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»

Студент А. М. Потапова

гр. 3530202/90202

Руководитель И. А. Селин

Санкт-Петербург

2022 г

**Содержание**

[Задание 1 3](#_Toc117682292)

[Задание 2 5](#_Toc117682293)

[Задание 3 7](#_Toc117682294)

[Задание 4 9](#_Toc117682295)

[Задание 5 12](#_Toc117682296)

[Задание 6 17](#_Toc117682297)

[Задание 7 21](#_Toc117682298)

[Задание 8 22](#_Toc117682299)

[Задание 9 24](#_Toc117682300)

# **Задание 1**

Загрузите данные из файла reglab1.txt. Постройте по набору данных регрессии, используя модели с различными зависимыми переменными. Выберите наиболее подходящую модель.

*Исходные данные:*

Изображение выглядит как текст

Автоматически созданное описание

Наборы переменных - x(y, z), y(x, z), z(x, y), x(y), z(x), y(z)

Для каждого набора зависимых переменных строим регрессионную модель. Воспользуемся функцией score модуля LinearRegression, которая возвращает коэффициент детерминации (чем коэффициент ближе к 1, тем более подходящим будет модуль):

Изображение выглядит как текст

Автоматически созданное описание

*Результат:*

Изображение выглядит как текст

Автоматически созданное описание

**Вывод**

Исходя из полученных результатов можно заметить, что наиболее подходящей моделью стала модель с зависимостью z(x, y), поскольку значение коэффициента детерминации оказалось очень близко к 1 (0,9686).

# **Задание 2**

Реализуйте следующий алгоритм для уменьшения количества признаков, используемых для построения регрессии: для каждого  выбрать подмножество признаков мощности , минимизирующее остаточную сумму квадратов *RSS*. Используя полученный алгоритм, выберите оптимальное подмножество признаков для данных из файла reglab.txt. Объясните свой выбор.

*Исходные данные:*

Изображение выглядит как текст

Автоматически созданное описание

Переберем все варианты зависимости переменной y от 1, 2 или 3. Выведем значения коэффициента детерминации для различных вариантов зависимости переменных:

Изображение выглядит как текст

Автоматически созданное описание

*Результаты:*

Изображение выглядит как текст

Автоматически созданное описание

**Вывод**

Исходя из полученных результатов можно заметить, что наиболее подходящими зависимостями переменных стали y(x1, x2), y(x1, x2, x3) и y(x1, x2, x4). Оптимальное подмножество признаков - y(x1, x2).

# **Задание 3**

Загрузите данные из файла cygage.txt. Постройте регрессию, выражающую зависимость возраста исследуемых отложений от глубины залегания, используя веса наблюдений. Оцените качество построенной модели.

*Исходные данные:*

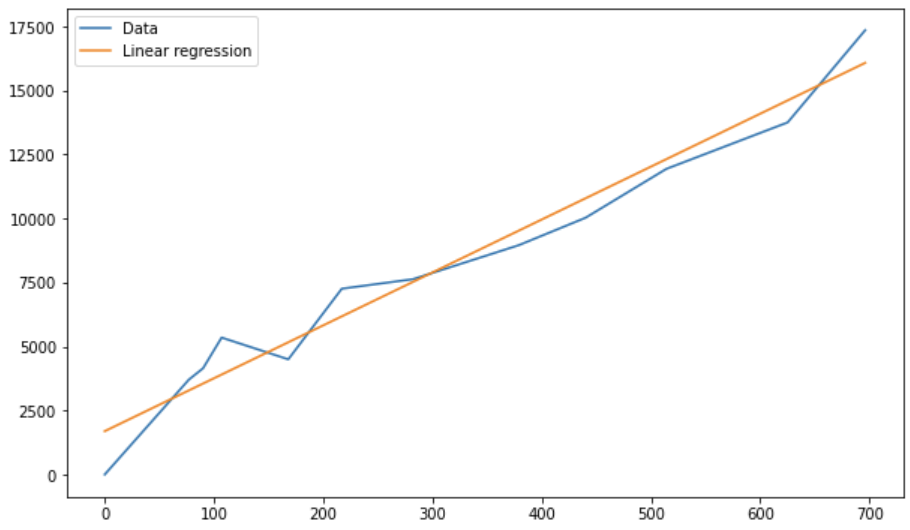


Для построения регрессии используем LinearRegression и веса наблюдений:

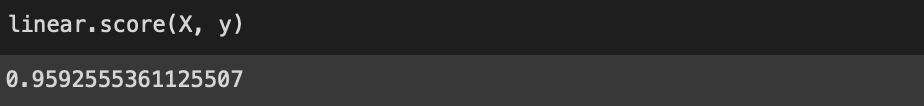
Изображение выглядит как текст

Автоматически созданное описание

*Результат:*



Оценим точность модели:



**Вывод**

Полученные результаты говорят о достаточно хорошем качестве построенной модели, поскольку коэффициент детерминации оказался равен 0,9592.

# **Задание 4**

Загрузите данные из файла longley.csv. Данные состоят из 7 экономических переменных, наблюдаемых с 1947 по 1962 годы (*n=16*). Исключите переменную Population. Разделите данные на тестовую и обучающую выборки равных размеров случайным образом. Постройте линейную регрессию по признаку Employed. Постройте гребневую регрессию для значений . Подсчитайте ошибку на тестовой и обучающей выборке для линейной регрессии и гребневой регрессии на данных значениях λ, постройте графики. Объясните полученные результаты.

