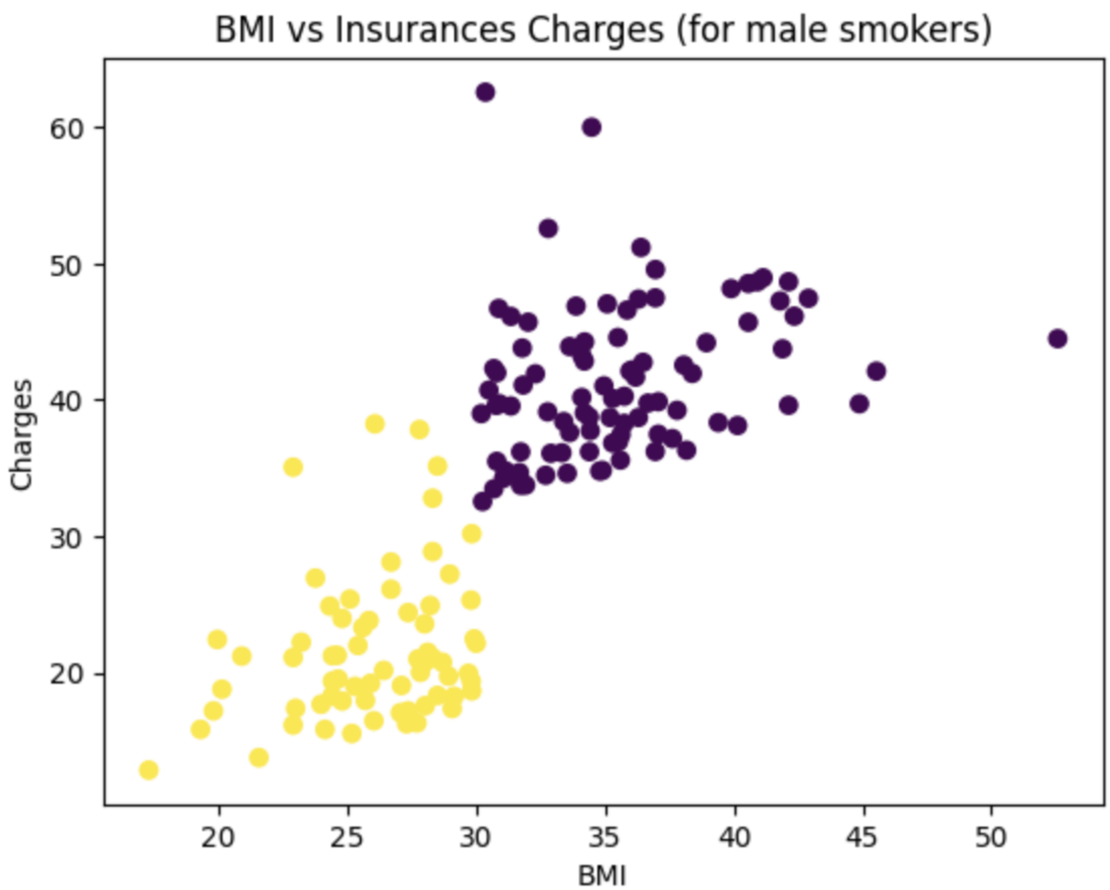
Swap Regression

Viral Chitlangia

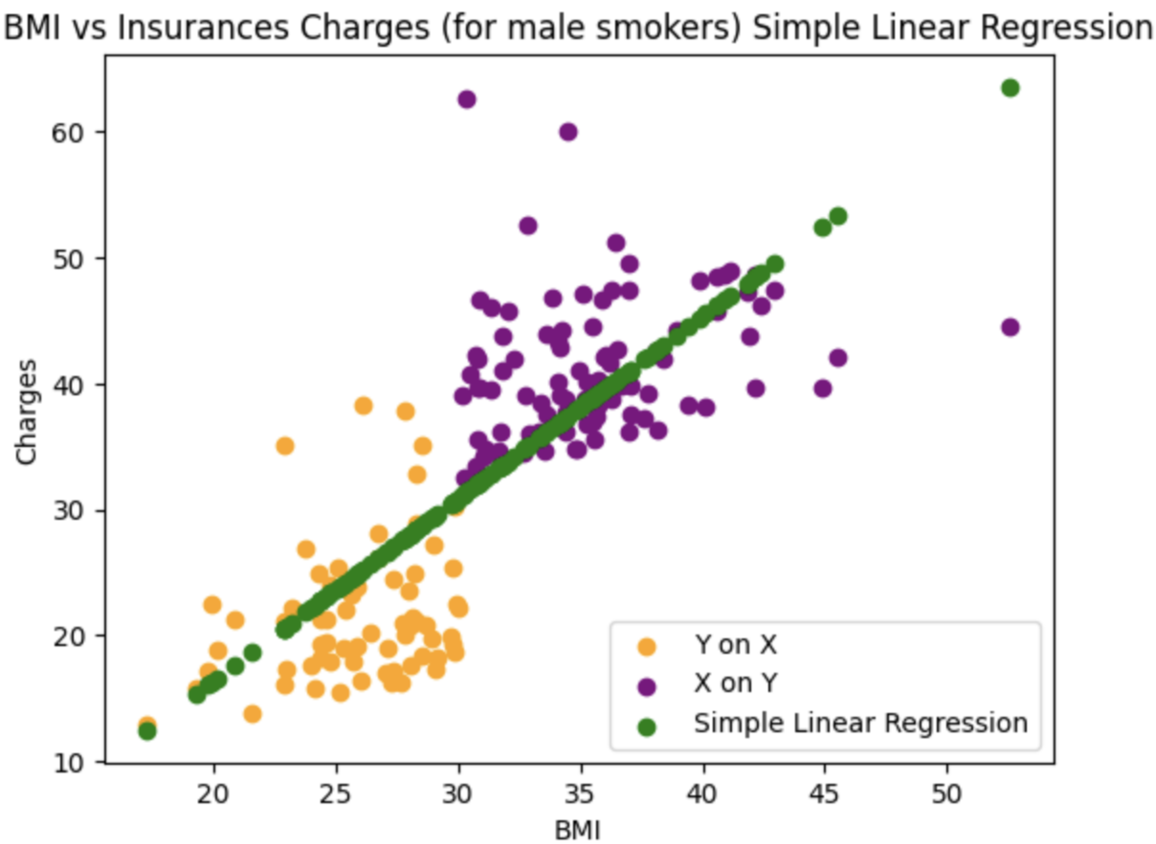
1 Data Used

I used the Kaggle dataset on US Health Insurance Dataset. We are considering the dataset of Male Smokers. Here, we are taking Charges to be regressed on BMI, if BMI ¡ 30, and BMI to be regressed on Charges, if BMI $>$ 30. If Y is regressed on X, Z = 1, if X is regressed on Y, Z = 0.



