Gordon Lu

STAT 1223

Professor Nelson

02 April 2021

STAT 1223: Assignment 3

I seek to employ a multiple regression, specifically, I will run a generalized ARMA(p,q) model on the data. Eventually, this model will be upgraded to a logistic regression model based on a modified ARMA(p,q) model. The model will be of the following form:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

It is important to note that each of the X_{t-i}'s represent the price of a given stock at a certain period of time. With this model in mind, I seek to find the optimal periods in time to predict the price of a stock. Furthermore, once the parameters for the ARMA(p,q) model are found, this will be combined with other exogenous variables such as the volume of the stock, and the market beta. I seek to explore the relationships between the current price of a stock and factors like current news, and the sector the stock is in.

The model is in no way set in stone. The list of improvements to the regression model in mind are endless. One notable change to be implemented is the type of regression, as there are drawbacks in using ARMA(p,q) models, and other candidates such as LSTM regression, SVM regression, and panel data regression are candidate models. One option to explore in the future is to use the ARMA(p,q) model as a benchmark, and form regression models using other choices,

as well as incorporate other exogenous variables into each respective model. Perhaps performing something akin to cross-validation, treating different models as hyperparameters, and tuning which model would be the best to use at each period of time would be optimal. What is notable is that once the regressors are selected through the respective models, I plan to use stepwise regression to determine which regressors are the most important. I will use adjusted R^2 and BIC, so to penalize models with more parameters, and to prefer more significant terms in the respective regression models.

I seek to use R to analyze the data. Beyond metrics, I will look a residual plots, ACF plots, time series plots, and diagrams of the roots of the lag polynomials generated from the ARMA(p,q) model. I plan to use a slice of a stock from January 2020 to April 2021. I will measure the predictive accuracy of the regression by testing the labels that the regression model generates at each point in time to the actual results. I will accordingly use a test set, training set, and a validation set, and across different regression models determined by cross-validation I will compare the predictive accuracies of each model, and report which of the models yielded the highest accuracy and lowest error.

What I seek to learn is what variables influence the price of a stock, and given such variables, put together in a regression model, how well do said variables influence buying or selling decisions. If the result of overall regression turns out to be successful in predicting buying or selling, people would be able to boot up their computers, run a program, and watch their pockets grow (at some rate) and enjoy the show. It's a win-win, both the general population would get to enjoy their wealth increase, and I would learn a great deal more about how stocks and time series work more closely. I seek to use the quantmod package in R in order to collect data regarding stocks. In the quantmod package, it retrieves data from a variety of online sources

such as Yahoo Finance, Google Finance and OandA, each stock records the current listing for its price, as well as its historic data and other logistics such as the PE ratio, beta value and volume. The amount of data collected for a given stock from the quantmod package will vary depending on how frequent the price is updated as well as the date it was listed on the stock market.

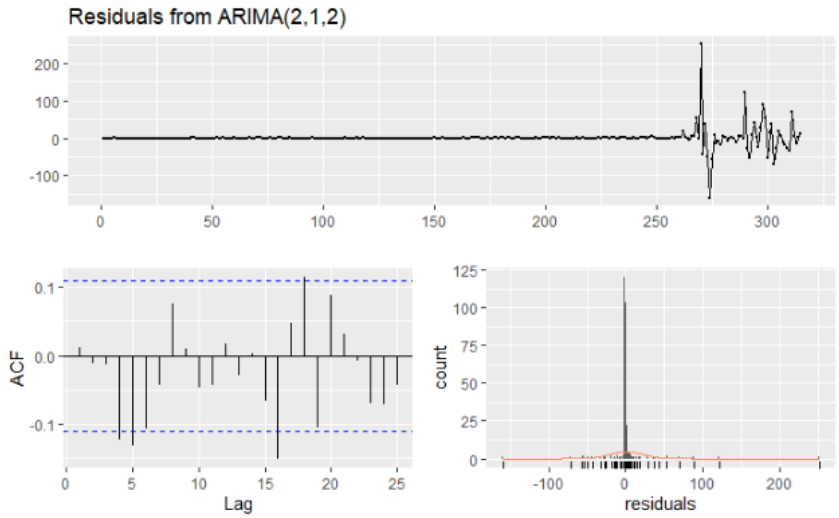
Overall, consolidating whether or not to buy a stock into a single response is the goal. The response, rather than being what the price of a stock could be, I seek to instead use a binary variable that will determine whether buying a stock at a certain day is worth the investment. I will call the response, "buy_stock". In an attempt to account for the potential influences on "buy_stock", I will consider the following predictor variables:

- stock_price: Quantitative variable determined by running the ARIMA model on the data, and will yield the optimal times to include for the price that will minimize volatility. The times will be based on days, so stock_price_1 would be yesterday's price.

