Sunrise Futures Financial Modeling Data Competition

Short Summary

University of Illinois at Urbana-Champaign

Yiwei Shen

Jiahao Zhu

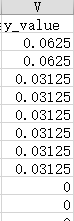
Donghan Liu

Pengyu Chen

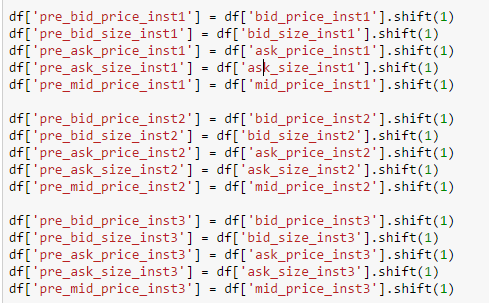
Ho Hin Jeremy Wong

1. First Glance of Data

Our goal is to optimize models to increase the prediction power of y\_values. At the first glance of data, we observed that y\_values are discrete. As shown in the picture, we aim to find the jump points of y\_values, trying to explain why an y\_value changes suddenly at a specific time point.



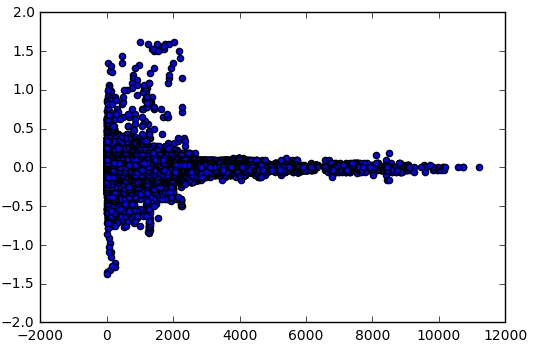
We believe that the status of the current time point depends only on the status of the time point 1 second before but not the past due to the nature of high frequency trading as we assumed. Therefore, we created 1 second time lag in our data set. We shifted the previous second trading price and size down to the current second, and then tried to predict y\_value based on the data of the previous second.



1. Relation Exploration

After obtaining our new shifted data set, we began to scatter plot relations between y\_value and other independent variables.

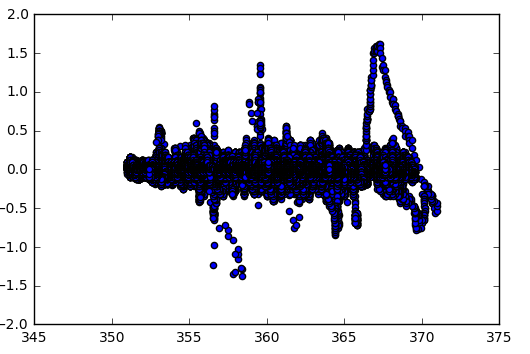
1. Size:



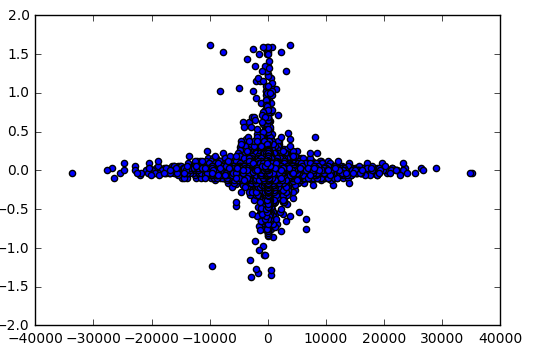
We observed that the y\_value v.s. size plots have the same pattern for all three instruments. Small sized observations tend to have higher variances in y\_values.

1. Price:

There are no obvious patterns in y\_value v.s. prices for all three instruments.



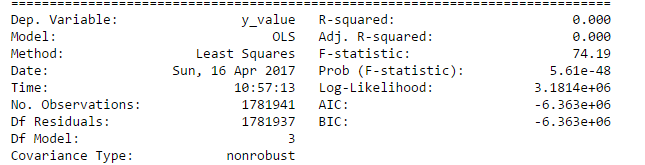
1. Net Trade



The plots of y\_values v.s. net trade indicating a pattern that when net traded volumes are close to 0, the variances of y\_values become larger. We inspired from these plots that the difference of trading sizes instead of sizes themselves are important to our models. Therefore, we created new columns representing price\_diff and size\_diff for three instruments.

1. Regression
2. Price

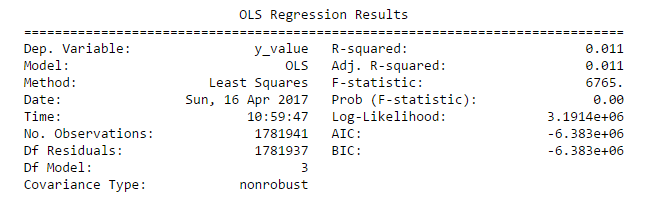
Since we are not sure whether there is any relationship between price and y\_value, we tried to regress y\_value on prices first.



The result turns out that the price is not so much related to y\_values.

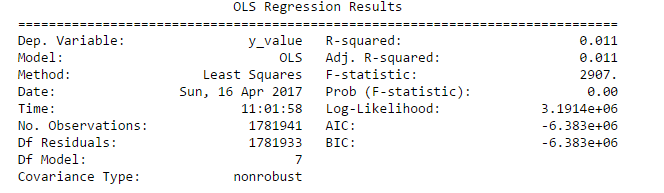
1. Size

As we have discovered before, the size differences could be more useful than sizes themselves. Therefore, we tried to regress y\_values on size differences.



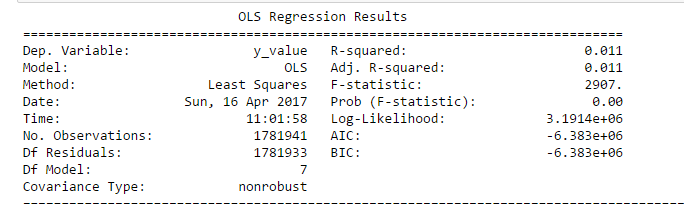
It turns out size differences does explain some of the behaviors of y\_values.

Further, we tried to include interactions between size\_diffs. It turns out that the R-square does not increase much. Therefore, simple model without intersection is more appropriate.



1. Size\*Price

Now we consider the model include intersections between size\_diff and price\_diff



The model is not improved as well.

1. **Final Model**

In conclusion, we choose **y\_value~ size\_diff1 +size\_diff2+size\_diff3** as our final model due to consideration of R-square and concise of our model.