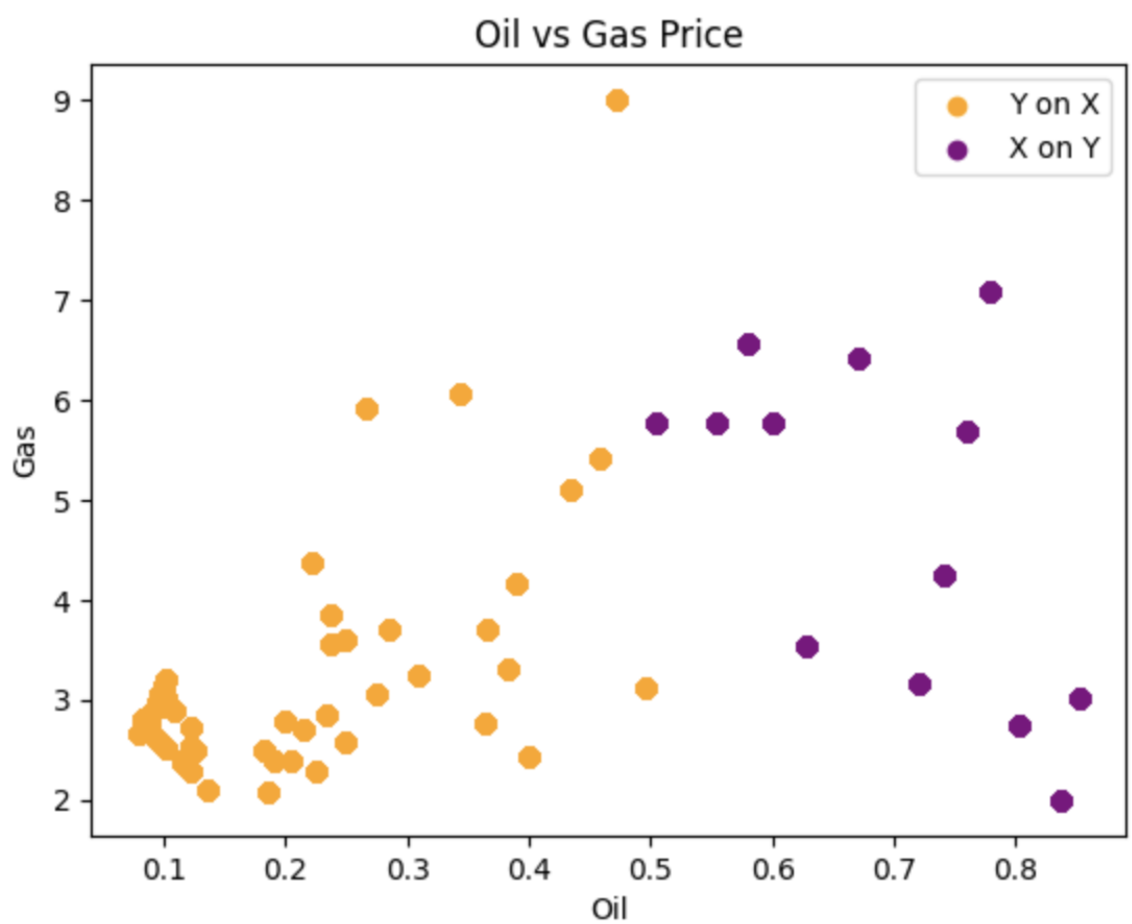
# Swap Regression

Viral Chitlangia

## 1 Data Used

I used the Kaggle dataset on Price of Natural Gas vs Oil from 1932-2014. Here, we take Gas on Oil. If Oil Price ¡ 50, then we regress Gas on Oil, and if Oil Price ¿ 50, we regress Oil on Gas. In the graph, we scale Oil $\rightarrow$ Oil/100.



## 2 Testing Goodness of Data

### 2.1 Testing Bi Causality

To test bi causality, we employed the Granger Causality Test. First, we tested the null hypothesis that Gas Prices does **NOT CAUSE** Oil Prices.

