Analysing and Re-Implementing SWAP Regression

Zehaan Naik

May 2024

Contents

1 Introduction

When we look at a standard linear regression problem on bi-variate data, the general approach used in most problems is to consider one variable as a stochastic variable (the regressor) and the other as a non-stochastic variable (the regressed).
The standard model that we fit in such situations is the following:

$$y = \beta X + \epsilon$$

where, ϵ $N(0, \sigma^2)$ And we know that the co-efficient matrix β is calculated with the formula :

$$\beta = ((X^T X)^{-1} X)y$$

The problem with this model lies in our fundamental assumption that the problem is classified as a strictly stochastic and non-stochastic variable.
As we will observe further in the report, this scenario is not very likely. When we consider experiments where a particular variable cannot be observed directly beyond a specific range and depends on the other variable, the assumption of a strict divide between the stochastic and non-stochastic variables feels moot.

In this report, we aim to understand a linear regression method that seeks to establish a model that can circumvent this assumption that the roles of predictor and regressed remain constant for the range of values in the experiment. We study a method that allows us to **"SWAP"** these roles for both variables in question.
This will allow us to better fit our model for the trends in the data and be better able to predict observations using the same model.

2 Motivation

Human's tolerance to heat and humidity (as measured by water sure) is a subject of considerable interest in human physiology. A common tool used in such studies is the psychometric chart, based on an alternating design where the response and the predictor trade places in one experiment. To our knowledge, statistical methods have not been available to handle this problem. Our new regression allows the response and the predictor to **"trade places."**

Our inquiry originated from a data set collected in an experiment in Kinesiology (Zeman (2001)). The study was concerned with epidemics of deaths in heat waves for older people, and its purpose was to determine the "Upper Limit of the Prescriptive Zone" (ULPZ) on a psychrometric chart of the ambient dry bulb temperature T versus the water vapor pressure P. The study performed a sequence of temperature-pressure tolerance experiments which were age- and sex-specific. Forty healthy subjects, including older men, older women, younger men, and younger women of average fitness, were recruited, with each of the four groups containing 9 to 11 subjects. The older subjects were aged between 63 and the younger subjects between 18-30. For each subject, six experiments were performed, among which three were under warm and humid conditions, to be called the P_{crit} experiments, and three were under hot and dry conditions, to be called the T_{crit} experiments.
In the three T_{crit} experiments, P was held constant at 12 mmHg, 16 mmHg, or 20 mmHg, and the temperature was increased to one °C every five minutes, starting from 28°C after a 30-minute equilibration period. This continued until a tolerance limit T was reached. In the three P_{crit} experiments, T was held constant at 34° C, 36° C, or 38° C while the pressure increased by one mmHg every five minutes, starting from 9 mmHg after a 30-minute equilibration period. This continued until a pressure tolerance limit P was reached. Thus, experiments always started at regions of pressure and temperature that were comfortable for the human subjects and gradually increased one variable. During each exper-

iment, the subjects walked continuously on a treadmill for up to 2.5 hours at a constant speed in an environmental chamber. One point on the ULPZ line was determined as the ambient conditions at which body core temperature was forced out of equilibrium.
Figure 1 shows the portion of the data set for the older males to illustrate the data. In the upper-left part, temperature acts as the predictor, and the pressure acts as the response, whereas in the lower-right part, the pressure acts as the predictor, and the temperature acts as the response. The solid curve is the ordinary least squares fit, treating pressure as the response and using a quadratic polynomial of the temperature as the regression function. This analysis is inadequate; for example, the observations lie almost entirely to the left of the curve at the bottom of the chart. Our goal is to combine the two parts of the data into a coherent regression analysis, where the regression curve passes through the center of the response variables, whether temperature or pressure. Since the regression is designed for Samples With Alternating Predictors, we call it **SWAP regression**.

3 Which assumption(s) do we break

We know that linear regression follows the follwoing 6 fundamental assumptions:

- **Linearity:** The relationship between the dependent and independent variables is linear.
  This assumption is broken in multiple polynomial models regularly and isn't of much interest to us in terms of discussion in this conversation, hence we shall ignore this topic by just mentioning that the relation that we wish to obtain is indeed non-linear.

- **Independence:** The observations are independent of each other.
  This assumption is followed.

- **Homoscedasticity:** The variance of the errors is constant across all levels of the independent variables.
  This assumption is followed.

