Project for High Frequency Trading

Jingtian Zhang, Romil Gopani, Xinyi Lin, Yuhao Wang

Summary

In this paper, we conduct a VAR pricing prediction and designed an arbitrage strategy based on our result. We first cleaned and reorganized our data, then studied the exact relationship between spot and future prices. We find that future price has a leading predictability towards spot price, and the lag term should be 2 in VAR model. After that, we fit our regression, the result turns out to be solid with R-square of 0.9675. Finally, we design and test our trading strategy, the strategy yield is promising and significantly out runs the yield of random simulation.

1. Data processing

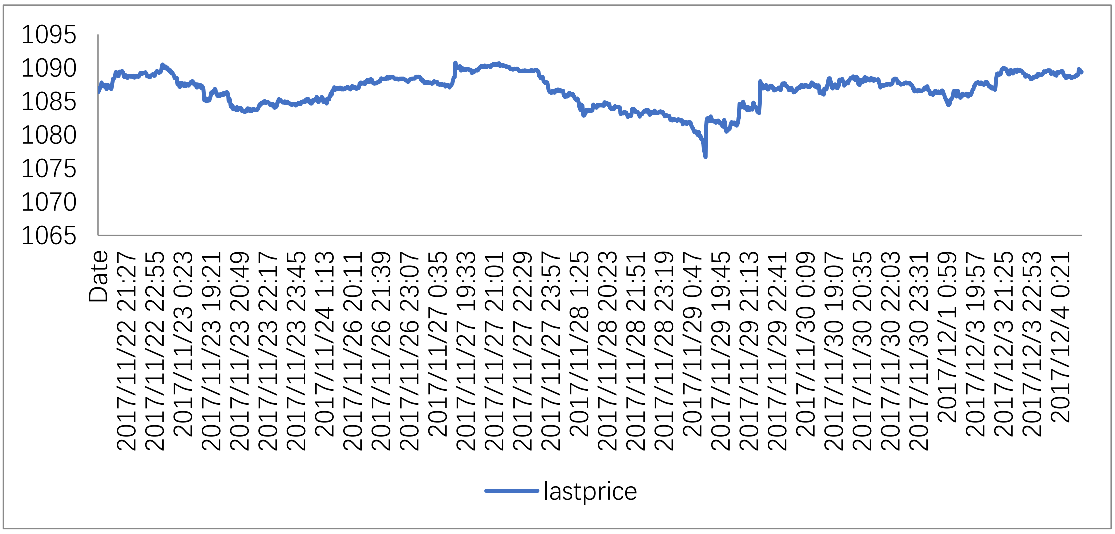
1.1 Data Cleaning:

We look through the data and found out that there are several numbers for some given minutes while some given minutes does not have matching data. Therefore, our first step is to import the data into R and generate our initial raw data matrix. To do that, we created two single-page csv files and used Read.CSV function to grasp the data. After that, we divide the raw data into 8 sub-matrixes according to the data using function Sliding. To process the data, we also need to convert the data into POSIXCT form. Interestingly, all the date went back to year 0017, so I adjusted the time accordingly. Our second step is to aggregate the data within same minute. Before the aggregation, we double checked if the ask and bid orders match for all times. After the checking process, we generated a new time grid so that all data have a more readable format. Using for loop to do the aggregation, we now come to our third step, which is to fill in the missing data. Most researchers use the lasted data or the mean of nearby data to simulate the missing ones. These two methods, however are based on different assumptions: when people believe the data are not changing (stationary) by time, it is desirable to use the latest data. Mean, on the other hand, are used when people believe the data are changing by time with a relative constant changing rate. After viewing the plot of USDKRW and KUZ7, we believe our data follows the second assumption. Thus, we use mean of nearest two data to fill in the blank. Function Filling did the job. The last step is to calculate return for both data, we can use for loop to generate the result. Also, we tried to plot our data using R. The plots do not look good so we use excel to plot the data instead.

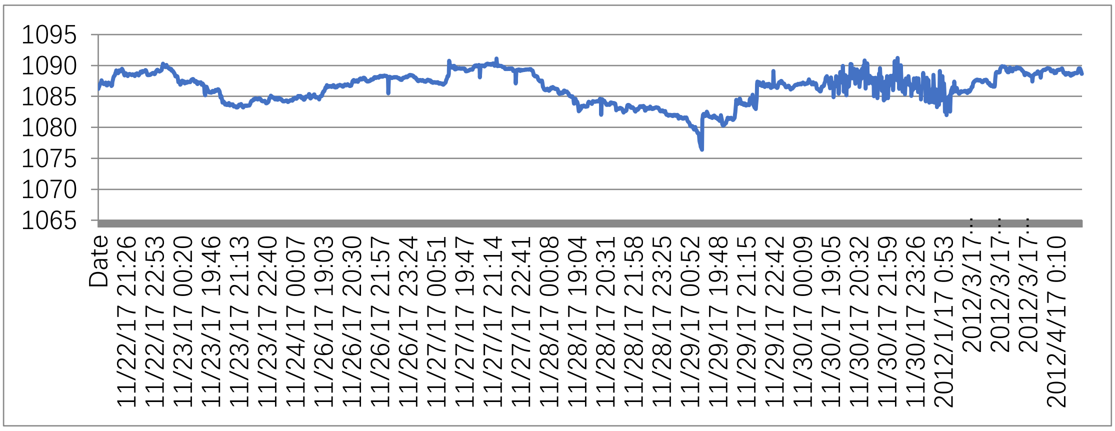
1.2 Data Description:

According to plot 1 and plot 2, both USDKRW and KUZ7 share a common change: the general shapes of these two curves are similar. This shows that the spot price and future price of USD/KRW are highly correlated. By setting the same x-label, we found that the change of future price often occurs before spot price. Therefore, from our intuition, future price is the leading factor of spot price. We also noticed that the average price of USDKRW is 1086.677 while the standard deviation is 2.43. For future price, the average is 1086.44 and the standard deviation is 2.47. As expected, the volatility of future is higher than the volatility of spot price.