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=64.1597 , p=0.0000  , df_denom=11216, df_num=1
ssr based chi2 test:   chi2=64.1769 , p=0.0000  , df=1
likelihood ratio test: chi2=63.9940 , p=0.0000  , df=1
parameter F test:         F=64.1597 , p=0.0000  , df_denom=11216, df_num=1
```

```
{1: ({'ssr_ftest': (64.15971510758934, 1.2606872486026923e-15, 11216.0, 1),
   'ssr_chi2test': (64.17687622967588, 1.1373584265559454e-15, 1),
   'lrtest': (63.99401538520033, 1.2479773507647798e-15, 1),
   'params_ftest': (64.159715107597, 1.2606872485979446e-15, 11216.0, 1.0)},
  [<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x1396c7200>,
   <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x139a898e0>,
   array([[0., 1., 0.]])])}
```

We get a P Value of the order of $10^{-15}$, so we can reject the NULL Hypothesis that Gas Price does not cause Oil Price.

Second, we tested the null hypothesis that Oil Prices does **NOT CAUSE** Gas Prices.

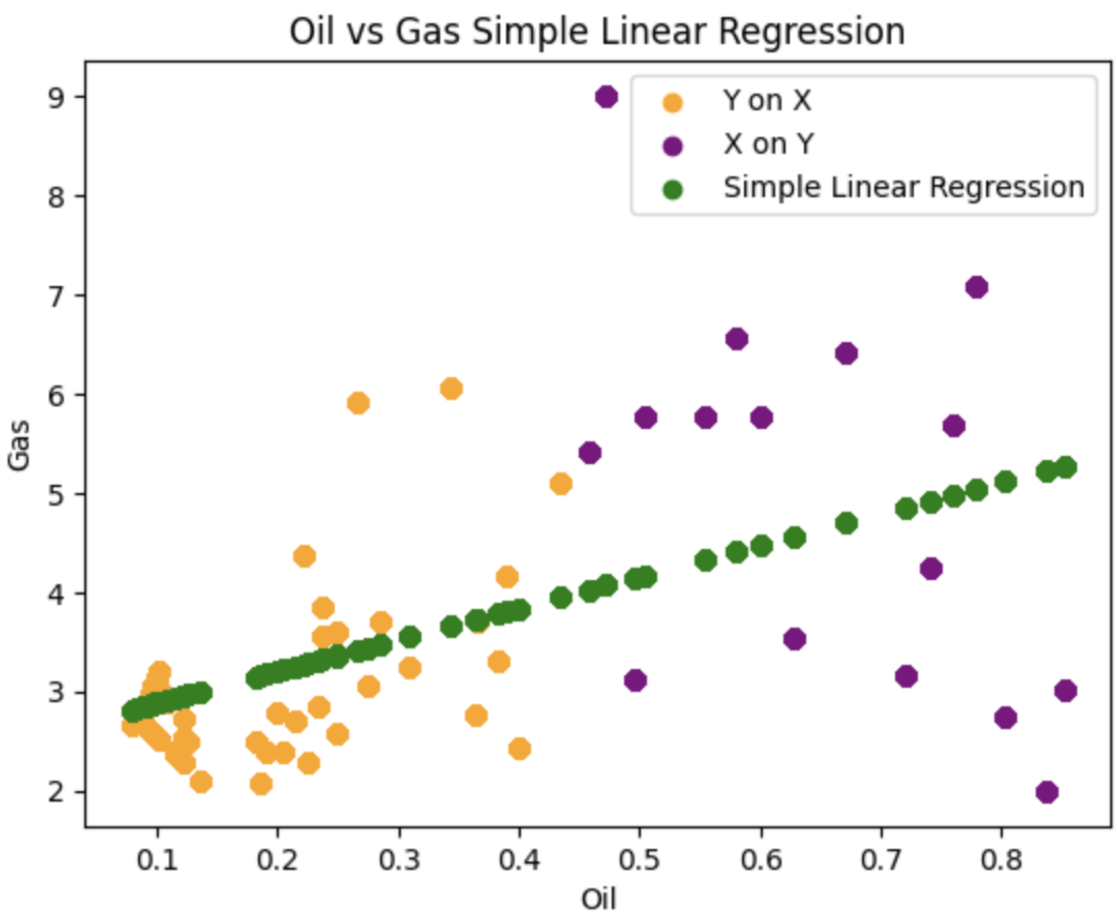
