ORIE 4741 Project Proposal: How does covid-19 affect stock index?

Jinyi Jiang(jj756), Qianqian Wu(qw273), Yujie Cao(yc2628)

Methods to avoid overfitting:

Given that we have significantly more columns than rows in the current dataset, we are not so worried about the overfitting. Still, we list some possible solutions in case of potential overfitting.

- Gaining access to more training data

In our case, since the Covid-19 is a public health issue started in the January 2020, we don't have excess data to obtain, therefore the current dataset is all we have.

- Cross-validation

We will partition our dataset and observe how the model fits observations that weren't used to estimate the model.

- Regularization

We are going to try several different regularizations such as quadratic regularizer, 11 and l2 regularizers.

Methods to avoid underfitting:

If the model underfits, we will try several different regression methods and will add more features if necessary.

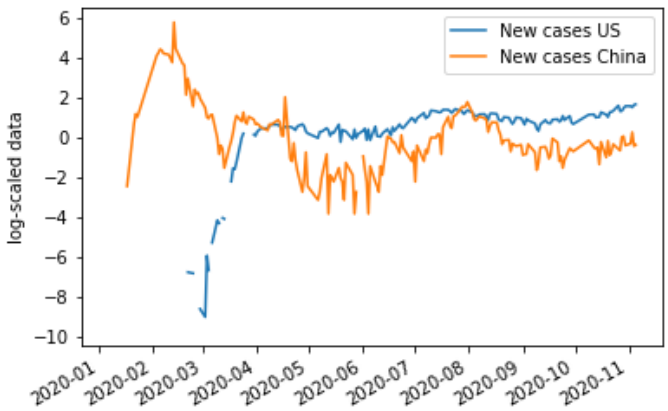
Current Features and Examples

Our dataset starts from January 3rd, 2020 till now November 5th, 2020 and is daily based.

Dependent variable:

Industry indicators	Stock market	Consumer products	Healthcare
China	CSI300	cproducts _ch	healthcare_ch
United States	SP500	DJI_Retail	DJI_Pharma

- We also take the log of all the variables above to measure the percentage change with the regression model.



Independent variable:

Covid-19's influence

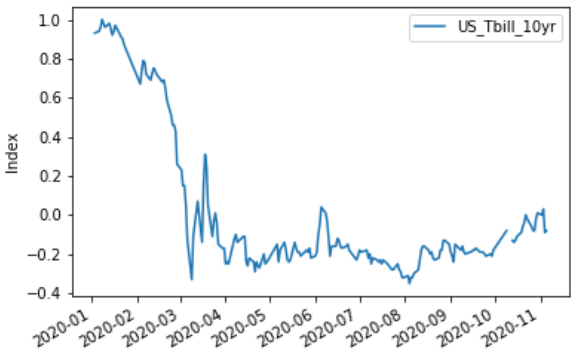
- Daily new cases: New_cases_CH, New_cases_US
- Cumulative cases: Cumulative_cases_CH, Cumulative_cases_US
- Daily deaths: New_deaths_CH, New_deaths_CH
- Cumulative deaths: Cumulative_deaths_CH, Cumulative_deaths_CH

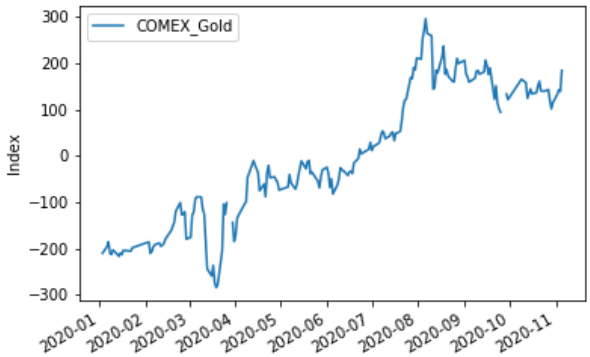
Covid-19's lag influence: Medical incubation period of Covid-19 is 14 days.

- Lag Daily new cases: lag14_New_cases_CH, lag14_New_cases_US

Black Swan measurement: Gold and 10-year treasury bill serves as safe haven assets.

- Gold: COMEX_Gold
- 10-year treasury bill: US_Tbill_10yr





Intermarket Influences:

The variance at US stock market may influence the performance of China's stock market and vice versa. We expected the lag to be one day, since the information flow spread incredibly fast.

Missing values:

Since the stock markets doesn't open on the weekend, we ignore the weekend data entry.

Testing for model effectiveness:

- R-squares: examine whether R-square is close to 1.
- Standardized Residuals Plot: check whether the plot seems to be random

Features and transformations

- We performed a log transformation on the Covid and index data, as we want to capture the percentage effect on the stock index if Covid cases change by 1%.
- We standardized the data using formula $Xnew = \frac{X-\mu}{\delta}$. Because in neural network model, we want to work with standardized data instead of decentralized data.

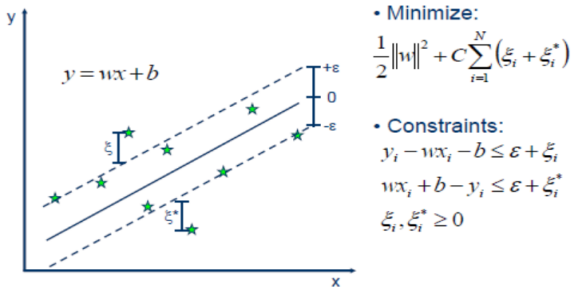
Supervised models

- Linear Regression

Linear regression is to find y = wx + b which minimizes the sum of the square errors. In linear model, we introduced regularization to avoid overfitting problem.

- SVR

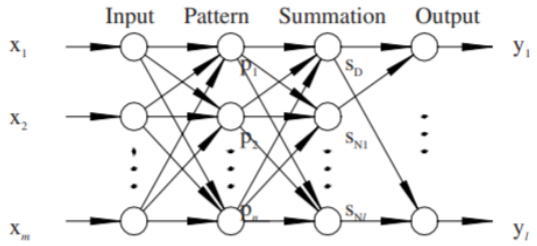
Support vector regression introduces non-linear part. SVR creates an interval band on both sides of the linear function. No loss will be calculated for all samples that fall into the interval band. And the optimized model is obtained by minimizing the total loss.



(source: Data Mining Map)[1]

- Neural Networks

Neural networks work like human brain. It has input layers for inputting data, hidden layers for improving prediction using backpropagation, and output layers which are y.



(source: Application of General Regression Neural Network to Vibration Trend Prediction of Rotating Machinery)[2]

Remains

We need to construct SVR model and neural networks model in the following week. And we also need to modify our linear model such as using regularization. What's more, we are going to test the efficiency of our models and adjust accordingly. Then we will use the data from 2020/11/06 to 2020/12/06 to test our models again and analyze the result.

Date	Plan
2020/11/08 - 2020/11/22	Finish building three supervised models
2020/11/22 - 2020/11/25	Testing efficiency and adjust overfit or underfit problems
2020/11/25 - 2020/12/06	Look for papers to study and enrich our report
2020/12/06 - 2020/12/10	Use new data to test models and analyze the result

[1] https://www.saedsayad.com/support_vector_machine_reg.htm
[2] https://link.springer.com/chapter/10.1007/978-3-540-28648-6_123