SWAP Regression: Full Report

Zehaan Naik

November 2024

Abstract

We propose a variation on the standard regression model called SWAP regression. Standard regression models work on the assumption that the regressor variables are non-stochastic. We propose a new regression where the predicted variable alternates between two correlated covariates linked by a bijective function. This method alleviates the assumption of non-stochasticity from the predictors and replaces it with a necessity for the covariates to be bi-directionally causal. The idea for this method was proposed in Chow et al. (2015), drawing motivation from an experiment in kinesiology.

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:

$$y = \beta X + \epsilon \tag{1}$$

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

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

Note: We use the assumption of gaussianity of errors later in Sec. 2.3.

The problem with this model lies in our fundamental assumption that the problem is classified as a strictly stochastic and non-stochastic variable. This assumption becomes hard to justify when the data need not clearly distinguish between the predictor and the predicted variable. We also consider data sets where the values of certain variables are hard or often impossible to directly observe in several situations. Either of these situations calls for a new regression method that does not require one of the variables to be non-stochastic for the entire range of the experiment.

As we will observe further in the report, this scenario is not very unlikely. Examples include kinesiology as illustrated in Chow et al. (2015) and several macro-economic indexes as illustrated in Kumari (2012). We observe that the standard assumptions of linear regression become hard to justify. In such situations, it's easier to check for a sense of bi-directional causality as a standard practice and work on a model based on this fundamental assumption.

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 The SWAP Method

The idea in Chow et al. (2015) is to consider an invertible function $f(x)$ such that:

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

and,

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

Here, X and Y are the two covariates under consideration, and Z is our latent variable that indicates which one of the two covariates is the regressor. Thus, when $Z = 0$, $E(Y \mid X, Z = 0)$ is of interest; when $Z = 1$, $E(X \mid Y, Z = 1)$ is of interest. Suppose that the sample space of (X, Y, Z) can be represented as the Cartesian product $\Omega_x \times \Omega_Y \times \{0, 1\}$.

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

$$\mathscr{G} := \{g \in L^2(P_x) \mid g \text{ is an injection from } \Omega_x \to \Omega_y \text{ and } g^{-1} \in L^2(P_x)\}. \tag{2}$$

Here, for any random variable U, particularly $U = X$, let P_U denote the distribution of U. Let $L_2(P_U)$ denote the class of functions square-integrable with respect to P_U. We want to minimize the quadratic loss function:

$$Q = E[(Y - g(X))^2 I(Z = 0)] + E[(X - g(Y))^2 I(Z = 1)]. \tag{3}$$

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,

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

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

$$Q(g_\theta) = E[\big(Y - aX^2 - bX - c\big)^2 I_{(Z=0)}] + E[\left(X + \frac{b}{2a} + \frac{\sqrt{b^2 - 4a(c - Y)}}{2a}\right)^2 I_{(Z=1)}]$$

Last question: How to find a value of θ to minimise Q_n?

In the next 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. We also note that Theorem 1. Chow et al. (2015) guarantees that g_{theta} would be unique almost surely.

2.1 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:

$$\begin{aligned}
\hat{Q_1} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c)^2 \mid Z = 0 \\
\hat{Q_2} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{b + \sqrt{b^2 - 4a(c - y_i)}}{2a})^2 \mid Z = 1 \\
\hat{Q} &= \hat{Q_1} + \hat{Q_2}
\end{aligned} \tag{4}$$

Remark 1. We can also show that the second derivative of the predictors is identically greater than 0. Hence, they will have a minimum.

$$\begin{aligned}
\frac{\partial \hat{Q_1}}{\partial a} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c)(-2x_i^2) & \frac{\partial^2 \hat{Q_1}}{\partial a^2} &= \frac{1}{n}\sum_{i=1}^{n}(2x_i^4) \\
\frac{\partial \hat{Q_1}}{\partial b} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c)(-2x_i) & \frac{\partial^2 \hat{Q_1}}{\partial b^2} &= \frac{1}{n}\sum_{i=1}^{n}(2x_i^2) \\
\frac{\partial \hat{Q_1}}{\partial c} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c)(-2) & \frac{\partial^2 \hat{Q_1}}{\partial c^2} &= \frac{1}{n}\sum_{i=1}^{n}(2) = 2
\end{aligned}$$

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:

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

where, l is an appropriate learning rate.

