GROUP WORK PROJECT # 1
GROUP NUMBER: 3997

MScFE 610: FINANCIAL ECONOMETRICS

FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Bharat Swami	India	bharatswami1299@gmail.com	
Ka Man Lui	United Kingdom	thomaslui.0924@gmail.com	
Daxin Niu	United States	daxinniu.work@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

Team member 1	Bharat Swami
Team member 2	Ka Man Lui
Team member 3	Daxin Niu

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
Note: You may be required to provide proof of your outreach to non-contributing members upon request.

Step 1

Skewness

Definition: One of the popular definitions of skewness is the one below:

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^3/n}{s_X^3}$$

Where $\overline{x}$ is the mean of X and s_X is the standard deviation of X

Sensitivity to outliers

Definition: To understand the sensitivity to outliers, we can use Cook's distance below to identify whether the outliers are influential

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{y}_j - \widehat{y}_{j(i)}\right)^2}{ps^2}$$

where $\widehat{y}_{j(i)}$ is the fitted value obtained when excluding observation i, p is the rank of the model and s^2 is the mean squared error. A data point with a Cook's D larger than 1 is sometime considered an influential point which has large impact to the regression model.

Over-reliance on the Gaussian Distribution

Definition: To understand the issue of over-reliance on Gaussian distribution, we must first understand what a Gaussian distribution is. By definition, Gaussian distribution is a probability distribution that is symmetric on the mean with a bell-shaped curve. It is usually written with the following function:
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ where μ is the mean and σ is the standard deviation.

Kurtosis / Heteroscedasticity

Kurtosis Definition: It is a property of a distribution which represents the shape of tail and peak height in the distribution. In more mathematical inclined definition, Kurtosis is fourth moment of the distribution, which is given by

$$Kurtosis = \frac{Fourth\ Moment}{Standard\ Deviationn^4} = \sum_{i=0}^{n} \frac{(X_i - \mu)^4/N}{\sigma^4}$$

where μ is mean of distribution with random variable X_i, N is number of observation of random variable and σ is standard deviation of distribution.

Heteroscedasticity Definition: If the variance of residual terms or error terms is not equal over the dataset then we can say there is heteroscedasticity present. In easier terms, if the shape of residual terms is cone or fan then we can say that heteroscedasticity is present (Adam Hayes).

Step 2

Skewness

Description: Skewness is the measure of the asymmetry of the probability distribution about its mean.

Demonstration: 1000 Data with skewness is generated using the skew-normal density function (Azzalini and Capitanio 6) below:

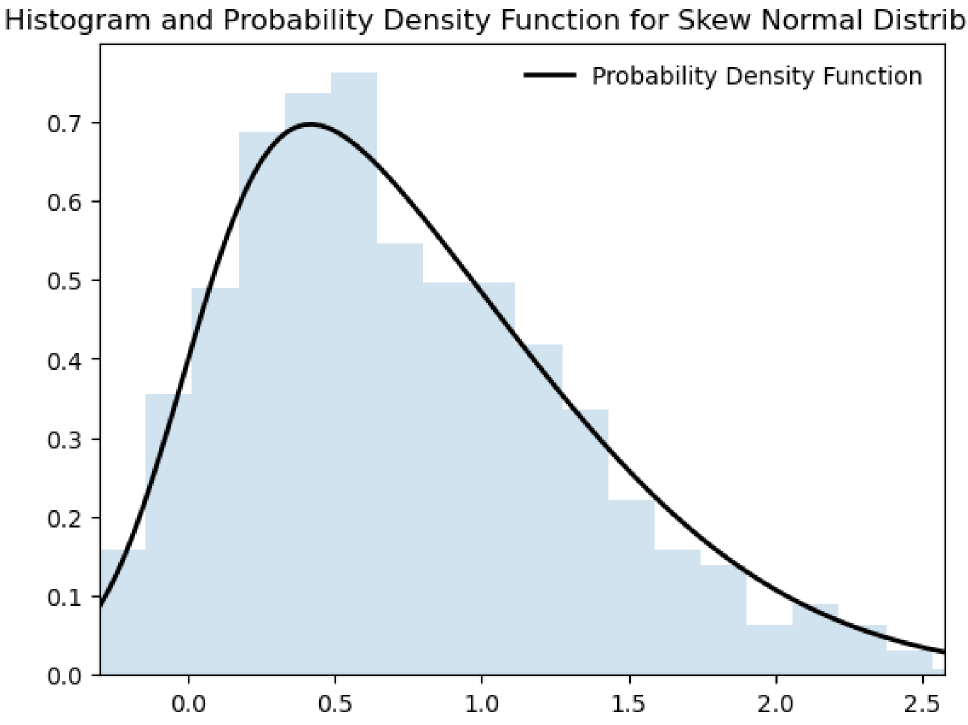
$$f(x) = 2\phi(X)\Phi(\alpha x)$$

where $\phi(x)$ is the standard normal density function and $\Phi(\alpha x)$ is the standard normal (cumulative) distribution function. α is a shape parameter.

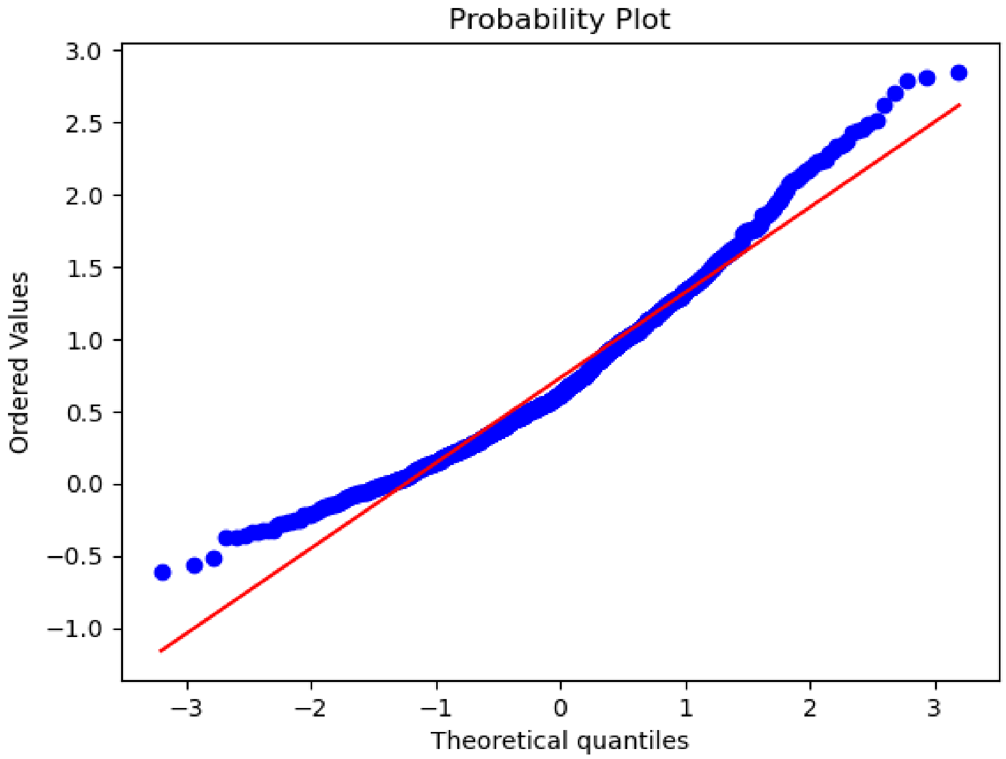
The statistical measurement of the simulated data is shown below:

Mean: 0.7293, Variance: 0.3585, Skewness: 0.7005, Kurtosis: 0.301

Diagram:



QQ Plot



Diagnosis: Histogram can be used to assess the skewness of dataset. A symmetric distribution has its peak at the center of the distribution. A right-skewed distribution has the long tail on the right side of the distribution and a left-skewed distribution has the long tail on the left side of the distribution.

We can also plot the QQ plot to assess the skewness. The right-skewed data appears as a concave upward curve in the QQ plot. The left-skewed data appears as a concave downward curve in the QQ plot.

Skewness can be computed using the formula with the aid of the statistical package to quantify the skewness effect.