- **Normality:** The errors follow a normal distribution.
  This assumption is followed.

- **No multicollinearity:** The independent variables are not highly correlated with each other.
  This is where our method majorly deviates from the standard.
  Multicolinearity is what we want to deal with in our paradigm. For reference, multicolinearity occurs when the independent variables show moderate to high correlation. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.

In our setting, we don't even assume that our predictor remains constant across the analysis. The advantage of avoiding this assumption is that we are able to construct a model that allows for the "trend" in our data to be affected by both contributing factors bringing it closer to the real world answer.
The problem with breaking this assumption is we can't simply use the closed form solutions to the linear models that we use and we need to figure out a different way to either analytically calculate or numerically approximate the "trend" coefficients. [The co-efficients of approximation function]

- **No endogeneity:** There is no relationship between the errors and the independent variables.
The current model does not break this assumption. However, the SWAP model has some potential to account for endogeneity. We shall discuss the same in the last part of this report.

4 The SWAP Method

The idea in the paper is to consider an invertible function $f(x)$ such that:

$$E(Y|X, Z = 0) = f(x)$$

and

$$E(X|Y, Z = 1) = f^{-1}(x)$$

To approximate such a function $f(x)$, we consider the set of functions :

$$\mathscr{G} = \{g \in L^2(P_x)|gisaninjectionfrom\Omega_x \to \Omega_yandg^{-1} \in L^2(P_x)\}$$

We want to minimize the quadratic loss function:

$$Q = E[(Y - g(X))^2 I(Z = 0)] + E[(X - g(Y))^2 I(Z = 1)]$$

To approximate the loss function for our data set, we can approximate g_θ to be:

$$g_\theta(x) = ax^2 + bx + c$$

Here, $\theta = (a, b, c)^T$
clearly,

$$g_\theta^{-1}(x) = -b/2a + \sqrt{b^2 - 4a(c - y)}/2a$$

Now, we need to minimize this approximator concerning the quadratic loss function Q discussed above:

$$Q_n(g_\theta) = E_n[(Y - aX^2 - bX - c)^2 I(Z = 0)] + E_n[X + (b/2a + \sqrt{b^2 - 4a(c - Y)}/2a)^2 I(Z = 1)]$$

Last question: How to find a value of θ to minimise Q_n
In the final stretch of this exercise, we plan to minimize the loss function presented above using a couple of known statistical methods. Since the above problem is now reduced to likelihood estimation, we are well within our domain of knowledge to employ classic MLE estimates to approximate the function g_θ. In the next section, we shall employ numerical methods to obtain this estimate.

5 Numerical Approximation

Since Q_n is an expectation value of a function of Random variables, we can approximate it using a simple Monte Carlo estimate from our available data. The closed-form expression of the function can be described as :

$$\hat{Q_1} = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c)^2 \mid Z = 0$$

$$\hat{Q_2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{b + \sqrt{b^2 - 4a(c - y_i)}}{2a})^2 \mid Z = 1$$

$$\hat{Q} = \hat{Q_1} + \hat{Q_2}$$

Remark: We can also show that the second derivative of the predictors is identically greater than 0. Hence, they will have a minimum.

$$\frac{\partial \hat{Q_1}}{\partial a} = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c) * (-2x_i^2)$$

$$\frac{\partial \hat{Q_1}}{\partial b} = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c) * (-2x_i)$$

$$\frac{\partial \hat{Q_1}}{\partial c} = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c) * (-2)$$

and

$$\frac{\partial^2 \hat{Q_1}}{\partial a^2} = \frac{1}{n}\sum_{i=1}^{n}(2x_i^4)$$

$$\frac{\partial^2 \hat{Q_1}}{\partial b^2} = \frac{1}{n}\sum_{i=1}^{n}(2x_i^2)$$

$$\frac{\partial^2 \hat{Q}_1}{\partial c^2} = \frac{1}{n}\sum_{i=1}^{n}(2) = 2$$

Lastly, We need to ensure that for the swap part of model fitting, the expectation value is only defined for values of y following :

$$y < c - \frac{b^2}{4a}$$

If the data points agree with this constraint, we can ensure that we can make sense of both the expectation value and the estimator. Keeping this constrain in mind, we can derive the expressions for:

$$\frac{\partial^2 \hat{Q}_2}{\partial a^2}$$
$$\frac{\partial^2 \hat{Q}_2}{\partial b^2}$$
$$\frac{\partial^2 \hat{Q}_2}{\partial c^2}$$

Use these in our standard gradient descent formula:

$$\beta_{k+1} = \beta_{k+1} - l * \beta_k$$

Where l is an appropriate learning rate.

