

Санкт-Петербургский Политехнический университет имени Петра Великого  
Институт компьютерных наук и технологий  
Высшая школа программной инженерии

## **ЛАБОРАТОРНАЯ РАБОТА №5**

**«РЕГРЕССИЯ»**

по дисциплине «Статистическое моделирование случайных процессов и систем»

Выполнил студент гр. 3530904/70103

Русаков Е.С.

Преподаватель

Селин И.

Санкт-Петербург  
2020

## Оглавление

Задание.....	3
Ход работы.....	4
1. Bagging.....	4
2. Бустинг.....	5
3. Стэкинг.....	6
Приложение.....	8
1. Баггинг.....	8
2. Бустинг.....	8
3. Стэкинг.....	9

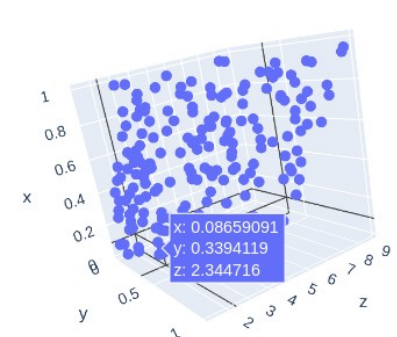
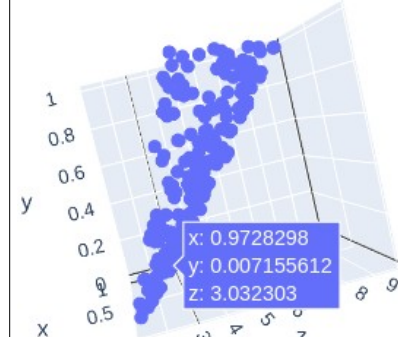
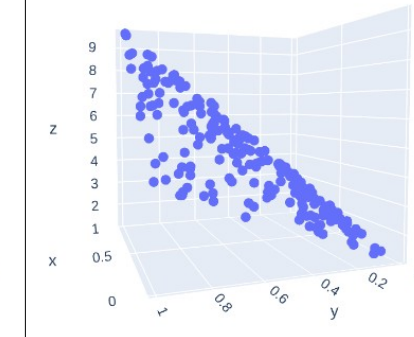
## Задание

1. Загрузите данные из файла reglab1.txt. Постройте по набору данных регрессии, используя модели с различными зависимыми переменными. Выберите наиболее подходящую модель.
2. Реализуйте следующий алгоритм для уменьшения количества признаков, используемых для построения регрессии: для каждого  $k \in \{0, 1, \dots, d\}$  выбрать подмножество признаков мощности  $k^1$ , минимизирующее остаточную сумму квадратов  $RSS$ . Используя полученный алгоритм, выберите оптимальное подмножество признаков для данных из файла reglab.txt. Объясните свой выбор.
3. Загрузите данные из файла cugage.txt. Постройте регрессию, выражающую зависимость возраста исследуемых отложений от глубины залегания, используя веса наблюдений. Оцените качество построенной модели.
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  $\lambda = 10^{-3+0.2i}$ ,  $i = 0, \dots, 25$ . Подсчитайте ошибку на тестовой и обучающей выборке для линейной регрессии и гребневой регрессии на данных значениях  $\lambda$ , постройте графики. Объясните полученные результаты.
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.
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.
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.
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  $\epsilon$ . Прокомментируйте полученный результат
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.

## Ход работы

### 1. Поиск правильной комбинации переменных для регрессии

В датасете reglab1.txt имеется трёхмерный массив точек. Выясним, регрессию для какой комбинации переменных строить рациональнее. В качестве регрессора используется класс LinearRegression из библиотеки Sklearn. Для оценки качества из выборки выделяется 30% тестовой части, а в качестве метрики используется коэффициент детерминации (R2\_score в реализации Sklearn).

X от Y и Z	Y от Z и X	Z от X и Y
		
0.933	0.961	0.975

Как видно по метрике и визуализациям, лучше всего рассматривать зависимость  $Z(X,Y)$ .