The returns for both prices share a common average of 0. The return of KUZ7 shows a big fluctuate over the last two days, the return of USDKRW, on the other hand, is relatively stable.



Plot1 last prices for cleaned spot



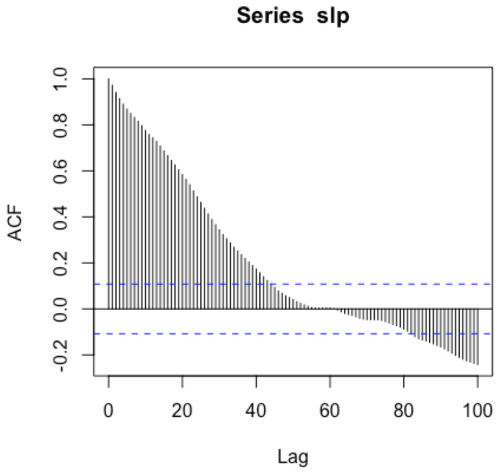
Plot2 last prices for cleaned future

2. Lead-Lag Relation

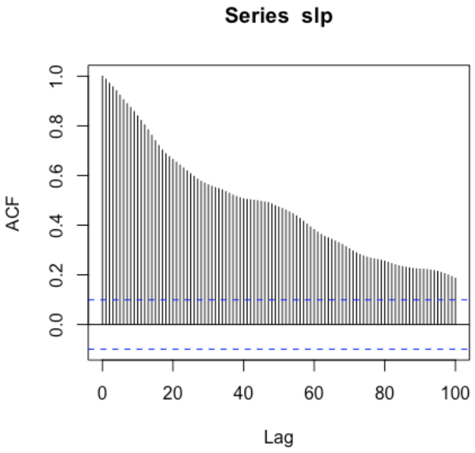
In this section, we study the relationship between spot prices and future prices. Specifically, we first conduct auto-correlation function (ACF) to see how one price itself can influence the later price. Then, we do Granger Causality test to see the relationship between different prices. The special point of our method is that we separated our data by into 8 sub-table dates and then run ACF and Granger Causality test for each of them. The reason for this data separation method is that we found prices for different days have significantly different features. The missile launch has dramatically influenced the underlying rules of price movement. Also by separating the data, we can better ensure the stationarity of data, and therefore avoid false regression. Our result shows that both prices have strong auto-correlations and the future price is a Granger Cause for spot price while the spot price is not a Granger Cause for future price.

2.1 ACF Test:

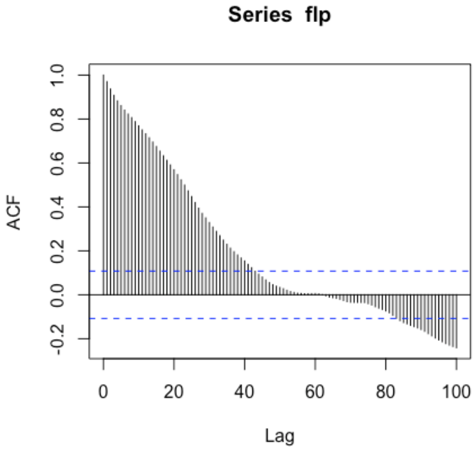
In this section, we focus on the result of ACF plot result. We witness a similar result for different data before and after missile launch. So here we only post the ACF of spot and future for November 22nd which is a typical day before missile launch and November 30th which is a typical day after missile launch. For more ACF plot, please check our appendix 1.2.



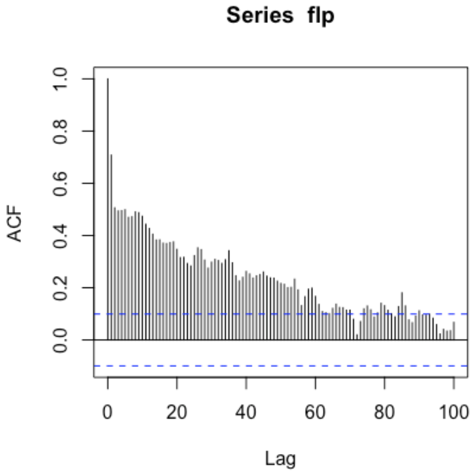
Plot3 November 22nd (Spot price)



Plot4 November 30th (Spot price)



Plot5 November 22nd (Future price)



Plot6 November 30th (Future price)