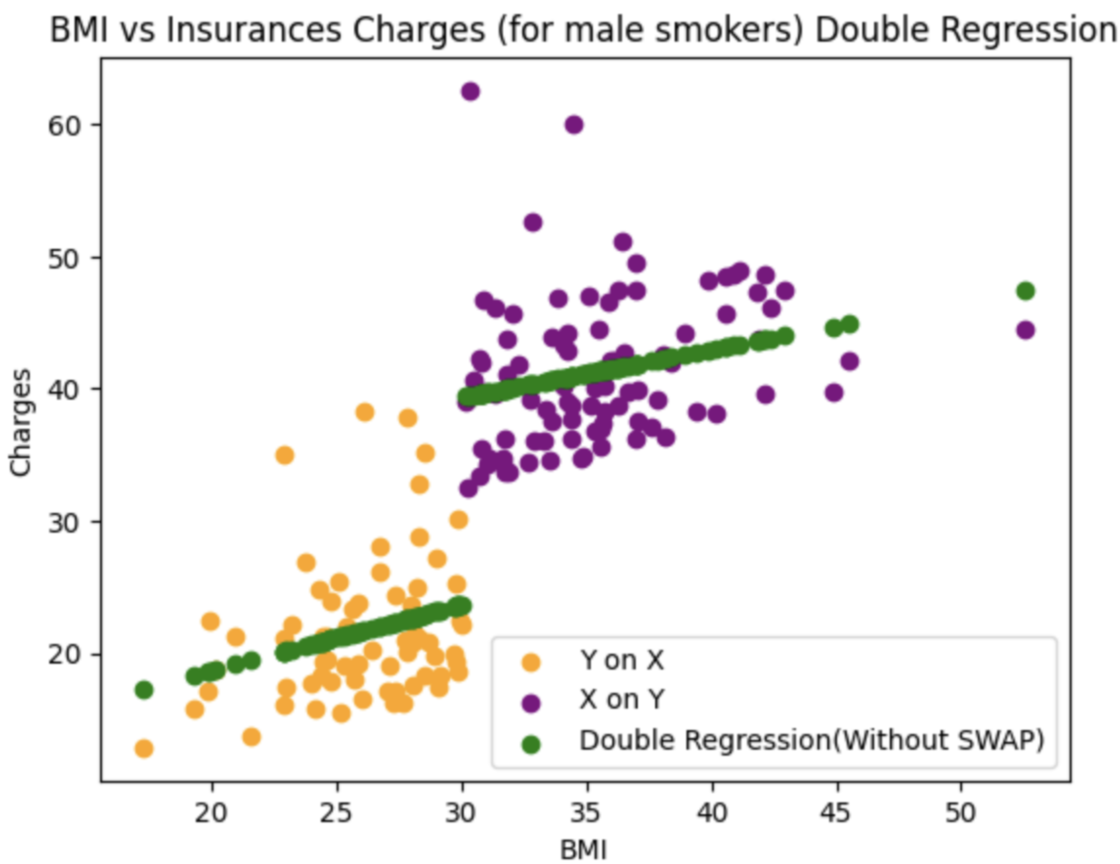
2 Simple Linear Regression

First, let us treat them as linearly related. So, we can draw a simple linear regression for the model, all together.
With that, we get the line as y = 1.4481x - 12.5776.



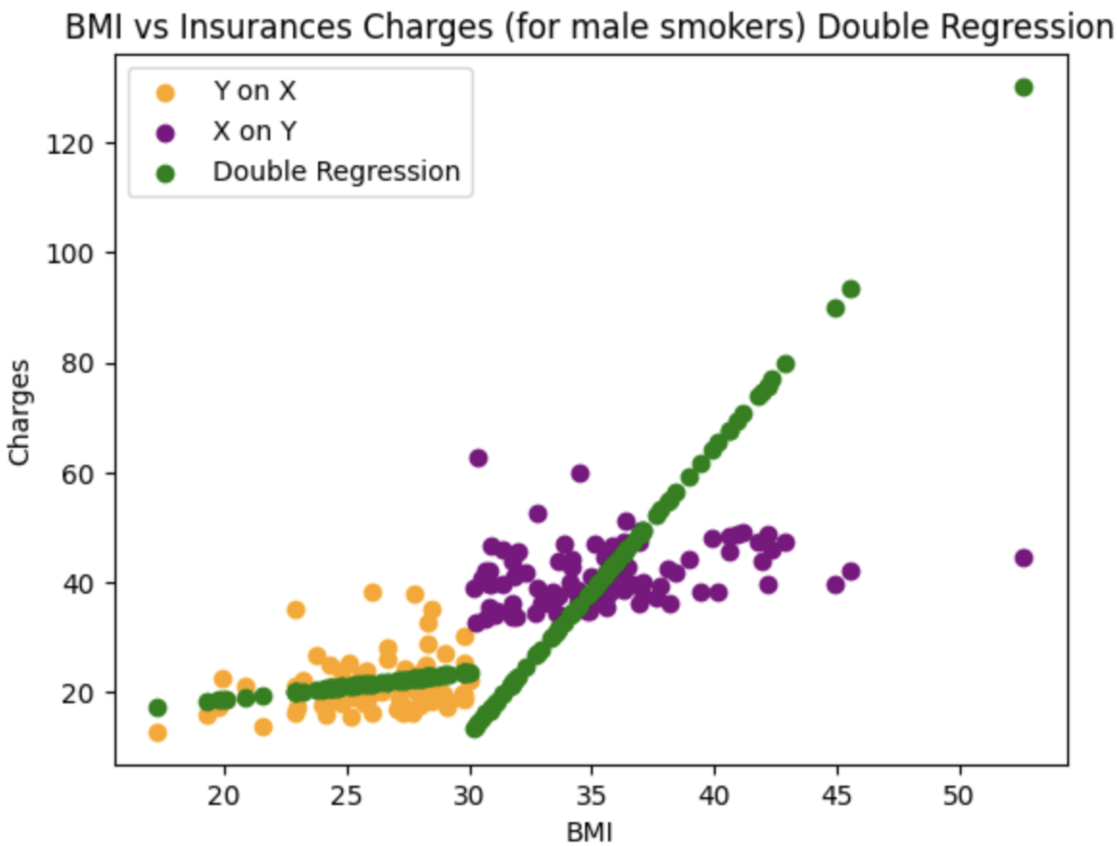
3 Two Simple Linear Regression

If we treat them as separate datasets, and create a model for them, separately, we get two linear models.
For Z = 1, we get line as y = 0.4986x + 8.6922
For Z = 0, we get line as y = 0.3628x + 28.4511


