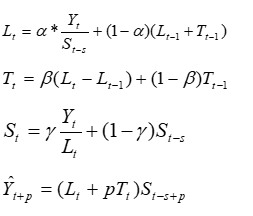
Лабораторная работа №

«Прогнозирование временного ряда в Python с помощью метода Хольта-Уинтерса»

**Цель работы:** получение навыков прогнозирования с помощью метода Хольта-Уинтерса и оценки результатов в Python.

**ТЕОРЕТИЧЕСКАЯ ЧАСТЬ**

Как мы знаем, временной ряд можно разложить на 3 составляющие: тренд, сезонность и случайная составляющая. Модель Хольта-Уинтерса использует идеи экспоненциального сглаживания, но является более сложной и может применяться к рядам, содержащим тенденцию и сезонность. Она работает по следующим формулам:



где

Lt – новая сглаженная величина;

α – постоянная сглаживания для данных (0 ≤ α ≥ 1);

Yt – новое наблюдение или реальное значение ряда в период t;

β - постоянная сглаживания для оценки тренда (0 ≤ β ≤1);

Тt - собственно оценка тренда;

(гамма) - постоянная сглаживания для оценки сезонности;

St – оценка сезонности;

s – длительность периода сезонного колебания;

р – количество периодов вперед, на которое делается прогноз;

Ŷ t+p - прогноз на р периодов вперед.

К счастью, модель реализована в одной из библиотек Python и не требуется программировать все формулы самостоятельно. Прогнозирование можно осуществить используя метод **statsmodels.tsa.holtwinters.ExponentialSmoothing()**.

В данном методе, в зависимости от наличия во временном ряде тренда и сезонности, добавляются или убираются параметры **trend** и **seasonal**, также нужно определить периодичность сезонности параметром **seasonal\_periods**.

**Пример 1**

Для начала нужно скачать необходимые компоненты, нам понадобятся следующие библиотеки: **pandas** для упрощения работы с табличными данными, **numpy** для математических операций, **statsmodels** для разбиения временного ряда на компоненты и последующего прогнозирования, а также **sklearn** для оценки построенной модели.

Перед началом работы нам потребуется установить данные библиотеки, это можно сделать с помощью следующих команд:

*!pip install pandas*

*!pip install numpy*

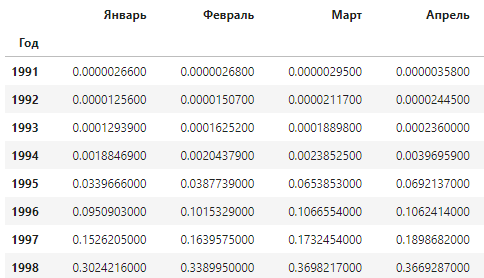
*!pip install statsmodels*

*!pip install sklearn*

В данной работе попробуем спрогнозировать среднюю месячную зарплату в Беларуси. Для этого понадобится скачать данные по [ссылке](https://www.belstat.gov.by/upload-belstat/upload-belstat-excel/Oficial_statistika/2022/nach_sr_zarplata-91-22-2202.xlsx) или с официального сайта Белстат. После скачивания необходимо будем провести действия по приведению данных в одинаковый вид (убрать комментарии, привести значения средней заработной платы в вид 2016 г. с учетом всех деноминаций). После данной операции, прочитаем данные с помощью метода **pandas.read\_csv()**, поставив первую колонку как индекс и сбросим 2022 год для упрощения последующих вычислений.

*df = pandas.read\_csv('salary.csv', index\_col=0).drop(2022)*

В итоге наша таблица примет следующий вид:



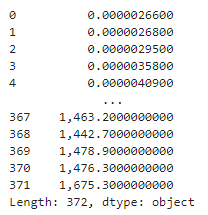
Сейчас данные представлены в табличном виде, однако для вычислений нам потребуется чтобы данные за каждый месяц шли один за другим.