```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=18.4690 , p=0.0000  , df_denom=11216, df_num=1
ssr based chi2 test:   chi2=18.4739 , p=0.0000  , df=1
likelihood ratio test: chi2=18.4587 , p=0.0000  , df=1
parameter F test:         F=18.4690 , p=0.0000  , df_denom=11216, df_num=1
```

```
{1: ({'ssr_ftest': (18.46898606619916, 1.741488374293125e-05, 11216.0, 1),
   'ssr_chi2test': (18.473926058905885, 1.7224472259469195e-05, 1),
   'lrtest': (18.458732559014607, 1.736234115978405e-05, 1),
   'params_ftest': (18.468986066202664, 1.741488374289679e-05, 11216.0, 1.0)},
  [<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x1399f2fc0>,
   <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x139930470>,
   array([[0., 1., 0.]])])}
```

We get a P Value of the order of $10^{-5}$, so we can reject the NULL Hypothesis that Oil Price does not cause Gas Price.

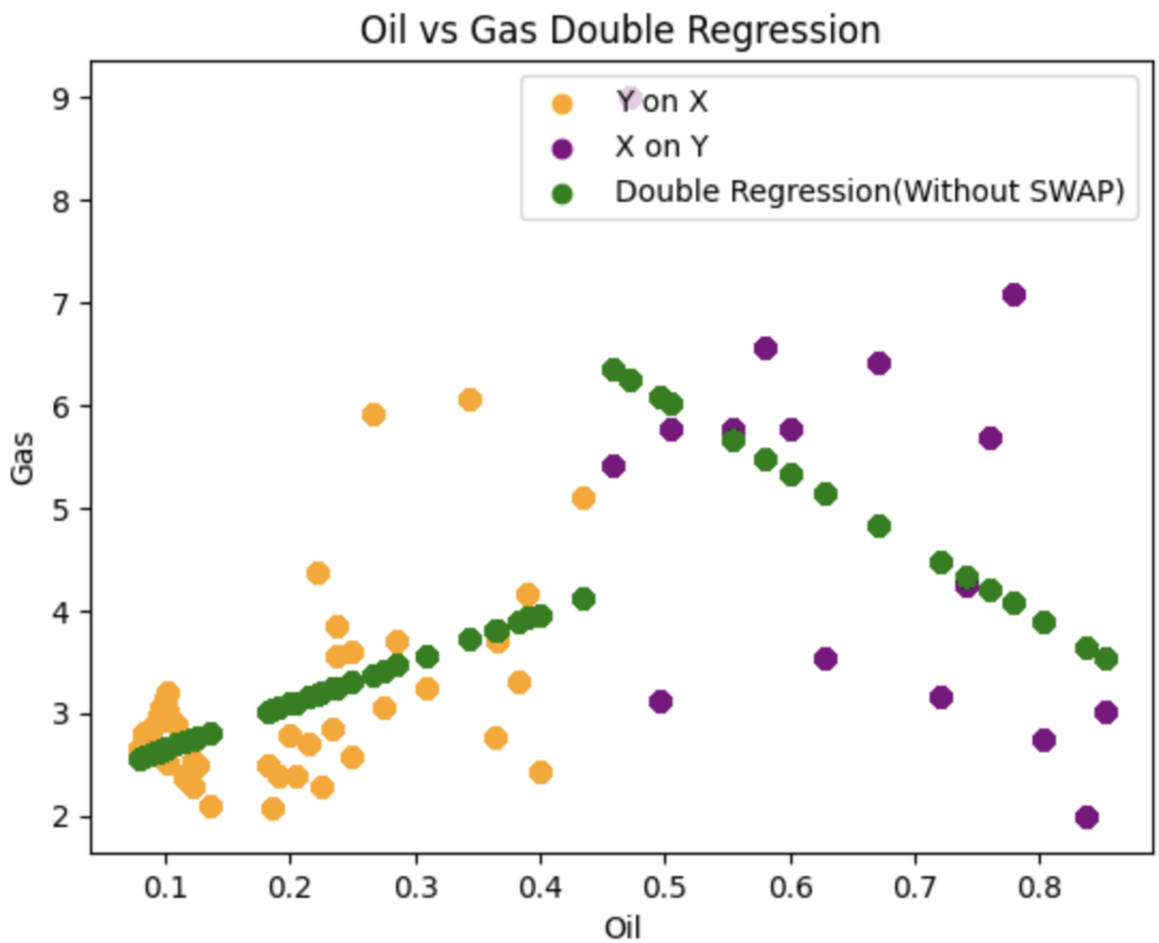
## 3 Simple Linear Regression

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 = 3.1874878879474853x + 2.566426674421489.