Исключим переменную Population:



*Исходные данные:*

Изображение выглядит как стол

Автоматически созданное описание

Разделим данные на тестовую и обучающую выборки равных размеров случайным образом:

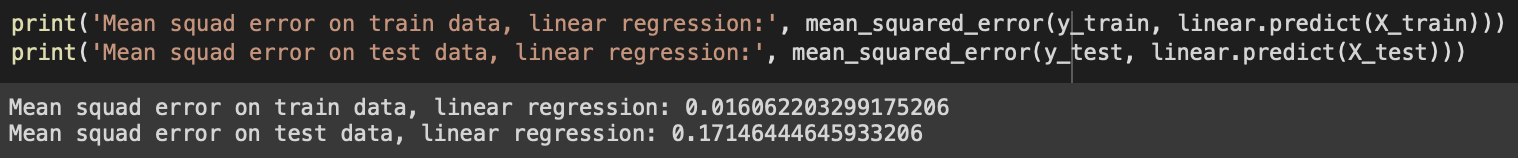


Построим линейную регрессию по признаку Employed:

Изображение выглядит как текст

Автоматически созданное описание

Посчитаем для нее среднеквадратичную ошибку на обучающей и тестовой выборке:



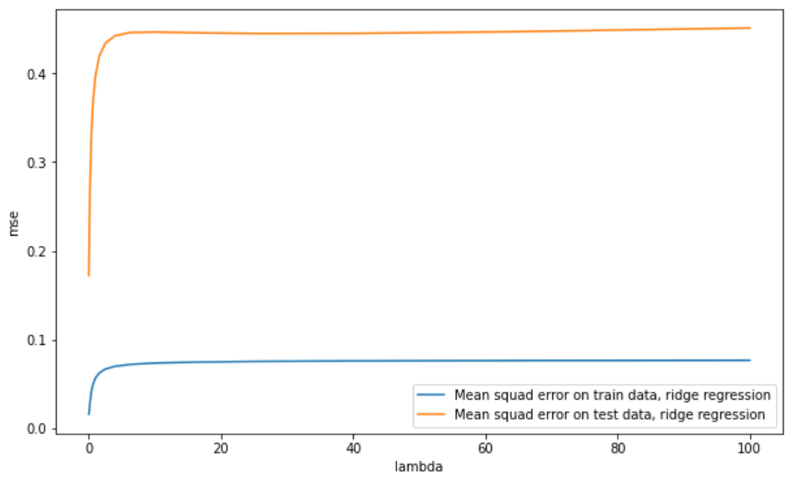
Построим гребневую регрессию для значений 

Изображение выглядит как текст

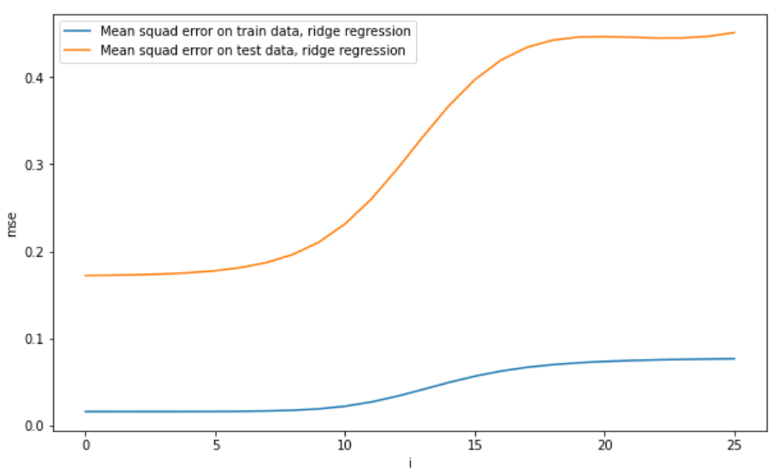
Автоматически созданное описание

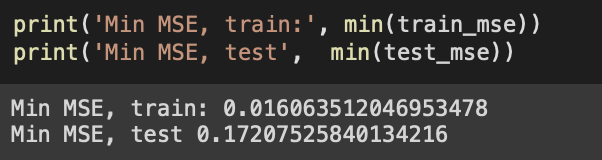
Посчитаем для них среднеквадратичную ошибку на обучающей и тестовой выборке:

Для λ:



Для i:





**Вывод**

Исходя из полученных результатов замечаем, что 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**

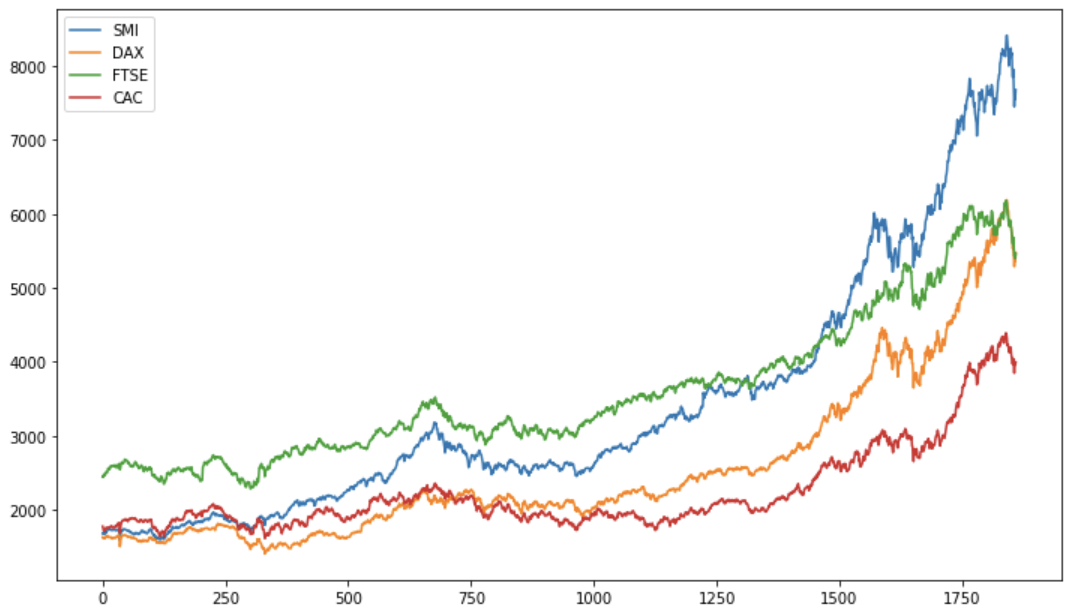
Загрузите данные из файла eustock.csv. Данные содержат ежедневные котировки на момент закрытия фондовых бирж: Germany DAX (Ibis), Switzerland SMI, France CAC, и UK FTSE. Постройте на одном графике все кривые изменения котировок во времени. Постройте линейную регрессию для каждой модели в отдельности и для всех моделей вместе. Оцените, какая из бирж имеет наибольшую динамику.