Remark 2. If we want to switch the loss function in Eq. 4 with:

$$\mathcal{L} = \sum_{i \in \mathbf{I}} I(Z_i = 0) \left| Y_i - g_\theta(X_i) \right| + I(Z_i = 1) \left| X_i - g_\theta^{-1}(Y_i) \right|. \tag{5}$$

We explore this idea in real data analysis in the last section of this project.

2.2 Breaking Point

Now, we need to establish a mechanism through which we make a division on the plane that gives us the two domains where CPI and SENSEX can take the role of being the predictors, respectively. One way to do this is through the Chow test, as explained in Lee (2008). Three problems that we observe with this approach are:

1. This method is very computationally intensive as it requires us to fit a regression model for each choice of the breaking point d, and its convergence is based on the assumption that the increments in d are not infinitely small.

2. The convergence is not stable for the L^2 model, and hence, we need to rely on some other method to figure out the breakpoint anyway.

3. The way the Chow test is structured forces us to have a hard decision boundary based on one of the coordinates in the data. For example, in our current data, the best Chow test can divide the data is all points such that $X < X_{\text{breakingpoint}}$. This means that we cannot have a cluster type of classification using this approach, something that the SWAP paradigm need not prohibit.

These three issues, coupled with the fact that the Chow test to determine the breakpoint is not problem agnostic, motivate us to consider an alternative method to find a natural breaking point.

2.3 Gaussian Mixture Model

We propose we use the Gaussian Mixture Model approach to cluster our data into 2 groups with a similar linear trend (i.e. classify $Z = 1$ or $Z = 0$ for all the observation points). To accomplish this, we use the standard GMM algorithm (Algo: 1) as it is characteristic of identifying linear patterns in data. A detailed explanation of this intuition comes from Ververidis and Kotropoulos (2008). The full code for this algorithm is available in the project repository at GMM.R.

The idea behind such clustering is to allow for the barrier between the two "regions" in our regression model to be not limited to one dimension (in our case, dependent on only one regressor). We also solve all the problems that we face with the Chow Test, as mentioned in Sec. 2.2.

3 Real Data Analysis

3.1 Motivation

In the SWAP model, as shown in Chow et al. (2015), we describe an intuition of two quantities being co-dependent on each other in a manner that their relationship can be modelled through a doubly stochastic model. This report aims to establish a sense of bi-directional causality between SENSEX and CPI in the Indian economy, which would be an example of such data.

The BSE SENSEX (the S&P Bombay Stock Exchange Sensitive Index or simply SENSEX) is a free-float market-weighted stock market index of 30 well-established and financially sound companies listed on the Bombay Stock Exchange. The 30 constituent companies, some of the largest and most actively traded stocks, represent various industrial sectors of the Indian economy. Recorded since 1 January 1986, the S&P BSE SENSEX is regarded as the pulse of the domestic stock markets in India.

A consumer price index (CPI) is the price index of a weighted average market basket of consumer goods and services purchased by households—changes in measured CPI track changes in prices over time. The CPI is calculated by using a representative basket of goods

Algorithm 1 Gaussian Mixture Model

Require: Data points (X_i, Y_i), for $i \in \{1, \ldots, N\}$, tolerance ϵ

Ensure: $Z_i \in \{0, 1\}$ for all $i \in \{1, \ldots, N\}$

1: Scale the data

2: Set μ_0 and μ_1 as random observations from the data

3: $\sigma_j \leftarrow I_2$ for $j = 1, 2$

4: weights $\leftarrow (0.5, 0.5)$

5: $Z_i \leftarrow 0$ for $i \in \{1, \ldots, N\}$

6: $ll \leftarrow 0$

7: **while** $\mid ll_{new} - ll \mid < \epsilon$ **do**

8: **E-Step:**

$$E(Z_i) = \pi_1^{(k)} \circ \text{dnorm}((X_i, Y_i), \mu_1^{(k)}, \sigma_1^{2(k)})$$