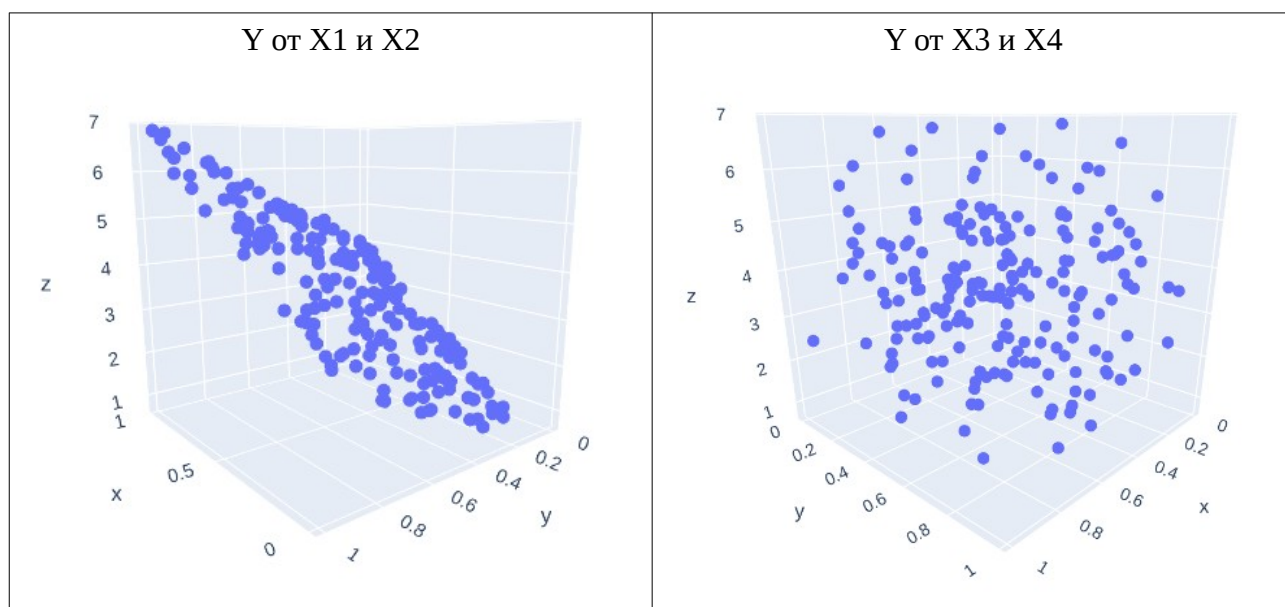
### 2. Уменьшение размерности признаков для регрессии

Датасет reglab.txt содержит пятимерные данные: Y, X1, X2, X3, X4. Требуется уменьшить их размерность построив регрессии. Будем предсказывать Y от X и измерять метрику правильности предсказаний.

Измерения				Метрика
X1	X2	X3	X4	
+	+	+	+	0.9996
+	+	+	-	0.9991
+	+	-	+	0.9993
+	-	+	+	0.5848
-	+	+	+	0.3192
+	+	-	-	0.9985
+	-	+	-	0.6074
-	+	+	-	0.3546
+	-	-	+	0.6035
-	+	-	+	0.3511

-	-	+	+	-0.1014
+	-	-	-	0.5552
-	+	-	-	0.3787
-	-	+	-	-0.0167
-	-	-	+	-0.0095

Из таблицы видно, что качество регрессии сильно падает, когда из выборки исключаются компоненты X1 и X2. В то же время, исключение остальных двух едва ли сказывается на качестве предсказаний.