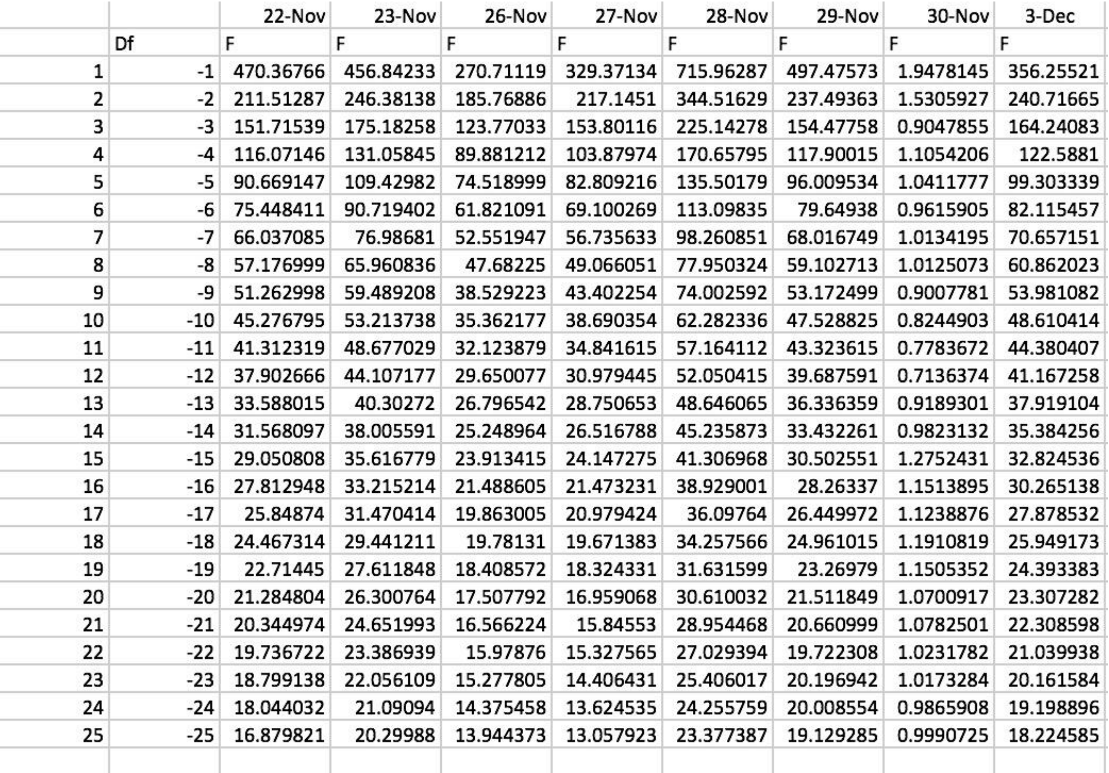
Looking at the spot price, we can see: before the missile launch, positive auto-correlation effect fades after about 40 minutes. This phenomenon can be explained as the information efficiency of foreign exchange market, that the market will react to a price change and the effect of a shock will goes away after 40 minutes of time. The negative part, interestingly, shows that after about 80 minutes, a price change will actual has a negative effect on the current price. Such effect can be explained as mean reversion effect. After launch, however, we see a whole different picture. The price has a continuously positive effect for more than 100 minutes. Our guess is that the effect of missile launch on the market is so big that it cannot be easily digested in a short time.

Similarly, we look at the plots for future prices. The ACF plot before missile launch is almost the same as the plot of spot. The plot after missile launch, on the other hand, is quite different. The autocorrelation is no longer significant in 70 minutes, which to some degree proves the fact that the future market is more volatile and therefore digests the shock in a shorter time.

2.2 Granger Causality Test:

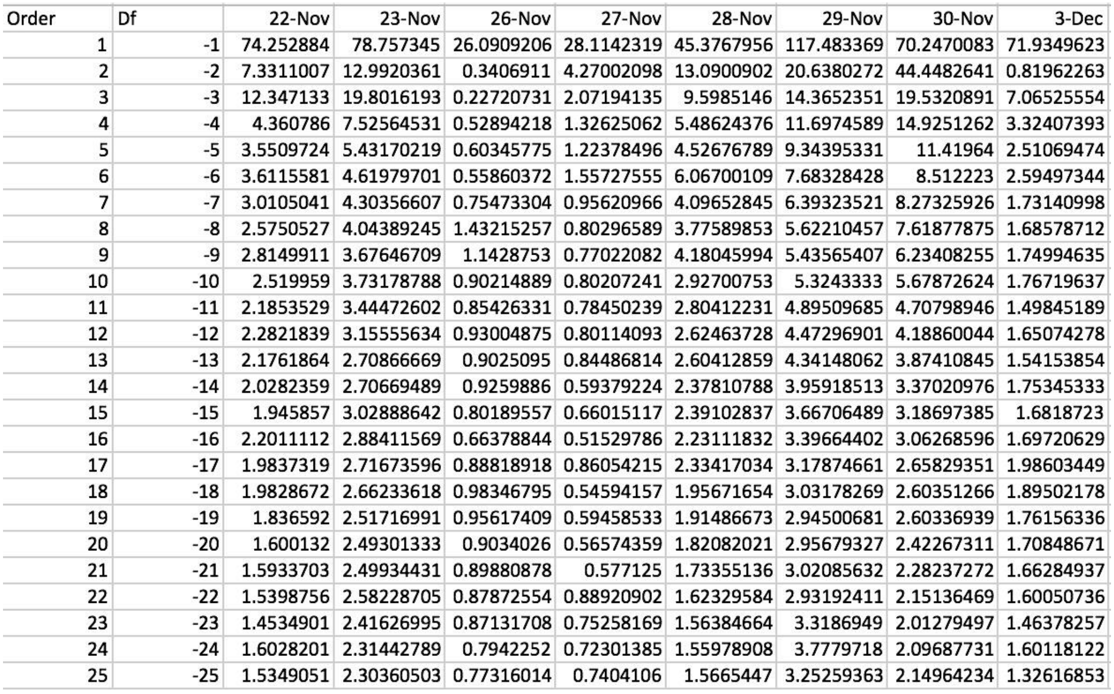
The idea of Granger Causality Test is a little similar to Bayesian theory: if the behavior of variable Y does not have a big change after we added variable Z to into the prediction (see equation (1)), we can think that the second variable has no effect to predict the first one. If not, then we say the variable Z is the Granger Cause of variable Y. The result of Granger Causality test goes as follows:

$$L(Y_t|Y_{t-1}, Z_{t-1}, Y_{t-2}, Z_{t-2} \dots) = L(Y_t|Y_{t-1}, Y_{t-2} \dots) \qquad (1)$$



