VE406

Group 3

Boqun Li
Shensong Zhao
Xinmiao Yu

December 1, 2020

Overview

Overview


Close Price Analysis (univariate time series)
Collect Data
MLR analysis
monthly data
collinearity
heteroskedasticity (time series)
time series
multi variable
single variable
GLMM
PCA
GARCH
single variable
multi variable
Test Train Splitting
Fit Model
Evaluate and Compare the models
Final Model Discussion

Close Price Analysis

Goal: Predict the Stock Price

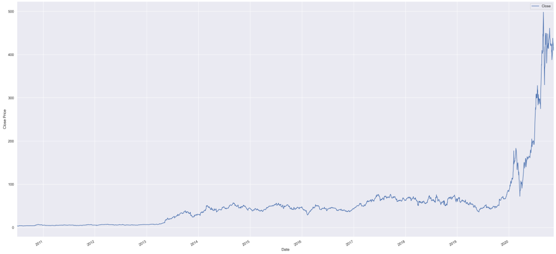

Figure 1: *Close Price* vs. *Date*

- Moving Average
- Smooth
- Correlation
- Year Trend and Seasonality
- Outliers

Moving Average

- Reduce noise
- Better understanding of underlying trend

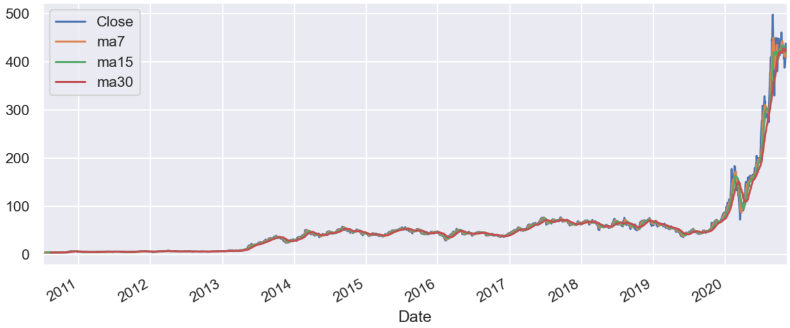

Figure 2: Moving Average Plot with $7, 15, 30$ Days

Smooth

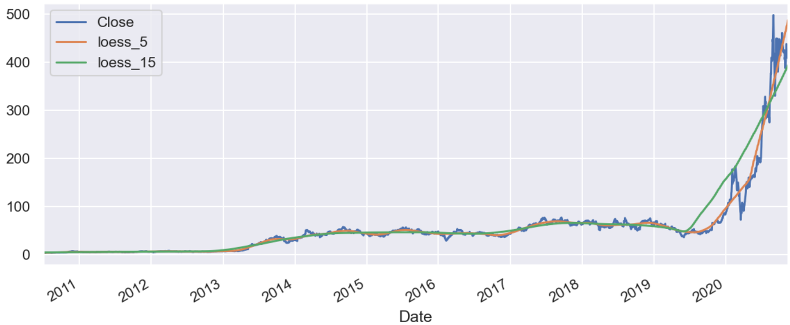

Figure 3: Smoothing with Fraction is 0.05 and 0.15

- Overall increasing trend
- Sharp gap between year 2020 and previous years

Correlation

Shift the *Close price* by x days, denoted as $Close_x$

Figure 4: Pearson Correlation Coefficient

- Show correlation, less time shifted, higher correlated

Correlation

- Highly correlated, need further discussed

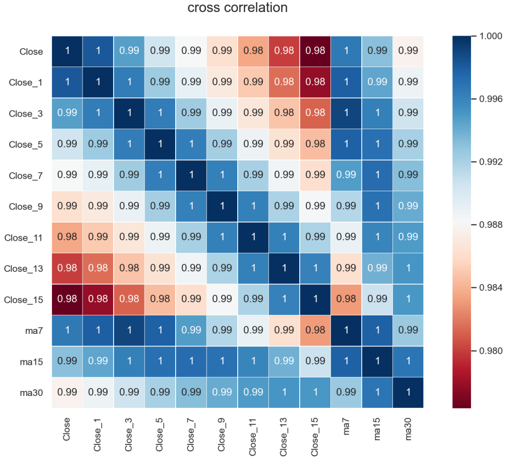

Figure 5: Pearson Correlation Coefficient

Year Trend and Seasonality

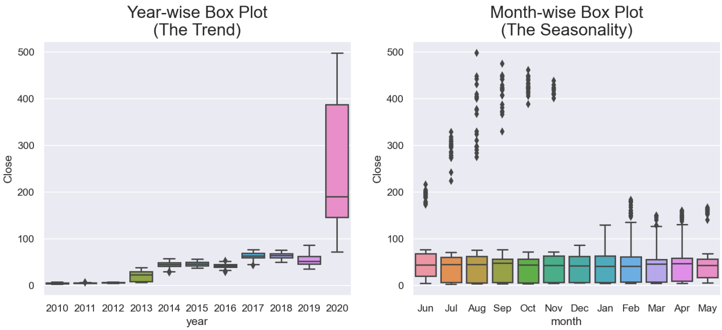

Figure 6: Year and Month Box Plot

- Clear gap and no seasonality

Outliers

Use K-means as a quick reference for outliers identification

Figure 7: Outlier Detection

- Year 2020 identified as outliers, as expected