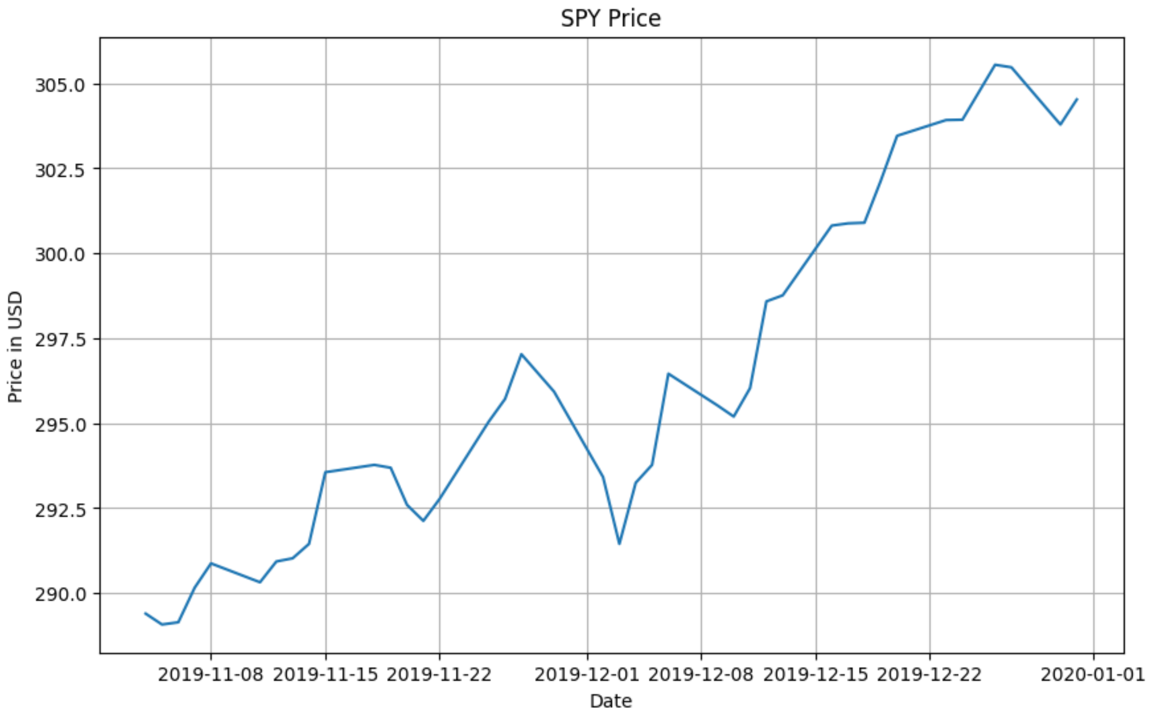
Damage: A lot of statistical methods assume the underlying distribution to be approximately normal. The results can be inaccurate or even wrong if the skewed data is fed into the statistical methods with assumption that the distribution is normal.

Over-reliance on the Gaussian Distribution

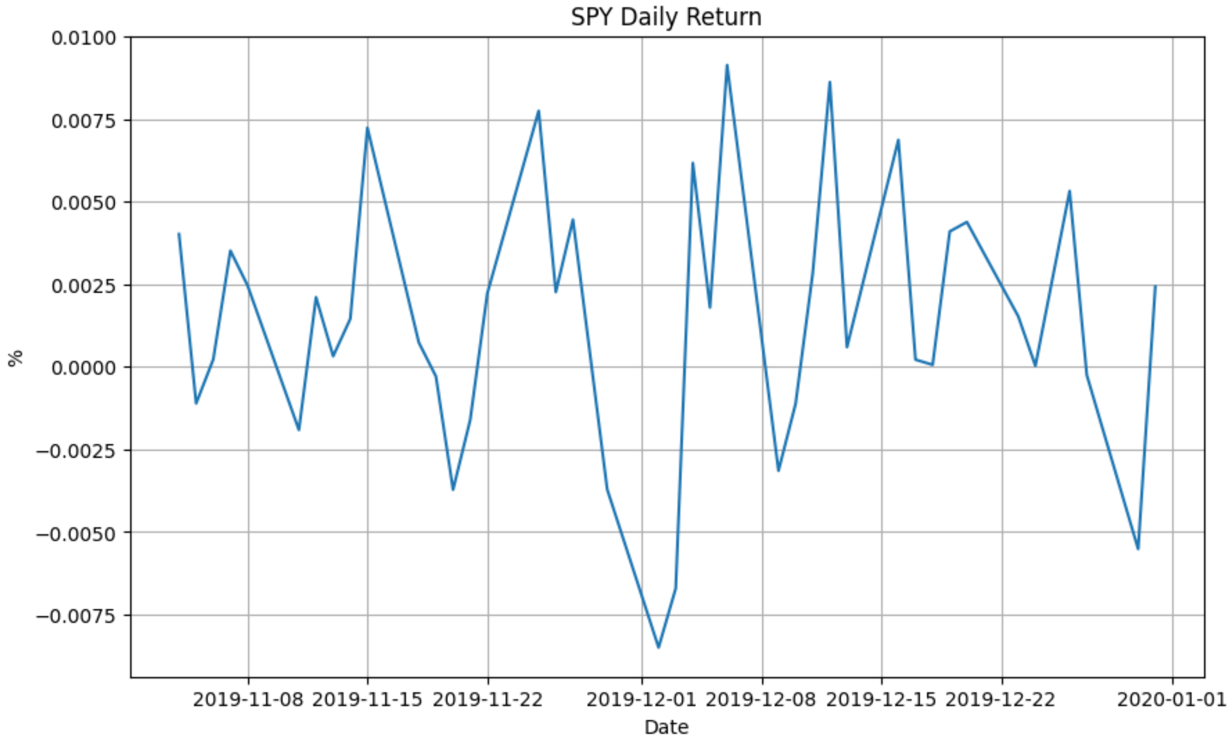
Description: Over reliance on Gaussian distribution could ignore extreme events and suffer from it. It also has systematic risk when the underlying distribution changes but the model still makes the assumption of Gaussian distribution.