6 Real Data Application

To understand this method further, we implement the logic discussed in the previous section on a real-world data set that fits the problem described in the report. Consider insurance.csv:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

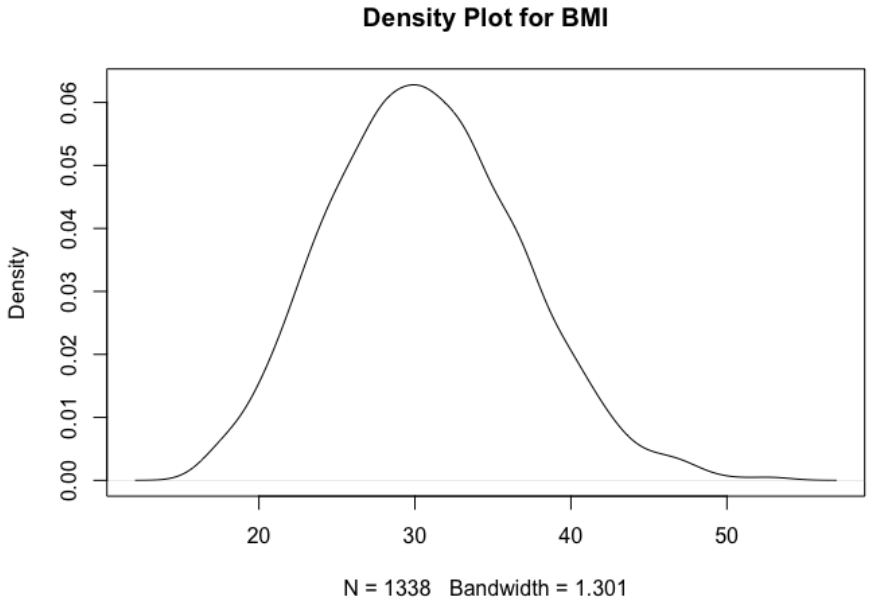
This dataset comprises several personal details about the customers of a particular bank and their respective health insurance charges. We observe two categorical variables in the data, namely, 'sex' and 'smoker' indicating the sex of the customer and whether or not they smoke, respectively.

The summary for the data is as follows:

```
> summary(dat)
      age            sex                 bmi           children        smoker             region             charges     
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000   Length:1338        Length:1338        Min.   : 1122  
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character   Class :character   1st Qu.: 4740  
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character   Mode  :character   Median : 9382  
 Mean   :39.21                      Mean   :30.66   Mean   :1.095                                         Mean   :13270  
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000                                         3rd Qu.:16640  
 Max.   :64.00                      Max.   :53.13   Max.   :5.000                                         Max.   :63770  
> 
```

For this problem, we want to establish a relationship between a customer's BMI (Body Mass Index) and their respective insurance charges. It is natural for one to assume that as a person gets progressively more overweight, their chances of contracting chronic diseases increase. Hence, as a bank, I'd impose a greater insurance charge on them to compensate for a greater risk of losing my money.
Additionally, BMI values above a certain threshold do not vary significantly with an increase in a person's weight. The reason for this is the fact that a BMI of over 30 is considered obese. Hence, in this range of values, even a slight increase in BMI would amount to a significant increase in the risk a person faces to incur severe health problems. Consequently, their insurance charges have increased by a substantial amount as well.

The plot gives us a visual for the distribution of BMI values for all individuals:



Hence, our model should regress insurance charges on BMI for the initial set of BMI values (18.0-30.0) and regress BMI on insurance charges for BMI values

greater than 30.0. This description perfectly fits our SWAP model, so we are trying to implement the abovementioned logic in this problem.