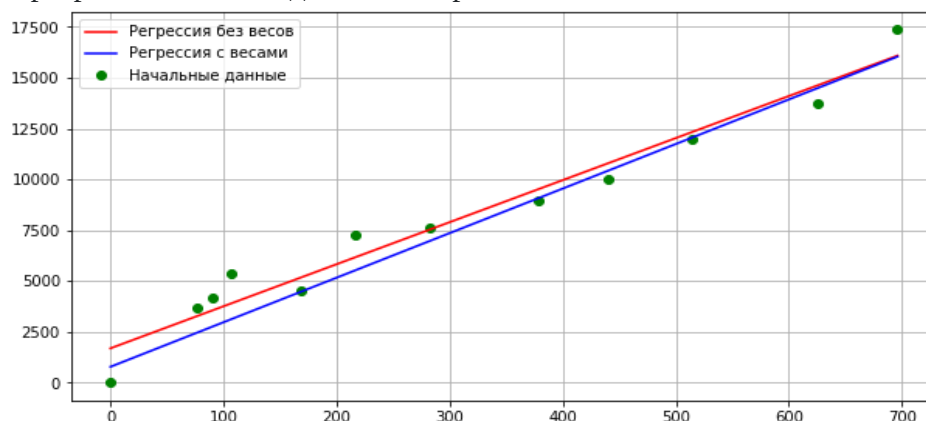


Это же следует из визуализаций зависимости Y от X. В левом случае зависимость очевидна, а в правом её совсем нет.

### 3. Исследование ископаемых отложений

Датасет `sygate.txt` содержит данные о возрасте и глубине ископаемых пород, а также весовые метки для каждого наблюдения.

Построим две регрессии по этим данным: с применением весов и без них.



Значения метрики:

- Без весов: 0.9593
- С весами: 0.9737

Регрессия с весами оказалась несколько лучше по значению метрики, чем без оных.

В общем случае весами стоит пользоваться, когда есть некоторые предположения о достоверности полученных точек относительно друг друга. В этом случае выбор весовых коэффициентов позволяет лучше подобрать решение для конкретной практической задачи.

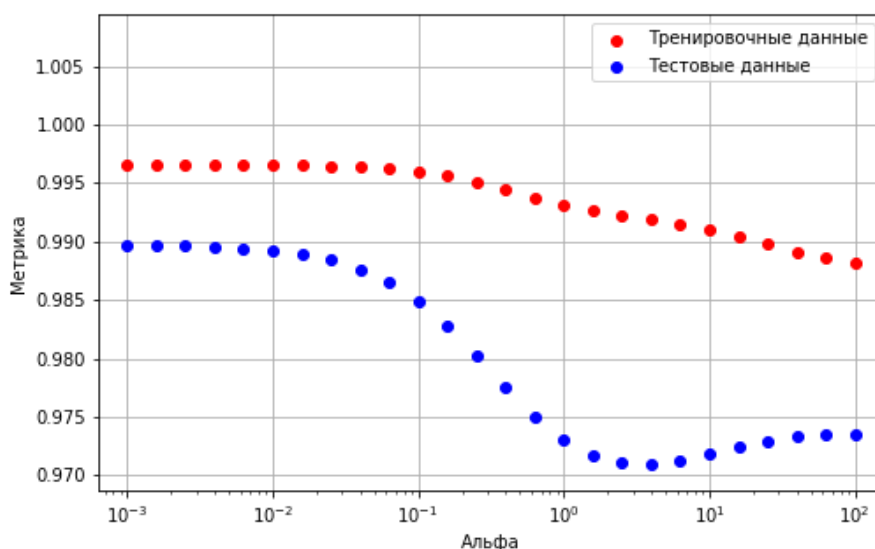
#### 4. Регрессия занятости

В предложенном датасете находятся данные по шести показателям по экономике некоторого государственного образования за 16 лет. Требуется, отбросив численность населения, построить линейную и гребневую регрессии для предсказания уровня занятости. Общая выборка делится пополам на обучающую и тестовую.