9: **M-Step:**

$$\begin{aligned}
\pi_1^{(k+1)} &\leftarrow \sum_{i \in \{1,\ldots,n\}} E(Z_i) \\
\pi_0^{(k+1)} &\leftarrow 1 - \pi_1 \\
\mu_0^{(k+1)} &\leftarrow \frac{1}{N\pi_0} \sum_{i \in \{1,\ldots,n\}} (1 - E(Z_i)) \circ Z_i \\
\mu_1^{(k+1)} &\leftarrow \frac{1}{N\pi_1} \sum_{i \in \{1,\ldots,n\}} E(Z_i) \circ Z_i \\
\sigma_0^{2(k+1)} &\leftarrow \frac{1}{N\pi_0} \sum_{i \in \{1,\ldots,n\}} (1 - E(Z_i)) \circ (Z_i - \mu_0^{(k+1)})^2 \\
\sigma_1^{2(k+1)} &\leftarrow \frac{1}{N\pi_1} \sum_{i \in \{1,\ldots,n\}} (1 - E(Z_i)) \circ (Z_i - \mu_1^{(k+1)})^2
\end{aligned}$$

10: **Set:**

$$\begin{aligned}
l_1 &\leftarrow \sum_{i \in \{1,\ldots,n\}} \log \left\{ \pi_0^{(k+1)} \left[\text{dnorm} \left((X_i, Y_i), \mu_0^{(k+1)}, \sigma_0^{2(k+1)} \right) \right] \right\} \\
l_2 &\leftarrow \sum_{i \in \{1,\ldots,n\}} \log \left\{ \pi_1^{(k+1)} \left[\text{dnorm} \left((X_i, Y_i), \mu_1^{(k+1)}, \sigma_1^{2(k+1)} \right) \right] \right\} \\
ll_{new} &\leftarrow l_1 + l_2
\end{aligned}$$

11: **end while**

12: Set $Z_i \leftarrow I(E(Z_i = 0) > 1)$ ▷ Assigning Clusters

and services. The basket is updated periodically to reflect changes in consumer spending habits. The prices of the goods and services in the basket are collected monthly from a sample of retail and service establishments. The prices are then adjusted for changes in quality or features. Changes in the CPI can be used to track inflation over time and to compare inflation rates between different countries.

From basic intuition, we can understand that SENSEX and CPI must be co-dependent. Extensive literature on the same claim can be found in many articles such as Kumari (2012), an empirical study of the same. The core fact that we base this intuition on is that the Consumer Price Index (CPI) is a measure of inflation, which can impact the stock market, including the BSE Sensex. Conversely, SENSEX is India's principal stock exchange, which affects the market and liquidity, ultimately affecting CPI.

3.2 Data

The data set that I use for this analysis is sourced from Kaggle.

We look at monthly data from October 2000 to August 2020. Table 1 contains the summaries for the two data sets.

Min.	36.73
1^{st} Quantile	45.19
Median	68.47
Mean	73.37
3^{rd} Quantile	101.37
Max	129.30

Min.	36.73
1^{st} Quantile	45.19
Median	68.47
Mean	73.37
3^{rd} Quantile	101.37
Max	129.30

Table 1: Summary for the Data (CPI and SENSEX respectively)

Along with their summaries, it also helps us to look at their time series plots and perform the Augmented Dickey-Fuller as described in Harris (1992) to test the hypothesis. The null hypothesis for this test claims that the time series under consideration has a unit root, implying that it is not stationary. Hence, we seek to find a small p-value to reject the hypothesis. Figure 1 and Table 2 demonstrate the plots and results for testing the hypothesis mentioned before.

Clearly, neither series is stationary.

