Algorithmic Trading & Quantitative Strategies
Homework #2: Impact Model

Ju Hyung Kang

jk8448

1. Introduction

In this paper and the script files, 'the paper' indicates Direct Estimation of Equity Market Impact by Robert Almgren. Likewise, 'the model' indicates the model in the paper. 'The lecture note' refers to the lecture note named Building Our Own Impact Model given by Prof. Maclin.

2. Model Clarification

In the paper, the model is given as

$$h = J - \frac{I}{2} = \sigma\eta\ sgn(X)\left|\frac{X}{VT}\right|^{\beta} + < noise >$$

where

$$I = \frac{S_{post} - S_0}{S_0}, J = \frac{\tilde{S} - S_0}{S_0}$$

According to the lecture note, σ is an average 1-day volatilities which is scaled from the average of 2-minute volatilities, $sgn()$ is a function that returns 1 if positive and -1 if negative, X is an average of imbalance values from 9:30 to 3:30, V is an average of total daily trading values, T is $\frac{6}{6.5}$, S_{post} is terminal price, S_0 is arrival price, and $\tilde{S}$ is vwap from 9:30 to 3:30.

Note that the absolute value of $\frac{X}{VT}$ is used to avoid numerical error when X is negative. To consider the sign of X, $sgn()$ function is used.

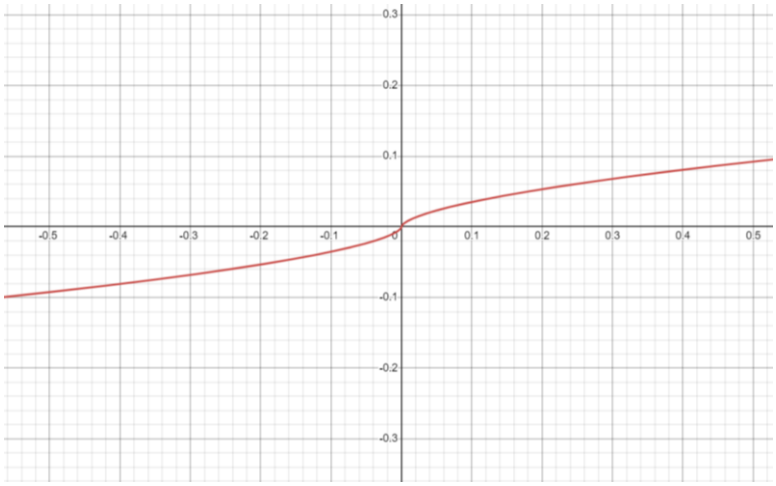
The model suggests that

$$\frac{h}{\sigma} = \eta sgn(X)\left|\frac{X}{VT}\right|^{\beta}$$

Put

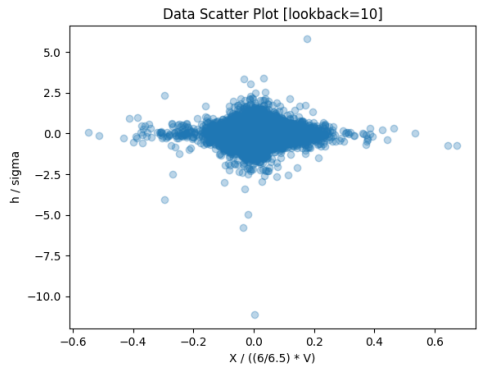
$$y = \frac{h}{\sigma}, x = \frac{X}{VT}$$

then, the graph should look as follows with $\eta = 0.14, \beta = 0.6$:

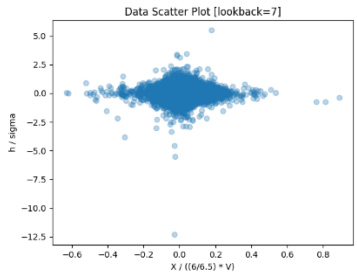


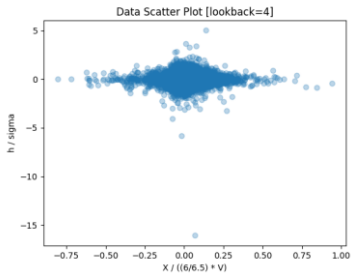
3. Data Exploration

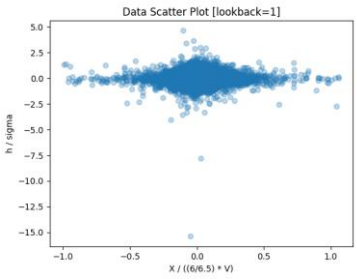
Using lookback period of 10 days to calculate average volatility, imbalance, and value and filtering out top 5% volatile days, the plot of data is given as follows:



The plot implies that h/sigma and X/V do not have any visible pattern. The data under other lookback periods do not showcase any visible pattern.







4. Directories

There are 7 directories for this Pycharm project: data, evaluation, impactUtils, model, preprocess, taq, and tests. In ‘data’ folder, there are mainly preprocessed matrices data, quotes data, and trades data. In ‘evaluation’ folder, there are codes that evaluate the impact model. In ‘impactUtils’ folder, there are classes that calculates ingredients for the model such as arrival prices, imbalance values, and vwap. These classes are used for preprocessing. In the ‘model’ folder, there are classes that build and fit the impact model using the preprocessed matrices. In ‘taq’ folder,

there are mainly quotes reader and trades reader. In 'test' folder, there are unit test codes for all classes and methods in the project. The parametric bootstrap test is also included in this folder.

5. Model Details

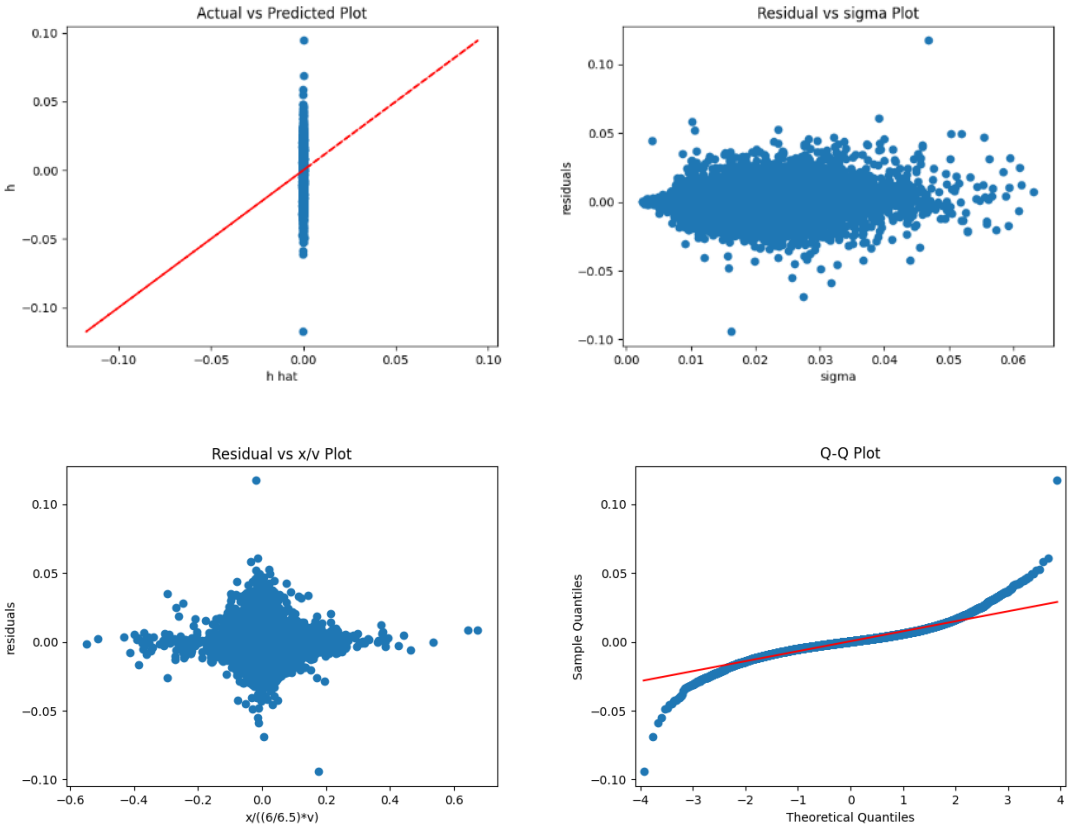
- The boundary of β is set as $0 < \beta < 1$ according to the paper (p. 17).
- The paper tries to find the model parameters that are applicable across universal stocks using cross sectional regression approach. Since it is unreasonable and unpractical to devise models for each day, all dates are treated equally, which collapses the date dimension of the data.
- The data are dropped if $\left|\frac{X}{\frac{6}{6.5}V}\right|$ is smaller than certain value (e.g., 0.0025%) to align with the assumption in the paper (p. 8).
- It is possible to filter out outliers based on h values, but outliers are not filtered for the models in this paper.