BMI vs Insurances Charges (for male smokers) Double Regression
Charges
BMI
60
50
40
30
20
20
25
30
35
40
45
50
Y on X
X on Y
Double Regression(Without SWAP)

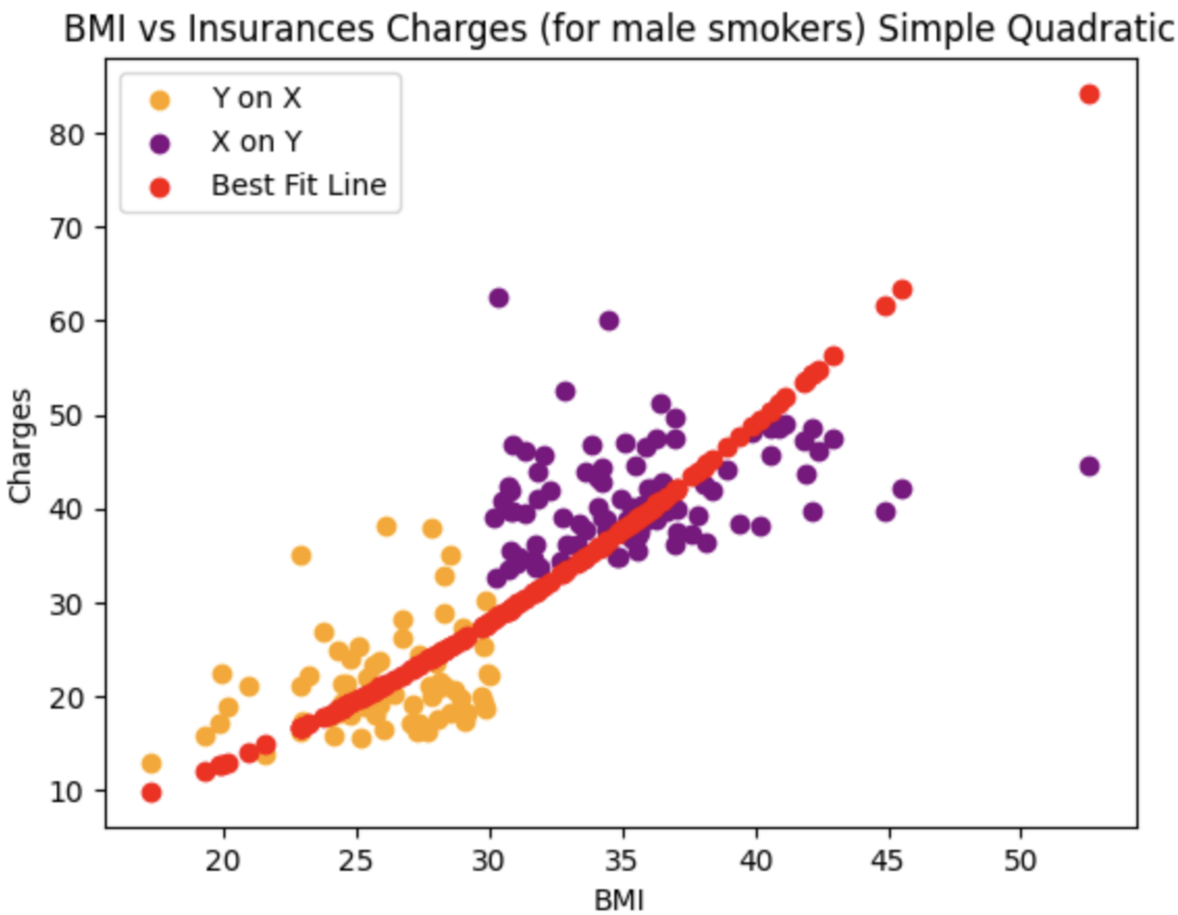
4 Two SWAP Linear Regression

If we treat them as separate datasets, and create a model for them, separately. However, for the dataset with Z = 0, we regress X on Y, instead of Y on X(Swapping the parameters)
For Z = 1, we get line as y = 0.4986x + 8.6922
For Z = 0, we get line as y = 5.2068x - 143.6114



5 Simple Quadratic Regression

If we try to get this in a Quadratic Model, we will get the quadratic as
y = $0.0304x^2$ - $0.0165x$ + 0.9996.