Линейная регрессия дала следующие значения метрики:

- 0.9999 для обучающей выборки
- 0.9754 для тестовой выборки

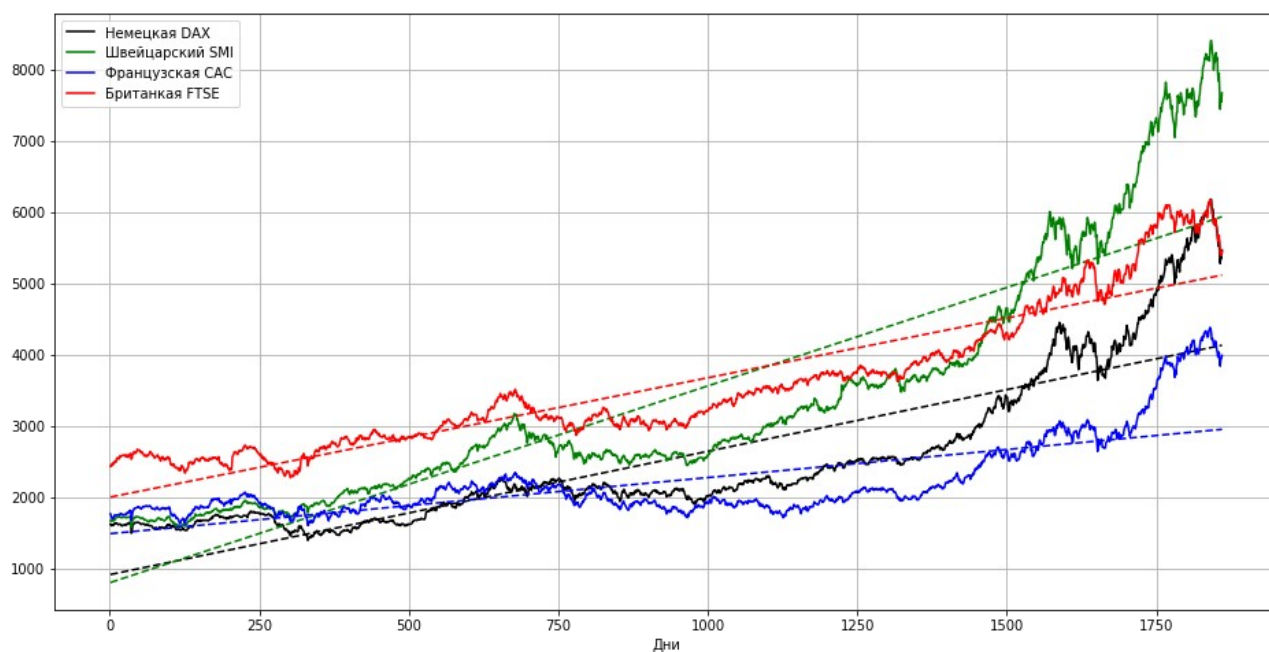
Для гребневого регрессора получены следующие показатели в зависимости от параметра альфа. Этот параметр управляет обусловленностью задачи, делая оценки более стабильными. По умолчанию альфа устанавливается равной единице, однако из графика видно, что для нашей задачи лучше всего значения альфа поближе к нулю.



#### 5. Анализ котировок фондовых бирж

В датасете имеются ежедневные показатели котировок четырёх фондовых бирж на протяжении 1850 дней.

По каждой построена и оценена линейная регрессия.



Метрики построенных регрессий:

- Биржа DAX: 0.7331
- Биржа SMI: 0.7944
- Биржа CAC: 0.5303
- Биржа FTSE: 0.8482

Также из графика очевидно, что наибольший прирост на данном периоде произошёл на швейцарской бирже.

