Task 2 : Теоретическая база

Временной ряд(time series).

Временным рядом будем называть совокупность наблюдений некоторой величины в различные моменты времени. Будем рассматривать временной ряд как выборку из последовательности случайных величин:

$$X_t, t \in [1, T]$$

Совокупность случайных величин $\{X_t, t \in [1, T]\}$ мы будем называть дискретным случайным(стохастическим) процессом.

Стационарность.

Случайный процесс является стационарным в широком смысле, если у него:

- существует не зависящее от времени математическое ожидание
- существует не зависящая от времени дисперсия
- автокорреляционная функция зависит только от $(t_1 - t_2)$

Любой стационарный в широком смысле случайный процесс может быть представлен в виде:

$$X_t - \mu_t = \sum_{\tau=0}^{\infty} \psi_\tau \varepsilon_{t-\tau} \quad — \quad \text{разложение Вольда,}$$

μ_t - математическое ожидание процесса, ψ_τ-весовые коэффициенты, ε_j-белый шум с конечным математическим ожиданием и дисперсией.
Примечание:

$$cov(X_{t_1}, X_{t_2}) = \iint (x_{t_1} - \mu)(x_{t_2} - \mu) f_2(x_{t_1}, x_{t_2})\, dx_{t_1}\, dx_{t_2},$$

$f_2(x_{t_1}, x_{t_2})$-функция плотности распределения.

Интеграл не изменяется при сдвиге времени. Следовательно, ковариация может рассматриваться как функция не двух переменных, а единственной переменной: разности $(t_1 - t_2)$. Совокупность значений ковариаций при всевозможных значениях расстояния между моментами времени называется автоковариационной функцией случайного процесса.

Автокорреляционная функция временного ряда(показывает,насколько статистически зависимы значения временного ряда):

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)},\ \gamma(\tau) - \text{всевозможные значения автоковариации}, \tau \in \mathbb{R}$$

Белый шум(white noise).

Процесс ε_t, удовлетворяющий условиям теоремы Гаусса-Маркова, называется белым шумом.

Свойства:

$$\mathbb{E}(\varepsilon_t) = 0$$
$$Var(\varepsilon_t) = \sigma^2$$
$$Cov(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j.$$

Данный процесс заведомо стационарен в широком смысле.

Тест Дики-Фуллера.

Назначение: проверка временного ряда на стационарность.

Идея: использование единичных корней. Временной ряд имеет единичный корень, или порядок интеграции один, если его первые разности образуют стационарный ряд.

При помощи этого теста проверяют значение коэффициента α в авторегрессионном уравнении первого порядка AR(1): $X_t = \alpha X_{t-1} + \varepsilon_t$.
Возможные ситуации:

- $\alpha = 1$ — ряд нестационарен, является интегрированным временным рядом первого порядка
- $|\alpha| < 1$ — ряд стационарный
- $|\alpha| > 1$ — процесс является «взрывным», не рассматривается

Проверка, используемая в программе:

Уравнение AR(1) перепишем в виде:

$$\Delta X_t = bX_{t-1} + \varepsilon_t,\ b = \alpha - 1, \Delta : \Delta X_t = X_t - X_{t-1}$$

Нулевая гипотеза $H_0 : b = 0$ – существует единичный корень, ряд нестационарный.

Альтернативная гипотеза $H_1 : b < 0$.

Статистика теста (DF-статистика) — это обычная t-статистика для проверки значимости коэффициентов линейной регрессии. Однако, распределение данной статистики отличается от классического распределения t-статистики. Распределение DF-статистики выражается через винеровский процесс и называется распределением Дики — Фуллера.

Существует три версии теста:

Без константы и тренда:

$$\Delta X_t = bX_{t-1} + \varepsilon_t$$

С константой:

$$\Delta X_t = b_0 + bX_{t-1} + \varepsilon_t$$

С константой и линейным трендом:

$$\Delta X_t = b_0 + b_1 t + bX_{t-1} + \varepsilon_t$$

Для каждой из трёх тестовых регрессий существуют свои критические значения DF-статистики, которые берутся из специальной таблицы Дики — Фуллера. Если значение статистики лежит левее критического значения (критические значения — отрицательные) при данном уровне значимости, то нулевая гипотеза о единичном корне отклоняется и процесс признается стационарным (в смысле данного теста). В противном случае гипотеза не отвергается и процесс может содержать единичные корни, то есть быть нестационарным (интегрированным) временным рядом.

