Отчет по практическому заданию 2.

Анализ временных рядов.

Команда:

Раева Анастасия - 312 группа

Богданова Елена - 312 группа

Чистяков Иван - 312 группа

Постановка задачи.

1) Считать данные из файла, проверить, является ли ряд стационарным в широком смысле.

2) Провести визуальную оценку, визуализировать ряд и скользящую статистику.

3) Оценить достоверность статистики.

4) Разложить временной ряд на тренд, сезональность, остаток в соответствии с аддитивной, мультипликативной моделями.

5) Проверить, является ли ряд интегрированным порядка k. Если да, то применить

модель ARIMA.

6) Визуализировать решение

Подход к решению задачи.

Временной ряд -последовательность значений(измерений), упорядоченных в неслучайные моменты времени. Анализ временных рядов основывается на предположении, что последовательные значения в файле данных наблюдаются через равные промежутки времени.

Цель анализа временных рядов - прогнозирование (предсказание будущих значений временного ряда по настоящим и прошлым значениям).

Отличительная особенность статистического анализа временных рядов состоит в том, что последовательность наблюдений рассматривается как реализация последовательности, вообще говоря, статистически зависимых случайных величин. Чтобы сделать задачу статистического анализы доступной для практического решения, приходится ограничивать класс рассматриваемых моделей временных рядов, вводя те ли иные предположения относительно структуры ряда и его вероятностных характеристик. Одно из таких ограничений предполагает стационарность.

Под *стационарностью* будем понимать, что у временного ряда некоторые свойства не меняются с течением времени. В анализе временных рядов стационарные ряды имеют постоянные по времени среднее, дисперсию и автокорреляции.

Ряд называется *строго стационарным* или стационарным в узком смысле, если совместное распределение наблюдений не зависит от сдвига по времени.

Ряд называется *слабо стационарным* или стационарным в широком смысле, если такие статистические характеристики временного ряда как его математическое ожидание (среднее), дисперсия (ср. кв. отклонение) не зависят от момента времени, а ковариационная функция зависит только от сдвига.
Если нарушается хотя бы одно из этих условий, то ряд называется не стационарным.

Отклонение - это разность числа и некоторой точки отсчета, чаще всего это среднее арифметическое или медиана.

Ковариация - в теории вероятности мера линейной зависимости двух случайных величин.

Дисперсия выборки - это среднее арифметическое квадратов отклонений

Анализ временных рядов предполагает,что данные содержат систематическую составляющую и случайный шум. Большинство регулярных составляющих временных рядов принадлежит к двум классам: они являются либо трендом, либо сезонной составляющей.

Тренд представляет собой общую систематическую линейную или нелинейную компоненту, которая может изменяться во времени. Под трендом временного ряда будем понимать изменение, определяющее общее направление развития ряда - рост, падение, неизменность.

Сезонность - периодические колебания уровней временного ряда внутри года.

Остаток – величина, показывающая нерегулярную (не описываемую трендом и сезонностью) составляющую исходного ряда в определенном временном интервале. Фактически, остатком называется разница между предсказанным и наблюдаемым значением.

Таким образом, каждый уровень временного ряда может формироваться из трендовой, сезонной компоненты и шума(ошибка).

T - тренд, S - сезонная компонента, E - ошибка.

Модель, где временной ряд представляется в виде суммы перечисленных компонент называется аддитивной. Если ряд представляется в виде произведения этих компонент, то мультипликативной.

Аддитивная модель: $Y = T + S + E$

Мультипликативная: $Y = T * S * E$

Скользящее среднее – семейство функций, значения которых в каждой точке равны среднему значению исходной функции за предыдущий период.

Стандартное отклонение – это показатель рассеивания значений случайной величины относительно её математического ожидания. Оно демонстрирует, на сколько в среднем отклонился ряд от своей средней вариации (то есть от среднего арифметического).

Скользящее среднее и стандартное отклонение называются скользящими статистиками.

Автоковариационная функция - совокупность значений ковариаций при всевозможных значениях T' (Обозначение y). Является четной функцией для строго стационарного ряда.

Коэффициент корреляции - Corr(T') = □(□′) □(0) График Corr(T') коэффициента корреляции носит название коррелограммы.

Визуальная оценка в работе проводится следующим образом. Обратим внимание на наличие у графика тренда. Тренд на изменения есть - ряд нестационарный, тренда нет - ряд стационарен. Стоит отметить, что эта оценка довольно грубая, и во многих случаях из внешнего вида графика нельзя сказать, стационарен ряд или нет.