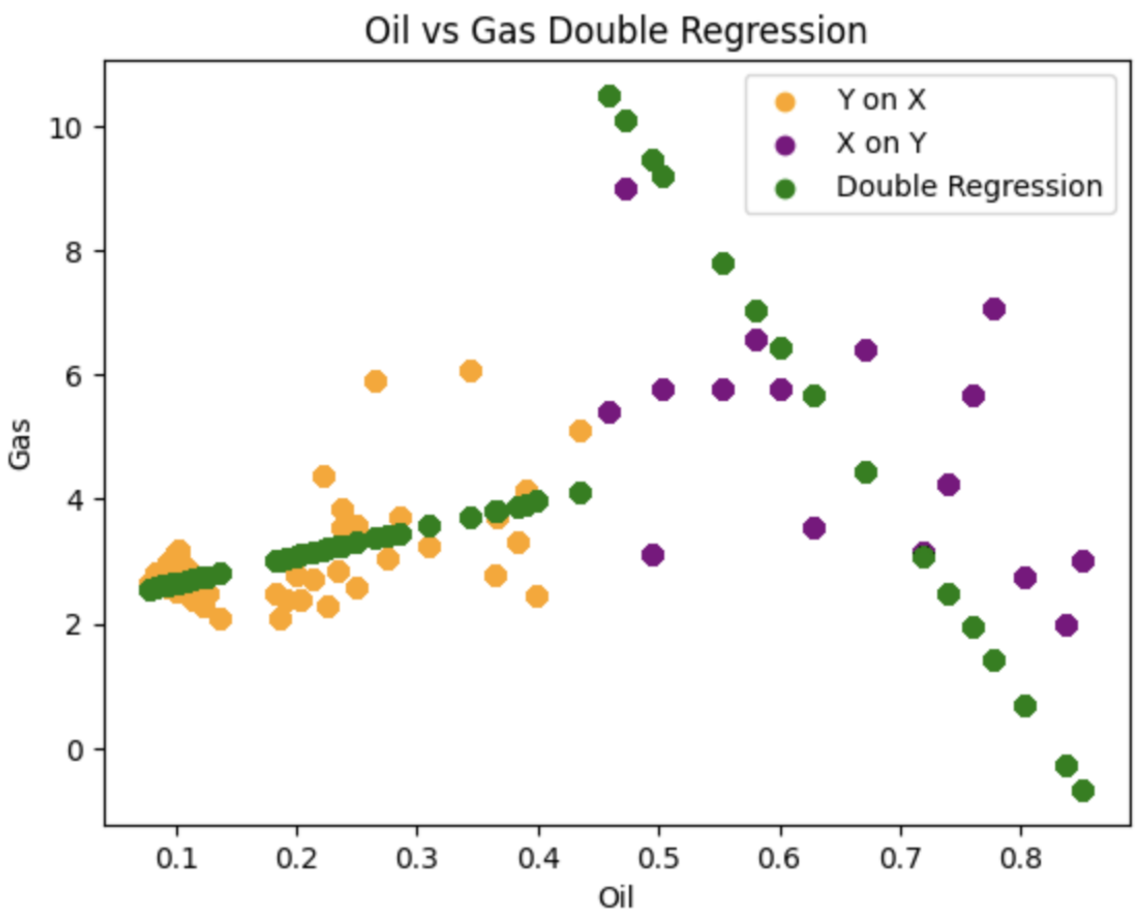

## 4 Two Simple Linear Regression

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 = 4.379099039459698x + 2.217083255904215
For Z = 0, we get line as y = -7.1232139653074595x + 9.61235437529847


Oil vs Gas Double Regression
Y on X
X on Y
Double Regression(Without SWAP)
Gas
Oil
9
8
7
6
5
4
3
2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