		22-Nov	23-Nov	26-Nov	27-Nov	28-Nov	29-Nov	30-Nov	3-Dec
	Df	F	F	F	F	F	F	F	F
1	-1	470.36766	456.84233	270.71119	329.37134	715.96287	497.47573	1.9478145	356.25521
2	-2	211.51287	246.38138	185.76886	217.1451	344.51629	237.49363	1.5305927	240.71665
3	-3	151.71539	175.18258	123.77033	153.80116	225.14278	154.47758	0.9047855	164.24083
4	-4	116.07146	131.05845	89.881212	103.87974	170.65795	117.90015	1.1054206	122.5881
5	-5	90.669147	109.42982	74.518999	82.809216	135.50179	96.009534	1.0411777	99.303339
6	-6	75.448411	90.719402	61.821091	69.100269	113.09835	79.64938	0.9615905	82.115457
7	-7	66.037085	76.98681	52.551947	56.735633	98.260851	68.016749	1.0134195	70.657151
8	-8	57.176999	65.960836	47.68225	49.066051	77.950324	59.102713	1.0125073	60.862023
9	-9	51.262998	59.489208	38.529223	43.402254	74.002592	53.172499	0.9007781	53.981082
10	-10	45.276795	53.213738	35.362177	38.690354	62.282336	47.528825	0.8244903	48.610414
11	-11	41.312319	48.677029	32.123879	34.841615	57.164112	43.323615	0.7783672	44.380407
12	-12	37.902666	44.107177	29.650077	30.979445	52.050415	39.687591	0.7136374	41.167258
13	-13	33.588015	40.30272	26.796542	28.750653	48.646065	36.336359	0.9189301	37.919104
14	-14	31.568097	38.005591	25.248964	26.516788	45.235873	33.432261	0.9823132	35.384256
15	-15	29.050808	35.616779	23.913415	24.147275	41.306968	30.502551	1.2752431	32.824536
16	-16	27.812948	33.215214	21.488605	21.473231	38.929001	28.26337	1.1513895	30.265138
17	-17	25.84874	31.470414	19.863005	20.979424	36.09764	26.449972	1.1238876	27.878532
18	-18	24.467314	29.441211	19.78131	19.671383	34.257566	24.961015	1.1910819	25.949173
19	-19	22.71445	27.611848	18.408572	18.324331	31.631599	23.26979	1.1505352	24.393383
20	-20	21.284804	26.300764	17.507792	16.959068	30.610032	21.511849	1.0700917	23.307282
21	-21	20.344974	24.651993	16.566224	15.84553	28.954468	20.660999	1.0782501	22.308598
22	-22	19.736722	23.386939	15.97876	15.327565	27.029394	19.722308	1.0231782	21.039938
23	-23	18.799138	22.056109	15.277805	14.406431	25.406017	20.196942	1.0173284	20.161584
24	-24	18.044032	21.09094	14.375458	13.624535	24.255759	20.008554	0.9865908	19.198896
25	-25	16.879821	20.29988	13.944373	13.057923	23.377387	19.129285	0.9990725	18.224585