6. Model Evaluation

The summary statistics for a base model are as follows:

Hyperparameters		
Lookback	**Volatility Filter**	**Order Filter**
10 days	95 percentiles	0.025%

Coefficients	**Value**	**p-value**
η	0.00904	0.349546
β	0.0001	0.9989439
r-squared	-0.003	
White's test p-value	8.315644813087616e-299	



It is observable that the model does not fit well. The r squared value is negative which indicates that the model is inferior than the mean of dependent variables in explanation power. The p-values are not significant, especially with β. The White's general test suggests that the residuals are highly likely to be heteroskedastic. The variance of the residuals increase as sigma increases, and then decays slightly afterward. The variance of the residuals is high around x/v=0. The normality assumption for residuals is relatively well preserved upon looking at the QQ plot.

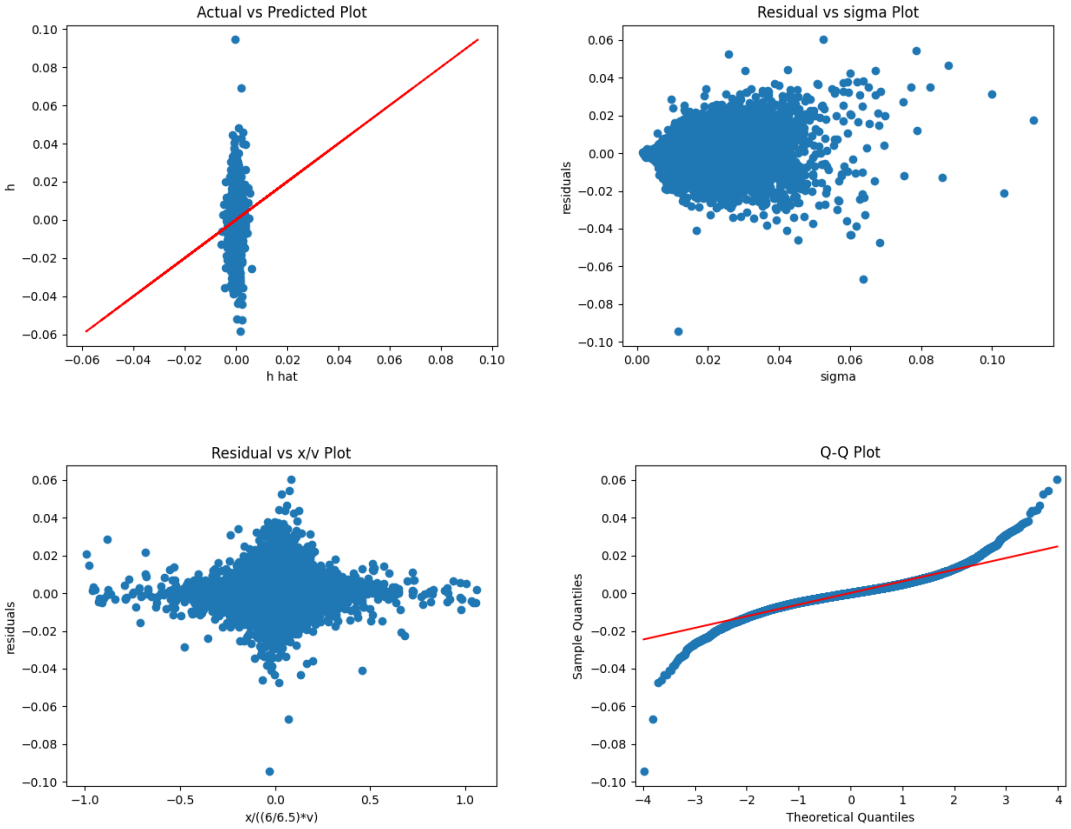
The malfunction of the model implies:

- The input values, especially the average imbalance value, do not explain the temporary cost.
- The model itself is not appropriate to capture the relationship between independent variables and a dependent variable.