Авторегрессионная модель (AR) - модель временного ряда, в которой значение ряда y(t) линейно выражается через предыдущие значения этого же ряда. Если зависимость происходит только от последних p значений, то говорят, что задан авторегрессионный временной ряд порядка p (или AR(p)).

В процессе решения данной задачи применяется *тест Дики-Фуллера*. Это метод анализа временных рядов на стационарность. Суть метода заключается в проверке ряда на наличие так называемых единичных корней.

Временной ряд имеет единичный корень (хотя бы один), если его первые разности образуют стационарный ряд. (Обозначение Y(t) ~ I(1), т.е. ΔY(t) = Y(t) - Y(t-1) ~ I(0), где Δ - разностный оператор, I(j) - означает, что ряд является интегрированным порядка j, I(0) - ряд стационарен)

Интегрированный временной ряд - нестационарный временной ряд, разности некоторого порядка от которого являются стационарным рядом.

Временной ряд называется *интегрированным порядка k*, если разности ряда k-го порядка `Δ^k x(t)` являются стационарными, а разности меньшего порядка (и сам временной ряд соответственно) не являются стационарными рядами.

- Δ Y(k) = Y(k+1) - Y(k)
- Δ^2 Y(k) = Δ Y(k+1) - Δ Y(k) = Y(k+2) - 2*Y(k+1) + Y(k)
- Δ^m Y(k) = Δ^m-1 Y(k+1) - Δ^m-1 Y(k), где Δ - разностный оператор Обозначают Y(k) ~ I(k) - интегрированный временной ряд порядка k

Фактически, тест Дики-Фуллера проверяет значение коэффициента `a` в авторегрессионном уравнении 1-го порядка - AR(1). Оно имеет вид: Y(t) = a * Y(t-1) + ε(t), где ε(t) - ошибка значения.

В результате работы метода возможны 3 исхода:

- a=1 => есть единичные корни => стационарности нет
- |a| нет единичных корней => есть стационарность
- |a|>1 => не свойственно для временных рядов, которые встречаются в реальной жизни - требуется более сложный анализ.

Для удобства проведем преобразование уравнения:

Y(t) = a * Y(t-1) + ε(t) =>

Y(t) - Y(t-1) = a * Y(t-1) - Y(t-1) + ε(t) =>

ΔY(t) = (a-1) * Y(t-1) + ε(t) =>

ΔY(t) = b * Y(t-1) + ε(t), где b = a-1

- Основная гипотеза: H0: b = 0 - процесс не стационарен
- Альтернативная гипотеза: H1: b

Уровень значимости - показывает (обычно в процентном выражении) степень отклонения от гипотезы. Грубо говоря, если наша достоверность превысила, скажем, 5% уровень значимости, то гипотеза отвергается, что говорит о том, что процесс не стационарен. Если достоверность статистики больше критических значений, тогда единичный корень есть и ряд не стационарен. Иначе ряд является стационарным.

Модель ARIMA

ARIMA (autoregressive integrated moving average model) - интегрированная модель авторегрессии скользящего среднего - модель анализа временных рядов. Это расширение моделей ARMA для нестационарных временных рядов. ARMA (AutoRegressive Moving Average model) - математическая модель, используемая для анализа и прогнозирования стационарных временных рядов. Объединяет 2 более простые модели: авторегрессии (AR) и скользящего среднего (MA). AR (autoregressive model) - авторегрессионная модель временных рядов, в которой значение временного ряда в данный момент зависит от предыдущих значений этого же ряда. MA (moving average model) - модель скользящего среднего, в которой моделируемый уровень временного ряда можно представить как линейную функцию прошлых ошибок, т.е. разностей между прошлыми фактическими и теоретическимиуровнями.

Алгоритм построения модели ARMA заключается в поиске коэффициентов p, q - порядков для моделей AR(p) и MA(q). Это позволит построить функцию автокорреляции и функцию частичной автокорреляции.

Таким образом, построение ARIMA зависит от 3 параметров: ARIMA(p, d, q), где p - порядок AR(p), d - порядок интегрированности, q - порядок MA(q). ARIMA(p, d, q) = AR(p) + MA(q) ~ I(d)

Описание подсчета коэффициента детерминации R2

Коэффициент детерминации (R2) - доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости. Более точно — это единица минус доля необъяснённой дисперсии (дисперсии случайной ошибки модели, или условной по факторам дисперсии зависимой переменной) в дисперсии зависимой переменной. Лучший результат, который может дать R2 score — это 1.0. Однако возможны как отрицательные значения, так и значения, превышающие 1.0. Это говорит об очень большой неточности предсказания модели.