Consider only use year 2020 data for our goal...

Year Trend and Seasonality Revisited

No year 2020 data involved!

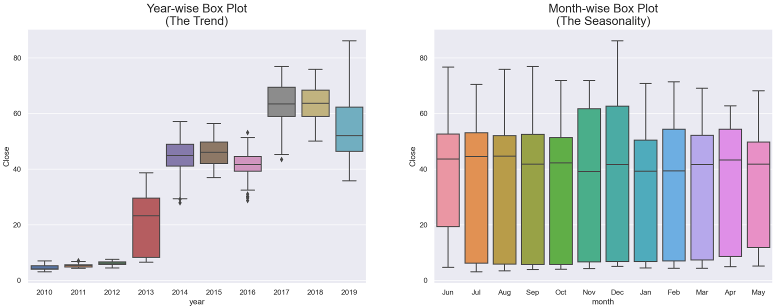

Figure 8: Without Year 2020 Trend

- No outliers anymore and still no seasonality

Decomposition

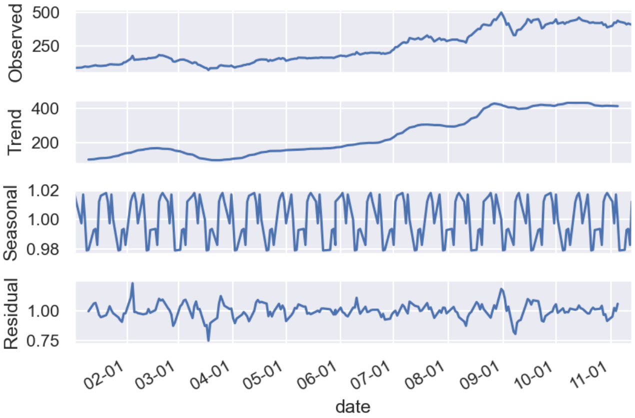

Figure 9: Year 2020 Decomposition

Decide to only use year 2020 data

Overview

Close Price Analysis (univariate time series)
Collect Data
MLR analysis
monthly data
collinearity
heteroskedasticity (time series)
time series
multi variable
single variable
GLMM
PCA
GARCH
single variable
multi variable
Test Train Splitting
Fit Model
Evaluate and Compare the models
Final Model Discussion

Data Collecting

variable	brief explanation
OilPrice	The daily price of oil in US
death	The number of death caused by the car of Tesla
DPRIME	daily Bank Prime Loan Rate
TOTALSA	total number of sales monthly data divided by 30
new-death	newly death numbers due to COVID-19
new-case	new cases of COVID-19
GoogleTrend	The number of people searching for TSLA on Google

- Research to find the related factors
- Choose data in different categories to reduce predictors' correlation
- Choose daily data

Overview


Close Price Analysis
(univariate time series)
Collect Data
MLR analysis
monthly data
collinearity
heteroskedasticity
(time series)
time series
multi
variable
single
variable
GLMM
PCA
GARCH
single variable
multi variable
Test Train Splitting
Fit Model
Evaluate and Compare
the models
Final Model Discussion

- First we need to select the related variable from the collected data.