If the model itself is not working, then the performance of the model under lookahead bias should be bad as well. The summary statistics of the model allowing lookahead bias are as follows:

Hyperparameters		
Lookback	**Volatility Filter**	**Order Filter**
0 days	95 percentiles	0.025%

Coefficients	**Value**	**p-value**
$\boldsymbol{\eta}$	0.2045434	0
$\boldsymbol{\beta}$	0.7264143	2.0872192862952943e-14
r-squared	0.00816	
White's test p-value	0.0	



The statistics indicates that the model parameters are highly significant. In addition, the eta as 0.2 and the beta as 0.72 are relatively aligned with the result given in the paper ($\eta = 0.142 \pm 0.0062, \beta = 0.6 \pm 0.038$) compared to the result under lookback period as 10. Furthermore, the errors are expected to be heteroskedastic under the model, so the p-value of the White's test as 0 is acceptable. Even though the r squared value is low, it is not negative at least. Thus, the inputs rather than the model itself may be the reason for the poor performance.

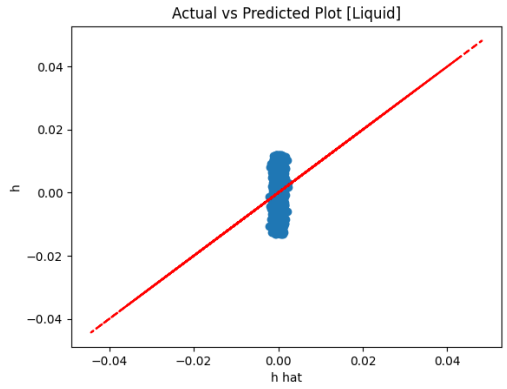
Liquid vs Illiquid

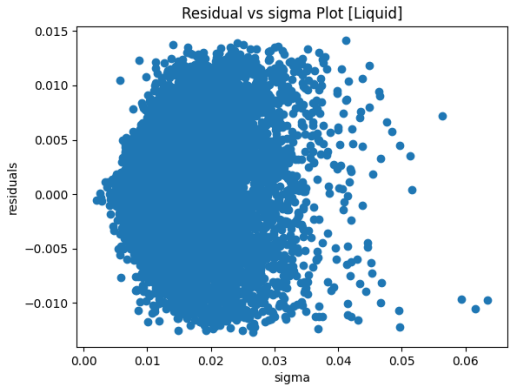
The liquidity of a stock has been measured with the mean of total daily volume. Since lookback inputs are not working, lookahead bias is allowed for the models.

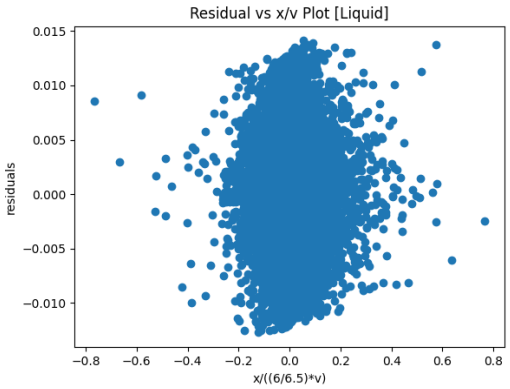
The summary statistics for liquid stocks (top 200 volume) are as follows:

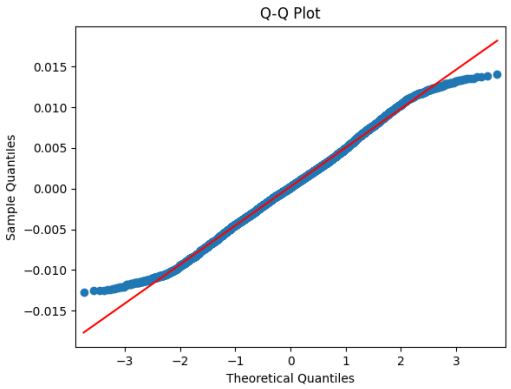
Hyperparameters		
Lookback	**Volatility Filter**	**Order Filter**
0 days	95 percentiles	0.025%

Coefficients	**Value**	**p-value**
η	0.1247664	0
β	0.5322811	0.24492875
r-squared	0.008243	
White's test p-value	2.48244093651956e-310	





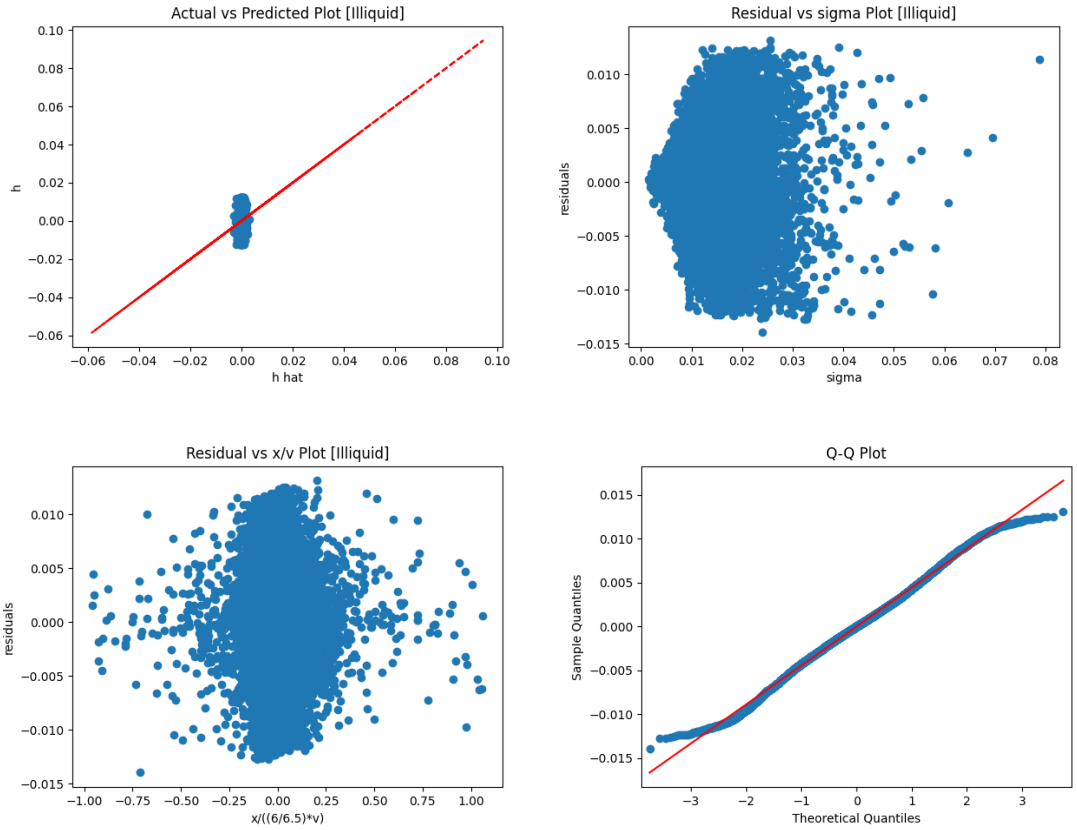




The summary statistics for illiquid stocks (bottom 200 volume) are as follows:

Hyperparemters		
Lookback	**Volatility Filter**	**Order Filter**
0 days	95 percentiles	0.025%

Coefficients	**Value**	**p-value**
η	0.1139878	0
β	0.6703227	0.45135616
r-squared	0.0063649	
White's test p-value	3.359754214236221e-287	



The r squared for liquid stocks (0.008) is slightly higher than that of illiquid stocks (0.006). The eta's are similar as 0.12 for liquid stocks and 0.11 for illiquid stocks. Both are significant with p value of 0. The beta p value of liquid stocks is lower than that of illiquid stocks though both are not significant at the level of 5%. The beta for liquid stocks is 0.53 while the beta for illiquid stocks is 0.67. This indicates that the temporal cost is lower for liquid stocks.

