ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

***«*САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО»**

Институт компьютерных наук и технологий

**Высшая школа программной инженерии**

**Изображение выглядит как текст

Автоматически созданное описание**

**ЛАБОРАТОРНАЯ РАБОТА №5**

по дисциплине «Машинное обучение»

4 Вариант



Студент гр. *3530202/80201* *Дац П.*

Руководитель Селин И.А.

Санкт-Петербург

2021 г

**Задание**

**Изображение выглядит как текст, газета, снимок экрана, документ

Автоматически созданное описание**

**Изображение выглядит как текст

Автоматически созданное описание**

**Задание 1**

Используем LinearRegression из библиотеки sklearn для построения регрессий.

Данные в файле reglab1.txt выглядят так:

Изображение выглядит как текст

Автоматически созданное описание

Выделим наборы переменных:

z(x, y), y(x, z), x(y, z), y(x), z(x), y(z)

Для каждого набора зависимых переменных построим регрессионную модель. Чтобы найти наиболее подходящую модель воспользуемся функцией score модуля LinearRegression. Эта функция выводит коэффициент детерминации. Чем ближе этот коэффициент будет к 1, тем более подходящей будет модель.

Выведем полученный результат:

Изображение выглядит как текст

Автоматически созданное описание

**Вывод**

Из полученных результатов можем понять, что наиболее подходящей моделью будет модель с зависимостью z(x,y), т.к. для этого набора зависимых переменных коэффициент детерминации оказался ближе всего к 1.

**Задание 2**

Выведем данные из файла reglab.txt:

Изображение выглядит как текст

Автоматически созданное описание

В файле 4 признака:x1,x2,x3,x4. Будем перебирать все варианты зависимости переменной y от 1,2 или 3 переменных из этих признаков. Вместо минимизации остаточной суммы квадратов RSS можно приближать коэффициент детерминации R2 к 1, т.к. имеет место формула R2 = 1-(RSS/TSS).

Выведем значения коэффициента детерминации для различных вариантов зависимости переменных:

Изображение выглядит как текст

Автоматически созданное описание

**Вывод**

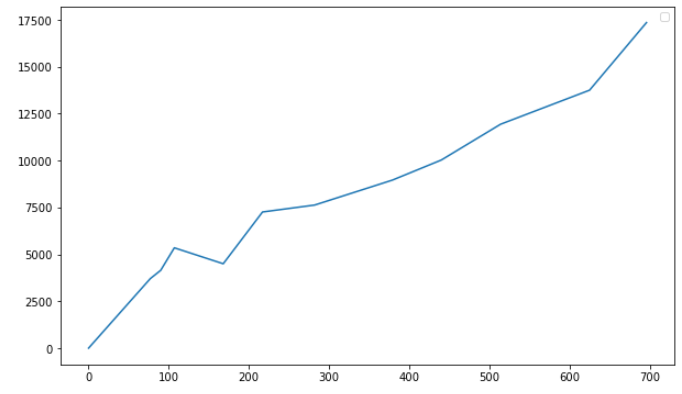
Лучшие зависимости переменных: y(x1, x2), y(x1, x2, x3) и y(x1, x2, x4). Но задача состоит в том, чтобы выбрать оптимальное подмножество признаков, уменьшив при этом их количество, поэтому выбираем y(x1,x2), т.к. она уступает остальным лишь на 0.001.

**Задание 3**

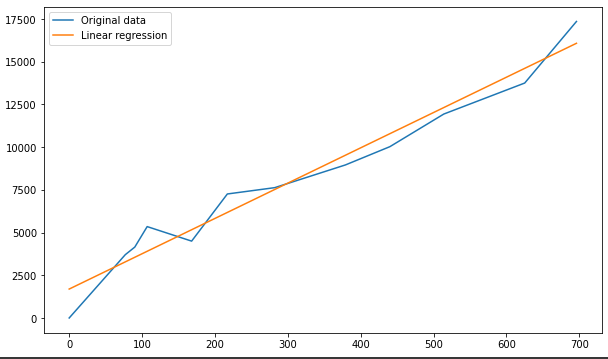
Данные из файла cygage.txt:



Построим график из данных:



Используем линейную регрессию:



Точность модели оценим при помощи коэффициента детерминации:

Изображение выглядит как текст

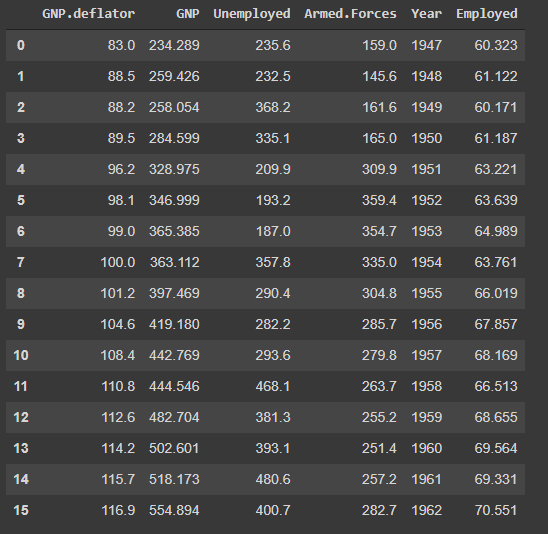
Автоматически созданное описание

**Вывод:**

Мы построили линейную регрессию, выражающую зависимость возраста исследуемых отложений от глубины залегания, используя веса наблюдений. Коэффициент детерминации для нашей модели оказался равен 0.959, что говорит о достаточно хорошем качестве построенной модели.

**Задание 4**

Выведем данные из Longley.csv, предварительно исключив переменную Population:



Разделим данные на обучающую и тестовую выборки случайным образом:



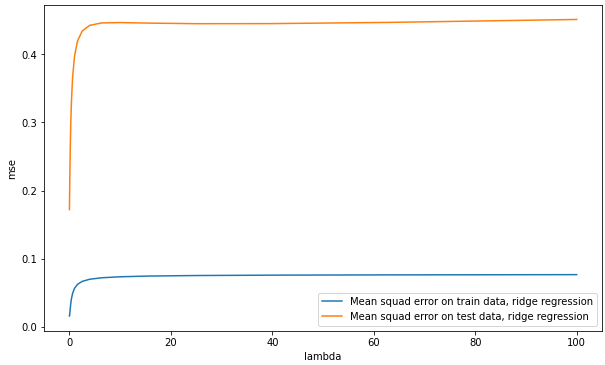
Построим линейную регрессию по признаку Employed.

Посчитаем для нее среднеквадратичную ошибку на обучающей и тестовой выборке:



Построим гребневую регрессию для .

Посчитаем для нее среднеквадратичную ошибку на обучающей и тестовой выборке:



**Вывод**

Видим, при увеличении параметра λ ошибка как на обучающей, так и на тестовой выборках для гребневой регрессии увеличивается. Линейная регрессия показала себя лучше как на обучающей, так и на тестовой выборках, т.к. ее среднеквадратичная ошибка была меньше.

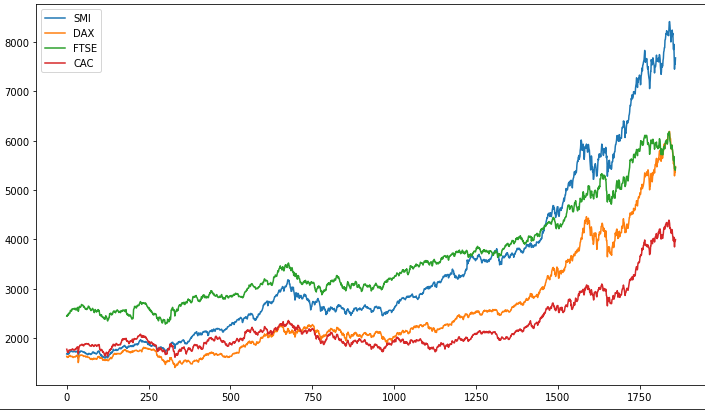
**Задание 5**

Загрузим и выведем данные из eustock.csv:

Изображение выглядит как текст, стол

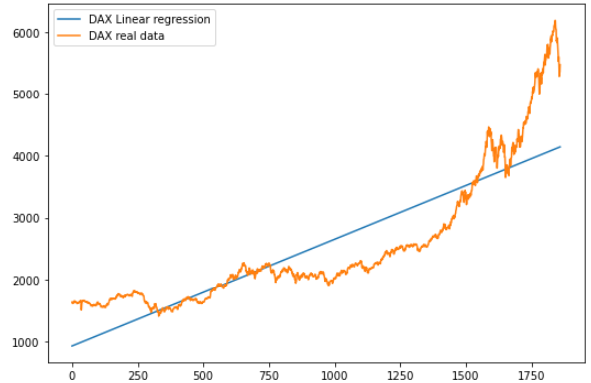
Автоматически созданное описание

Построим на одном графике все кривые изменения котировок во времени:

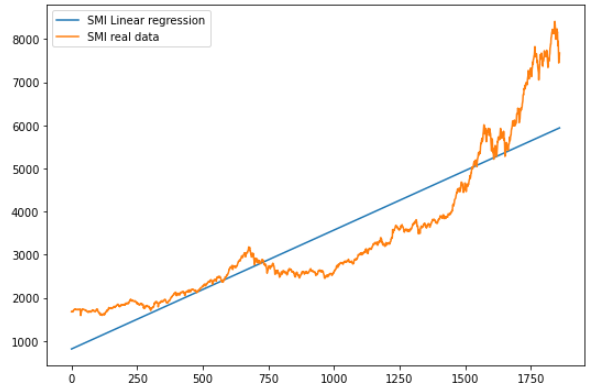


Построим линейную регрессию для каждой модели.

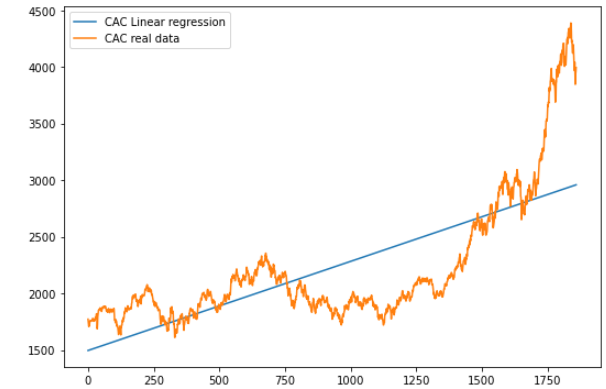
Линейная регрессия для модели DAX:



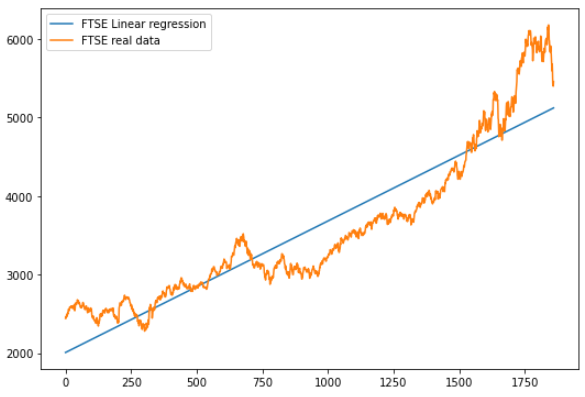
Линейная регрессия для модели SMI:



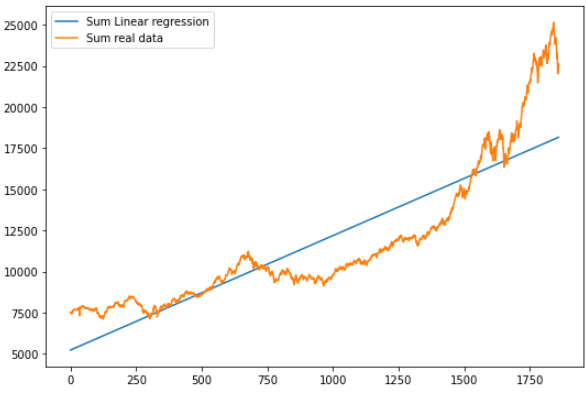
Линейная регрессия для модели CAC:



Линейная регрессия для модели FTSE:

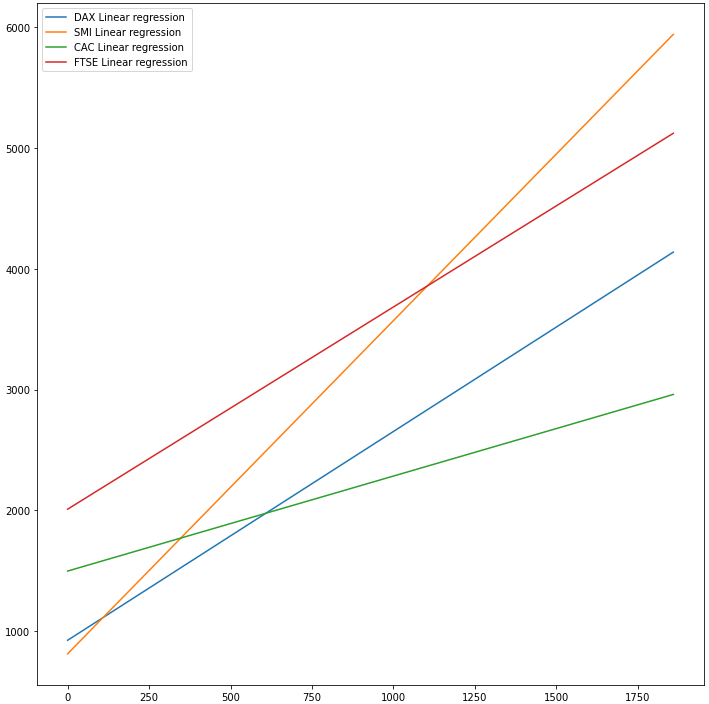


Линейная регрессия для всех моделей вместе:



**Вывод**

Исходя из наклонов прямых, расположенных на графиках линейных регрессий, можно утверждать, что биржа SMI имеет наибольшую динамику. Чтобы увидеть это, изобразим на одном графике линейные регрессии всех моделей:

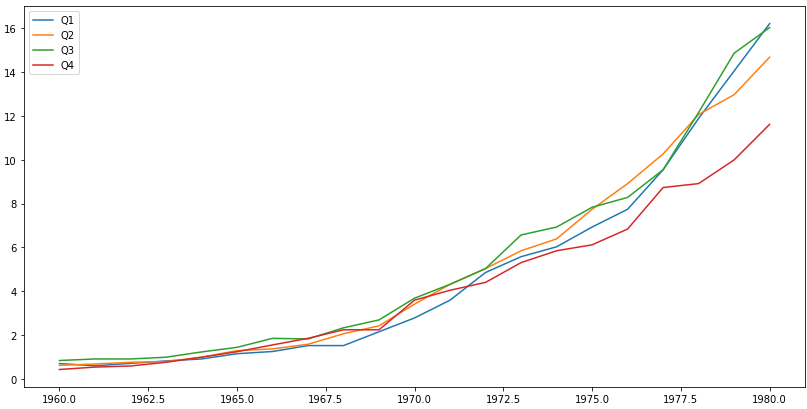


**Задание 6**

Загрузим и выведем данные из JohnsonJohnson.csv:

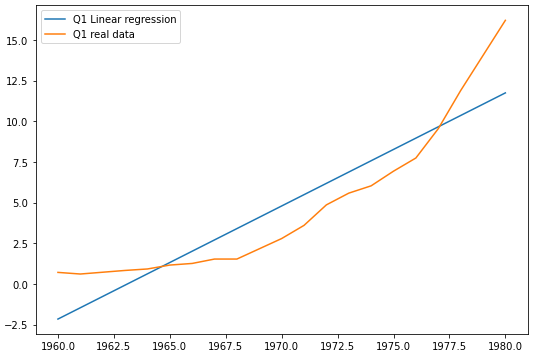


Построим на одном графике все кривые изменения прибыли во времени:

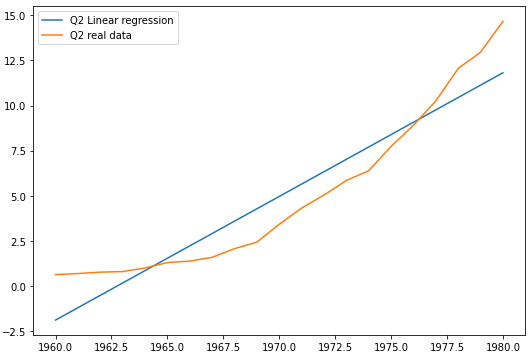


Построим линейную регрессию для каждого квартала в отдельности.