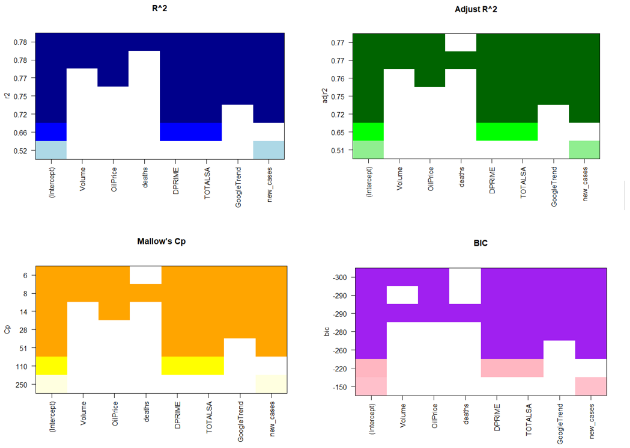

Figure 10: Variable selection

- Based on the acf plot and the residual plot, the errors are correlated

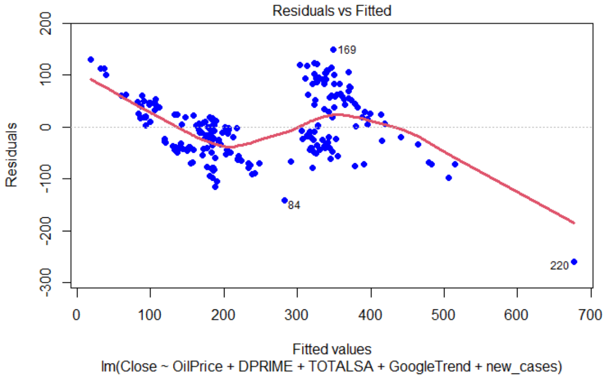

Figure 11: Residual plot

- From the pattern of the residuals, we can see that the residuals series is not a white noise.

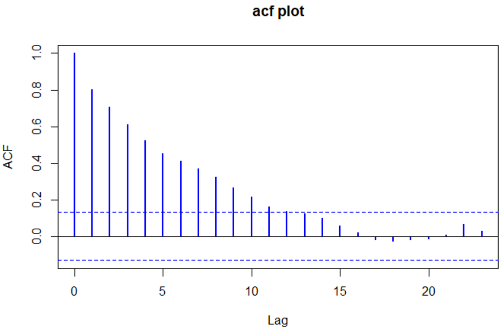

Figure 12: ACF plot

Overview

Close Price Analysis (univariate time series)
Collect Data
MLR analysis
monthly data
collinearity
heteroskedasticity (time series)
time series
multi variable
single variable
GLMM
PCA
GARCH
single variable
multi variable
Test Train Splitting
Fit Model
Evaluate and Compare the models
Final Model Discussion

Problem Addressing

Based on the problems

- Monthly Data
- Collinearity
- Heteroskedasticity
- Time Series

Different Methods will be used respectively

Monthly Data

The monthly collected data are mainly *TOTALSA* and *DPRIME*, which is as following,

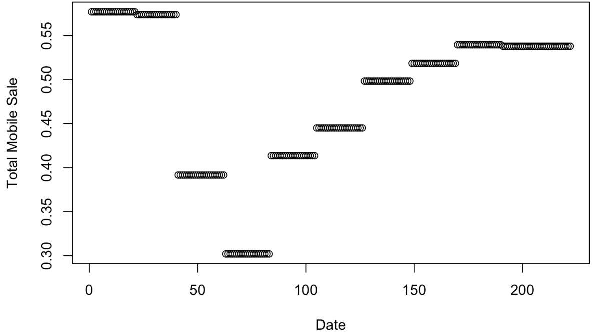

Figure 13: Scatter plot for TOTALSA

Monthly Data

Generalized Linear Mixed Model

- Fixed Effects
  Fixed across the date
  Oil price that is daily collected

- Random Effects
  Random across the date
  Total mobile sale collected monthly – average to daily basis

Monthly Data

We can have the random effect either affect intercept or the slope of the model. Here we assume the random effect only contributes to the intercept.