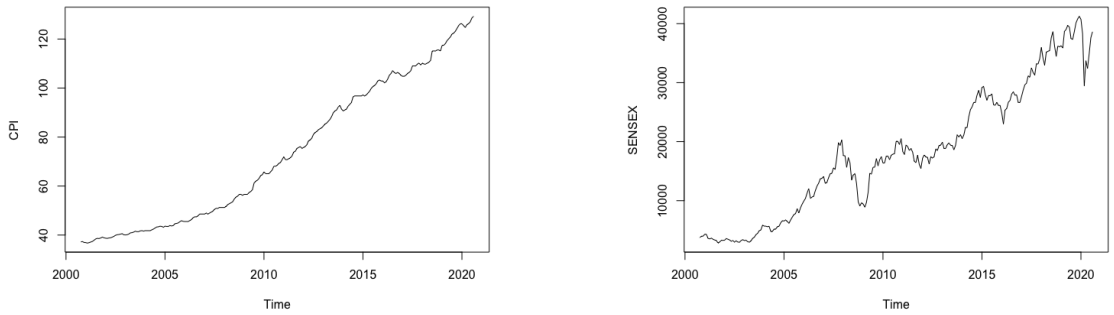

Figure 1: CPI and SENSEX Time Series Plot

CPI	0.5031
SENSEX	0.1628

Table 2: Augmented Dickey-Fuller Test p values

Finally, we conduct a simple Granger test on our two time series to see if they can be used as good predictors for each other. Figure 2 and Figure 3 give us the results.

We see that CPI can be a great predictor of SENSEX, but the opposite is not trivially true. This motivates us to consider a lagged causality model on the data set.

```
Granger causality test

Model 1: SENSEX.dat ~ Lags(SENSEX.dat, 1:1) + Lags(CPI.dat, 1:1)
Model 2: SENSEX.dat ~ Lags(SENSEX.dat, 1:1)
  Res.Df Df      F  Pr(>F)
1    235
2    236 -1 7.3841 0.00707 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2: Granger Test | CPI on SENSEX

3.3 Causality Analysis

Consider the following model:

$$Y \leftarrow \text{CPI}$$