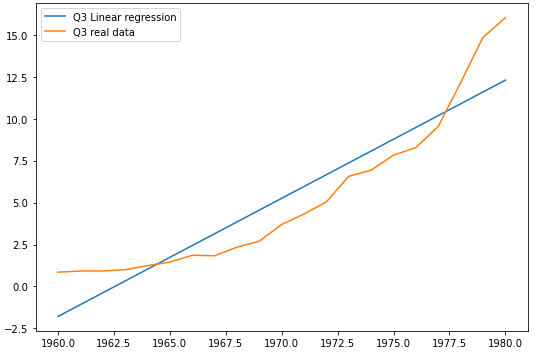
Для квартала Q1:



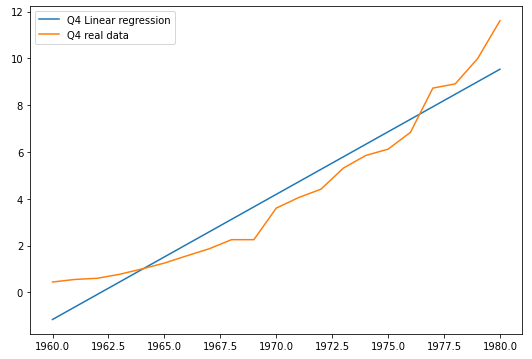
Для квартала Q2:



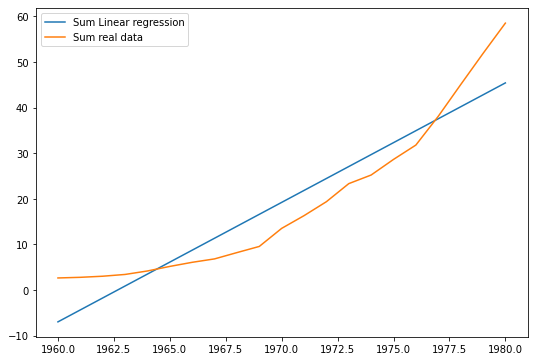
Для квартала Q3:



Для квартала Q4:



Для всех кварталов вместе:



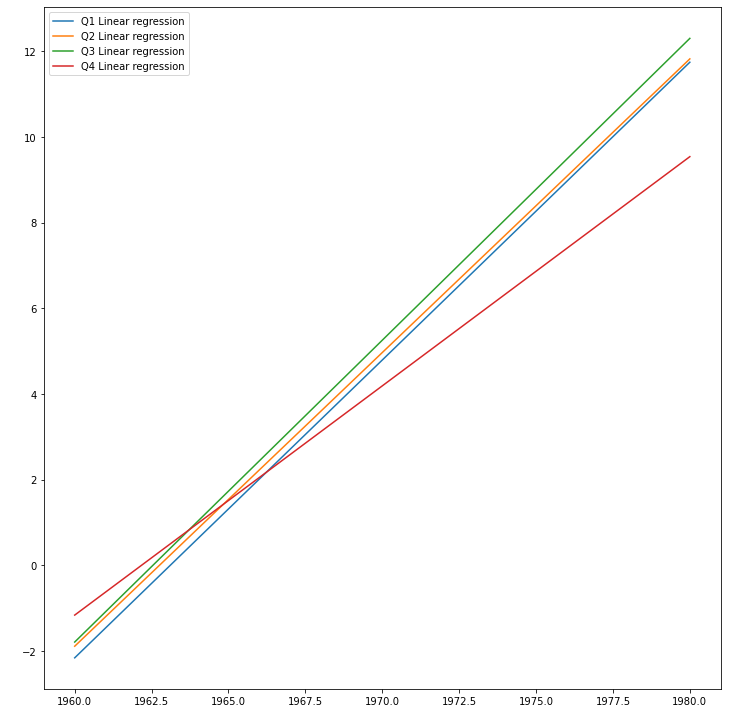
Сделаем прогноз на 2016 год:

Изображение выглядит как текст, клавиатура

Автоматически созданное описание

**Вывод**

Исходя из наклонов прямых, расположенных на графиках линейных регрессий, можно утверждать, что в квартале Q3 компания имеет наибольшую динамику доходности, а в квартале Q4 – наименьшую. Чтобы увидеть это, изобразим на одном графике линейные регрессии всех кварталов:



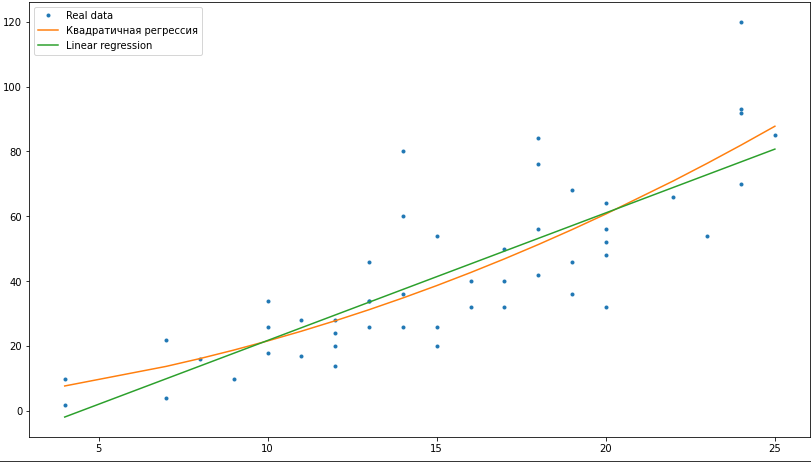
**Задание 7**

Загрузим и выведем данные из cars.csv:

Изображение выглядит как текст, часы

Автоматически созданное описание

Построим линейную регрессию:



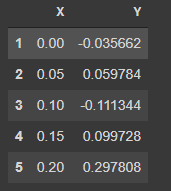


**Вывод**

С помощью регрессионной модели мы оценили длину тормозного пути при скорости 40 миль в час. Она равна 139.71 футов.

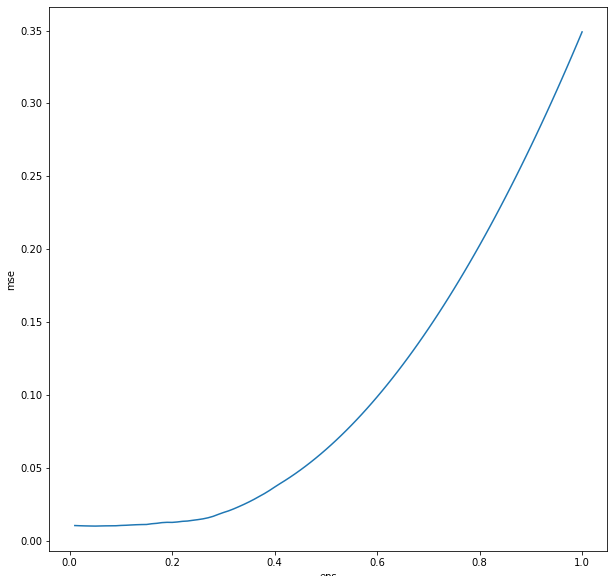
**Задание 8**

Загрузим и выведем дынные из файла svmdata6.txt:



Построим регрессионный алгоритм метода опорных векторов с параметром C=1 и ядром “rbf”.

Построим график зависимости среднеквадратичной ошибки на обучающей выборке от значения параметра:



**Вывод**

Видим, что чем меньше значение параметра ε, тем меньше среднеквадратичная ошибка. Это происходит из-за того, что параметр ε отвечает за допустимое отклонение. Чем оно больше, тем больше отклонение, значит и хуже регрессия.

**Задание 9**

Загрузим и выведем данные из nsw74psid1.csv:

Изображение выглядит как текст

Автоматически созданное описание

Построим регрессионное дерево для признака re78:

Его качество оценим с помощью метода score (коэффициент детерминации):



Построим линейную регрессионную модель и SVM-регрессию для этого набора данных. Их качество также оценим с помощью метода score:





**Вывод**

Оптимальной моделью является линейная регрессионная модель, т.к. ее значение коэффициента детерминации ближе к 1.