We fit three models in total.

- DPRIME + TOTALSA
- TOTALSA
- DPRIME

Then we compare the three models using anova table. We mainly focus on AIC and BIC criteria across the model.

```
Data: data.frame(tesla.training)
Models:
tesla.TOTALSA: Close ~ OilPrice + GoogleTrend + new_cases + (1 | TOTALSA)
tesla.DPRIME: Close ~ OilPrice + GoogleTrend + new_cases + (1 | DPRIME)
tesla.glmm: Close ~ OilPrice + GoogleTrend + new_cases + (1 | DPRIME) + (1 |
tesla.glmm:     TOTALSA)
              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
tesla.TOTALSA    6 1968.4 1988.2  -978.2   1956.4
tesla.DPRIME     6 2329.4 2349.3 -1158.7   2317.4   0.00  0
tesla.glmm       7 1970.4 1993.5  -978.2   1956.4 361.04  1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Monthly Data

We further check the summary of the model fitted only with *TOTALSA*.

```
Formula: Close ~ OilPrice + GoogleTrend + new_cases + (1 | TOTALSA)
   Data: data.frame(tesla.training)

REML criterion at convergence: 1963.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2321 -0.4571  0.0706  0.5182  4.3036

Random effects:
 Groups   Name        Variance Std.Dev.
 TOTALSA  (Intercept) 15105.5  122.90
 Residual               711.9   26.68
Number of obs: 202, groups:  TOTALSA, 10

Fixed effects:
             Estimate Std. Error t value
(Intercept) 1.298e+02  4.520e+01   2.870
OilPrice    1.360e+00  4.106e-01   3.313
GoogleTrend 7.110e-01  1.594e-01   4.460
new_cases   3.365e-04  3.193e-04   1.054

Correlation of Fixed Effects:
            (Intr) OilPrc GglTrn
OilPrice    -0.409
GoogleTrend -0.231 -0.018
new_cases   -0.302  0.197  0.173
```

Monthly Data

We make the prediction on the testing dataset and compare it with the real close price.

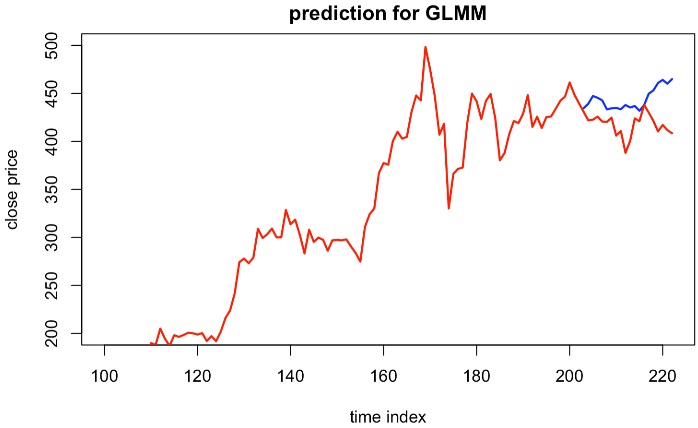

Figure 14: Predict close prices for GLMM model

Collinearity

To address the collinearity problem, Principal Component Analysis is used. We first center and scale the data. Then we plot the total variance proportion explained by the principal components.

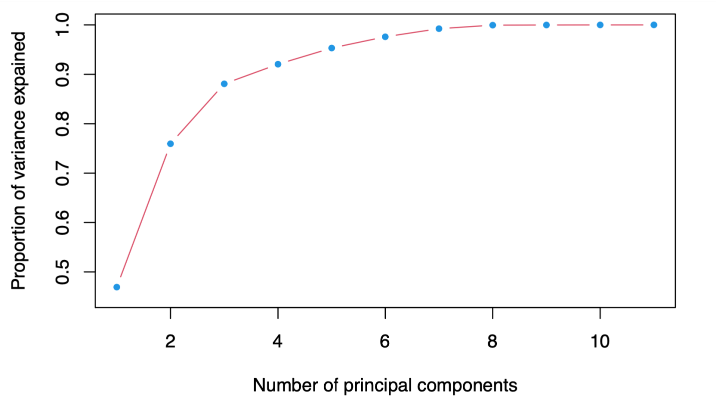

Figure 15: Variance proportion explained by different principal components

Collinearity

Besides, we also give the numeric value for the cumulative proportion of total variance explained by each component.

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     2.2719 1.7868 1.1550 0.6609 0.59955 0.50071 0.42512 0.28016 0.05867
Proportion of Variance 0.4692 0.2903 0.1213 0.0397 0.03268 0.02279 0.01643 0.00714 0.00031
Cumulative Proportion  0.4692 0.7595 0.8808 0.9205 0.95313 0.97593 0.99236 0.99949 0.99980
                          PC10    PC11
Standard deviation     0.04045 0.02259
Proportion of Variance 0.00015 0.00005
Cumulative Proportion  0.99995 1.00000
```