*monthly\_salaries = []*

*for row in range(len(df)):*

*monthly\_salaries.extend(list(df.iloc[row]))*

Получим следующий вид:



# Для того чтобы избавиться от вероятности применения неправильной модели - логарифмируем наши данные (данная операция позволит без опасений применять аддитивную модель разложения временного ряда) - это можно сделать с помощью метода log() библиотеки numpy. Далее произведем предобработку данных:

* Заменим индекс на Datetime
* Определим параметр frequency как месячный

*monthly\_salaries = pandas.Series(monthly\_salaries, index=pandas.date\_range(start='1/1/1991', periods=len(monthly\_salaries), freq='M'))*

# Всего у нас есть 372 наблюдений, для тренировки оставим 350, а для тестирования оставшиеся 22.

*train = monthly\_salaries [:350]*

*test = monthly\_salaries [350:]*

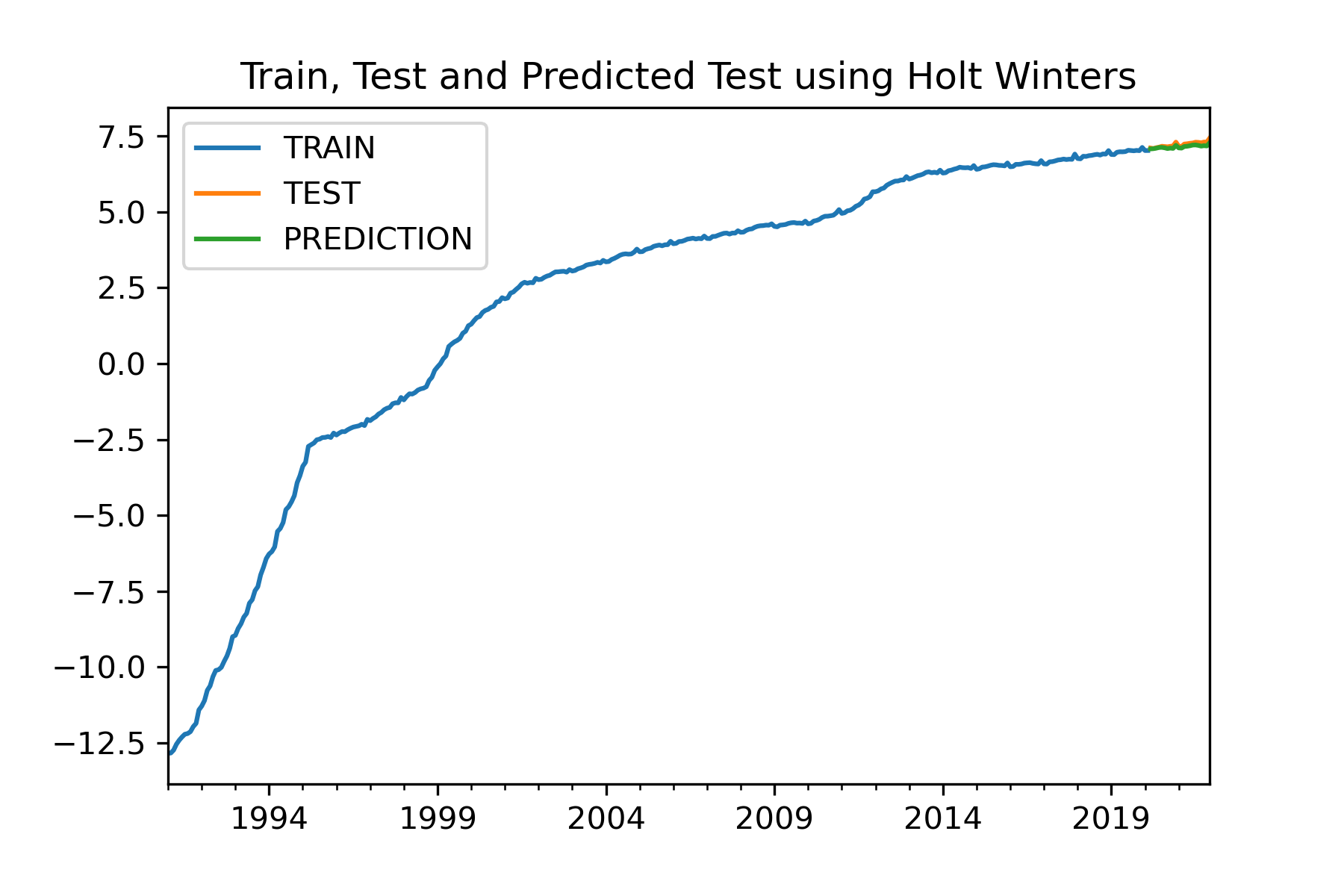
# Теперь тренируем нашу модель

# *fitted\_model = statsmodels.tsa.holtwinters.ExponentialSmoothing(train,trend='add',seasonal='add',seasonal\_periods=12).fit()*

Делаем прогноз

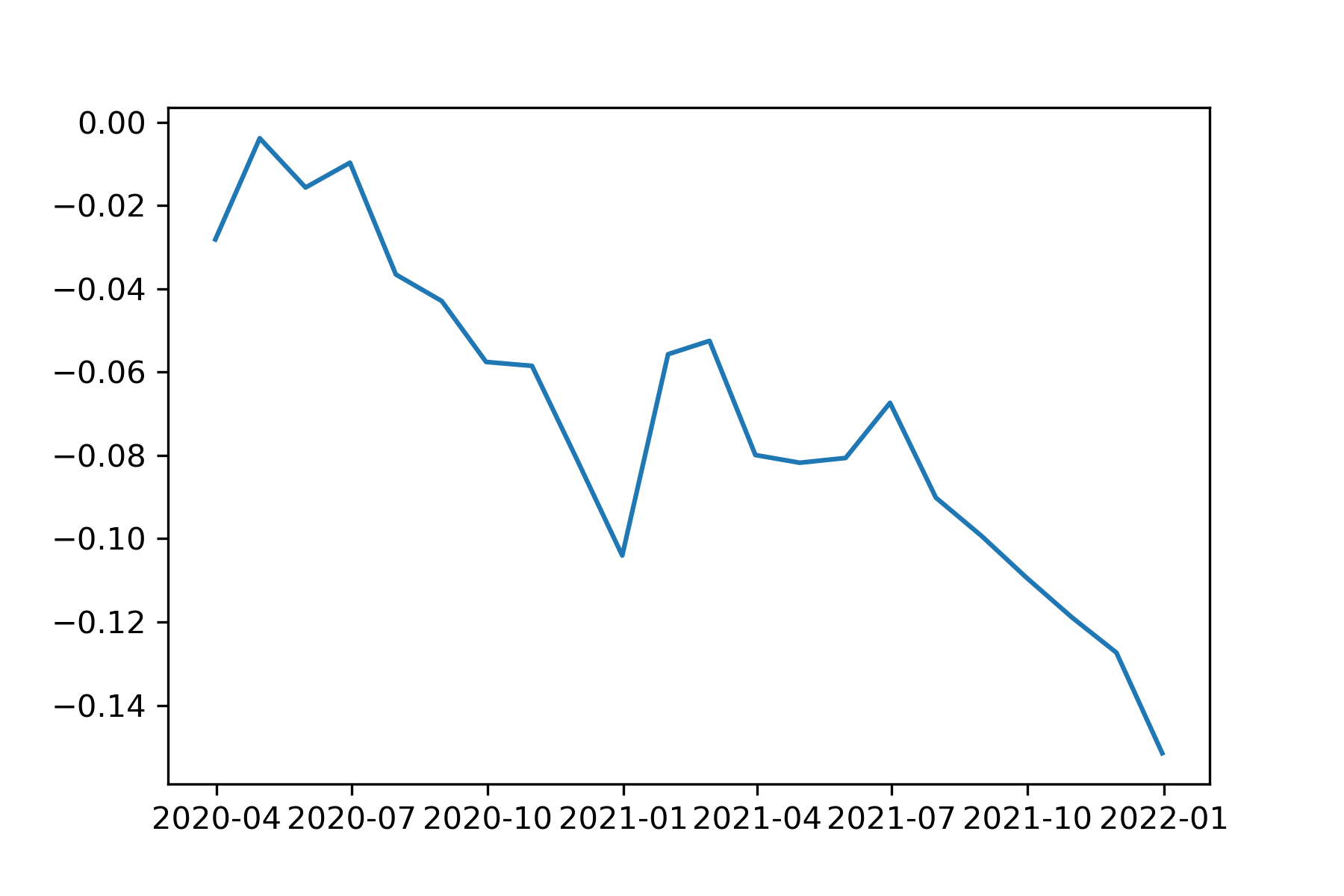
*test\_predictions = fitted\_model.forecast(len(test))*

