Time Series analysis with Python

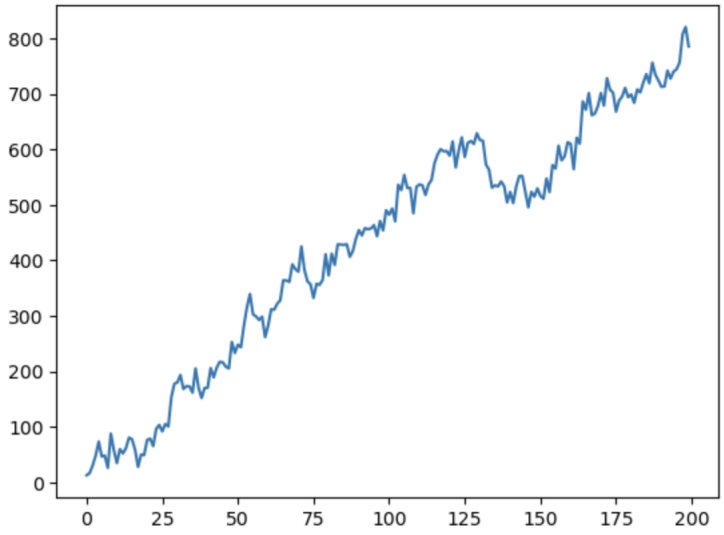
Вариант 3. Бакушкин Даниил

30 марта 2023 г.

Содержание

Определение типа наблюдаемого ряда с помощью процедуры Доладо-Дженкинса-Сосвилла-Риверо

Анализируемый ряд:



Шаг №1

Рассмотрим модель:

$X_t = \mu + bt + \alpha_1 X_{t-1} + \varepsilon_t$

Гипотезы:

$H_0 := \alpha_1 = 1$

$H_A := \alpha_1 < 1$

Согласно тесту Дикки-Фуллера, статистика ADF превышает все критические значения при уровнях значимости 1%, 5% и 10%. Поэтому мы не можем отвергнуть нулевую гипотезу о том, что ряд имеет единичный корень и является нестационарным. Это указывает на то, что серия, вероятно, представляет собой случайное блуждание с дрейфом.

```
ADF Statistic: -0.915124
p-value: 0.782951
Critical Values:
        1%: -3.465
        5%: -2.877
        10%: -2.575
```

Шаг №2

Рассмотрим модель:

$\Delta X_t = \mu + bt + \varepsilon_t$

Гипотезы:

$H_0 := \alpha_1 = 1$

$H_A := \alpha_1 < 1$

Применим критерий Стьюдента. Считаем статистику.

H_0 не отвегается $=>$ переходим к следующему шагу.

Шаг №3

Рассмотрим модель:

$X_t = \mu + \alpha_1 X_{t-1} + \varepsilon_t$

Гипотезы:

$H_0 := \alpha_1 = 1$

$H_A := \alpha_1 < 1$

Используем ADF с распределением DF_μ

H_0 не отвегается $=>$ переходим к следующему шагу.

Шаг №4

Рассмотрим модель:

$\Delta X_t = \mu + \varepsilon_t$

Гипотезы:

$H_0 := \mu = 0$

$H_A := \mu 0$

Используем критерий Стьюдента.

H_0 не отвегается $=>$ переходим к следующему шагу.

Шаг №5

Рассмотрим модель:

$X_t = \alpha_1 X_{t-1} + \varepsilon_t$

Гипотезы:

$H_0 := \alpha_1 = 1$

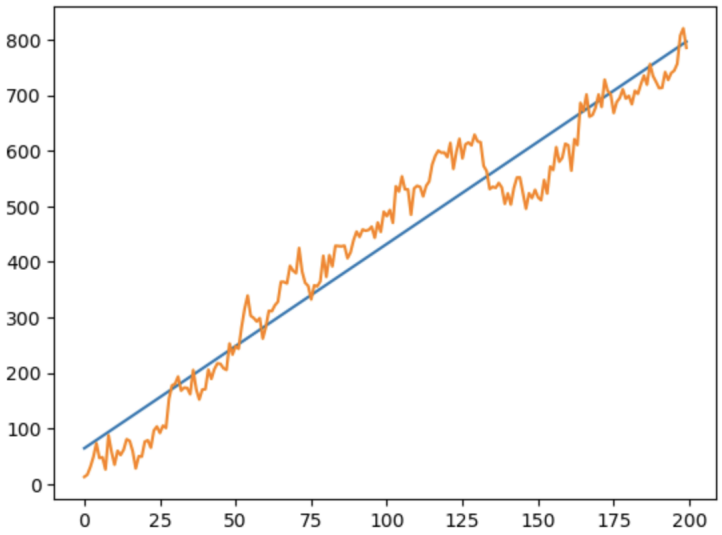
$H_A := \alpha_1 < 1$

Используем ADF с DF_0 распределение

Гипотеза H_0 не отвергается $=>$ выбираем модель: случайное блуждание с дрейфом.

Оценка по МНК детерминированной составляющей ряда

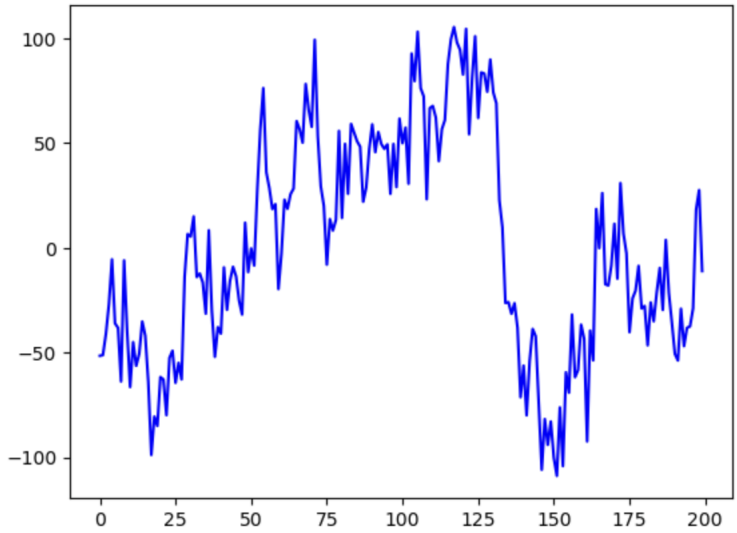
С помощью МНК приблизим наш график и оценим $r^2 score$.



$r^2 score = 0.942$ – высокий скор означает, что линейный тренд хорошо аппроксимирует наш ряд.