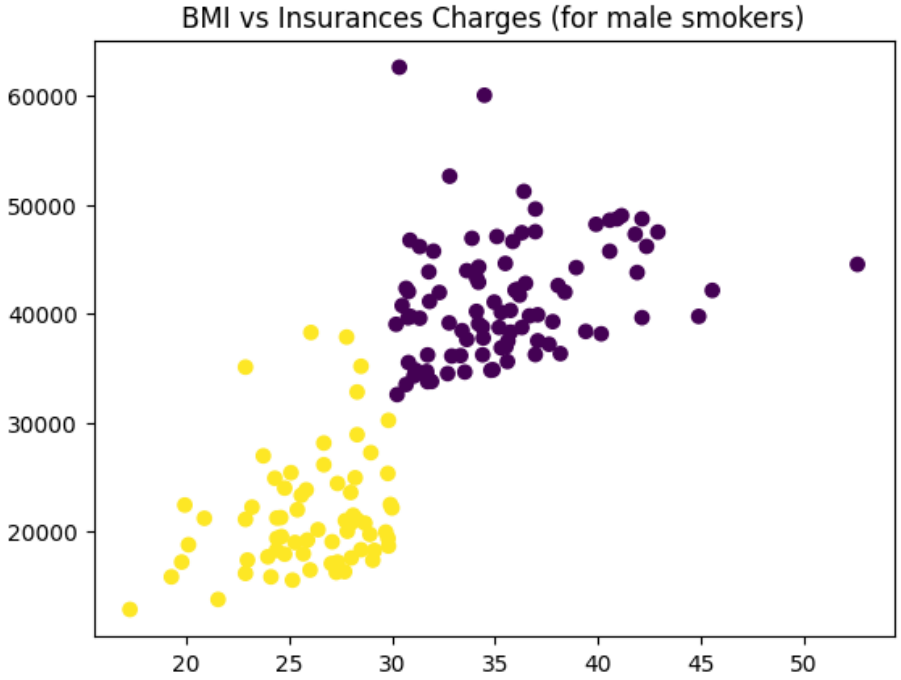
7 Implementation

Let's filter our data into four categories:

- Male Smokers
- Male Non-Smokers
- Female Smokers
- Female Non-Smokers

We make the following categories to avoid any confounding variables in our data.

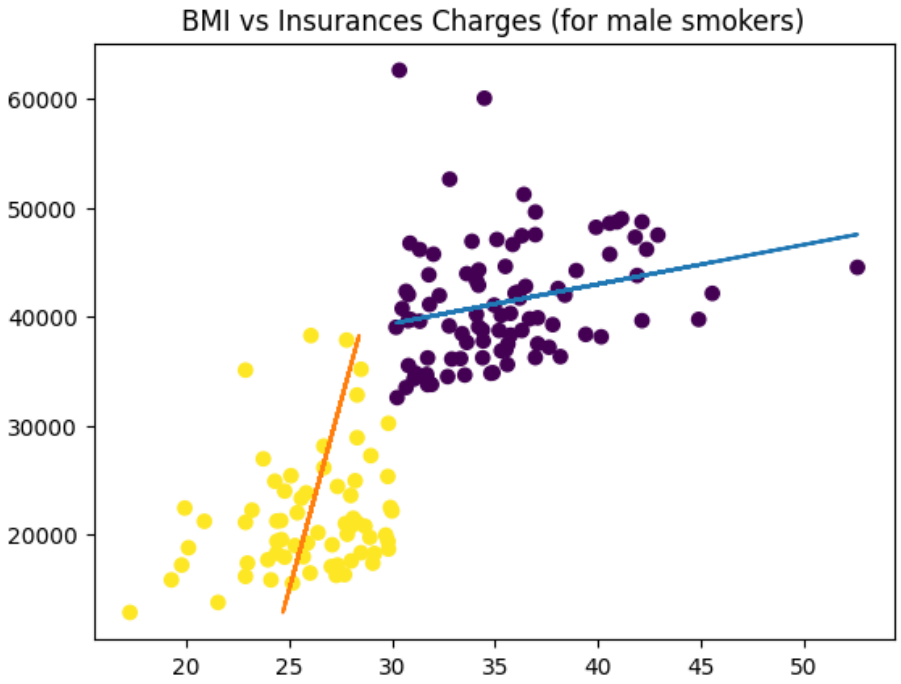
Now consider the category of Male Smokers. When we plot the BMI against the insurance charges, we get the following plot:



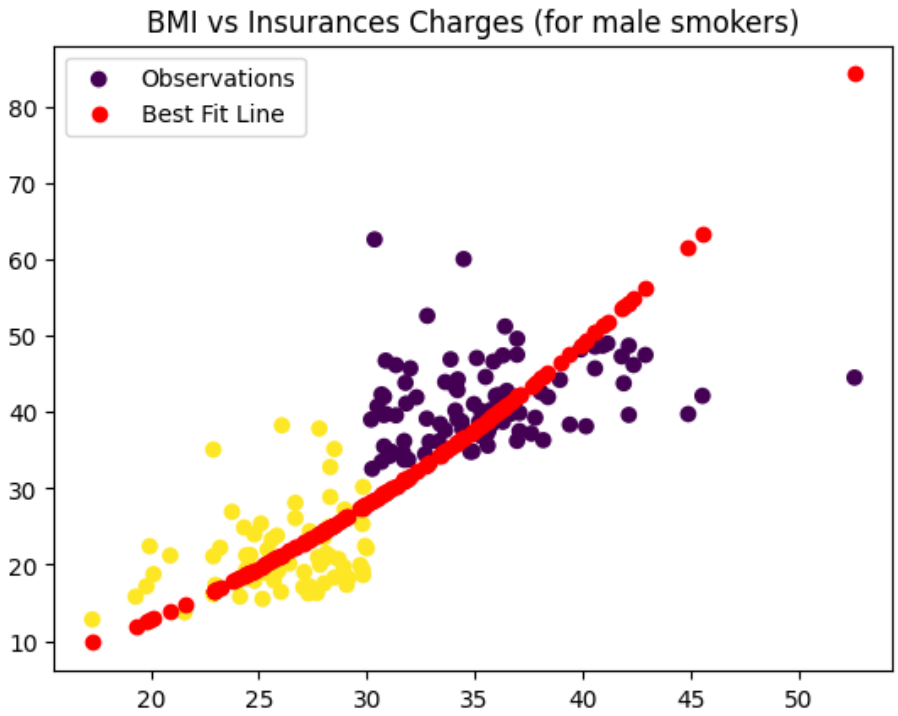
The yellow points indicate the customers whose BMI is under 30.0, and the purple dots represent the customers whose BMI is over 30.0.

It is evident that the two sections in the data conform to two different linear trends, and our model should be able to account for them.

For a comparative study, let's try to fit two different values of β, say β_{yellow} and β_{purple} on our data. For this, we get the following plot:



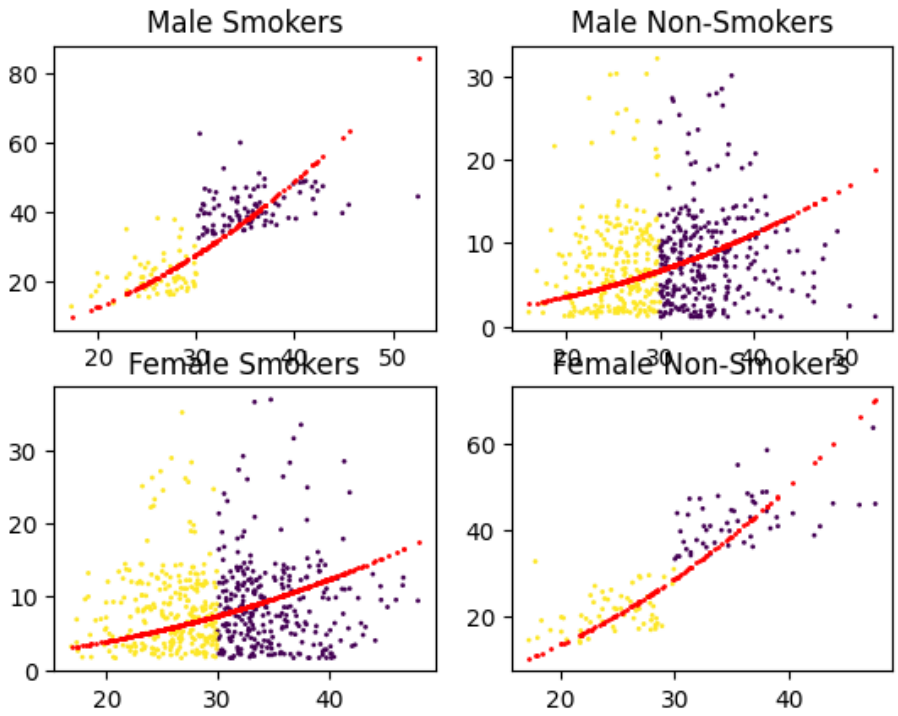
With this model, the observed average L^2 norm loss over the data is : **1197.69173132**
Now, we fit our SWAP regression model on the same data while treating the points having BMI¡30.0 with Z=0 and the other data points as having Z=1. For this, we get the following plot:
Note: We've scaled the insurance charges values by 1000 for the following plot. This scaling is due to the computational complexity of computing large numbers' squares and square roots. This scaling was considered while computing the model's loss value.