Then we try to explore whether PCA contributes to addressing collinearity. We plot the correlation pair plot.

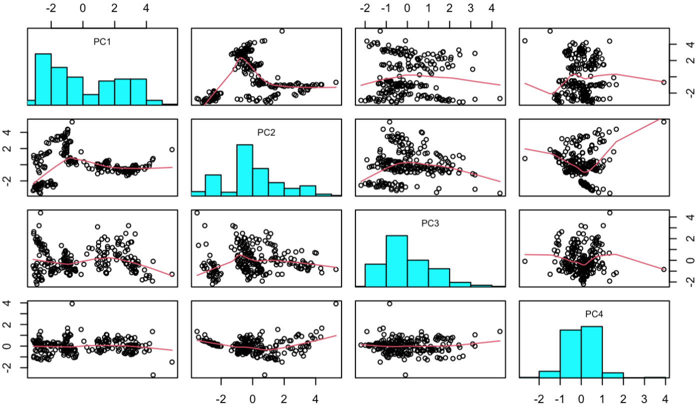

Figure 16: Pair plot for correlation between PCA components

Time Series

Single *Close Price* Variable

- Simple Exponential Smoothing (SES)
- Holt's Method
- Holt-Winter exponential trend
- Year Trend and Seasonality
- Seasonal Autoregressive Integrated Moving Average (SARIMA)
- Autocorrelation(AR)

Multiple Variables

- GLMM
- ARIMA
- VectorAutoRegression (VAR)