Table1. Testing if future price is a Granger Cause of spot price



Order	Df	22-Nov	23-Nov	26-Nov	27-Nov	28-Nov	29-Nov	30-Nov	3-Dec
1	-1	74.252884	78.757345	26.0909206	28.1142319	45.3767956	117.483369	70.2470083	71.9349623
2	-2	7.3311007	12.9920361	0.3406911	4.27002098	13.0900902	20.6380272	44.4482641	0.81962263
3	-3	12.347133	19.8016193	0.22720731	2.07194135	9.5985146	14.3652351	19.5320891	7.06525554
4	-4	4.360786	7.52564531	0.52894218	1.32625062	5.48624376	11.6974589	14.9251262	3.32407393
5	-5	3.5509724	5.43170219	0.60345775	1.22378496	4.52676789	9.34395331	11.41964	2.51069474
6	-6	3.6115581	4.61979701	0.55860372	1.55727555	6.06700109	7.68328428	8.512223	2.59497344
7	-7	3.0105041	4.30356607	0.75473304	0.95620966	4.09652845	6.39323521	8.27325926	1.73140998
8	-8	2.5750527	4.04389245	1.43215257	0.80296589	3.77589853	5.62210457	7.61877875	1.68578712
9	-9	2.8149911	3.67646709	1.1428753	0.77022082	4.18045994	5.43565407	6.23408255	1.74994635
10	-10	2.519959	3.73178788	0.90214889	0.80207241	2.92700753	5.3243333	5.67872624	1.76719637
11	-11	2.1853529	3.44472602	0.85426331	0.78450239	2.80412231	4.89509685	4.70798946	1.49845189
12	-12	2.2821839	3.15555634	0.93004875	0.80114093	2.62463728	4.47296901	4.18860044	1.65074278
13	-13	2.1761864	2.70866669	0.9025095	0.84486814	2.60412859	4.34148062	3.87410845	1.54153854
14	-14	2.0282359	2.70669489	0.9259886	0.59379224	2.37810788	3.95918513	3.37020976	1.75345333
15	-15	1.945857	3.02888642	0.80189557	0.66015117	2.39102837	3.66706489	3.18697385	1.6818723
16	-16	2.2011112	2.88411569	0.66378844	0.51529786	2.23111832	3.39664402	3.06268596	1.69720629
17	-17	1.9837319	2.71673596	0.88818918	0.86054215	2.33417034	3.17874661	2.65829351	1.98603449
18	-18	1.9828672	2.66233618	0.98346795	0.54594157	1.95671654	3.03178269	2.60351266	1.89502178
19	-19	1.836592	2.51716991	0.95617409	0.59458533	1.91486673	2.94500681	2.60336939	1.76156336
20	-20	1.600132	2.49301333	0.9034026	0.56574359	1.82082021	2.95679327	2.42267311	1.70848671
21	-21	1.5933703	2.49934431	0.89880878	0.577125	1.73355136	3.02085632	2.28237272	1.66284937
22	-22	1.5398756	2.58228705	0.87872554	0.88920902	1.62329584	2.93192411	2.15136469	1.60050736
23	-23	1.4534901	2.41626995	0.87131708	0.75258169	1.56384664	3.3186949	2.01279497	1.46378257
24	-24	1.6028201	2.31442789	0.7942252	0.72301385	1.55978908	3.7779718	2.09687731	1.60118122
25	-25	1.5349051	2.30360503	0.77316014	0.7404106	1.5665447	3.25259363	2.14964234	1.32616853

Table2. Testing if spot price is a Granger Cause of future price

Looking through the F-distribution data, we found that the threshold value for given degree of freedom should be 2.85. We notice that in table1, all numbers are far greater that this threshold. Therefore, we believe the future price is the Granger Causality of spot price. From the table2, we see that most numbers after 17 minutes are no longer significant. November 29th is a special case since a big event took place on that time. Thus, we believe that spot price is not a Granger Cause of future price. Considering these two facts, we can conclude that the future price should be the independent variable and the spot price should be the dependent variable in our prediction model.

3. Vector Auto-Regression Model

Vector Auto-Regression (VAR) Model is the simplest model when people study the relationships between different prices. The general form of VAR is equation (2). We can also expand the form into equation (3). To simulate such model, our first step is to get the most reasonable number of lag terms. Then, after we settled our model structure, we did regression based on our model. For here we used first 2000 data as training data, and use the rest as testing data.

$$C_0 y_t = \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + H x_t + \varepsilon_t, t = 1, 2, \ldots, T \qquad (2)$$

$$C_0 = \begin{bmatrix} 1 & -c_{12} & \cdots & -c_{1k} \\ -c_{21} & 1 & \cdots & -c_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ -c_{k1} & -c_{21} & \cdots & 1 \end{bmatrix}, \Phi_i = \begin{bmatrix} \Phi_{11}^{(i)} & \Phi_{12}^{(i)} & \cdots & \Phi_{1k}^{(i)} \\ \Phi_{21}^{(i)} & \Phi_{22}^{(i)} & \cdots & \Phi_{2k}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{k1}^{(i)} & \Phi_{k2}^{(i)} & \cdots & \Phi_{kk}^{(i)} \end{bmatrix}$$

$$i = 1, 2, \ldots, p \quad \varepsilon = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \vdots \\ \varepsilon_{kt} \end{bmatrix} \qquad (3)$$

3.1 How many lag terms should we use: AIC and BIC:

People decide how many terms to use based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), the lower the both criterion is, the more reliable the model is. The formula of AIC can be written as equation (4), Where, k is the number of variables in the system, and $\sum$u is the estimation of covariance matrix and can be expressed as $\sum_u(n) = T^{-1}\sum_{t=1}^{T}\widehat{u_t u_t'}$. The idea of AIC is that R-square always gets larger when we add more terms into our function. To solve this problem, AIC imports a penalize for every term we add. Similarly, BIC shares that common idea (see equation (5)). The main difference is however, is that BIC has a tougher penalize for terms. We conduct several literature reviews and figured out that AIC are more widely used in high frequency data than BIC, and therefore we valued more on AIC in our research.

$$\mathrm{AIC(n)} = \ln\det(\textstyle\sum_u(n)) + \frac{2}{T} * n * k^2 \qquad (4)$$