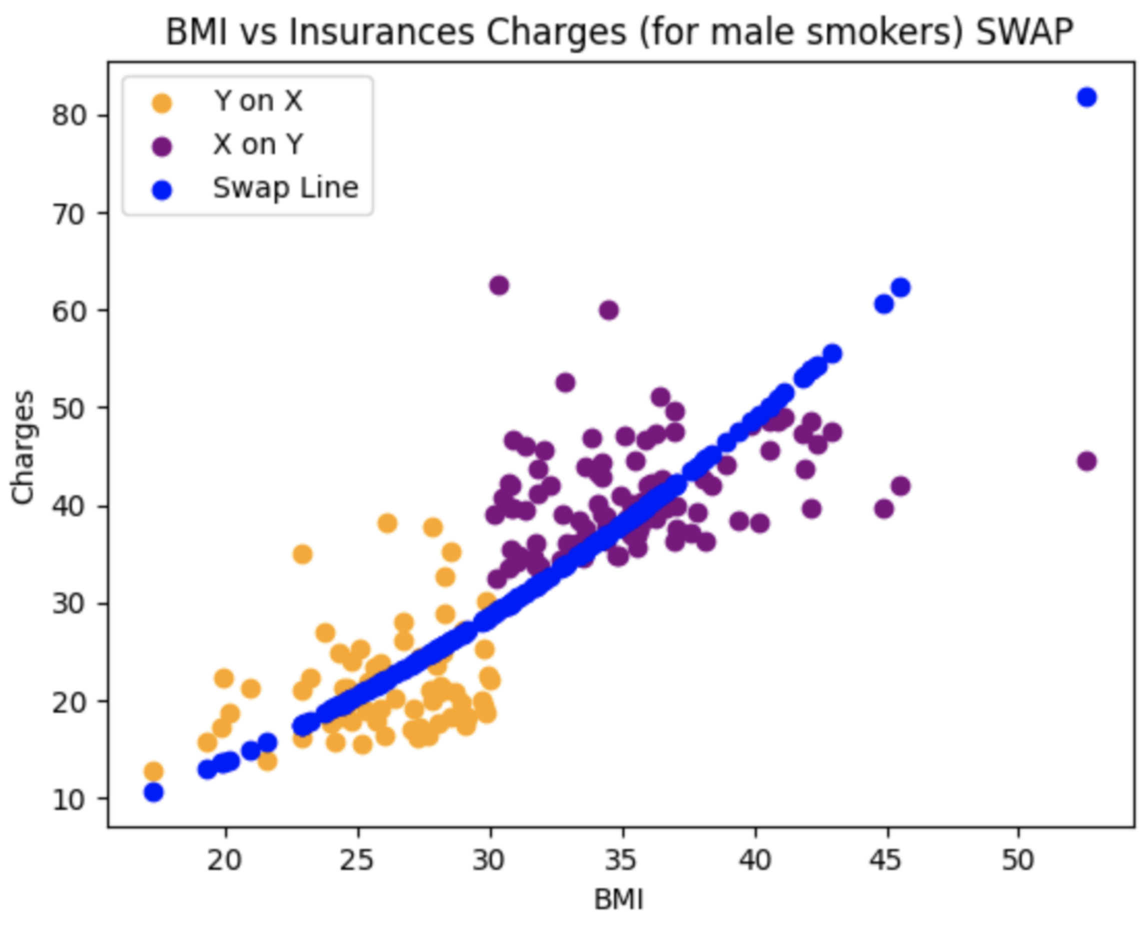

BMI vs Insurances Charges (for male smokers) Simple Quadratic
Y on X
X on Y
Best Fit Line
Charges
BMI
10
20
30
40
50
60
70
80
20
25
30
35
40
45
50