Информационный критерий - применяемая в эконометрике мера относительного качества эконометрических моделей, учитывающая степень "подгонки" модели под данные с корректировкой на используемое количество оцениваемых параметров. Т.е. критерии основаны на некотором компромиссе между точностью и сложностью модели. Критерии различаются тем, как они обеспечивают этот баланс. информационные модели используются исключительно для сравнения моделей между собой, без содержательной интерпретации значений этих критериев. Обычно чем меньше значения критериев, тем выше относительное качество модели.

AIC (an information criterion) - информационный критерий Акаике – критерий, применяющийся исключительно для выбора из нескольких статистических моделей. AIC = 2k - 2ln(L), где где k — число параметров в статистической модели, и L — максимизированное значение функции правдоподобия модели.

Описание решения задачи

1. С помощью функции read_exel считываем данные из файла

2. Визуализация.

а) Выводим ряд

б) С помощью функции rolling и mean считаем скользящие статистики. Среднее, с помощью функции std - стандартное отклонение. Выводим их, делаем выводы о стационарности ряда и достоверности статистики.

На всякий случай пррверяем себя функцией DF_test

3. С помощью функции decompose_time_series декомпозируем временной ряд

rolling - вычисление скользящего окна, mean возвращает среднее значение для этого окна. Получаем плавающее среднее=тренд.

Считаем detrend в соответствии с моделью - вычитаем тренд, если модель аддитивная, делим на тренд, если модель мультипликативная. * axis = 0 => рассматриваем столбцы

С помощью функции seasonal_mean нормализуем весь ряд

Shape возвращает количество строк

Title создает массив из повторений первого аргумента, количество повторений = 2-му аргументу

С их помощью (shape и title) получаем сезонность

Ищем шум(resid) - аддитивная - вычитаем тренд и сезонность, иначе делим на все

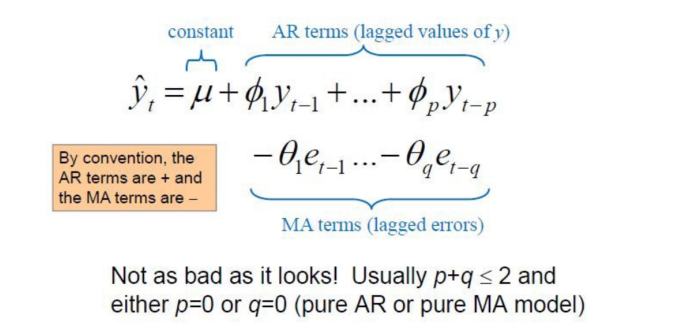
Выводим полученные ряды, проверяем на стационарность с помощью seasonal_decompose,

Сначала декомпозируем по аддитивной модели, потом все то же самое, но для аддитивной. Применяем decompose_time_series, затем проверяем на тренд, сезонность, шум. Dropna удаляет требуемое. Далее проводим тест Дики-Фуллера с помощью Diki_Fuller _test

4. С помощью функции k_integrability считаем порядое интегрируемости.

*Автокорреляция - корреляция между исходным рядом и его версией, который сдивнут на некий лаг **t** (лаг автокорреляции) или сила линейной взаимосвязи между последовательными показателями, упорядоченными во времени.

С помощью output_acf_and_pacf выводим автокорреляционную(acf) и частную автокорреляционнуюю(pacf) функции. Acf помогает определить порядок скользящего среднего(MA), т. к. по ее коррелограмме можно определить количество автокорреляционных коэффициентов сильно отличных от 0 в модели MA. PACF поможет определить порядок авторегрессии(AR), т. к. по ее коррелограмме можно определить



максимальный номер коэффициента сильно отличный от 0 в модели AR.

Делаем предсказание с помощью модели ARIMA

Выводим все.

Вывод

Для данного нам временного ряда мы построили графики скользящей статистики и по нему определили, что ряд нестационарен, так как его мат.ожидание и дисперсия зависят от времени. Нестационарность данного ряда подтверждает и тест Дики-Фуллера. Значение статистики лежит правее критического значения при уровне значимости 5%, поэтому нулевая гипотеза о единичном корне не отклоняется, то есть процесс может содержать единичные корни. Достоверность статистики больше уровня значимости. После разложения временных рядов на тренд, сезональность и остаток, мы заметили, что для обоих моделей оригинальный ряд и трендовая составляющая нестационарны, так как их скользящая статистика зависит от времени. Однако так же для обоих моделей справедлива стационарность сезонной составляющей и остатка, их скользящая статистика не зависит от времени. Наилучшее значение r2score = 1. Если оно будет отклоняться вправо или влево, то велика вероятность неточности построения модели. Более удачной моделью будет модель с меньшим AIC. PEP8 был соблюден автокорректором кода (autoformatted).