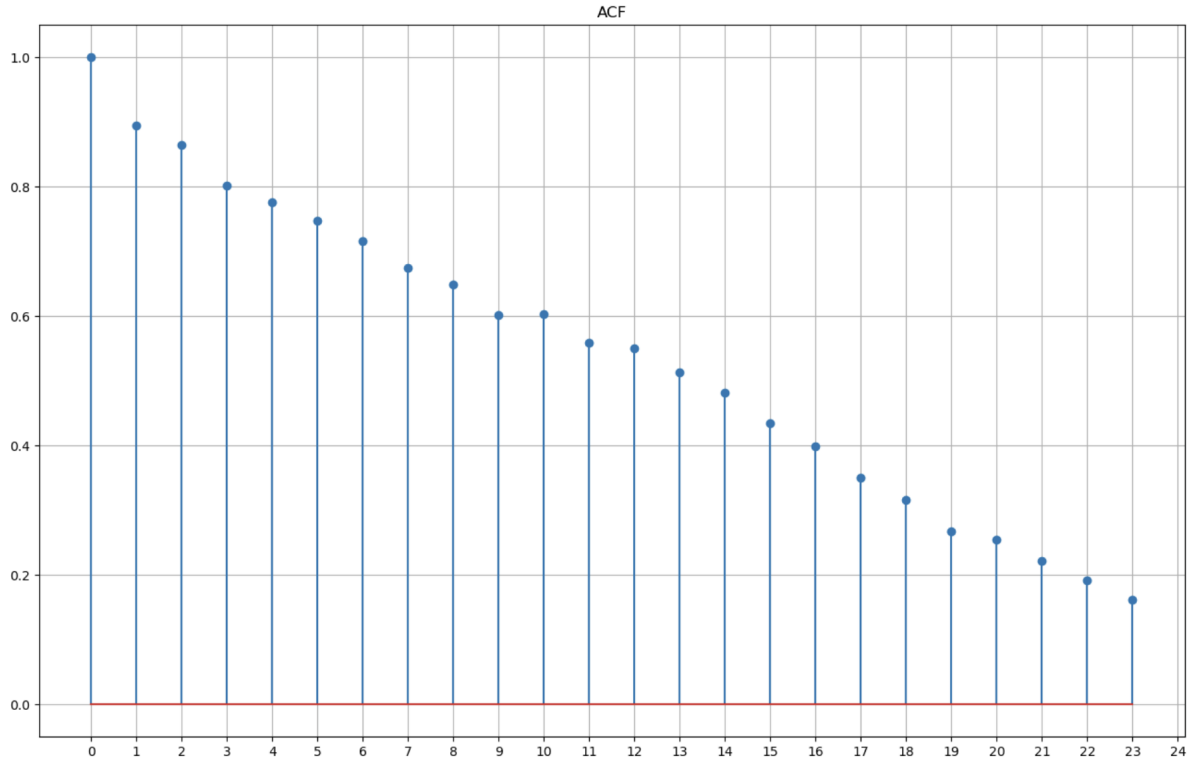
Удаление тренда из ряда

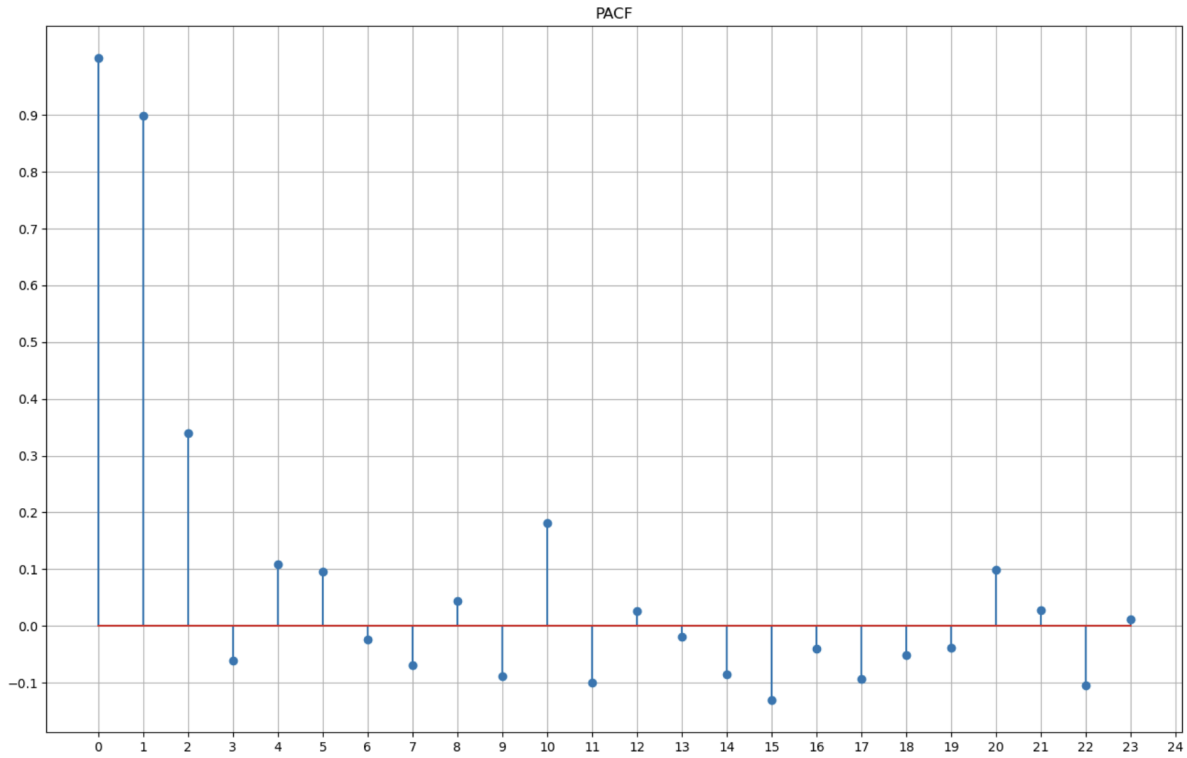
Удаляем тренд и отображаем на графике остатки.