Lookback optimization

The statistics for lookback periods from 1 to 13 are as follows:

Lookback	**0**	**1**	**2**	**3**	**4**	**5**
η	0.204	0.051	0.033	0.043	0.035	0.010
η **p-value**	0	0	0	0	0.0017	0.334
β	0.726	0.276	0.238	0.305	0.254	0.045
β **p-value**	0	0.81	0.961	0.939	0.967	0.998
r-squared	0.008	0.002	0.001	-0.000305	-0.00133	-0.00051
White pval	0	0	0	0	0	0
Lookback	**6**	**7**	**8**	**9**	**10**	**11**
η	0.007	0.010	0.017	0.010	0.009	0.017
η **p-value**	0.473	0.690	0.595	0.284	0.470	0.709
β	0.0001	0.162	0.269	0.0009	0.0001	0.282
β **p-value**	0.999	0.999	0.998	0.9984	0.9992	0.998
r-squared	-0.00035	-0.001132	-0.00131	-0.00118	-0.00273	-0.00292
White pval	0	0	0	0	0	0

(*volatility filter: 95 percentile, order filter 0.025%)

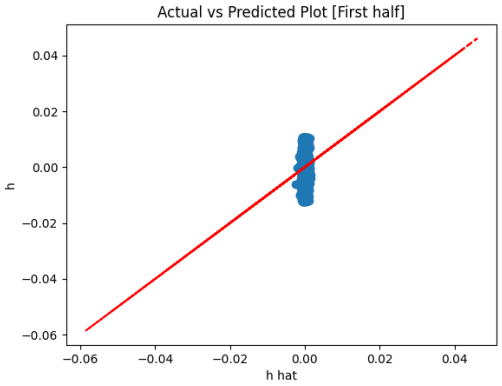
It is observable that the r squared is positive for short lookback periods of 1, 2, and 3, while the other lookback periods result in negative r squared values. Also, the p values of eta and beta increases as lookback period increases. Thus, it is optimal to use lookback period of 1 rather than 10.

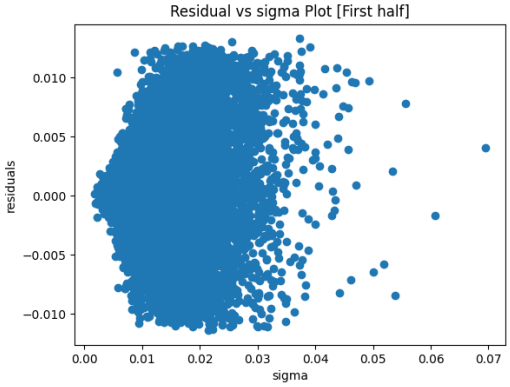
Temporal Component: First Half vs Second Half

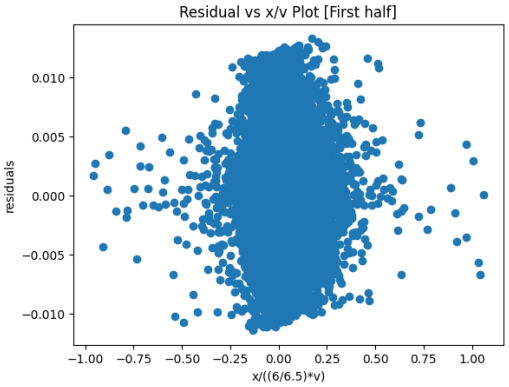
The statistics for the first half period are as follows:

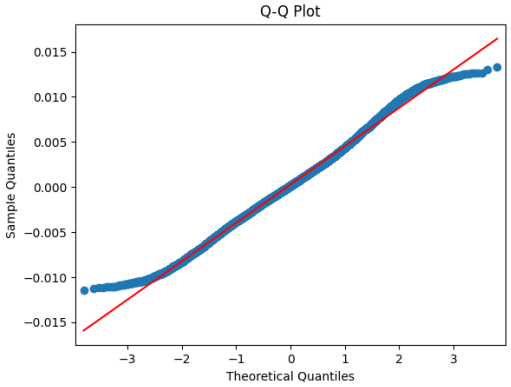
Hyperparemters		
Lookback	**Volatility Filter**	**Order Filter**
0 days	95 percentiles	0.025%