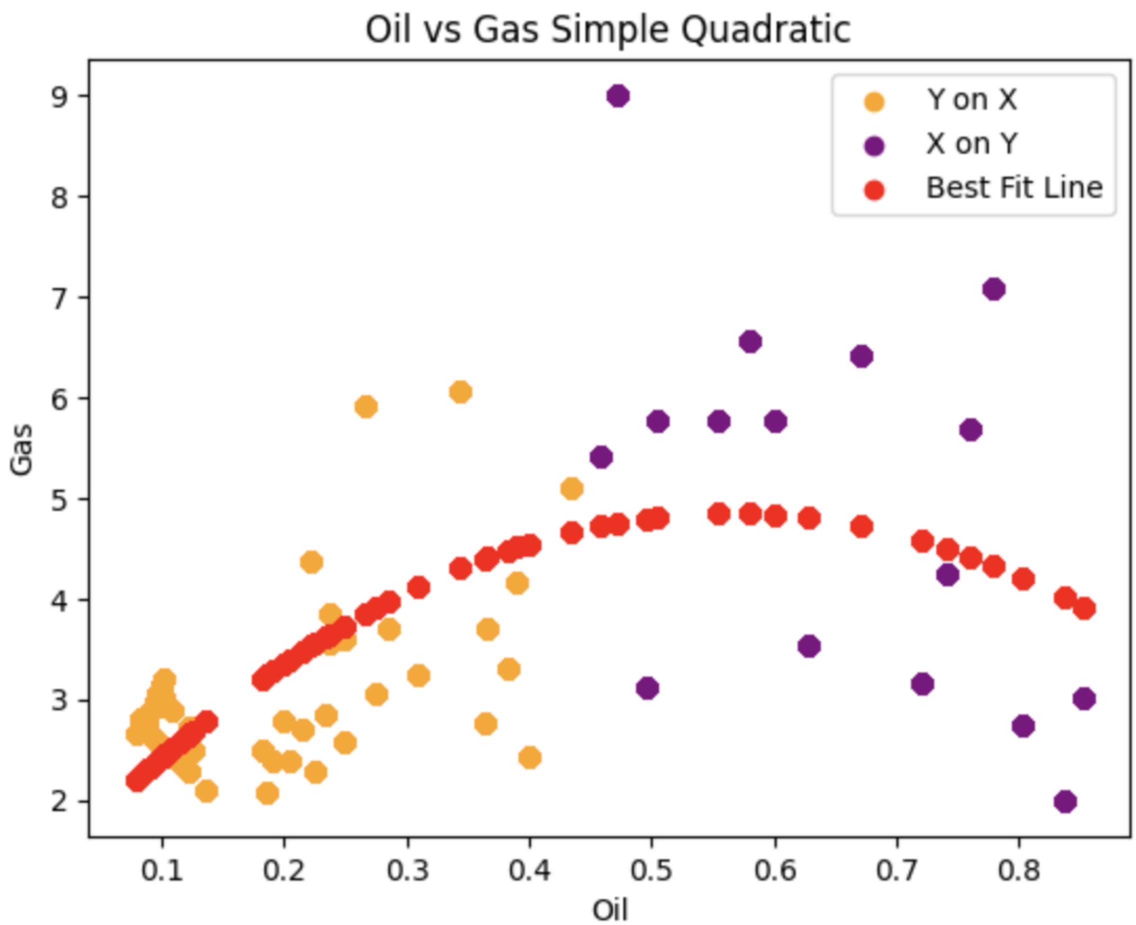
# 5 Two SWAP Linear Regression

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 = 4.379099039459698x + 2.217083255904215
For Z = 0, we get line as y = -28.362517447646457x + 23.489404298441745




# 6 Simple Quadratic Regression

If we try to get this in a Quadratic Model, we will get the quadratic as
y = -11.269789872036315$x^2$ + 12.715293502772141$x$ + 1.2649019521387417.