Проведение идентификации случайной составляющей ряда

Построим ACF и $PACF$ для временного ряда.





Анализ ACF and PACF

На вероятностном интервале 95% найдем порог отсечения.

$n = 200$
$\frac{1.96}{\sqrt{(n)}} = 0.139$

Теперь будем смотреть, с какого лага корреляция будет входить в этот диапозон.

Рассмотрим 3 модели:

1. Белый шум
2. Moving Average
3. Auto Regression

Анализ:

1. Это не белый шум, так как автокорреляционная функция ненулевая.
2. Можно аппроксимировать с помощью Moving Average очень высокого порядка > 23, но так как мы должны выбирать наиболее простую модель, то перейдем к AR
3. Если посмотреть на PACF, то видно, что все значения входят в наш диапозон, начиная с 3. Будем считать, что значение 10 – выброс.

Основной сделаем модель: **Auto Regression порядка 3**.

Оценка параметров выбранных моделей

Рассчитаем коэффициенты для:

AU порядка 3

	coef
const	−0.0134
ar.L1	0.6196
ar.L2	0.3693
ar.L3	−0.0652

MA порядка 25

	coef
const	−0.0090
ma.L1	0.6311
ma.L2	0.7108
ma.L3	0.4719
ma.L4	0.4642
ma.L5	0.4914
ma.L6	0.4629
ma.L7	0.4190
ma.L8	0.3603
ma.L9	0.1595
ma.L10	0.3523
ma.L11	0.2118
ma.L12	0.2767
ma.L13	0.3044
ma.L14	0.4531
ma.L15	0.4425
ma.L16	0.5567
ma.L17	0.6234
ma.L18	0.6381
ma.L19	0.4477
ma.L20	0.5037
ma.L21	0.3841
ma.L22	0.2120
ma.L23	0.0419
ma.L24	−0.1079
ma.L25	−0.0863

Выбор модели с помощью критериев Акаике и Шварца

Показатели AIC и BIC для AR(3) : 1815.07, 1831.56.

Показатели AIC и BIC для MA(25) : 1834.16, 1923.22.

Оба показателя для MA $>$ AR, следовательно, модель AR(3) лучше.

Диагностика остатков

Диагностику будем проводить с помощью статистики Льюнга-Бокса.

	lb_stat	lb_pvalue
1	0.009724	0.921446
2	0.167396	0.919709
3	2.518092	0.472030
4	2.735424	0.603030
5	3.957009	0.555622
6	4.395111	0.623369
7	4.422533	0.730024
8	4.487758	0.810657
9	9.131706	0.425207
10	13.443451	0.199923
11	13.943806	0.236119
12	15.734568	0.203698
13	16.807774	0.208241
14	17.708791	0.220367
15	17.713385	0.278033
16	17.871691	0.331464
17	17.909612	0.394565
18	17.912837	0.461408
19	22.392726	0.265149
20	23.294954	0.274545

Проверим гипотезу о белошумности остатков (независимости) - H_0.

Против альтернативной гипотезы $H_A := \sum p_\varepsilon^2(i) > 0$

Статистика критерия:

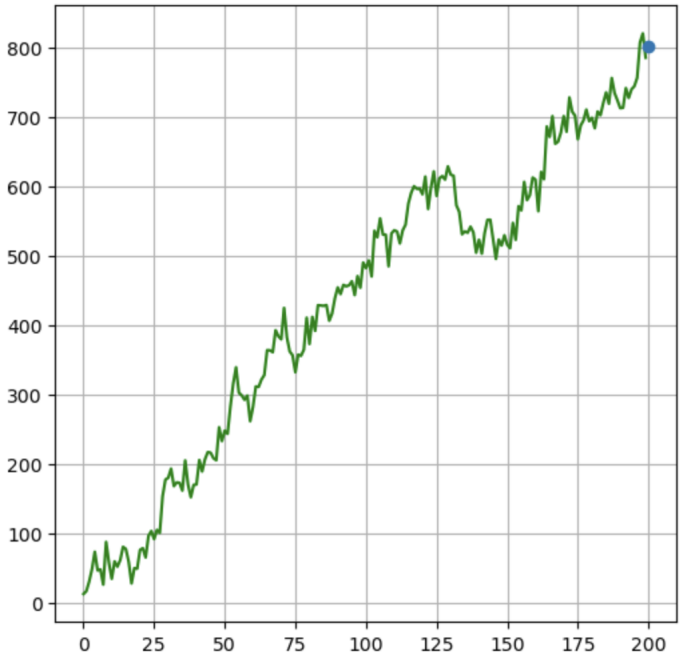
$T^* = n \sum_i^m = 1 \hat{p}_\varepsilon^2(j)$

Считаем.

Мы не можем отвергнуть H_0, т.к. у нас нет значений > 31.4.

Построение прогноза

Прогноз с помощью AR(3) вместе с трендом дает для 201 точки значение: 800.06372. Ниже ее график.



Приложение

Код, использовавшийся для проведения анализа:

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt

def read_file():
    # Вариант 4
    path = "/kaggle/input/200samples/() 48 .txt"
    with open(path, 'r',encoding='windows-1251') as f:
        content = f.read().split('\n\n')[3].split('\n')[-1]
        content = list(map(np.float32, content.split(',')))
        return pd.Series(content)

data = read_file()
plt.plot(data)
plt.show()

# Определение типа наблюдаемого ряда с помощью процедуры Доладо-Дженкинса-Сосвилла-Риверо
from statsmodels.tsa.stattools import adfuller
adfuller(data)

# Оценка по МНК детерминированной составляющей ряда
x = np.arange(0, 200)
y = data

from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
lr = LinearRegression()
lr.fit(x.reshape(-1, 1), y)
preds_linear = lr.predict(x.reshape(-1, 1))

print(r2_score(y, preds_linear))
plt.plot(preds_linear)
plt.plot(y)
plt.show()

# Удаление тренда из ряда
```

```
residuals = y - preds_linear
plt.plot(residuals_linear, color='blue')
plt.show()


# Проведение идентификации случайной составляющей ряда
# строим ACF, PACF
from statsmodels.tsa.stattools import acf, pacf

plt.figure(figsize=(16, 10))
plt.xticks(np.arange(0, 25, 1))
plt.yticks(np.arange(round(pacf_data.min(), 1), round(pacf_data.max(), 1), 0.1))
plt.grid()
plt.stem(acf_data)
plt.title('ACF')
plt.show()

pacf_data = pacf(residuals)
plt.figure(figsize=(16, 10))
plt.xticks(np.arange(0, 25, 1))
plt.yticks(np.arange(round(pacf_data.min(), 1), round(pacf_data.max(), 1), 0.1))
plt.grid()
plt.title('PACF')
plt.stem(pacf_data)
plt.show()

# Оценка параметров выбранных моделей

from statsmodels.tsa.arima.model import ARIMA
modelAR = ARIMA(residuals, order=(3,0,0)).fit() # AR 3
modelMA = ARIMA(residuals, order=(0,0,25)).fit() # MA 25

# коэффициенты AR
modelAR.summary()
# коэффициенты MA
modelMA.summary()
# Выбор модели с помощью критериев Акаике и Шварца

print(modelAR.aic, 'Акаике')
```

```
print(modelAR.bic, 'Шварца')
print(modelMA.aic, 'Акаике')
print(modelMA.bic, 'Шварца')

# Диагностика остатков

from statsmodels.stats.diagnostic import acorr_ljungbox as lb
import pandas as pd
lb(modelAR.resid, lags=20)

# Построение прогноза
plt.figure(figsize=(6, 6))
pred_next = modelAR.predict(200) + lr.predict(np.array([200]).reshape(1, -1))[0]

plt.grid()
plt.plot(data, color='green'),
plt.plot(200, pred_next, 'o')
plt.show()
```