Визуализируем результаты



В целом наше предсказание соответствует действительности, теперь займемся его оценкой:

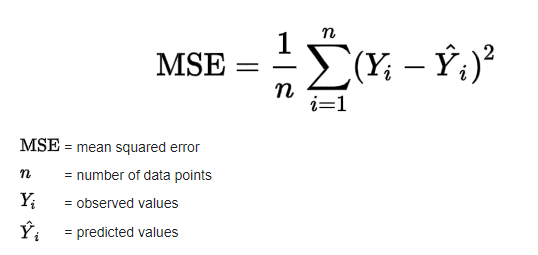
В начале глянем на остатки, ими является разница между нашим предсказанием и реальными данными



Лучшим результатом являются стационарные остатки около нуля, в нашем случае они растут со временем и не являются стационарными, что является очень плохим признаком.

**MSE (Mean Squared Error)**

Следующая мера оценки - средняя квадратичная ошибка, она, в отличие от средней ошибки, наказывает за сильное отклонение от истинного значения. Она считается по следующей формуле:



Но, так как в нашем случае мы имеем ошибки от 0 до 1 - будет правильно взять **RMSE (Root Mean Squared Error)** для интерпретации результатов. Данный показатель вычисляется, беря квадрат из **MSE**.

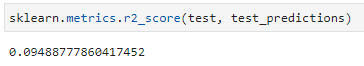
Для вычислений используется метод **mean\_squared\_error()** с параметром **squared=False** (для получения **RMSE**), пакета **sklearn.metrics**.



В нашем случае он равняется **0.08** или около 1% от последних значений, что являлось бы приемлемым значением, если бы наш ряд был стационарен и следующий показатель тоже имел приемлемое значение.

**R-квадрат (Коэффициент детерминации)**

Она показывает, насколько условная дисперсия модели отличается от дисперсии реальных значений Y. Если этот коэффициент близок к 1, то условная дисперсия модели достаточно мала и весьма вероятно, что модель неплохо описывает данные. Если же коэффициент **R-квадрат** сильно меньше, например, меньше 0.5, то, с большой долей уверенности модель не отражает реальное положение вещей. Вычисление можно произвести с помощью метода **r2\_score()**, пакета **sklearn.metrics**.



В нашем случае коэффициент детерминации **~0.095**, что говорит о полном отличии дисперсии предсказанных значений от дисперсии реальных значений.

**Самостоятельно задание:**

Скачайте в интернете набор данных временного ряда и сделайте прогноз на

ближайшее время с помощью метода Хольта-Винтерса, производите оценку своего прогноза и интерпретируйте её.

**Контрольные вопросы:**

1. Какие еще есть методы прогнозирования временных рядов?
2. В каком случае высокий коэффициент детерминации не говорит о низком отличии реальной и предсказанной дисперсии?
3. Перечислите способы оценки методов классификации в машинном обучении и объясните как их интерпретировать.