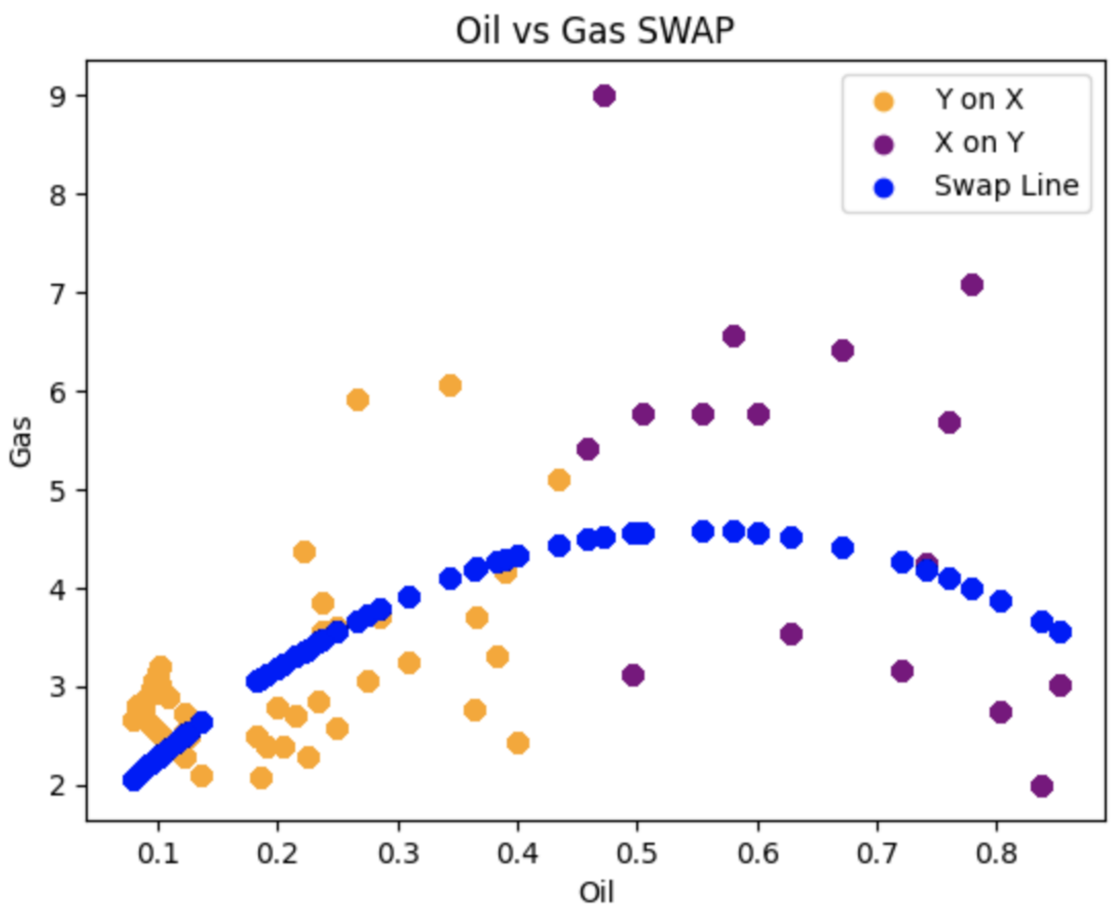

Oil vs Gas Simple Quadratic
Y on X
X on Y
Best Fit Line
Gas
Oil
9
8
7
6
5
4
3
2
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

# 7 SWAP Regression

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