```
Granger causality test

Model 1: CPI.dat ~ Lags(CPI.dat, 1:1) + Lags(SENSEX.dat, 1:1)
Model 2: CPI.dat ~ Lags(CPI.dat, 1:1)
  Res.Df Df      F Pr(>F)
1    235
2    236 -1 0.8949 0.3451
```

Figure 3: Granger Test | SENSEX on CPI

$$X \leftarrow \text{SENSEX}$$

Model-1:

$$Y_t = \alpha_0 X_t + \alpha_1 X_{t-1} + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + \epsilon$$

We observe that the p-value for the co-efficient of X in this model is **0.00203**.

Model-2:

$$X_t = \alpha_0 Y_t + \alpha_1 Y_{t-1} + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + \beta_4 X_{t-4} + \epsilon$$

We observe that the p-value for the co-efficient of Y in this model is **0.00110**.

Both these values suggest that we have **bi-directionally causal data** that fits well into the SWAP paradigm.

3.4 Residual Analysis

The last thing we wish to check to ensure that we are not getting a spurious fit is the stationarity of residuals. Figure 4. Table 3 gives us the P-values for the ADF test.

CPI	0.01
SENSEX	0.01

Table 3: Augmented Dickey-Fuller Test p values

Both residuals are clearly stationary, so the data is fit for a SWAP-type model.

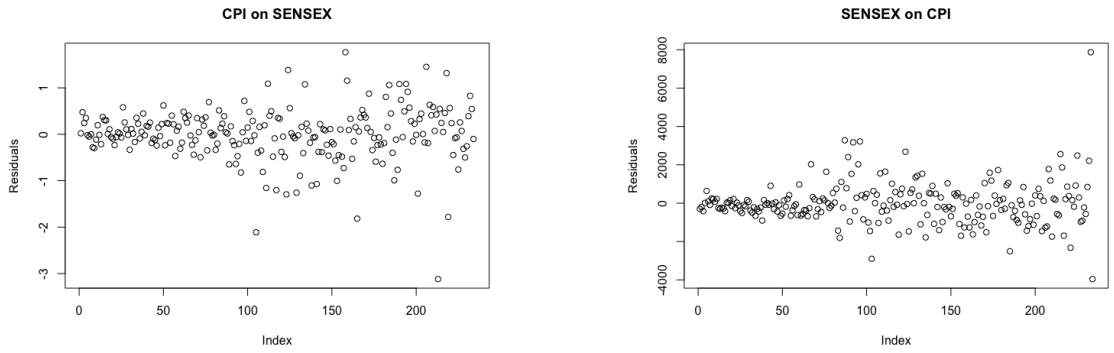

Figure 4: Residual Plots for Causal Model

3.5 Standard Linear Regression Model

With bi-directional causality established, we first aim to check if a SWAP model is even required for the data at hand. We consider a standard linear and a standard quadratic model to check if we get "good" fits.

Consider the model:

$$Y \leftarrow \text{SENSEX} \qquad\qquad X \leftarrow \text{CPI}$$

$$\text{Model: } Y = \beta_0 + \beta_1 X + \epsilon$$

The plot for the fitted data is shown in Figure 5.

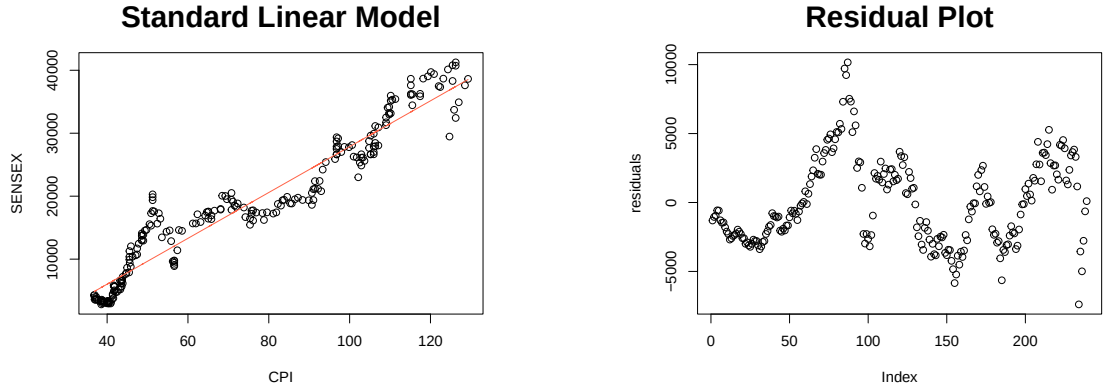

Figure 5: Standard LR model fit

Finally,

- The R^2 value for the model is **0.9222418**. This shows a decent fit for the model.