Demonstration: To illustrate this example, we purposely chose a small interval of the SPY data for some testing. We chose SPY data from 11/01/2019 to 01/01/2020 which is a very short timeframe. From the short time period, we have plotted the daily price. Furthermore, we have calculated the daily return so that we can run some test for Gaussian distribution. After generating the return data, we decided to run Shapiro test to see whether our daily return is normally distributed. From the test, we received a p-value of 0.826 which indicates that we failed to reject the null hypothesis that the data is normally distributed. With the data collected above and the test we ran, one might think this is efficient for us to use the Gaussian assumption and make strategy out of it.

Diagram:
Given the above data, we have plotted two diagrams. First one below is the price of SPY in the corresponding timeframe.



The image can help us to better understand the movement of the price. The next one is the daily return plot.



From the above plot, we can see the daily return in percentage.

Diagnosis: In this section, we can dive into the diagnosis. Although our data passed the Shapiro test, the statement doesn't hold true forever. If we simply expand our timeframe by one month and look at data from 11/01/2019 to 02/01/2020, we can run the Shapiro test again but we get a p-value of 0.0024. In this case, it is smaller than the 0.05 threshold and we have to reject the null hypothesis that the daily return of SPY is normally distributed.

Damage: Suppose we have a strategy that relies on the assumption that SPY's daily return is gaussian distributed, it could work within the first timeframe we picked since during that specific timeframe, our return data are indeed normally distributed. Nonetheless, the strategy would stop working when time goes on because even if we just extend the timeframe by one more month, our daily return data is no longer normally distributed. If we stick to the assumption that does not holds true anymore, the strategy will not only not make money, but also have potential to loss a lot of money for us. The damage could be devastating.

Kurtosis / Heteroscedasticity

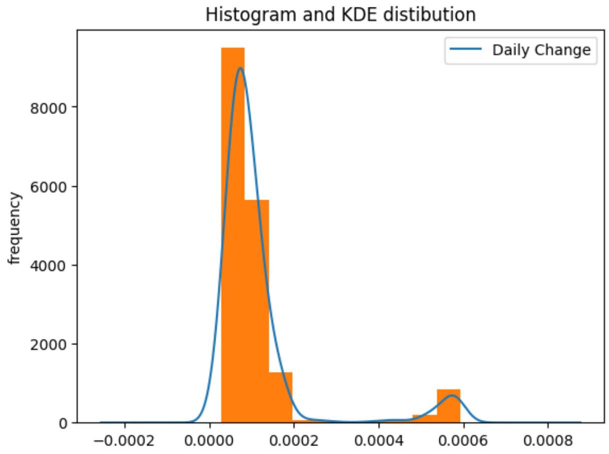
Description :

