# Code Assignment01

1. We chose to look at the S&P 500 daily price dataset recorded from 1986 to 2018 [[link](https://www.kaggle.com/datasets/pdquant/sp500-daily-19862018)]. The dataset only contains daily unadjusted close prices (in dollars) and corresponding dates.

Here is how the head of the dataset looks like:

Text

Description automatically generated

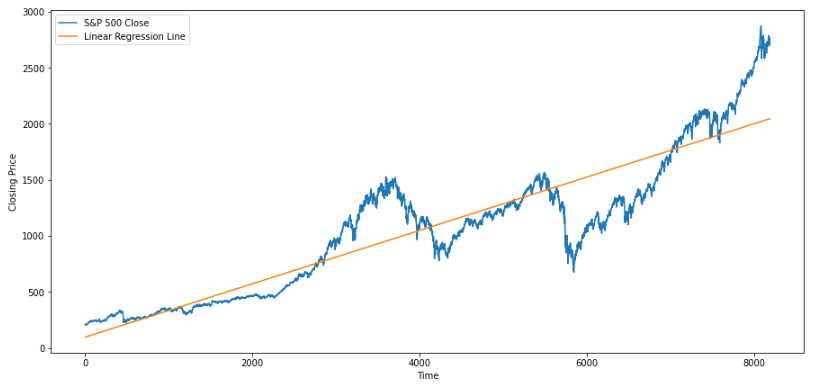
Chart, line chart

Description automatically generatedHere is a graphic representation:

We chose to look at this dataset because we know that S&P 500 index can approximately tell us how the US economy is doing and thus analyzing it also helps us understand how that economy moved and might move in the future.

1. We will fit three types of models to our dataset: linear model, quadratic model, and a more flexible model with sine and cosine features. Then, based on (Adj.) R-adjusted and MSE, we will decide which model to pick for later residual analysis.

**Chart, line chart

Description automatically generatedLinear model**: R-squared = 0.832 | MSE = 63875.54

Chart, line chart

Description automatically generated**Quadratic Model**: Adj. R-squared = 0.843 | MSE = 59693.88

**Curved Regression Model:** Adj. R-squared = 0.871 | MSE = 48875.21

We can see that the last model (Curved Regression model) best describes our data compared to other estimated models since it has the highest Adj. R-squared and lowest MSE. So, we will do residual analysis on that model below.

1. **Residual Analysis**

Graphical user interface, chart, line chart

Description automatically generatedHere is the graph of standardized residuals of our curved regression model:

Chart, line chart

Description automatically generatedWe can see that the distribution of residuals clearly does not resemble that of white noise since consecutive residuals seem to be highly correlated with each other.

Chart, histogram

Description automatically generated

By looking at the above histogram and Q-Q plot, we can say that residuals are not normally distributed (although the distribution is symmetric around the mean) since occurrence of residuals above one standard deviation from the mean is much higher than expected.

The assumption of independence can be checked by looking at the correlogram shown below:

Chart

Description automatically generated

We can see that the autocorrelation for all shown lags is out of range of standard errors, and thus the residuals can not be seen as white noise.

So, our estimated trend model generates residuals that violate the assumptions of normality and independence, and thus that trend model might not be considered as a correct model.