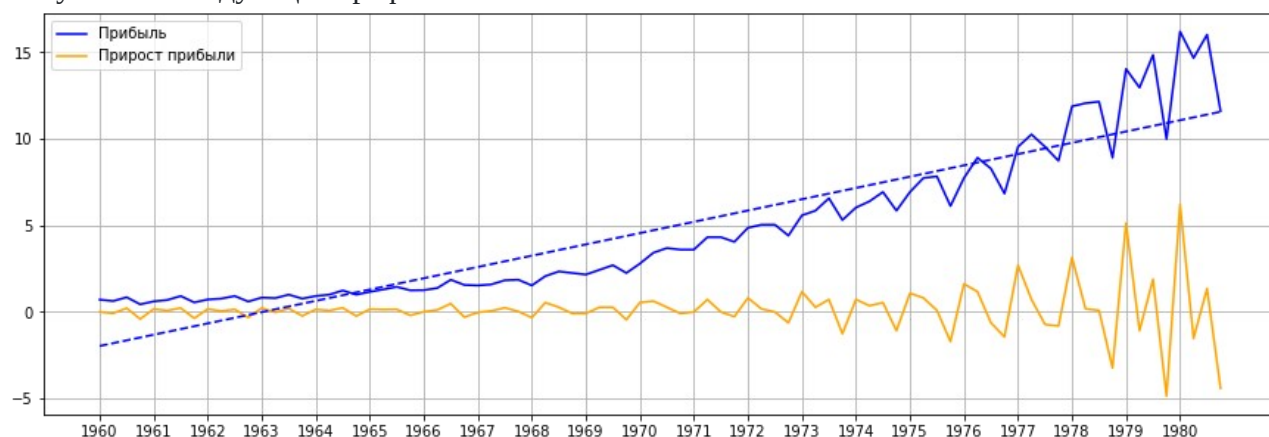
Что касается построения регрессора для всех графиков сразу, то это равносильно построению четырёх отдельных регрессоров.

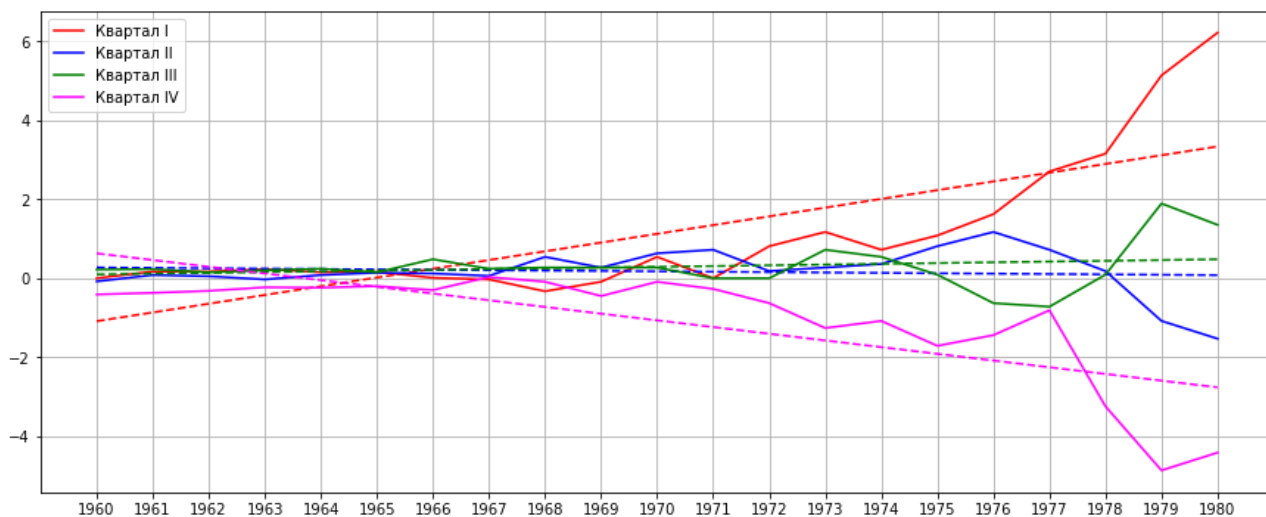
В данном случае метрика комбинированного регрессора равна 0.7265.

## 6. Джонсон-и-Джонсон

В данном задании предстоит выполнить анализ прибыли компании за 20 лет с 1960 по 1980 года. Требуется построить линейную регрессию по каждому кварталу и по всему периоду в целом, и с их помощью предсказать прибыль за 2016 год в целом и в каждом его квартале.