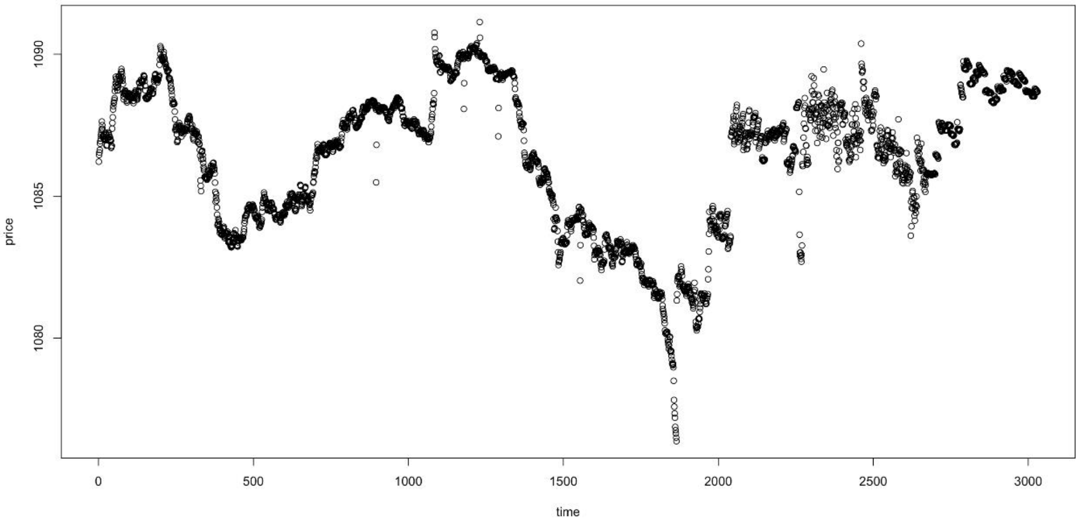
$$\mathrm{BIC(n)} = \ln\det(\textstyle\sum_u(n)) + \frac{pln(T)}{T} * k^2 \qquad (5)$$

The result shows that the differences between AIC values for different periods are relatively small. Besides, different criterions return results that are not consistent. For AIC, the best lag term is 2.

3.2 VAR Regression and testing:

In this part, we first conduct the VAR using lag term of 2 and get the predicted prices based on our training data. After we get the parameters, we predict the future USDKRW price using the KUZ7 prices. Comparing the predicted data with real data in our testing data set, we can calculate the MSE and get predictability of our model.

The result show that R-square of our model reaches 0.9675 and all parameters are significant for bid prices. See plot (7) our predicted spot prices.



Plot7 predicted spot prices

4. Vector Auto-Regression Model

4.1 Trading assumptions:

Our trading strategy is based on assumptions as below:

(1) No initial values: we intend to design an arbitrage strategy, thus, our initial capital should be 0. However, our strategy does allow us to borrow since that is the only way we can trade.

(2) No short selling: to better simulate the market situation in some emerging market, we do not allow people to short sell, which means the only way to earn money is to buy at low and sell at high.

(3) Based on one unit of stock: Since the initial value of our portfolio is 0, any positive return generates a yield rate of infinite. Thus, we set that whenever there is a buy order we just borrow the money and buy 1 quantity, and whenever a sell order is executed, all the quantities are sold. At the end of period, all positions are closed and the result is measured by the real dollar amount that we made.

4.2 Trading strategies:

Based on the predicted values, we have made a trading model which uses moving average as the standard parameter for trading. We compute the 5-min standard moving average and the standard deviation from the market data. In order to specify the trading direction, we use the upper and lower band which are basically moving average plus/minus one standard deviation. If the predicted price is greater than the upper band then the instrument is overpriced and if it is lower than the lower band then it is underpriced.

If the predicted price touches the upper band and is still going up, then we set a limit buy order at the predicted price. Only if our predicted price is higher than the market bid, will it be executed. Similarly, when the predicted price touches the lower band, we place our limit sell order at predicted price if our current position is greater than 0. Only when our predicted price is lower than the ask price will such order be executed.

At the end of our 8 days period, if there is any remaining holding positions, we liquidate them all using market orders.

4.3 Trading result:

By applying this strategy to predicted stock price, we get profit of 5.273589 US. Dollars in 8 days. We compare this strategy to random trading strategy. The random strategy chooses random number between -1 and 1 for all the data points in the trading period. If the data point is above 0.7 then a buy order is sent and if it is less then -0.7, a sell order is sent. All the positions at the end are liquidated. We ran a monte Carlo simulation on the output and so as to get average of all the output.

The random simulation on 10000 trails of the stock data always showed loss of average loss of 1194.853.

It clearly demonstrated that our strategy build above is better than random trading.

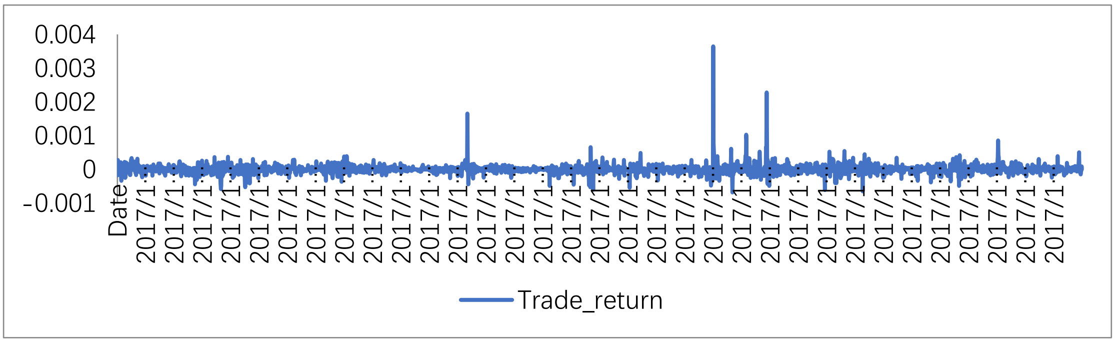
5. Conclusion

In this paper, we conduct a VAR pricing prediction and designed an arbitrage strategy based on our prediction model. We first cleaned and reorganized our data, then studied the real relationship between spot and future prices. After that, we tried to fit our regression, the result turns out to be solid. Finally, we run our trading strategy and the yield significantly out runs the yield of random simulation.