Coefficients	**Value**	**p-value**
η	0.0832181	0
β	0.5501049	0.71610041
r-squared	0.0025745	
White's test p-value	0	





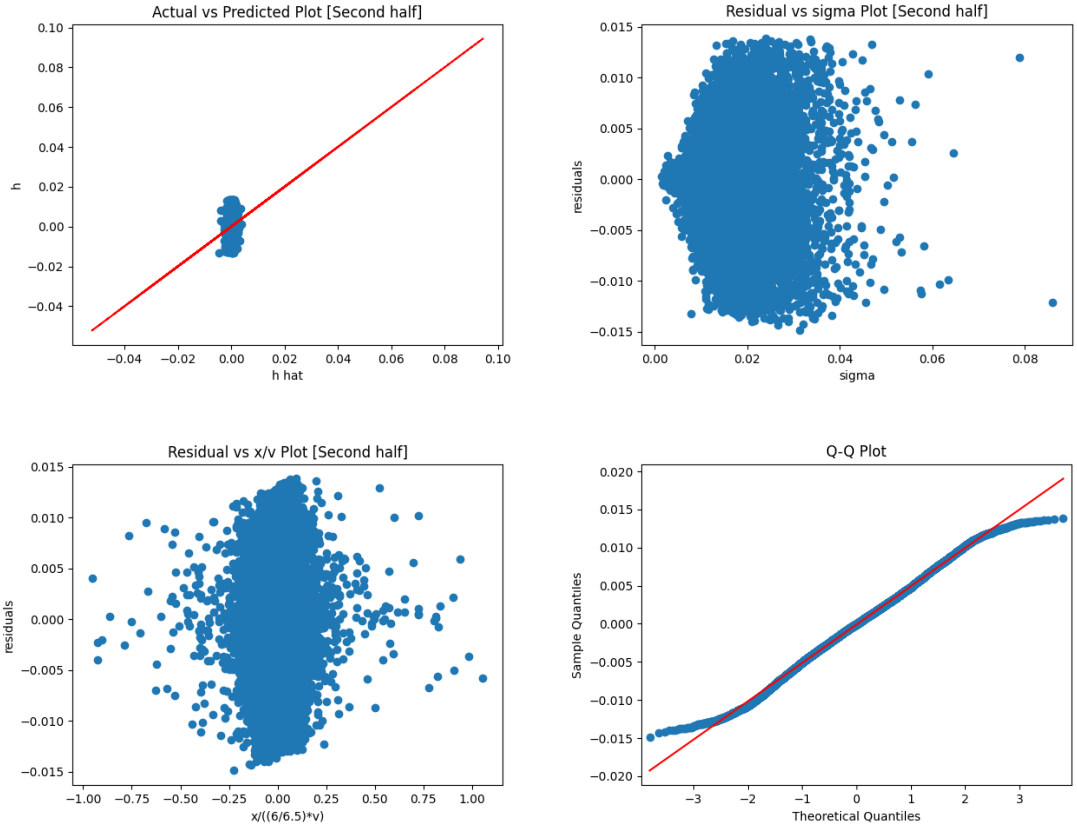




The statistics for the second half period are as follows:

Hyperparemters		
Lookback	**Volatility Filter**	**Order Filter**
0 days	95 percentiles	0.025%

Coefficients	**Value**	**p-value**
η	0.1476257	0
β	0.6325316	0.03579339
r-squared	0.0089528	
White's test p-value	0	



It is observable that the eta and beta are different between the first half and the second half. This indicates that there is a temporal component in the fit of this data. The beta and eta are both larger for the second half. This means that the temporary cost is higher for the second half. Also, it is notable that the p-value of beta is significant for the second half while that of the first half is not significant. Furthermore, the r squared increased largely from 0.0025 to 0.0089, which means that the model explains the second half data better.