Получились следующие графики





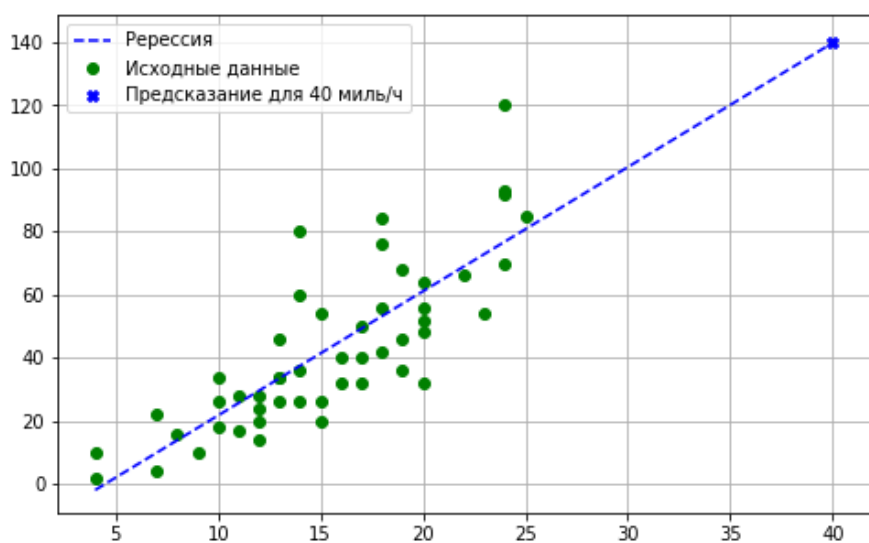
Согласно этим линейным моделям в 2016 году компания будет иметь следующую прибыль:

Квартал I	Квартал II	Квартал III	Квартал IV	Год
11.28	-0.2702	1.165	-8.860	34.56

Разумеется, говорить о правдоподобности этих предсказаний не приходится. Как минимум потому, что зависимость, как видно из графиков, у показателей не линейная.

## 7. Тормозной путь

Построим регрессию данных тормозного пути и узнаем тормозной путь при скорости 40 миль/ч.



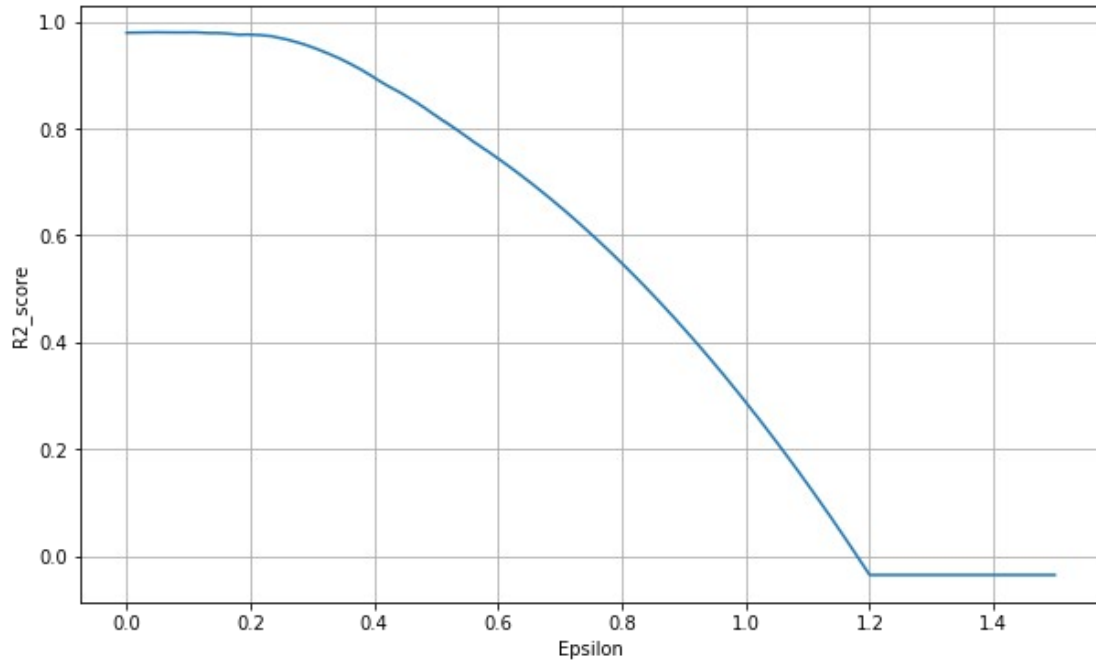
Построенная линейная регрессионная модель оценивает тормозной путь в 139.7 футов.