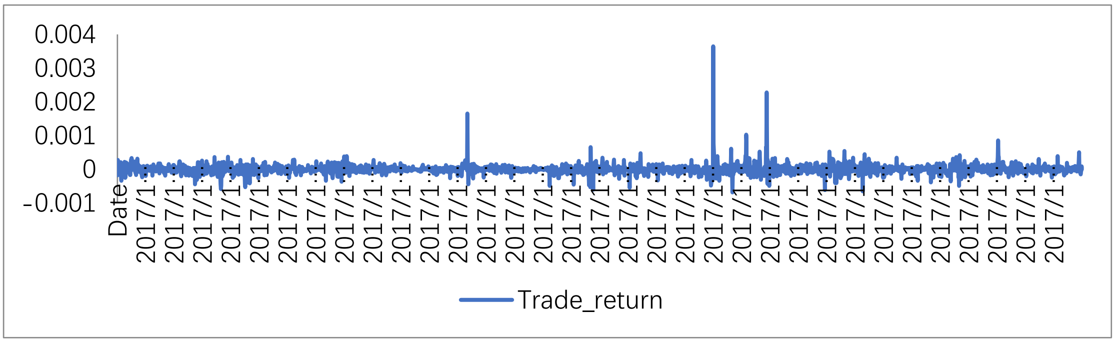
6. Appendix:

1.1 Plots for data

1.1.1 USDKRW Return

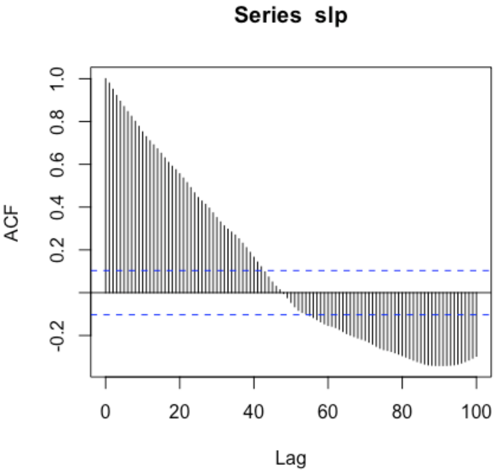


1.1.2 KUZ7 Return

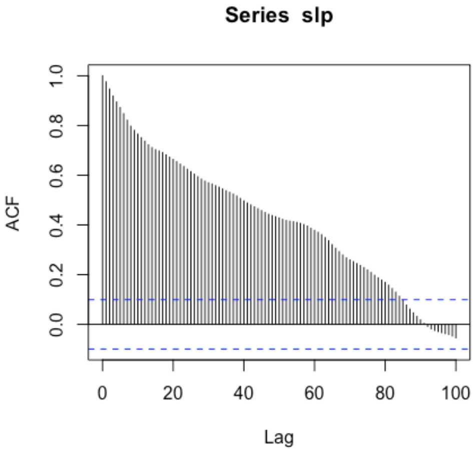


1.2 Plots of ACFs

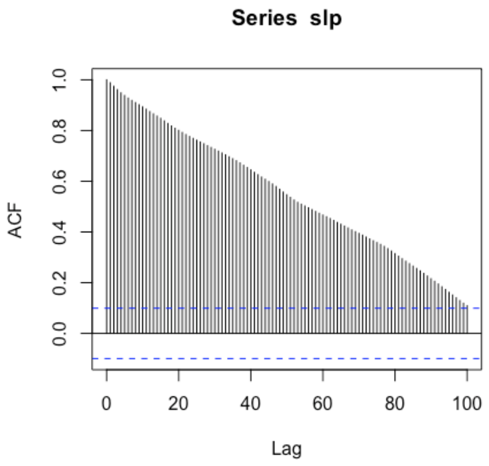
1.2.1 ACF Spot November 23rd:



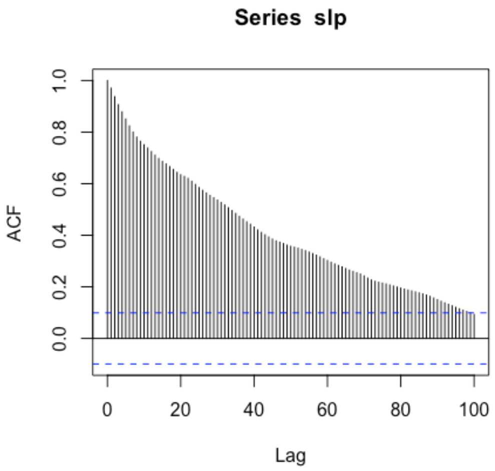
1.2.2 ACF Spot November 26th:



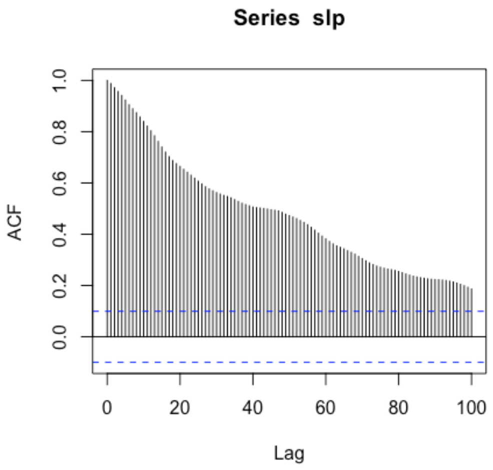
1.2.3 ACF Spot November 27th:



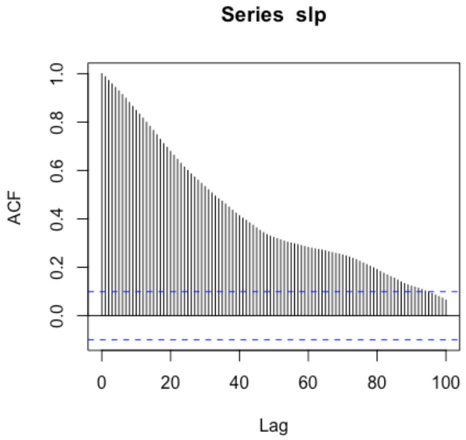
1.2.4 ACF Spot November 28th:



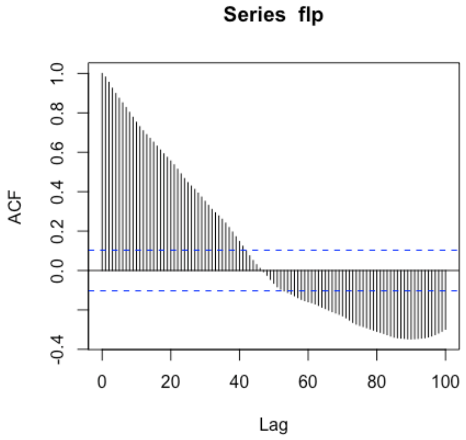
1.2.5 ACF Spot November 29th:



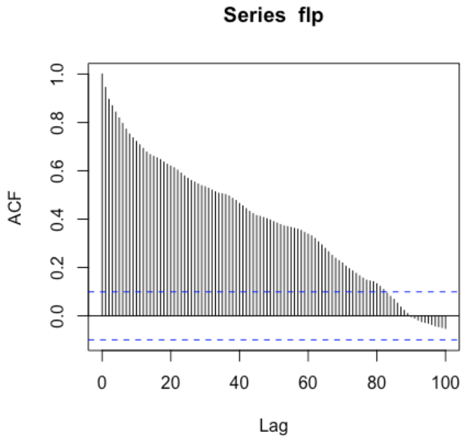
1.2.6 ACF Spot December 3rd:



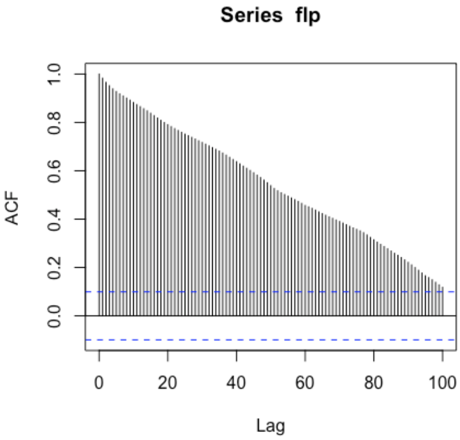
1.2.7 ACF Future November 23rd:



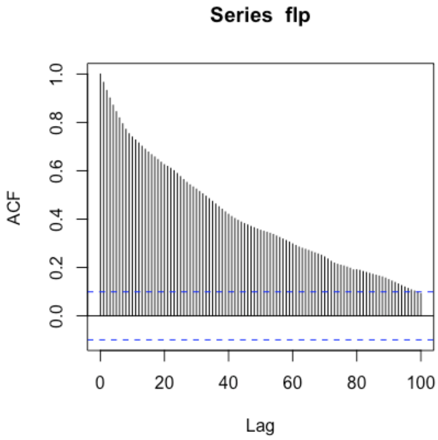
1.2.8 ACF Future November 26th:



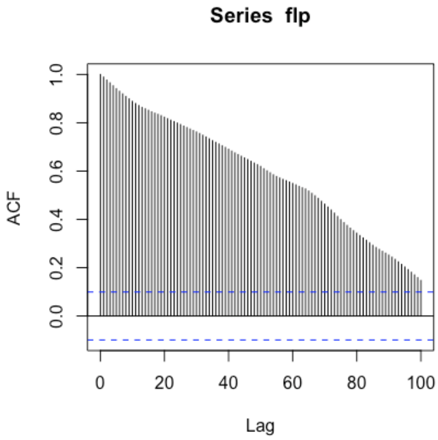
1.2.9 ACF Future November 27th:



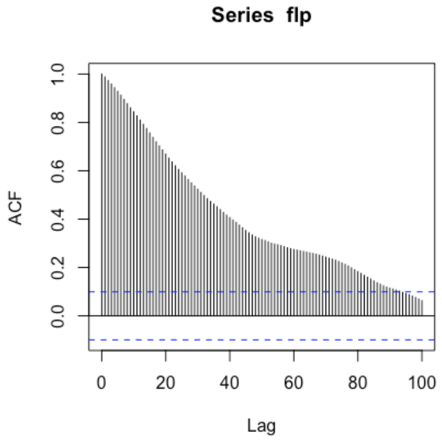
1.2.10 ACF Future November 28th:



1.2.11 ACF Future November 29th:



1.2.12 ACF Future December 3rd:



1.3 Confirmation

We, Jingtian Zhang, Romil Gopani, Xinyi Lin, Yuhao Wang thereby confirm that all team members contributed equally to this project and every team member has coded in order to finish the research.