- The p value for the Augmented Dickey-Fuller Test is **0.3822**. This means that the residuals are not stationary.
- Sum of squared errors = **2252413332**
- Sum of absolute errors = **612724.3**

Now, consider the flip model:

$$Y \leftarrow \text{CPI} \qquad\qquad X \leftarrow \text{SENSEX}$$

$$\text{Model: } Y = \beta_0 + \beta_1 X + \epsilon$$

The plot for the fitted data is shown in Figure 6.

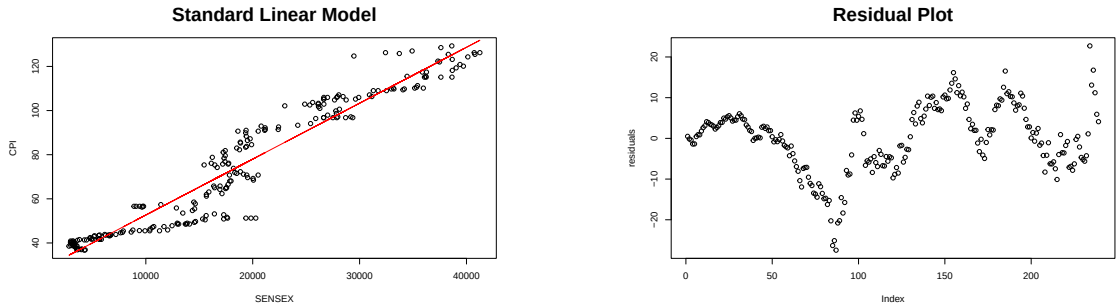

Figure 6: Standard LR model fit

Finally,

- The R^2 value for the model is **0.9222418**. This shows a decent fit for the model.
- The p value for the Augmented Dickey-Fuller Test is **0.4052**. This means that the residuals are not stationary.
- Sum of squared errors = **15678.07**
- Sum of absolute errors = **1510.16**

Note: These values are comparatively lower as the range of values of CPI is lesser than that for SENSEX.

3.6 Standard Quadratic Regression Model

Consider the model:

$$Y \leftarrow \text{SENSEX} \qquad\qquad X \leftarrow \text{CPI} \tag{6}$$

$$\text{Model: } Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

The plot for the fitted data is shown in Figure 7.

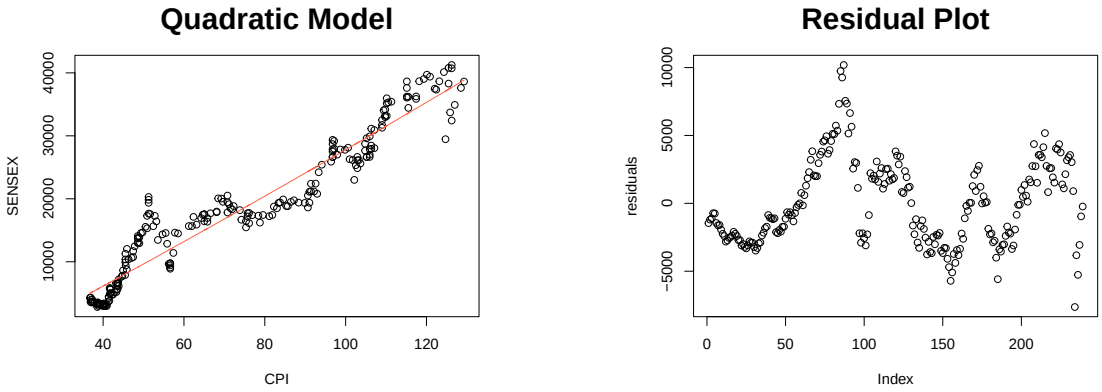

Figure 7: Quadratic LR model fit

Finally,

- The R^2 value for the model is **0.9223582**. This shows a decent fit for the model. It is slightly better than the standard LR case; however, it is not significant.
- The p value for the Augmented Dickey-Fuller Test is **0.3797**. This means the residuals are still not stationary, but we are progressing on the right track.
- Sum of squared errors = **2249041328**
- Sum of absolute errors = **614258.1**

Now, consider the flip model:

$$Y \leftarrow \text{CPI} \qquad\qquad X \leftarrow \text{SENSEX} \tag{7}$$

$$\text{Model: } Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

The plot for the fitted data is shown in Figure 8.

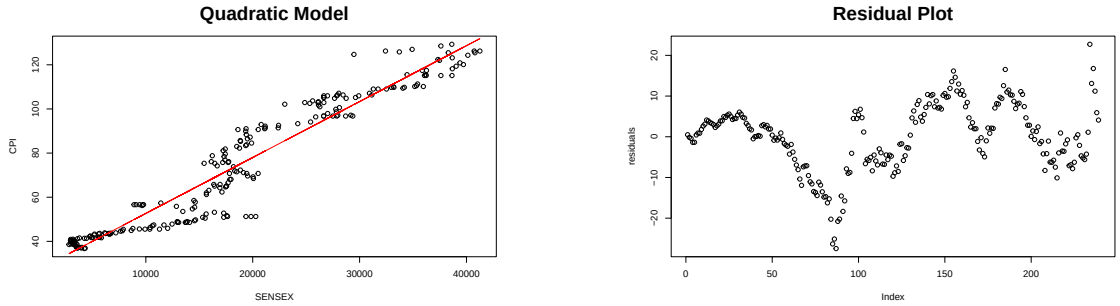

Figure 8: Quadratic LR model fit

Finally,