All the models are fitted using train-test spilt

Simple Exponential Smoothing

Figure 17: SES Model Forecast

Holt model

Holt's Method with linear and exponential trend

Figure 18: Holt's Method Model Forecast

Holt-Winter exponential trend, addition-addition and addition multiplication

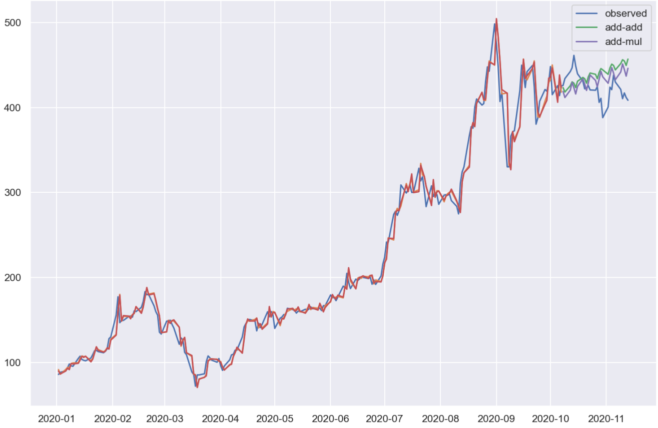

Figure 19: HWES Model Forecast

Seasonal Autoregressive Integrated Moving Average

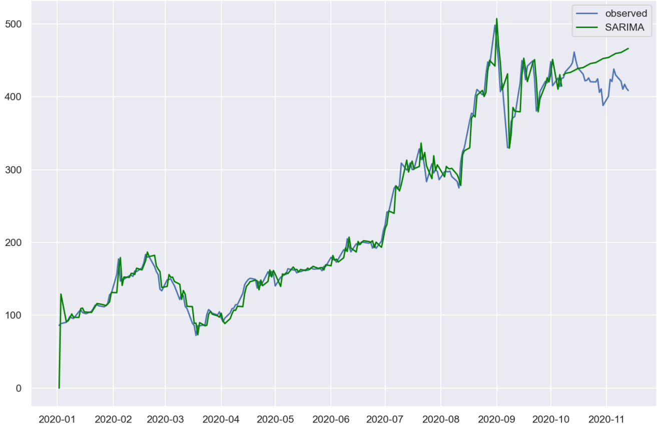

Figure 20: SARIMA Model

AR model

- First we seek AR model for help
- We tried AR(10) to solve the correlated errors

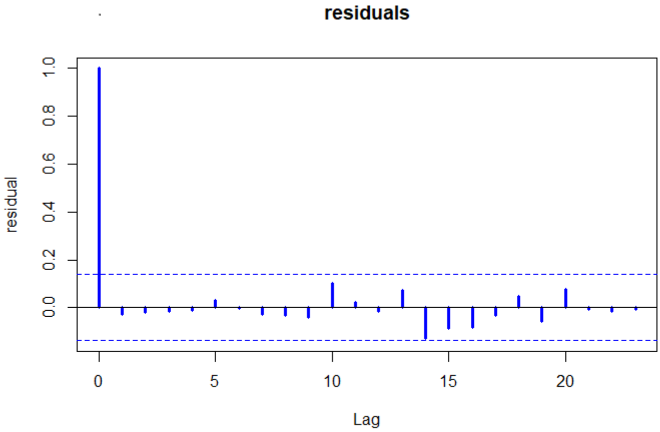

Figure 21: ACF plot

AR model

- The test using data spliting

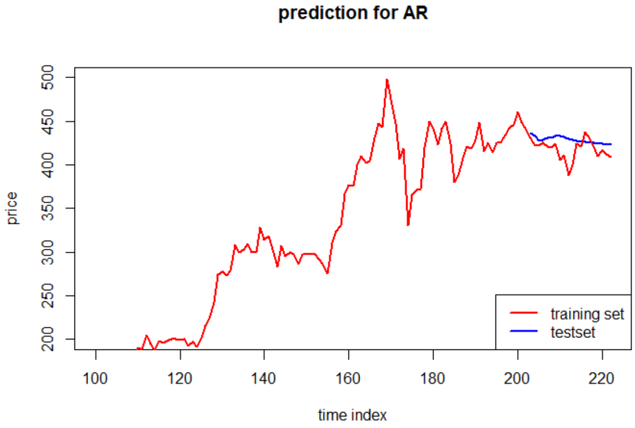

Figure 22: prediction vs real price

GARCH model

- The residuals of the mean model of a time series is a_t, The ARCH model is used to solve the heteroscedasticity of the time series. Though we solved correlated errors, the variance of the residuals is still not a constant.

$$\begin{aligned} a_t =& \sigma_t \varepsilon_t \\ \sigma_t^2 =& \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2 \end{aligned} \tag{1}$$

- The GARCH model is the generalized ARCH model, the residuals of mean model of time series a_t follows

$$a_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2 \tag{2}$$

GARCH model

- By boxtest, the residuals for $close - mean(close)$ shows that there is ARCH effect