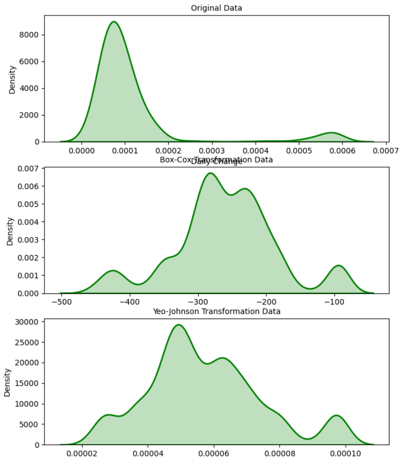
- Kurtosis of normal distribution is 3. If the kurtosis of a distribution is lower than 3 then we can say that distribution has a thin tail as compared to normal distribution while if kurtosis is greater than 3 then tail is fatter than normal distribution tail.
- heteroskedasticity says that there are unequal variances in residuals of a predicted model and original value. It can be detected by complicated mathematical models and calculation or roughly it can be seen in graph of residual distribution, if its cone or fan shape then we can say there is heteroskedasticity.

Demonstration: We are taking the adjusted close of NSE NIFTY50 from *"2012-02-01"*to*"2022-02-01"*. We are going to transform the data into rolling variance with a window size of *150 days* . First, we are going to convert the adjusted close to daily percentage adjusted close and then we treat the values as time series. And, then we are going to apply a rolling window to convert the dataset in variance of that window size.

After transforming the dataset in the desired way, we will plot the histogram with KDE (Kernel density estimation) distribution (Kernel density estimation). Then we plot the QQ plot and easily see that our data is not normally distributed. After that for confirmation we use two types of test to check whether our data is normally distributed or not. First is the Jarque–Bera test which results in p-value 0, and because of that we drop the null hypothesis that our data is normally distributed. Second test is the Shapiro–Wilk Test and which also results in p-value 0, making us drop the null hypothesis. This proved our data is not normally distributed and we have Kurtosis and skewness present.