- The R^2 value for the model is **0.9222418**. This shows a decent fit for the model. It is slightly better than the standard LR case; however, it is not significant.
- The p value for the Augmented Dickey-Fuller Test is **0.4052**. This means the residuals are still not stationary, but we are progressing on the right track.
- Sum of squared errors = **15678.07**
- Sum of absolute errors = **1510.16**

We notice that the main problem with these models is that they do not give us stationary residuals. This motivates us to look at our SWAP alternative for the bi-directionally causal data set.

3.7 The SWAP Model

Now, we consider the SWAP model to our data and contrast it with a standard linear regression model and a standard quadratic model to see which gives us the best fit. We show a scatter plot of the data in Figure 9.

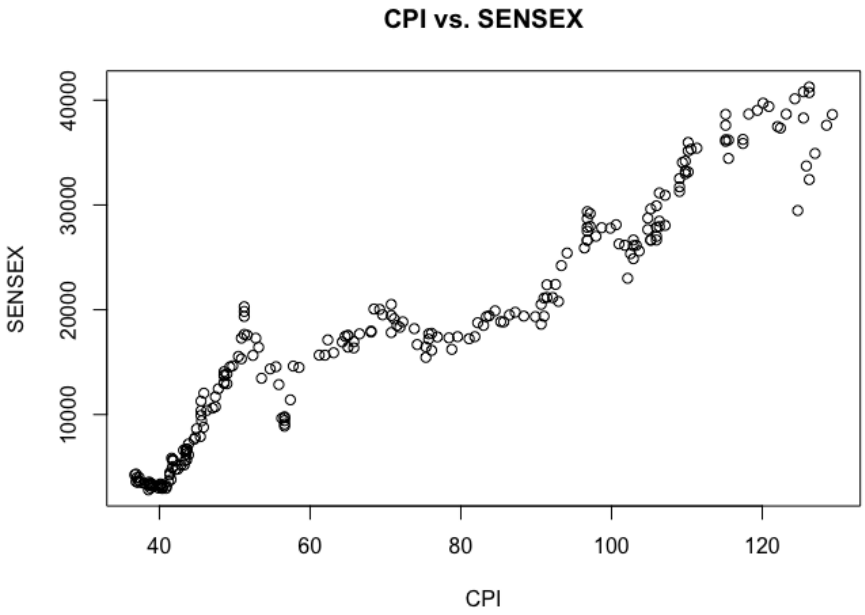

Figure 9: Scatter Plot for the full data

The first problem that we solve is to figure out a natural breaking point in the data that will tell us the $Z = 1$ and $Z = 0$ zones. We apply a Gaussian mixture model and the Chow test to the same and get the following results in Fig. 10. We observe that the results for GMM are better, as discussed in Sec. 2.2; we use the same breaking point for the rest of our analysis.

We use a Gaussian Mixture Model with Mahalanobis distance to ensure that we can capture the two separate linear trends in the data. This allows us to come to a natural breaking point for the SWAP paradigm and give us good fits.

Consider the model:

$$Y \leftarrow \text{SENSEX} \qquad X \leftarrow \text{CPI}$$

$$g(X) = aX^2 + bX + c \qquad g^{-1}(Y) = \frac{-b \pm \sqrt{b^2 - 4a(c-y)}}{2a}$$

3.7.1 L1 optimisation

The plot for the fitted data is shown in Figure 11.

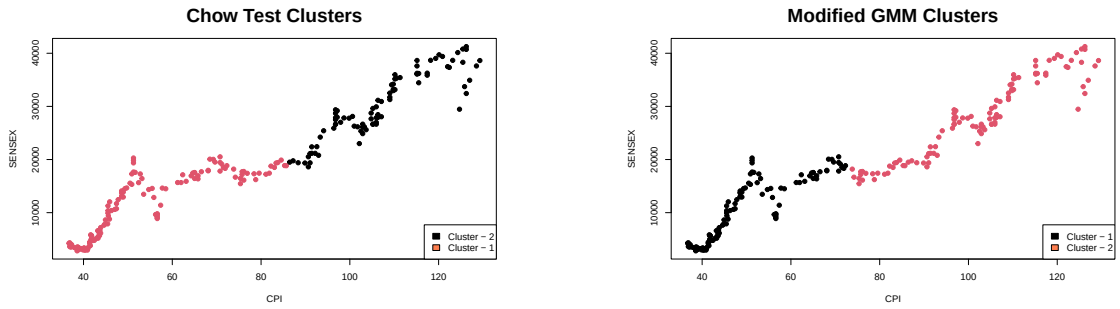

Figure 10: Breaking Point Analysis

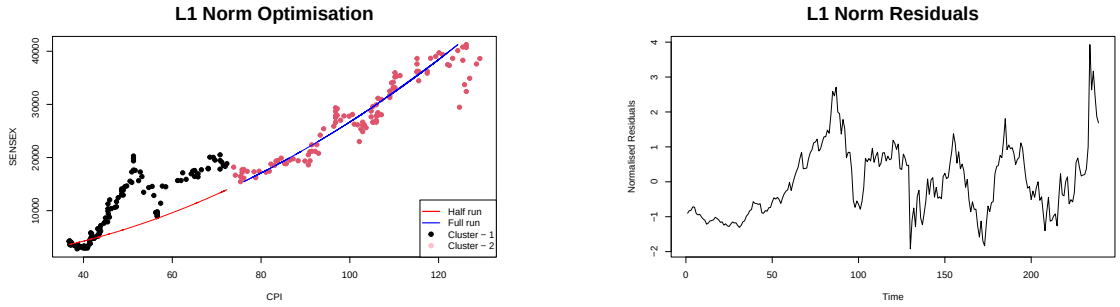