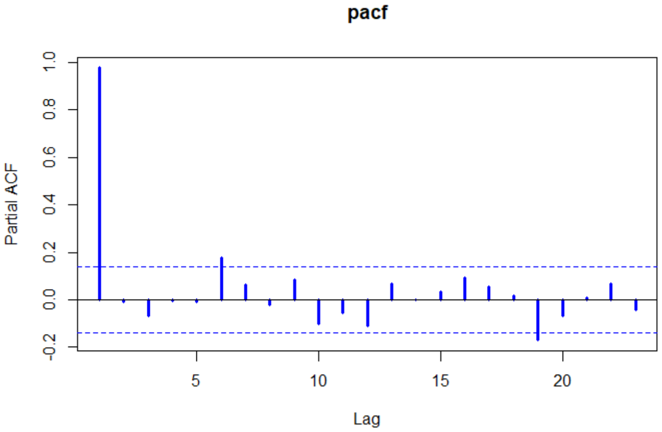

Figure 23: Pacf plot of a_t^2

GARCH model

- From the pacf plot, we can set the order in GARCH to be 1 and construct the model.

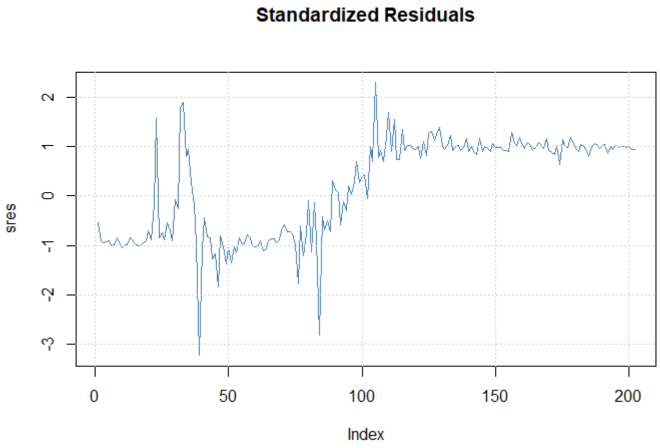

Figure 24: Residuals plot for garch model

ARIMA model

- The time series of the close price may not be stationary, so we need to further check if ARIMA model is necessary.