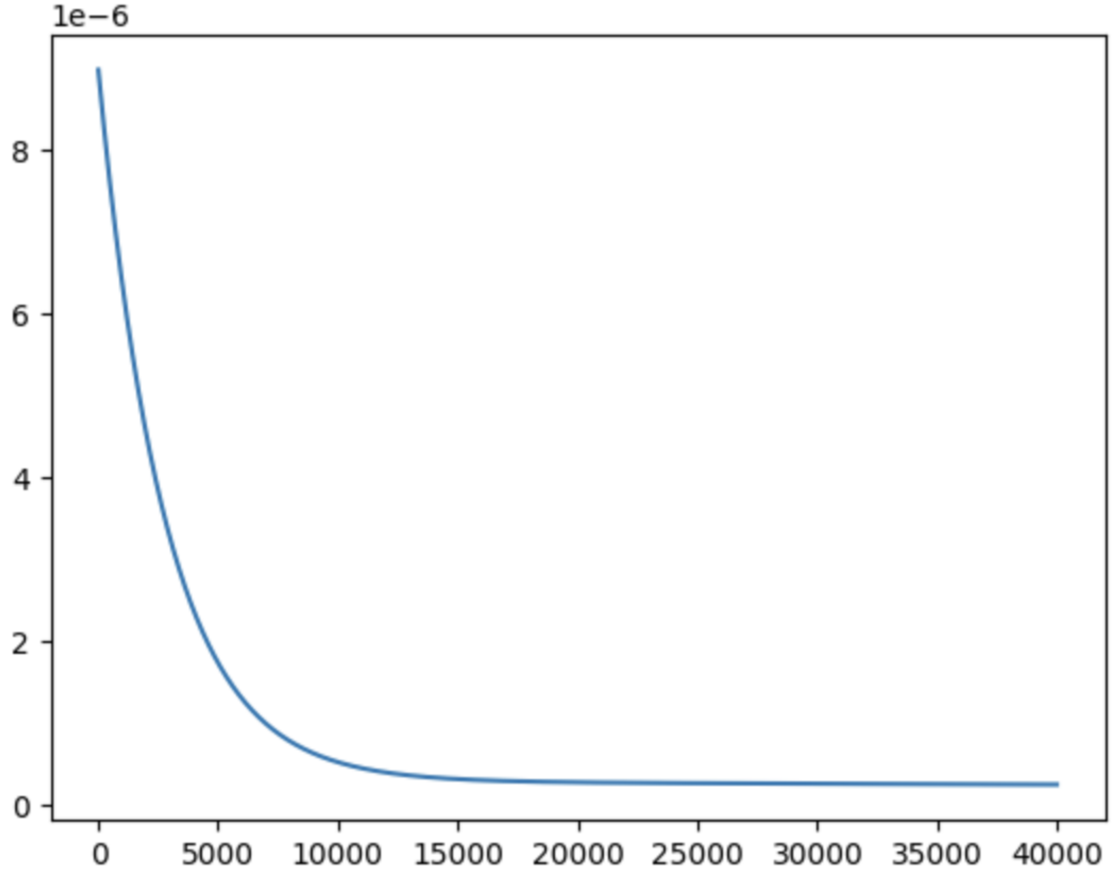
The inverse of the Quadratic over here is taking as such because we know the nature of the function. We can reverse the sign if the data set is shifted in another direction.