With this model, the observed average L^2 norm loss over the data is : **70.144267**

Just visually, we observe that this model is a much better fit for the problem at hand. This fact is further justified by the fact that the average loss for the model is 70.144267, which is over 17 times better than compared to the 2-line model illustrated above.

We can repeat the same process with our other categories and realize that the SWAP model is a robust fit for this problem. For this, we get the following plot:



8 Conclusion

In the analysis above, we've seen that for bi-variate data where neither of the variables can be considered non-stochastic, SWAP regression gives us a much better estimate fit than any other linear model. Additionally, we can get a single continuous model for all the data instead of multiple discontinuous models for various dataset partitions.

9 Future Aspects

In our SWAP method, we make a few assumptions:

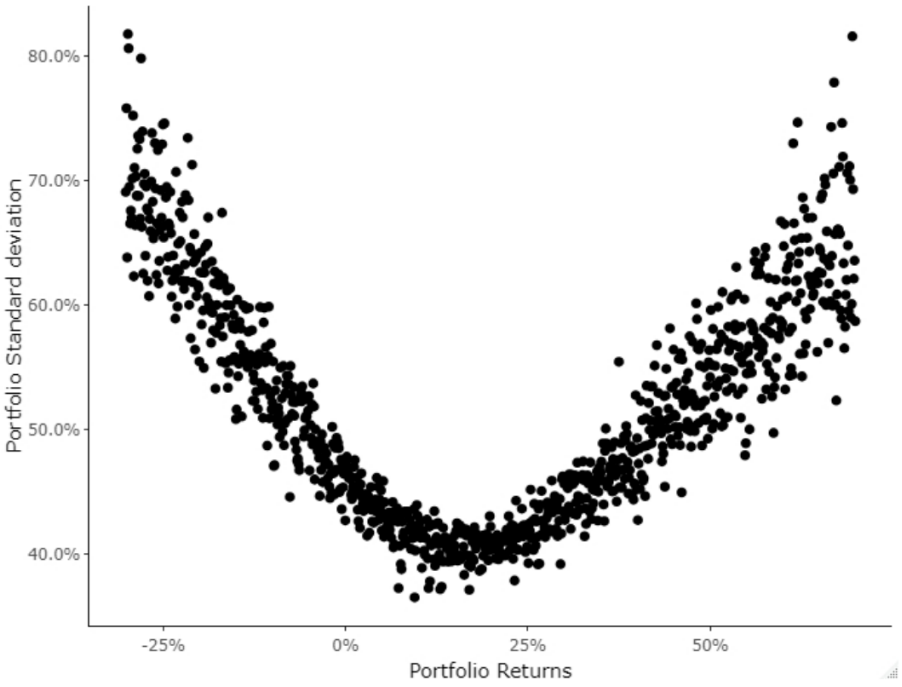
- The loss function is L^2:
  We know that L^2 norm is not robust for high deviation values. Hence, we can consider the L^1 norm loss function for data with high variance to obtain a more robust model.

- The model is quadratic:
  We need not consider the model always to be a quadratic one. The logic behind SWAP remains constant no matter the model function as long as

it is invertible. A small remark to consider is that we can have the model account for at most n trends in the data for an n degree model function.

- Multivariate Data:
  All of the arguments we've made in the SWAP model do not rely on the fact that X and Y are real numbers. We can adjust our model to accommodate both our variables being random vectors. The only constraint we face is that our underlying model should be an invertible function that transforms the vector space spanned by the range of X to that of Y.