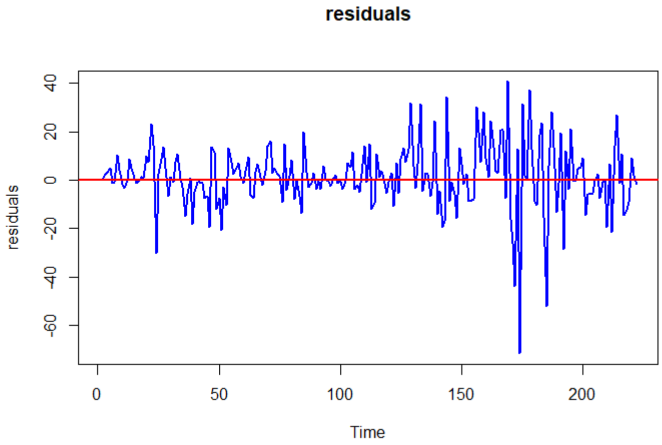

Figure 25: Residuals for arima model

ARIMA model

- The acf plot shows the correlated errors problem is solved

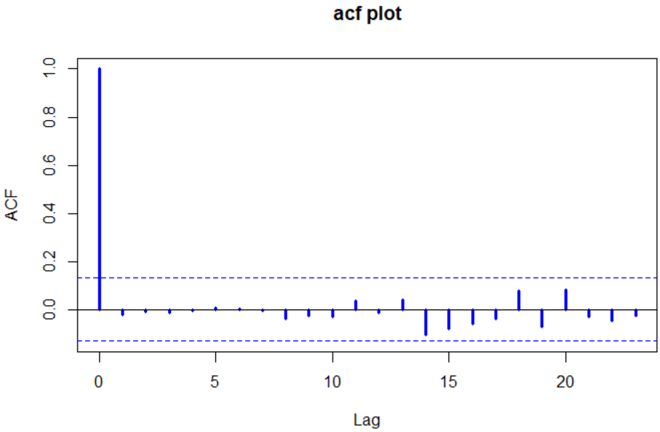

Figure 26: Acf plot for arima model

ARIMA model

- The test using data spliting

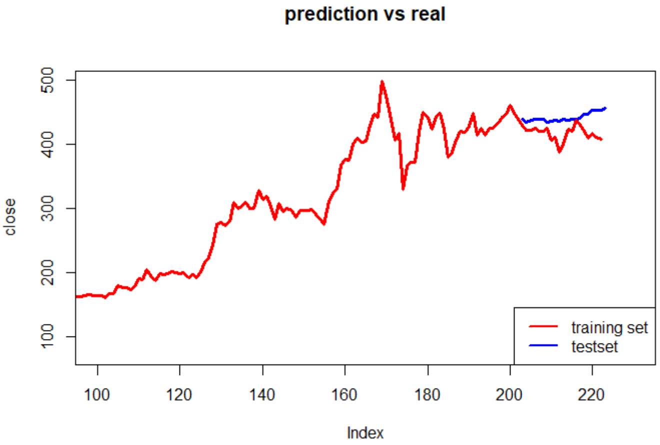

Figure 27: prediction vs real price

VAR model

- The test using data spliting

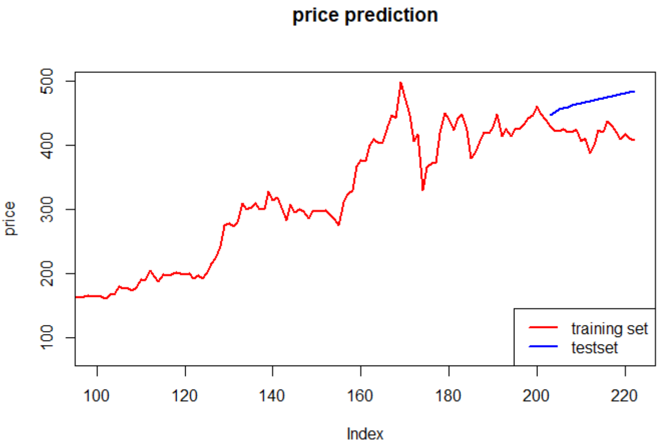

Figure 28: prediction vs real price

Overview

Close Price Analysis (univariate time series)
Collect Data
MLR analysis
monthly data
collinearity
heteroskedasticity (time series)
time series
multi variable
single variable
GLMM
PCA
GARCH
single variable
multi variable
Test Train Splitting
Fit Model
Evaluate and Compare the models
Final Model Discussion

Model Comparison

All the models following

- **Method**: Train/Test Spilt
- **Criteria:**

$$y_{test} = \text{original test set}$$
$$y_{pred} = \text{predicted values from fitted models for the test set}$$
$$score = \sum (y_{pred} - y_{test})^2$$

Model Score

Final Model: AR!

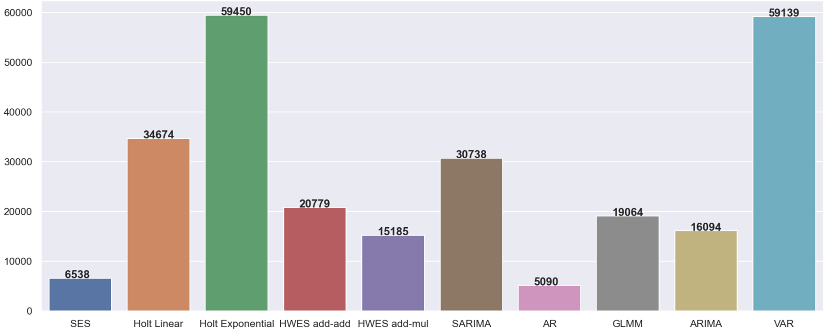

Figure 29: Different Model Score

Discussion of Final Model

Potential issues for large datasets.

- Relative small sample size
- Stationary violation
- Fitting time

Reference

1. Peking University, https://www.math.pku.edu.cn/teachers/lidf/course/fts/ftsnotes/html/_ftsnotes/fts-var.html#var-mod
2. Kaggle, https://www.kaggle.com/jutrera/stanford-car-dataset-by-classes-folder
3. PennState, Eberly College of Science https://online.stat.psu.edu/stat501/lesson/
4. nwfsc-timeseries, https://nwfsc-timeseries.github.io/atsa-labs/sec-tslab-moving-average-ma-models.html

Thanks for your listening