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 = -11.33183060777661$x^2$ + 12.49207752587935$x$ + 1.1443830629343317



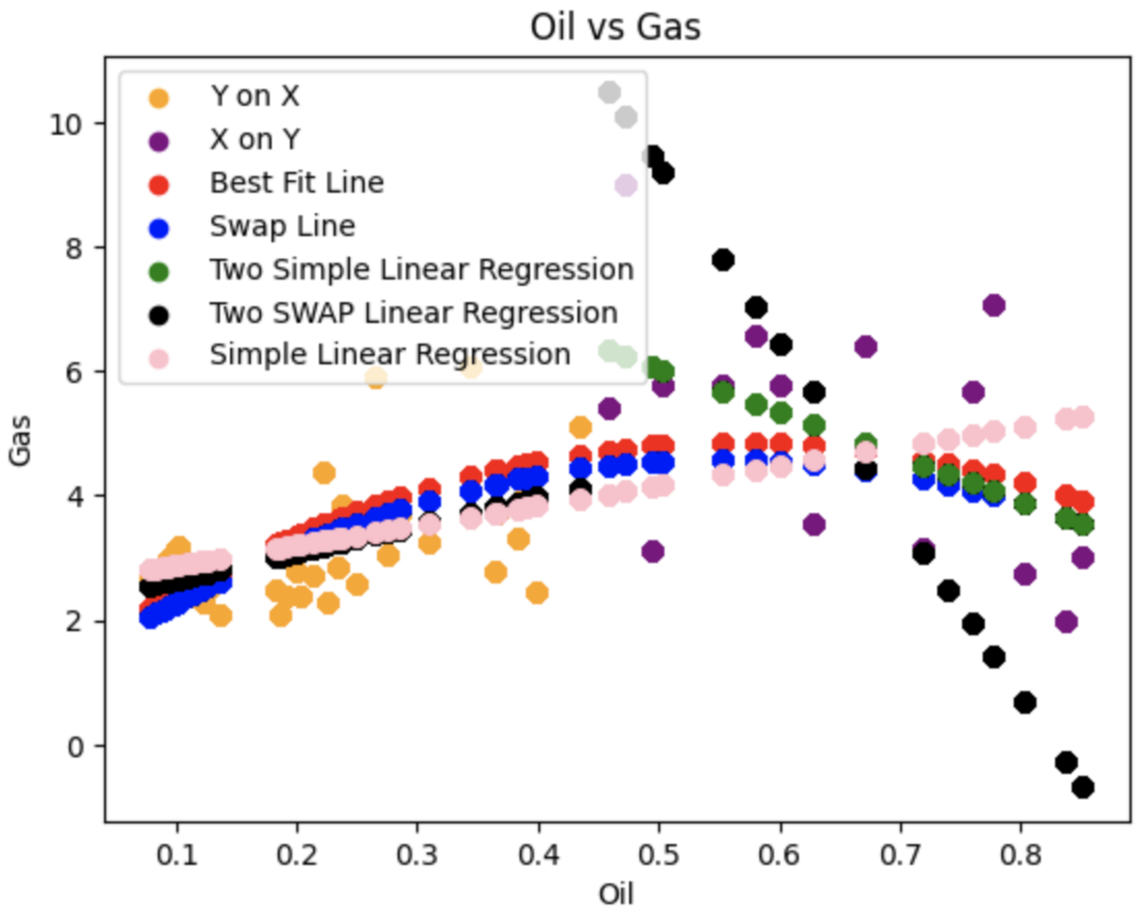
## 8 Error Analysis

The Gradient of the error, i.e. $(\frac{\partial L}{\partial a})^2 + (\frac{\partial L}{\partial b})^2 + (\frac{\partial L}{\partial c})^2$ is decreasing, and appears to tend to 0, as iterations increases.
Finally, the gradient is at $2.3610921850719418 * 10^{-7}$.




## 9 Comparison and Analysis

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



Oil vs Gas
Y on X
X on Y
Best Fit Line
Swap Line
Two Simple Linear Regression
Two SWAP Linear Regression
Simple Linear Regression
Gas
Oil
0
2
4
6
8
10
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

| Model | Error |
|---|---|
| Simple Linear Regression | 1.2620226968139898 |
| Two Simple Linear Regression | 1.5311459235576887 |
| Two SWAP Linear Regression | 1.521036549380825 |
| Simple Quadratic Regression | 0.6635223190363686 |
| SWAP Regression | 0.6306609524418897 |

## 10 Final Analysis

Here, we can see, in this type of data, we can see that SWAP is working better than the other Linear Models for this Bi Directional Causal Dataset.