*Исходные данные:*

Изображение выглядит как текст

Автоматически созданное описание

Построим на одном графике все кривые изменения котировок во времени:

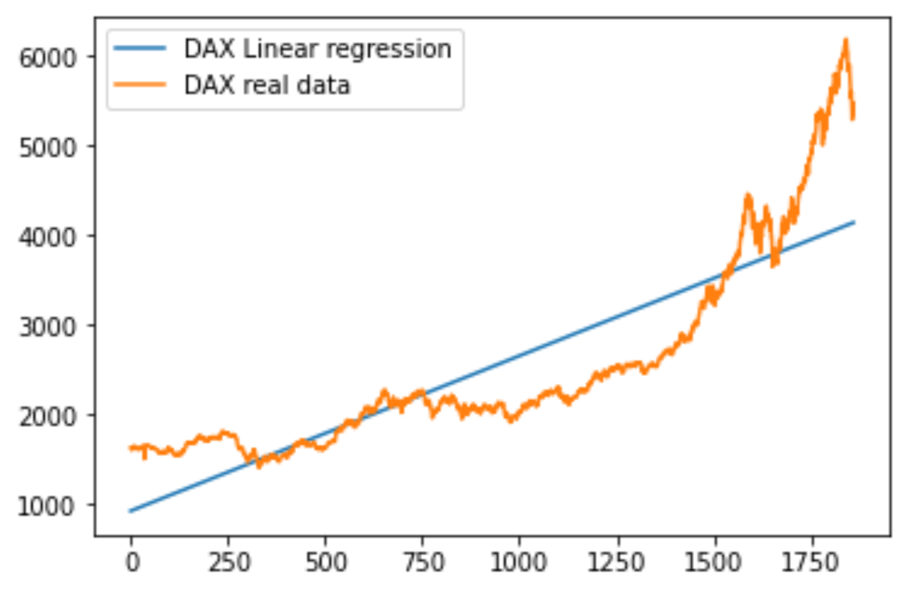


Построим линейную регрессию для каждой модели в отдельности и для всех моделей вместе:

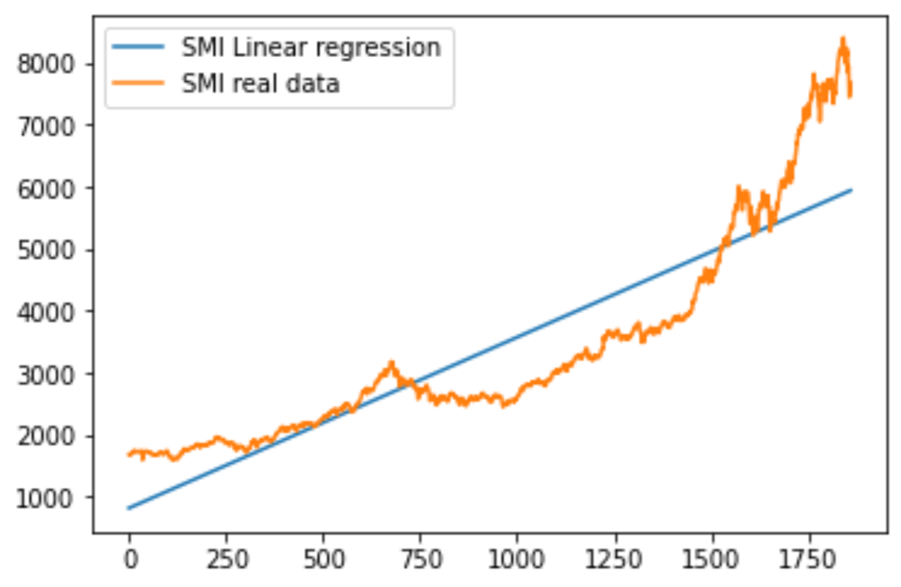
Изображение выглядит как текст

Автоматически созданное описание

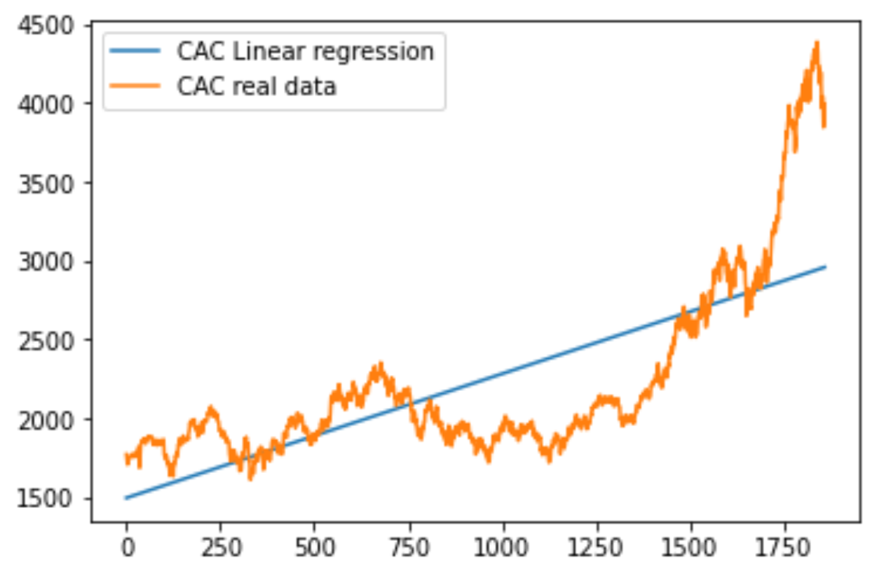
* Линейная регрессия для DAX



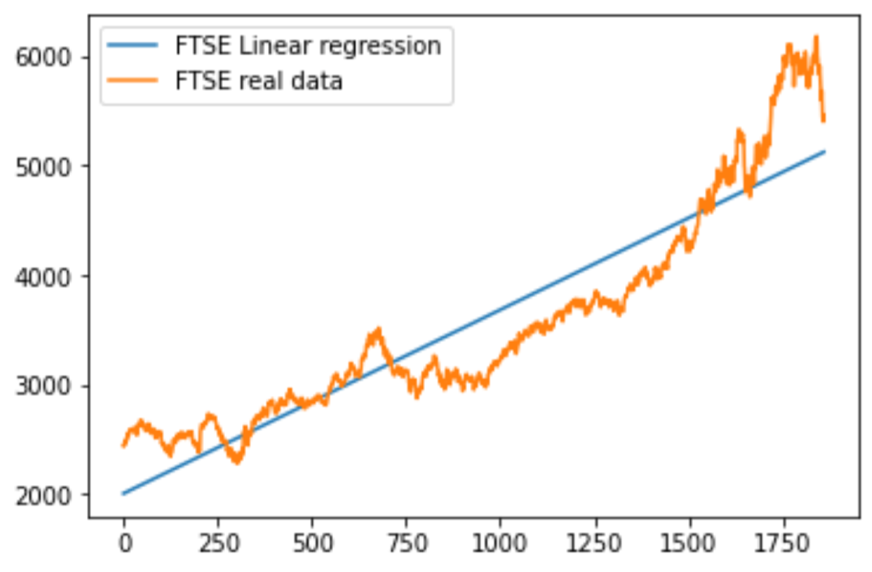
* Линейная регрессия для SMI



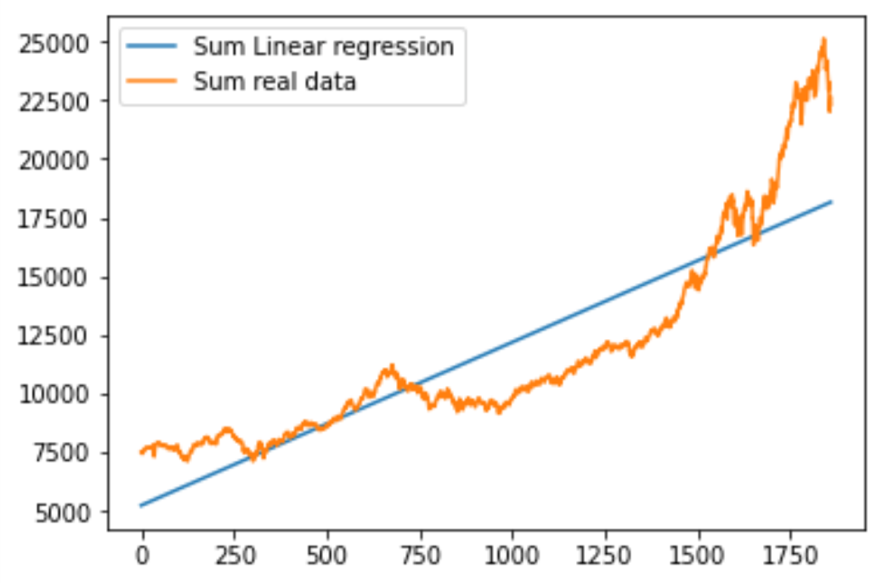
* Линейная регрессия для CAC



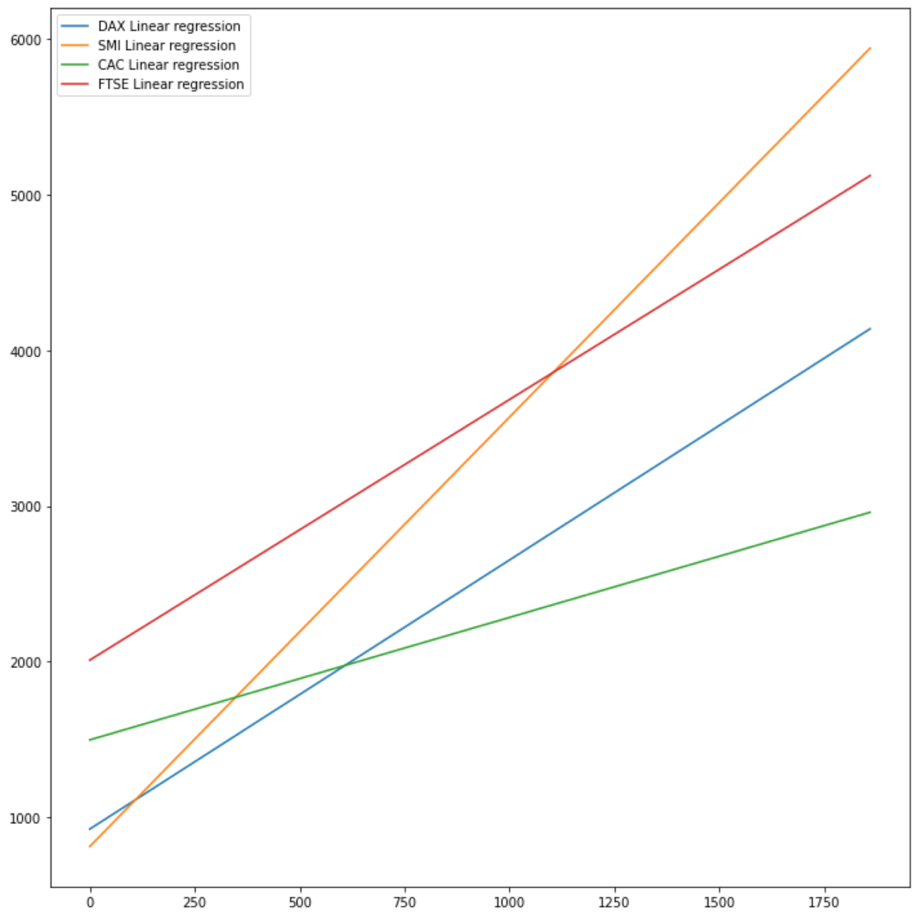
* Линейная регрессия для FNSE



* Линейная регрессия для всех моделей:



Отобразим линейные регрессии всех моделей по-отдельности:



**Вывод**

Исходя из полученных результатов замечаем, что биржа SMI имеет наибольшую динамику.