6 SWAP Regression

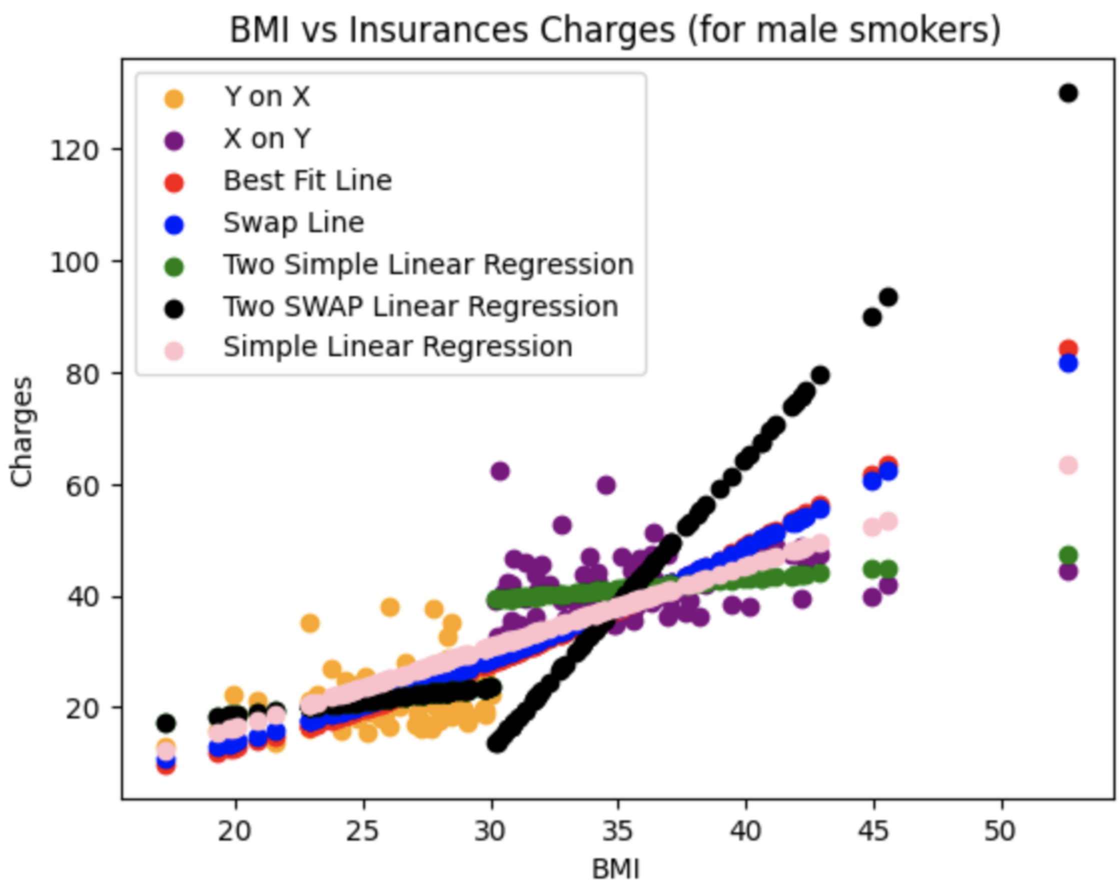
$L(a,b,c) = \Sigma_i(Y_{1i} - aX_{1i}^2 - bX_{1i} - c)^2 + \Sigma_i(X_{0i} + \frac{b}{2a} - \frac{\sqrt{b^2+4ac-4aY_{0i}}}{2a})^2$

The inverse of the Quadratic over here is taking as such because we know the nature of the function. We can reverse the sign if the data set is shifted in another direction.

We can optimize this using Gradient Descent. However, the problem arises due to the square root. Over here, we can use the trick of only considering the points where the square root is defined. For considering the final gradient, we can average out the sum with the total number of points which were defined. We get the final quadratic as y = $0.02687x^2$ + $0.13945x$ + 0.23846.



7 Comparison and Analysis



Like in the paper, we have defined the error as
$L(a,b,c) = \Sigma_i(Y_{1i} - aX_{1i}^2 - bX_{1i} - c)^2 + \Sigma_i(X_{0i} + \frac{b}{2a} - \frac{\sqrt{b^2+4ac-4aY_{0i}}}{2a})^2$

Model	Error
Simple Linear Regression	19100.889068546632
Two Simple Linear Regression	23517.073635955887
Two SWAP Linear Regression	4850.3253712304595
Simple Quadratic Regression	4076.7171547854573
SWAP Regression	4084.200145245626

8 Final Analysis

Here, we can see, in this type of data, we can see that SWAP is working better than linear models. In Quadratic Models, it is very close to the Simple Quadratic Regression model.