Diagrams





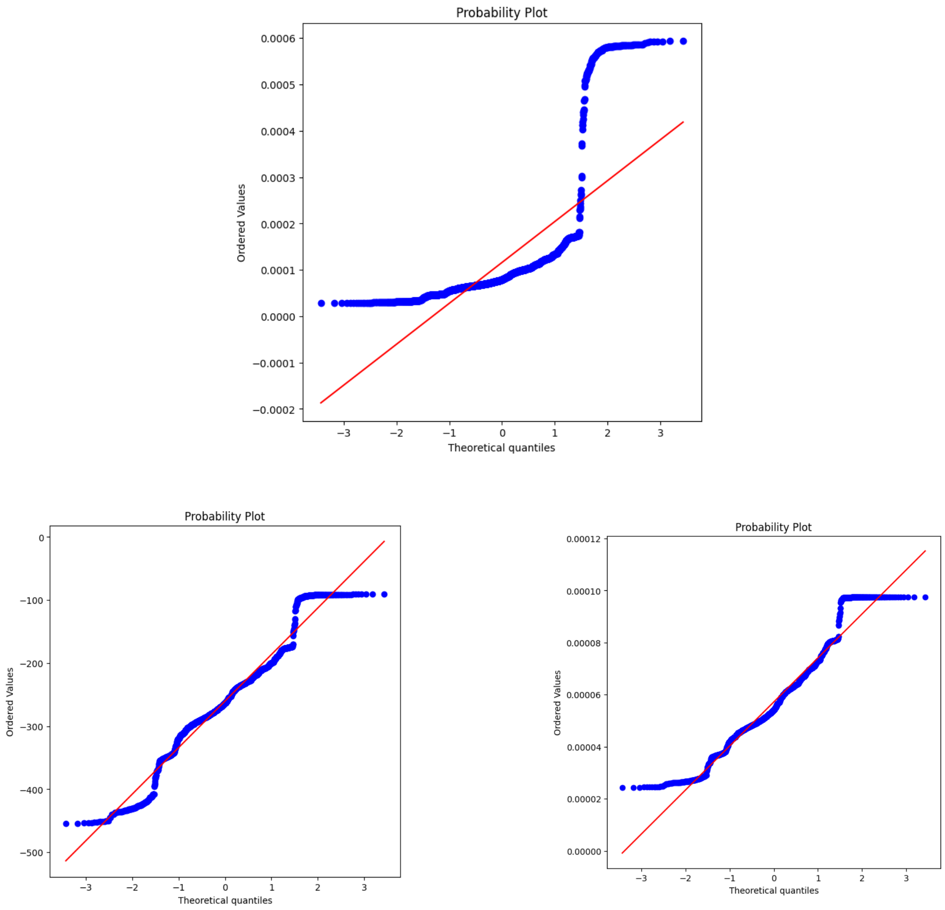


Figure : a) Histogram and KDE plot of variance of data with rolling window of 150 days, b) QQ plot of original data, c) Three subplots with original data, box-cox transformed data and yeo-johnson transformed data, d) and e) QQ plots for box-cox data and yeo-johnson data

For calculation, Excess Kurtosis of original data is **8.6625** (for normal distribution excess kurtosis is 0) and from different tests we get p-value 0 which means our data is not normally distributed.

After box-cox transformation and yeo-johnson transformation (Tamil Selvan S) we get p-values greater than 0.05 which mean we can accept the null hypothesis that our data is normally distributed with excess kurtosis **0.5443** and **0.0763,** respectively. Which means we transformed our data to nearly normal.

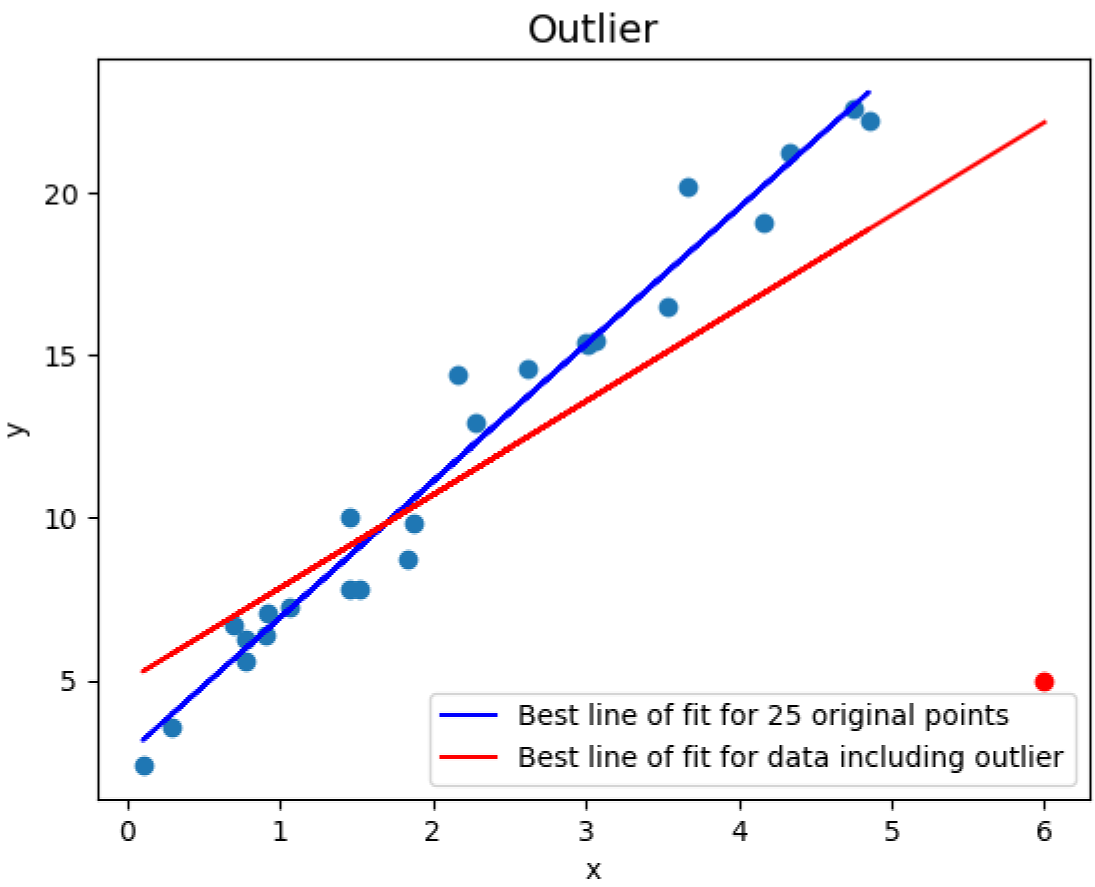
<u>Damage</u>: As we all know most of our models for calculating the price of stock, options or any derivatives are based on normal distribution, and it's also earlier to calculate things if our data is normally distributed. If we don't check the distribution of our data and apply these models to them, we may get incorrect or highly variable values which in the case of investment make our PnL very bad. And in the worst case we may lose most of the investment.