# **Задание 6**

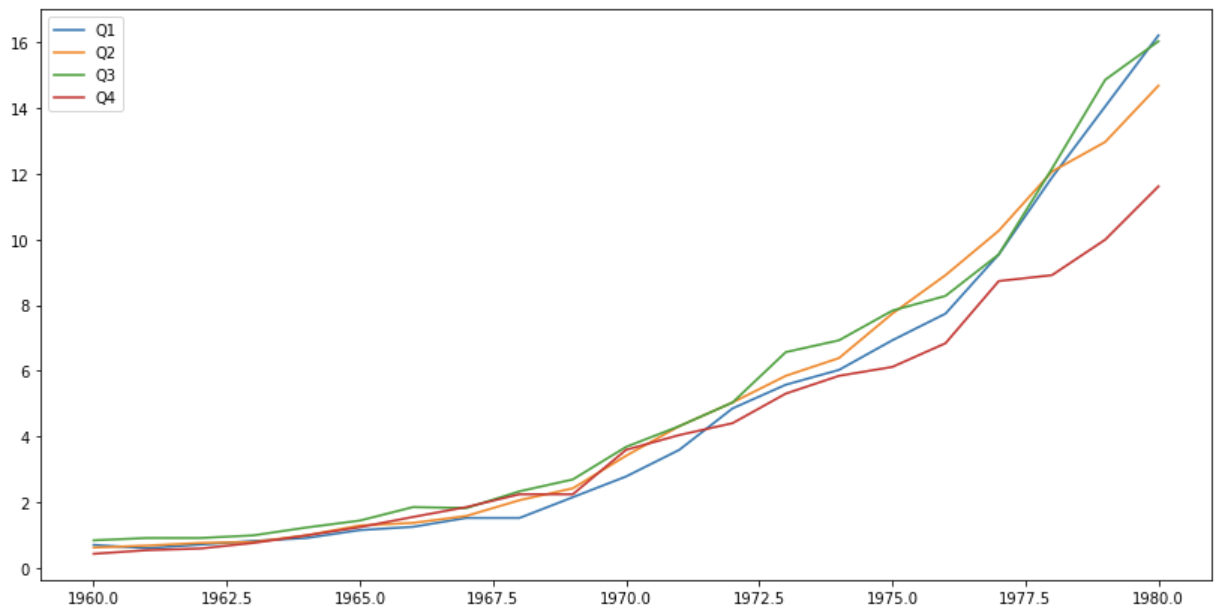
Загрузите данные из файла JohnsonJohnson.csv. Данные содержат поквартальную прибыль компании Johnson & Johnson с 1960 по 1980 гг. Постройте на одном графике все кривые изменения прибыли во времени. Постройте линейную регрессию для каждого квартала в отдельности и для всех кварталов вместе. Оцените, в каком квартале компания имеет наибольшую и наименьшую динамику доходности. Сделайте прогноз по прибыли в 2016 году во всех кварталах и в среднем по году.

*Исходные данные:*

Изображение выглядит как текст, табло, внешний

Автоматически созданное описание

Построим на одном графике все кривые изменения прибыли во времени:

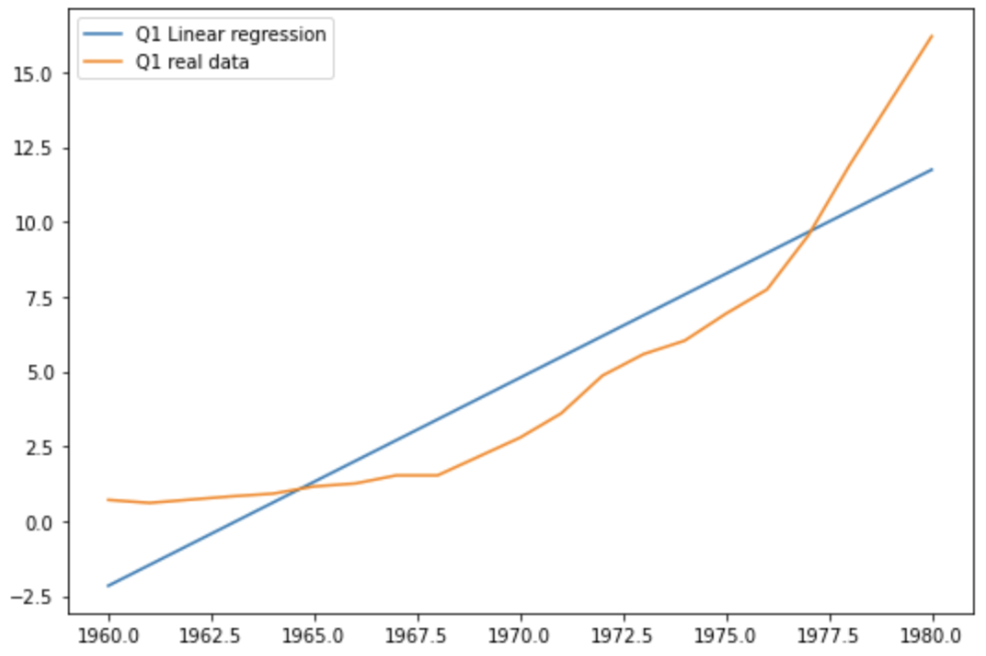


Построим линейную регрессию для каждого квартала в отдельности и для всех кварталов вместе:

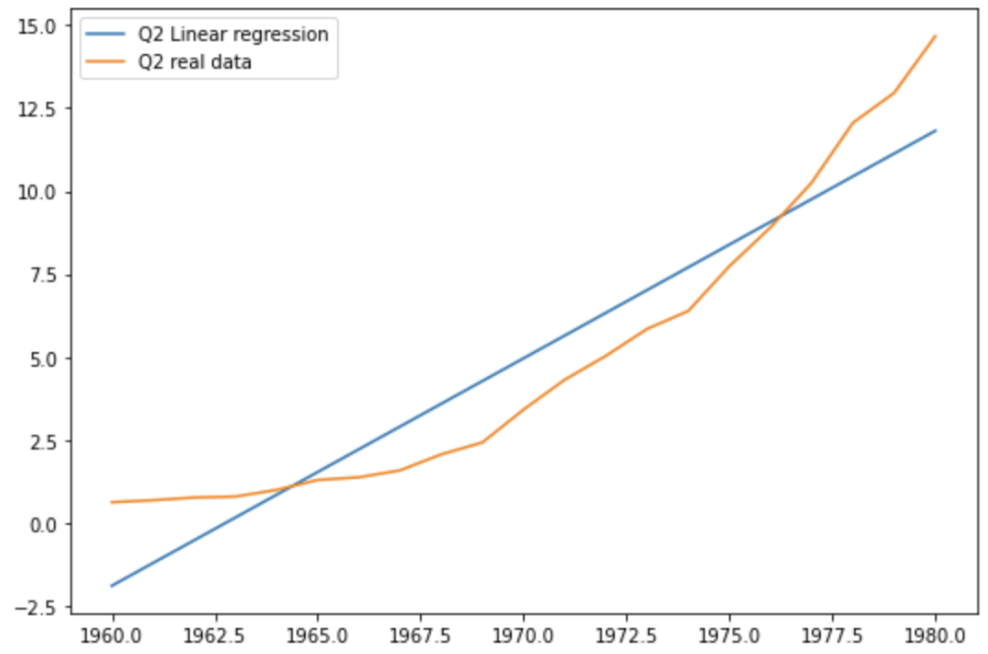
Изображение выглядит как текст

Автоматически созданное описание

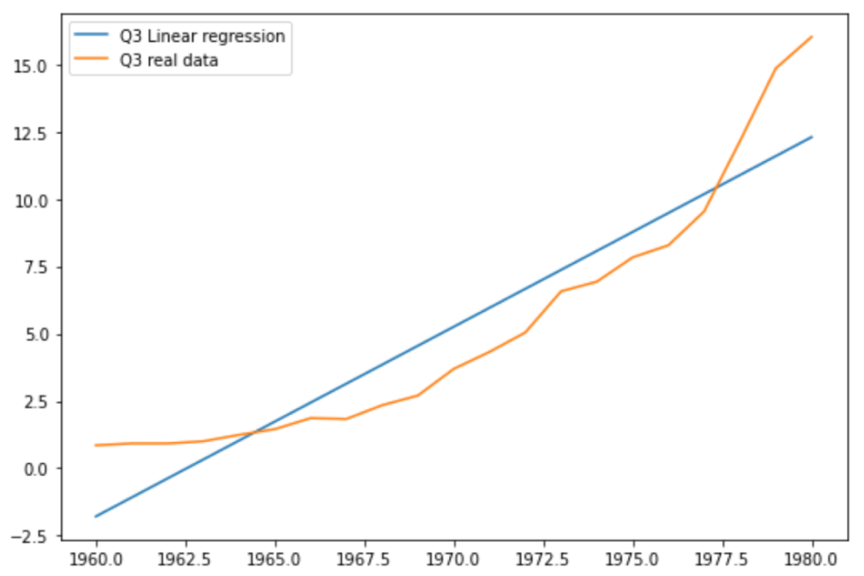
* Линейная регрессия для Q1:



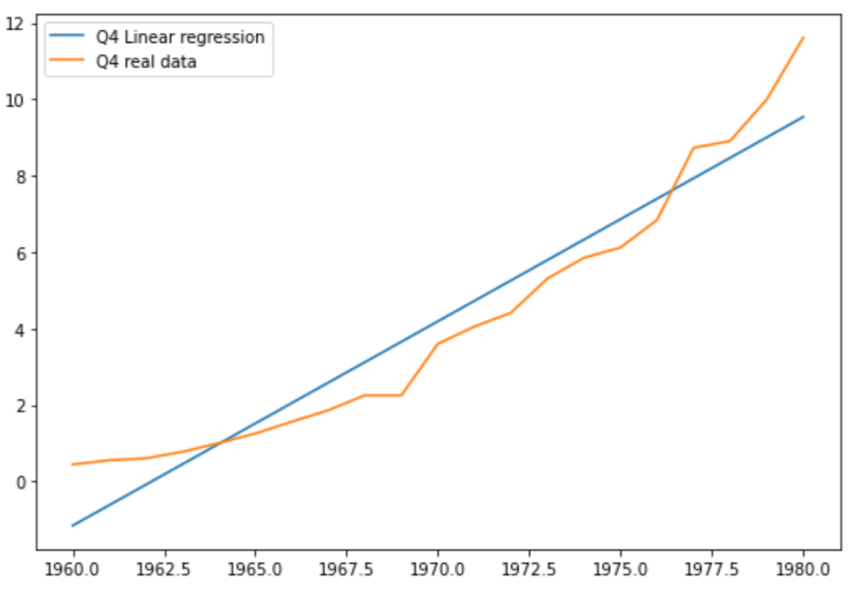
* Линейная регрессия для Q2:



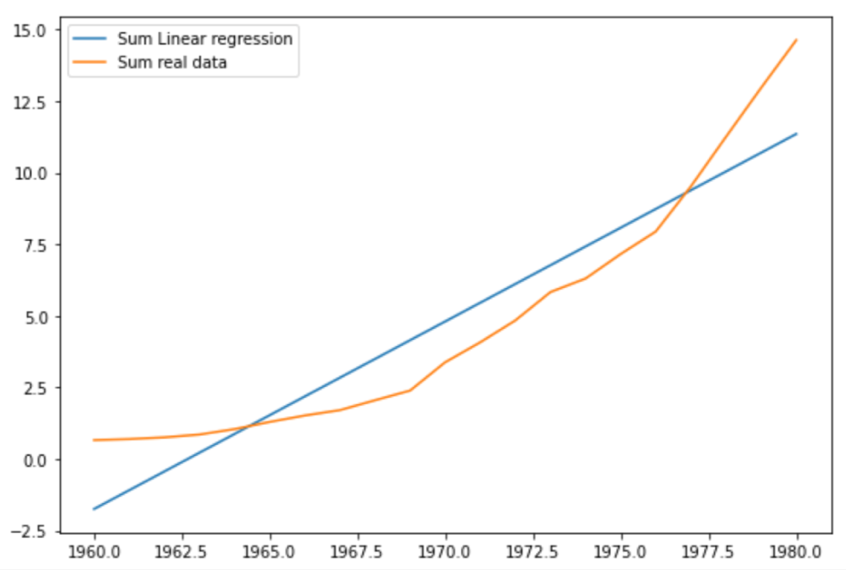
* Линейная регрессия для Q3:



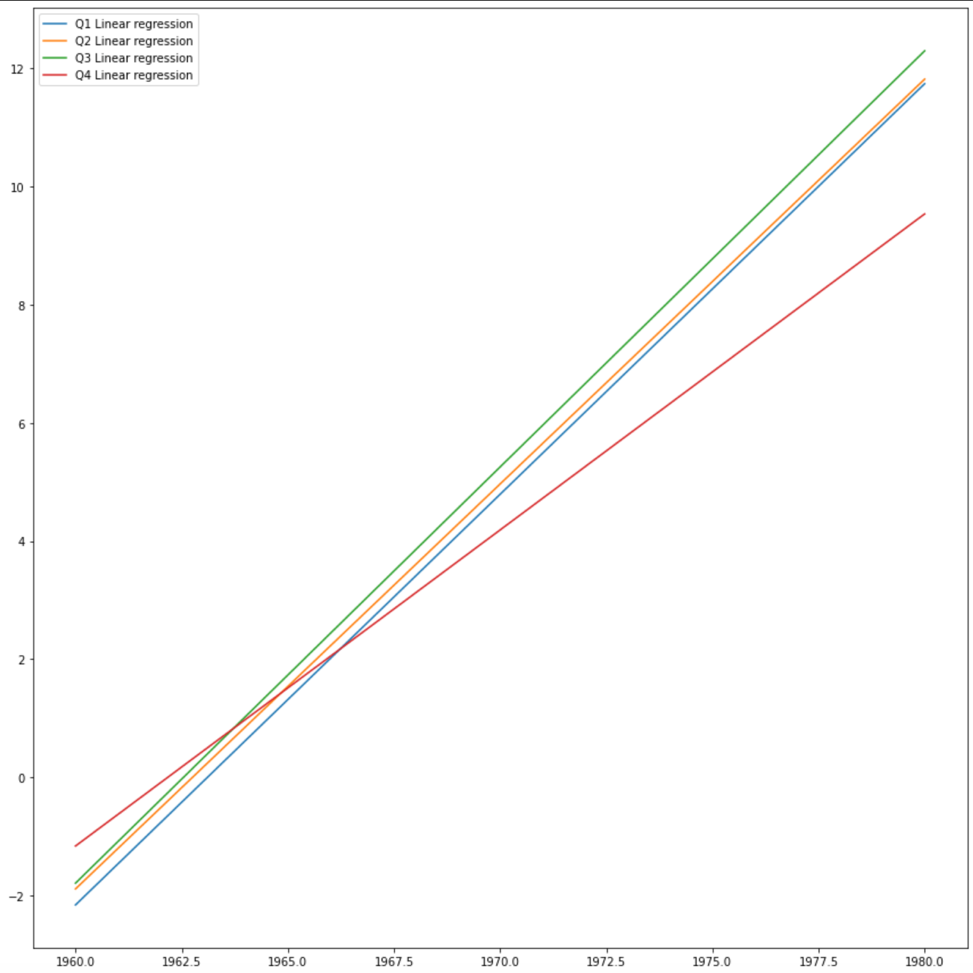
* Линейная регрессия для Q4:



* Линейная регрессия для всех кварталов вместе:



* Линейная регрессия для всех кварталов по-отдельности:



Сделаем прогноз по прибыли в 2016 году во всех кварталах и в среднем по году:

Изображение выглядит как текст

Автоматически созданное описание

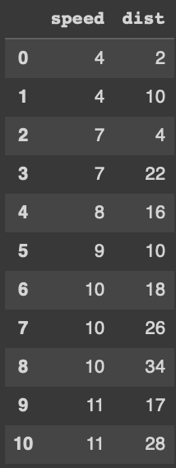
**Вывод**

Исходя из полученных результатов замечаем, что в квартале Q3 компания имеет наибольшую динамику доходности, а в Q4 – наименьшую. Прогноз соответствует увиденной динамике.

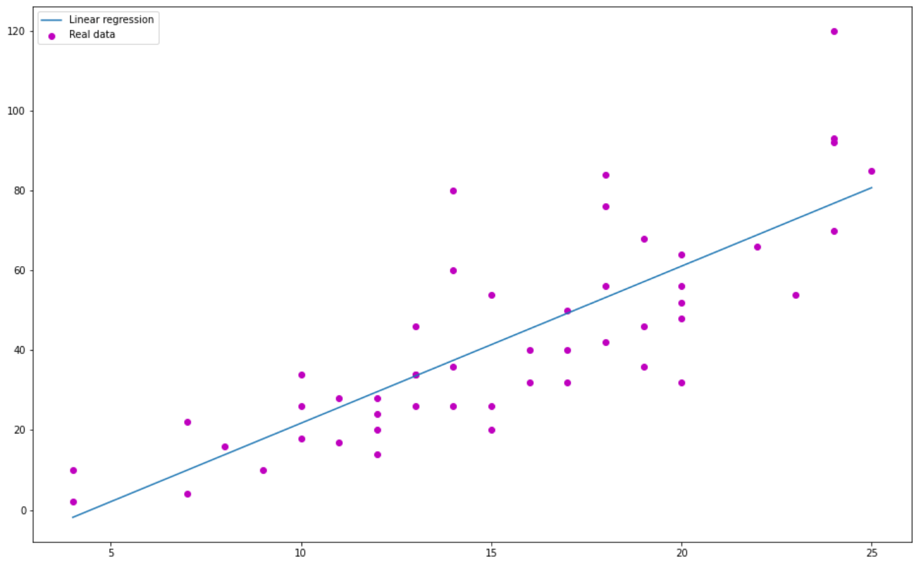
# **Задание 7**

Загрузите данные из файла cars.csv. Данные содержат зависимости тормозного пути автомобиля (футы) от его скорости (мили в час). Данные получены в 1920 г. Постройте регрессионную модель и оцените длину тормозного пути при скорости 40 миль в час.