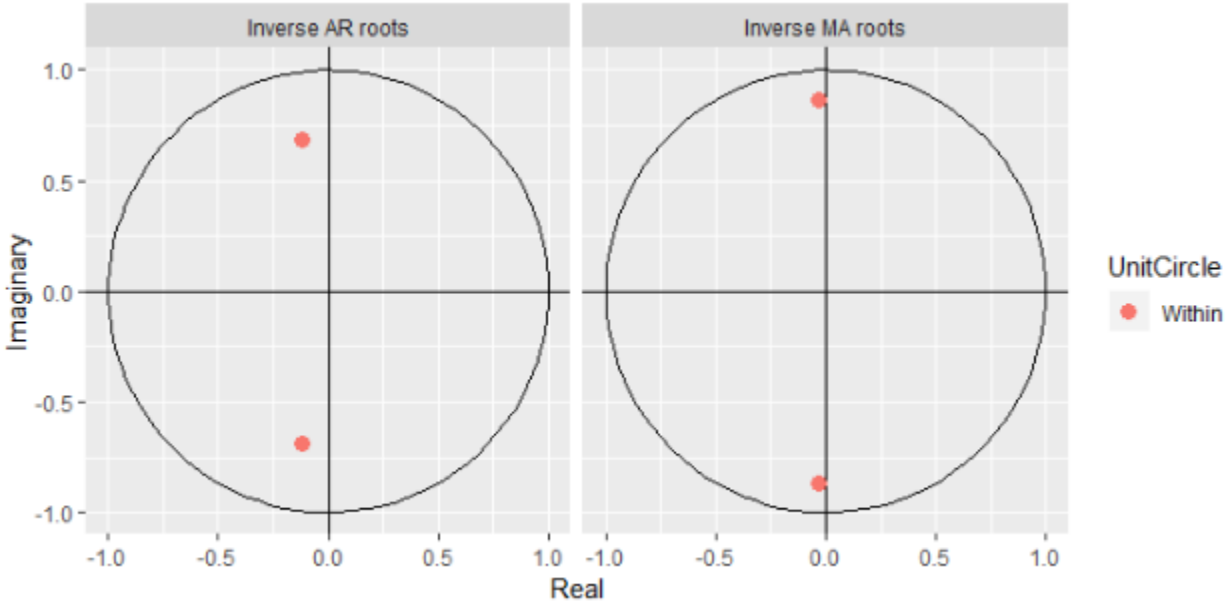
Currently, the big struggle is incorporating exogenous variables while maintaining the nice properties of the ARMA(p,q) model.

Using a thorough process of determining the optimal values, it was deduced that using an ARIMA(2,1,2) was optimal in prediction for a given stock, GameStop from January 1st, 2020 to the current date, April 4th 2021. Furthermore, a Ljung-Box test was conducted to determine whether regressors were autocorrelated. The Ljung-Box test yielded a p-value of 0.00736, which allows us to conclude that the data is not independently distributed, in other words, they exhibit serial correlation, in other words, the data can be expressed using a time series, as the errors will

be heteroscedastic rather than the typical assumption of OLS regression of the errors to be homoscedastic. Another important result is seen from looking at the below residual plots.



Based on the ARIMA errors, it is apparent that the errors represent a white noise series, and the residuals are not significantly different from white noise, thus this adds more to the point that a time series is feasible to use. This implies that the ARIMA(2,1,2) is a stationary process. Additionally, looking at the roots of the lag polynomial from the ARIMA(2,1,2)



it is apparent that all the roots lie within the unit circle, thus the process is stationary. Upon using an infinite geometric series to estimate the coefficients of the lag polynomial, the ratio will converge, and yield a number in the unit circle from the MA and AR parts of the ARIMA model. Thus, the process is stationary and can be approximated using a linear time series. In the future, I will likely use panel data regression and possibly a recurrent neural network to predict and determine the accuracy and error of each respective regression model. Furthermore, I will use benchmarks to determine how to formulate a reasonable way to assign labels to a day in time. In other words, I will try to determine a good way to use logistic regression to predict whether to buy or sell a stock based on several factors such as previous prices and exogenous variable independent of time such as PE ratio and sector.