В процессе выполнения программы так же фигурирует p-значение (p-value)– величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы (ошибки первого рода). Проверка гипотез с помощью p-значения является альтернативой классической процедуре проверки через критическое значение распределения.

Тренд, сезонность, остаток. Аддитивная и мультипликативная модели.

Тренд — тенденция изменения показателей временного ряда. Параметрические методы оценки тренда рассматривают временной ряд как гладкую функцию от t: $X_t = f(t, \theta),\ t = 1, \ldots, n$; затем различными методами оцениваются параметры функции θ, например, методом наименьших квадратов.

Сезонность – строго периодические и связанные с календарным периодом отклонения от тренда. Перед выделением сезонных колебаний необходимо вычислить период сезонности. Если период не известен заранее, то его можно найти с помощью автокорреляционной функции.

Общий вид аддитивной модели следующий:

$$Y = T + S + E$$

Общий вид мультипликативной модели выглядит так:

$$Y = T \cdot S \cdot E\cdot$$

T – трендовая компонента, S – сезонная компонента, E – остаток

Процесс MA(q)(скользящего среднего – moving average).

Случайный процесс является процессом скользящего среднего порядка q, если в разложении Вольда присутствует только q слагаемых:

$$x_t = \sum_{\tau=0}^{q} \psi_\tau \varepsilon_{t-\tau}, \quad x_t = X_t - \mu_t$$

Процесс AR(p)(авторегрессии – autoregressive).

Случайный процесс является процессом авторегрессии порядка p, если значение случайного процесса определяется линейной комбинацией конечного числа его предыдущих значений и добавлением белого шума:

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \varepsilon_t \ , \text{ где } \varepsilon_t - \text{белый шум.}$$

Интегрированный временной ряд.

Интегрированным временным рядом порядка k называется нестационарный временной ряд, разности k-ого порядка от которого являются стационарным временным рядом. При этом разности меньшего порядка не являются стационарными.

Процесс ARIMA(p,d,q)(авторегрессии-интегрированного скользящего среднего – autoregressive integrated moving average).

ARIMA(p, d, q):

- p – параметр AR-части
- q – параметр MA-части
- d – степень интеграции
- $\alpha_p(L)(1-L)^d X_t = \beta_q(L)\varepsilon_t$

Примечание:

Оператор сдвига(лага):

$$L : LX_t = X_{t-1}$$
$$L^p X_t = X_{t-p}$$

Операторный полином:

$$\alpha_p(L) = 1 - \alpha_1 L - \ldots - \alpha_p L^p$$
$$\beta_q(L) = (L^0 + \beta_1 L + \ldots + b_q L^q)\varepsilon_t$$

Пример:

ARIMA$(0, 1, 0)$— процесс случайного блуждания.

Применение модели ARIMA.
Информационный критерий Акаике (AIC):

AIC(Akaike's information criterion) – критерий выбора из класса параметризованных регрессионных моделей. Лучшая модель соответствует минимальному значению критерия Акаике. Абсолютное значение критерия не несет в себе полезной информации.
Коэффициент детерминации(R^2):

Коэффициент детерминаци – это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью. Более точно — это единица минус доля необъяснённой дисперсии (дисперсии случайной ошибки модели, или условной по признакам дисперсии зависимой переменной) в дисперсии зависимой переменной. Коэффициент детерминации для модели с константой принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется как соответствие модели данным. Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 50%. В общем случае коэффициент детерминации может быть и отрицательным, это говорит о крайней неточности модели.
Фильтр Калмана:

Фильтр Калмана - мощнейщий инструмент фильтрации данных. При фильтрации используется информация о физике самого явления. Возвращает объект ARIMAResults - содержит также и результаты предсказаний, которые можно выводить и использовать в дальнейших прогнозах