Sensitivity to outliers

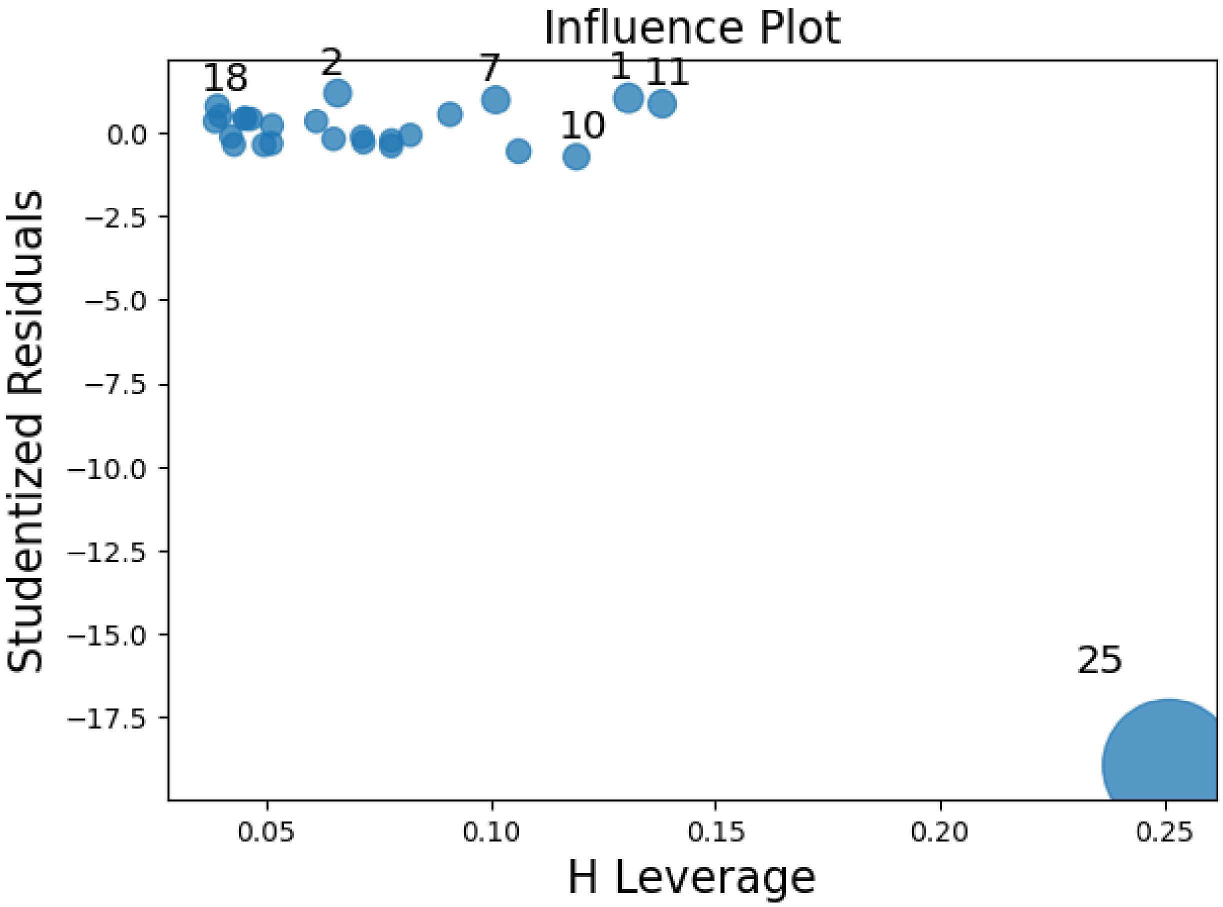
Description: Sensitivity to outliers is the impact of the outlier to the regression model

Demonstration: 25 random points are generated with the fitted line y = 4x + 3 with some random noise from N(0,1) added to the independent variable y and also one outlier point (6,5)

Diagram:



Influence Plot



Diagnosis: From the graph, it is clear to see that the point (6,5) is the outlier. We can run the regression analysis without the outlier and one with the outlier to compare the difference.