## 8. Опорновекторная регрессия



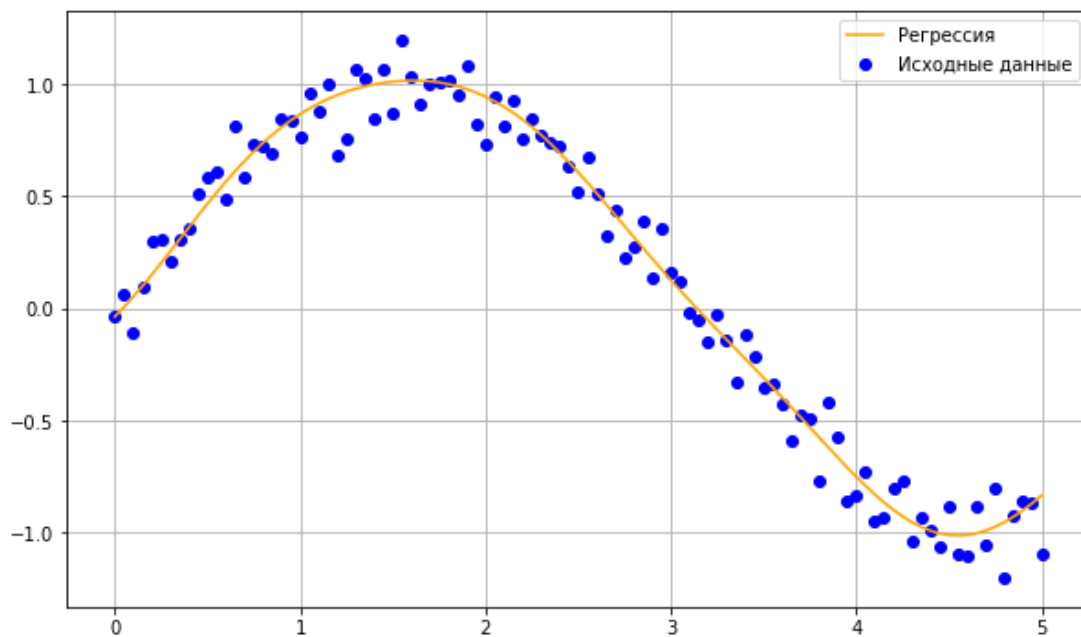
Исследуем зависимость успешности построенного опорновекторного регрессора от заданного параметра  $\epsilon$ . Этот параметр задаёт ширину области вокруг получающейся кривой, где регрессор не получает штрафов за точки, в ней оказавшиеся. Также задаётся ядро регрессора «rbf» и параметр  $C$ , равный единице.

По итогу имеем следующий график



Из него видно, что лучшее значение метрики достигается при малых допустимых эпсилон-окрестностях.

Рассмотрим регрессор с нулевым эпсилоном.



Его значение метрики: 0.979.

## 9. Регрессионное дерево

Построим по данным три модели: регрессионное дерево, линейный и опорновекторный регрессоры и сравним их по метрике  $r^2\_score$ . Сначала с параметрами по умолчанию.

Результаты:

- дерево: 0.20
- линейный: 0.65
- опорновекторный: -0.01

Попытаемся добиться лучших результатов, поменяв параметры. Увеличим штрафной параметр  $\gamma$  у SVR и ограничим максимальную глубину у дерева четырьмя уровнями (при таком параметре удалось добиться максимума от него).

Итоговые результаты следующие:

- дерево: 0.53
- линейный: 0.65
- опорновекторный: 0.13

Большого, скорее всего, от них уже не добиться.

Остаётся заключить, что в данном испытании лучшие результаты показал линейный регрессор.

## Приложение

Код программы на языке Python3

1.

2.

3.