*Исходные данные:*



Линейная регрессия:



Оценка длины тормозного пути при скорости 40 миль в час:

Изображение выглядит как текст

Автоматически созданное описание

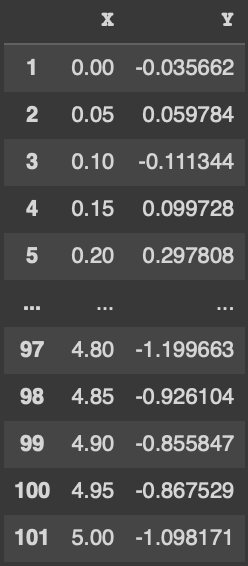
**Вывод**

При помощи регрессионной модели мы оценили длину тормозного пути при скорости 40 миль в час. Она составила 139,7172 фута.

# **Задание 8**

Загрузите данные из файла svmdata6.txt. Постройте регрессионный алгоритм метода опорных векторов (sklearn.svm.SVR) с параметром C = 1, используя ядро "rbf". Отобразите на графике зависимость среднеквадратичной ошибки на обучающей выборке от значения параметра ε. Прокомментируйте полученный результат.

*Исходные данные:*

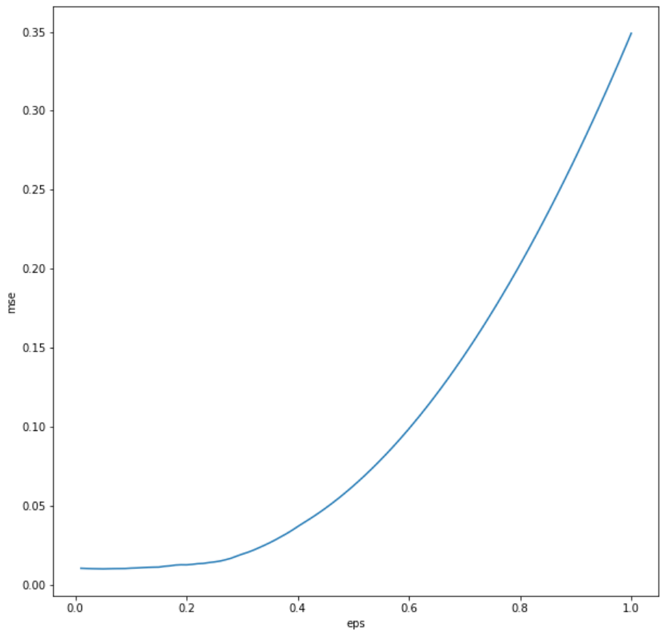


Построим регрессионный алгоритм метода опорных векторов (sklearn.svm.SVR) с параметром C = 1, используя ядро "rbf".

Изображение выглядит как текст

Автоматически созданное описание

Графике зависимости среднеквадратичной ошибки на обучающей выборке от значения параметра ε:



**Вывод**

Исходя из полученного результата мы видим, что чем меньше значения параметра ε, тем меньше среднеквадратичная ошибка. Это происходит из-за того, что данный параметр отвечает за допустимое отклонение. Следовательно, чем больше отклонение, тем хуже регрессия.

# **Задание 9**

Загрузите набор данных из файла nsw74psid1.csv. Постройте регрессионное дерево (sklearn.tree.DecisionTreeRegressor) для признака re78. Постройте линейную регрессионную модель и SVM-регрессию для этого набора данных. Сравните качество построенных моделей, выберите оптимальную модель и объясните свой выбор.

*Исходные данные:*

Изображение выглядит как текст, электроника, клавиатура

Автоматически созданное описание

Построим регрессионное дерево (sklearn.tree.DecisionTreeRegressor) для признака re78:

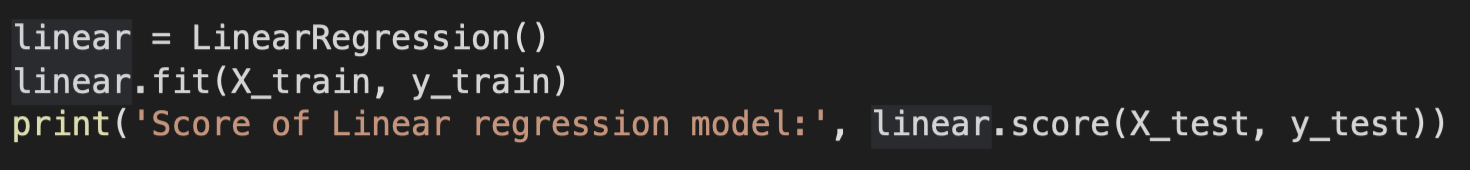
Изображение выглядит как текст

Автоматически созданное описание

Получили следующее качество:

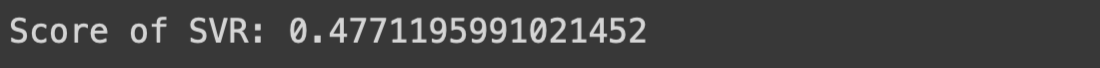
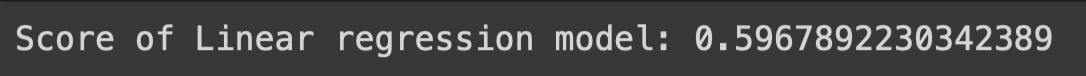


Построим линейную регрессионную модель и SVM-регрессию для этого набора данных:

Изображение выглядит как текст

Автоматически созданное описание

Получили следующее качество:



**Вывод**

Исходя из полученных результатов мы видим, что оптимальной моделью можно считать регрессионную модель, т. к. ее значение коэффициента детерминации ближе к 1.