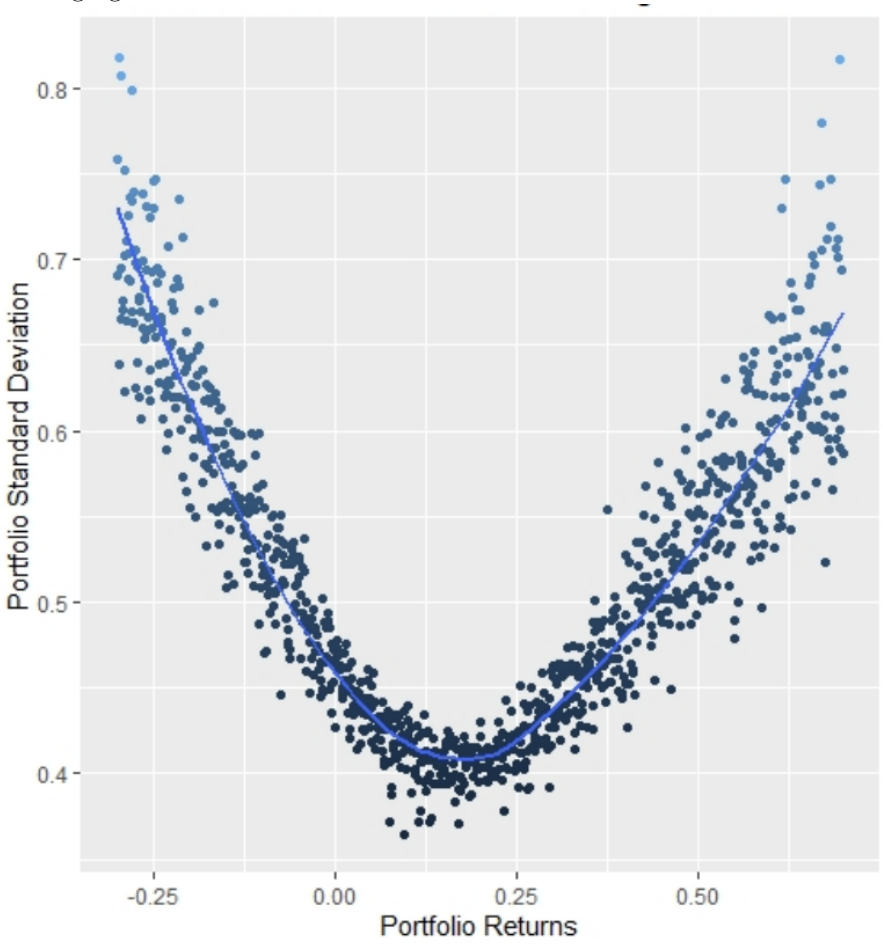
Besides the possible theoretical extensions to SWAP, I also explored a new application for the technique.
For any investment portfolio, we know that the following relationship is followed between the standard deviation in the amount of returns one can genereate from a portfolio. Commonly summed up by the sentence, "High risk, high rewards", we have the following relation:



Calculated for NIFTY100

Assuming that there are no costs for short selling, it is possible to obtain infinite returns and risks through portfolios. Hence, the horizon of the scatterplot was extremely vast due to exorbitant returns. I have restricted both the axis to 80% for the purpose of plotting the scatter plot.
With the first look of this plot it is evident that we can use our SWAP method to approximate the trend using a bijective function as can be seen done in the

following figure



What we can do with such a fit is start establishing potential connections between several portfolios or individual assets to determine their relative growth or decline. What this means is that we should be able to predict a potential connect between several assets that grow and deteriorate in relationship with each other.

A fundamental connection that we should be able to establish is that between **Gold-Oil-Dollar** prices and how they rise and fall relative to each other. Using this relationship as the base of comparison, we can start to figure out potential relationships between assets across several international markets as well.

The most important application of this analysis, if we are successfully able to establish one, is for us to solve the age-old problem of traders not being able to buy stocks in international markets. In the status quo, an equity trader in India cannot buy shares in Tesla. Hence, even if Tesla is a great hedging stock for his current portfolio to minimise the risk he has in his portfolio, he cannot help the situation. However, if we are able to establish a potential relationship

between Tesla stocks and a set of stocks available in the Indian market (say, Tata Electrics, Indian Capacitors and some other company) that have approximately the same growth and fall trends in relation to the portfolio that we want to hedge, our problem is solved.

10 Bibliography

This report is the re-implementation of:
ON REGRESSION FOR SAMPLES WITH ALTERNATING PREDICTORS AND ITS APPLICATION TO PSYCHROMETRIC CHARTS
by Mosuk Chow, Bing Li and Jackie Q. Xue
Source: Statistica Sinica, Vol. 25, No. 3 (July 2015), pp. 1045-1064