For the regression analysis without the outlier, the result is y = 4.2x + 2.7. For the one with outlier, the result is y = 2.9x + 5.0. The coefficient estimate for the regression has changed a lot.

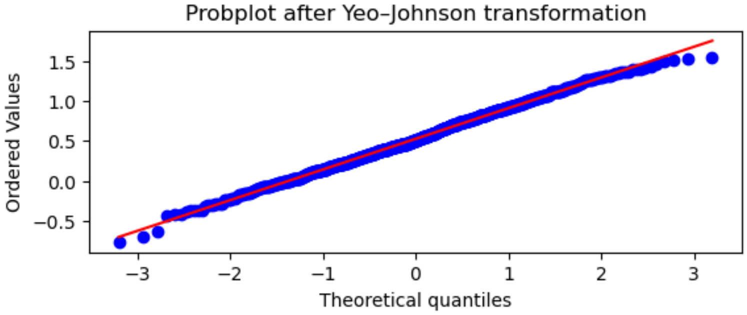
Also, we can calculate the Cook's distance to check the impact of each point to the regression point. From the Cook's distance graph above, we can see that the point 25 which is point (6,5) has a really large residual. Also, the Cook's distance of that particular point is 3.78 which is greater than 1. So we can consider that as a influential point.

Damage: Outliers and Influential points have impact to the regression model and they may be erroneous data or genuine data with valuable information. They may distort the regression result with inaccurate or wrong coefficient estimates. Outliers and influential points need to carefully examined to assess the impact to the model.

Step 3

Skewness

Directions: Variable transformation can be used to transform the skewed data to normal. Examples of the transformation are log transformation, square root transformation, Box and Cox transformation and Yeo–Johnson transformation. Yeo–Johnson transformation was chosen to transform the simulated data.



From the QQ plot, the transformed variable becomes approximately normal and it can be used in different statistical methods.

If the variable transformation is undesirable and we want to retain the original data, skew-normal distribution and skew-t distribution can be used to directly handle skewed data directly for analysis according to Azzalini and Capitanio's book (Azzalini and Capitanio).

Over-reliance on the Gaussian Distribution

Directions: To prevent these issue from happening, one possible direction is to check assumption frequently. If the team decided to check the distribution of daily return very frequently, they might have some loss at the beginning when the data are no longer normally distributed, but they can stop the strategy early and cut losses.

Sensitivity to outliers

Directions: Cook's Distance can be used to assess the whether there's any outlier or influential points in the data set for the regression model. If outliers or influential points are found and they are erroneous data, it can be removed to make sure the regression result is not distorted. If we confirm we can't remove the outliers/influential points, we can use the weighted least square in the regression model to reduce their impact on the regression result.

Kurtosis / Heteroscedasticity

Directions: First of all, we should check the distribution of the dataset as a rule of thumb and check every time we are modeling. And if the dataset is not normally distributed we should transform the data set into one. We can use log transformation, square root transformation, box-cox transformation,Yeo-Johnson Transformation, etc.

*** For calculation please refer to python code.*

References

1) Hayes,Adam. “Heteroscedasticity Definition: Simple Meaning and Types Explained”,2022, https://www.investopedia.com/terms/h/heteroskedasticity.asp#:~:text=Investopedia%20%2F%20Joules%20Garcia-,What%20Is%20Heteroskedasticity%3F,periods%2C%20are%20non%2Dconstant.
2) A. Azzalini and A. Capitanio (1999). Statistical applications of the multivariate skew-normal distribution. J. Roy. Statist. Soc., B 61, 579-602. https://arxiv.org/abs/0911.2093
3) Kernel density estimation, “Wikimedia Foundation”,2023, https://en.wikipedia.org/wiki/Kernel_density_estimation
4) Selvan S, Tami. “Types Of Transformations For Better Normal Distribution”, Towards Data Science, 2020 https://towardsdatascience.com/types-of-transformations-for-better-normal-distribution-61c22668d3b9
5) Adelchi Azzalini, and Antonella Capitanio. The Skew-Normal and Related Families. Cambridge University Press, IMS Monographs Series, 2014.