем все большие колебания.

На практике мы чаще наблюдаем не просто случайные блуждания, а случайные блуждания, комбинированные с некоторым стационарным процессом с краткосрочной памятью.
Соединяя их в одну модель, **модель unit root**, мы получаем: $a(L)Y_t = b(L)\varepsilon_t$, где $a(L)$ и $b(L)$ - это многочлены от оператора лага L, возможно разной степени.

Требование к этой модели - возможность переписать ее в регрессионном виде, т.е. чтобы можно было записать уравнение таким образом, что ненаблюдаемая величина ε_t была выражена через наблюдаемые величины. И таким образом мы могли провести регрессию и оценить соответсвующие коэффициенты.
Для того, чтобы это было возможно, полином $a(L)$ должен иметь один корень равный единице, а остальные - меньше единицы (отсюда и название - unit root).

5.2 Проблемы, возникающие в связи со случайными блужданиями

Если наши данные имеют тренд, связанный со случайным блужданием, то у нас могут возникать следующие проблемы:
- Если мы будем оценивать коэффициаент модели ARMA, то в этом случае распределение этих коэффициентов не симметрично и зависит от временных параметров, количества наблюдений, и точность оценки оказывается невысокой.
- Наличие данных случайного блуждания изменяет свойства оценок, в частности оценки корреляции. Если мы возьмем два случайных блуждания, то оценка корреляции будет высокой несмотря на то, что они изначально были независимыми. Таким образом, если мы будем опираться на эти оценки, то прийдем к неверным выводам. То есть случайные блуждания порождают фиктивные зависимости, которые никаких реальных оснований не имеют.
- У процесса случайных блужданий нет никаких сходимостей к средним величинам. Сходимость к среднему означает, что если в данных присутствуют какие-то случайные колебания, то они имеют свойства уменьшаться и таким образом в долгосрочном периоде они нивелируется. Для процесса случайного блуждания это не так. Более того, процесс случайного блуждания сам порождает отклонения от любого значения. Таким образом, здесь не может быть схожимости к какому-то долгосрочному уровню.

5.3 Как проверить данные на наличие случайного блуждания?

Для этого используется критерий Дики-Филлера (Augment Dickey-Fuller или просто ADF),
где нулевая гипотеза состоит в том, что данные представляют собой случайное блуждание плюс какой-то стационарный компонент.

Этот критерий устроен следующим образом.
Строится регрессия, где в качестве зависимой переменной берется первая разность $(Y_t - Y_{t-1})$, а в качестве независимых переменных берется некоторая детерминированная часть, дающая тренд.
Также в качестве независимых переменных берутся лагированные разности.
Лагированные разности нужны, чтобы описать стационарную часть данных.
Детерминированная часть компенсирует тренд.
Критерий проверяет, является ли остаток процессом случайного блуждания.

В применении критерия возникают вопросы о том, что является трендовой частью, и сколько лагов разности брать. Рекомендации такие: число лагов разности должно быть таким, чтобы описать любую возможную стационарность, пристутствующую в данных. Если об этом ничего не известно, то количество этих разностей выбирается с помощью информационного критерия Акаике (AIC).
Детерминированная часть берется в полиномиальном виде и далее с помощью критерия Стьюдента проверяется значимость коэффициентов, начиная со старшего. Первый значимый коэффициент и определяет степень детерменированной части.
Если определить детерменированную часть неправильно, то мощность критерия упадет, т.е. повысится шанс того, что нулевая гипотеза не будет отвергнута тогда, когда она должна быть отвергнута.

Если с помощью критерия мы определили, что в наших данных есть случайные блуждания, то вместо исходных данных мы рассматриваем первую разность, т.е. $Y_t - Y_{t-1}$. Далее с помощью критерия можно подтвердить, что мы избавились от случайного блуждания.

Почему удается избавиться от блужданий с помощью первой разности?
Посмотрим на то, как выглядит случайное блуждание: $Y_t = Y_{t-1} + \varepsilon_t$.
Мы видим, что $Y_t - Y_{t-1} = \varepsilon_t$, т.е. первая разность равна белому шуму - процессу, стационарному в широком смысле.

Если мы возьмем два независимых случайных блуждания и попробуем применить регрессию к одному случайному блужданию по другому, то очень часто критерий Стьюдента будет показывать, что этот коэффициент является значимым. В то же время, блуждания независимы, и, следовательно, между ними не может быть никакой зависимости. Это называется ложная зависимость (spurious regression).

Почему так происходит?
Одно из основных предположений регрессии состоит в том, что остатки ε_t являются независимыми и одинаково распределенными. Если это предположение не выполняется, а в случае, когда мы берем 2 случайных блуждания, оно не выполняется, критерий Стьюдента не работает.
Что делать в этом случае?
Нужно избавляться от нестационарности в исходных данных, а именно брать первые разности.