Figure 11: SWAP Regression with L1 norm optimisation

Finally,

- The R^2 value for the model is **0.8665867**. This shows a decent fit for the model.
- The p value for the Augmented Dickey-Fuller Test is **0.6076**. This means that the residuals are not stationary.

3.7.2 L2 optimisation

The plot for the fitted data is shown in Figure 12.

Finally,

- The R^2 value for the model is **0.8665867**. This shows a decent fit for the model.
- The p value for the Augmented Dickey-Fuller Test is **0.6076**. This means that the residuals are not stationary.

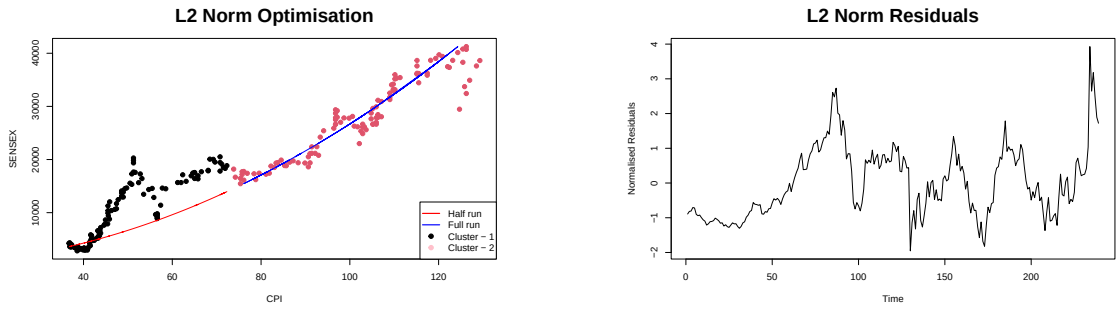

Figure 12: SWAP Regression with L2 norm optimisation

3.8 Comparing the Loss Functions

Now that we have two fits, we wish to compare the two residuals on some standard tests for regression. We compare the models based on the normalised residual errors:

1. **Squared Errors:**
   - L2 norm: 237
   - L1 norm: 237
1. **Absolute Errors:**
   - L2 norm: 197.4567
   - L1 norm: 197.9957

As we can observe, the two models perform extremely closely on our metrics, and hence, we can see that SWAP is a stable fit.

References

Chow, M., Li, B., and Xue, J. Q. (2015). On regression for samples with alternating predictors and its application to psychrometric charts. *Statistica Sinica*, 25(3):1045–1064.

Harris, R. (1992). Testing for unit roots using the augmented dickey-fuller test: Some issues relating to the size, power and the lag structure of the test. *Economics Letters*, 38(4):381–386.

Kumari (2012). Stock returns and inflation in india: An empirical analysis. *The IUP Journal of Monetary Economics*, IX:39–75.

Lee, H. (2008). Using the chow test to analyze regression discontinuities. *Tutorials in Quantitative Methods for Psychology*, 4.

Ververidis, D. and Kotropoulos, C. (2008). Gaussian mixture modeling by exploiting the mahalanobis distance. *IEEE Transactions on Signal Processing*, 56(7):2797–2811.