Unit 8 Case Study

MSDS Fall ‘19

7333 Quantify the World

Daniel Serna, Bruce Granger, and Brandon de la Houssaye

October 20, 2019

**Abstract**

In this paper we provide an in-depth time-series analysis on the four-year historical price of Walmart stock. We use both an intuition based approached and brute-force approach to determine optimal parameters for an ARIMA model.

**1 Introduction**

We have pulled four years of stock price data for Walmart (WMT). Walmart is a publicly traded company incorporated within the United States.

The company is a retailer with retail operations in well over 35 countries throughout the world from both a retail and services standpoint. The company is also globally active in its supply chain footprint and directly or indirectly responsible for vast quantities of imports/exports in a number of countries.

Walmart is a Fortune 100 company and the largest private employer in the world that operates in all 50 US states. The company first went public on Oct 1, 1970 with a stock price of $16.50 per share.

The company is often regarded as a bellwether of US financial health and growth amongst other things. For example, the 'Walmart effect' can refer to patterns utilized by the CDC for tracking flu outbreaks as well as macro consumer habits.

However, Walmart's stock price is clearly not singularly impacted by the company's performance or the US economy overall. Factors such as investor sentiment and competitor impacts may have a severe impact on pricing. As a result, it is useful to consider Walmart's stock price change over time in the context of whether it is stationary, cyclical, random, etc.

To begin, we will consider 10 years of financial data (Figure 1) in order to fully assess the stock market price since the last major recession; the rapid rise of key competitors (e.g., Amazon); and other PR related impacts.

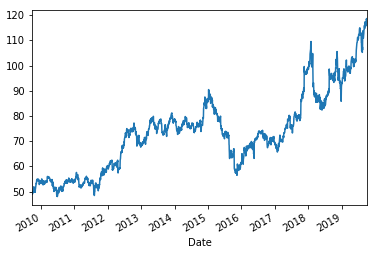


Figure - Ten Year Closing Price

**2 Data**

The team pulled four years of Walmart stock data from Yahoo Finance for analysis. While the data contained many interesting data points such as high, low, volume, close, adjusted close, the team focused on closing price for analysis. As can be seen in Figure 2, the four-year historical close price shows a positive linear trend with a spike in first quarter 2018.

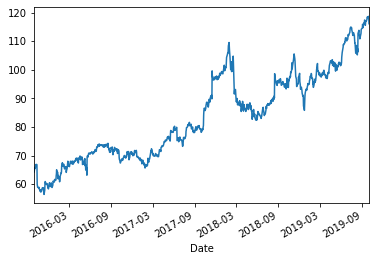


Figure - Four Year Closing Price

The data is well formed and did not contain missing data points. When tested for stationarity, we can see the test-statistic is greater than the critical value as seen in Figure 3.

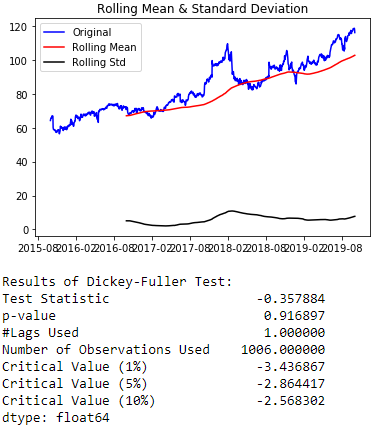


Figure - Dickey-Fuller Test

The above information shows us something very basic. The Walmart data is not stationary. There is an upward trend. There is clearly some cyclicality, or randomness in the data. The rolling mean information overlays quite well with the actual mean information (showing patterns as opposed to extreme outliers) of trend. While the std deviation results appear flat, there are some jumps. It is not perfectly smooth which would indicate that the results carry some randomness. Thus, the data isn't stationary or purely seasonal.

**3 Data Cleansing**

Fortunately, the data is well formed and did not contain missing values which eased the data cleansing process. As mentioned, the data contained numerous data points, but the team only focused on the closing price for this time series analysis. The team created a data frame consisting of closing price and date for analysis. While no NA’s were noticed in the four-year data, the team decided to add “dropna()” function calls when interacting with the data to futureproof the analysis if an expanded time window was desired.

**4 Methods**

The team decided to adopt a two-prong approach when analyzing the data. First, the team used intuition in an attempt to determine optimal autoregressive (p), integrated (d), and moving average parameters (q) for an ARIMA model. A rule-based approach was followed as part of this analysis. Second, the team applied a brute force approach to determine these optimal parameters. Every combination of parameter was attempted and observation of the Residual Sum of Squares and Akaike Information Criteria for each attempt was used to determine the optimal parameters. Ultimately, the team combined analysis from both of these approaches to arrive at the optimal choice of p, d, and q parameter values.

**5 Results**

The team first looked at the autocorrelation and partial autocorrelation plots of the data as seen in Figure 4 and Figure 5 respectively. Autocorrelation is a form of statistical correlation (i.e., the measured strength in relationship between two variables) for time series data (which is what we are managing here). In effect, for the autocorrelation function, we are trying to measure the statistical relationship for time series data against time series data from a previous time step (or periods); which are called 'lags'. That is, the time series of data is correlated over (and over, depending on defined 'lags') against the same series data from a previous period(s).

Whereas, with partial autocorrelation, we get a summary of the relationship between time series sets from different times steps - same as for autocorrelation - but without observations at intervening (or non-compared) time steps. In other words, partial autocorrelation attempts to get a more complete picture of the entire population of time series data.

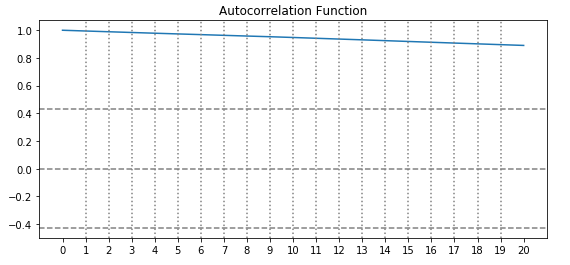


Figure – Autocorrelation

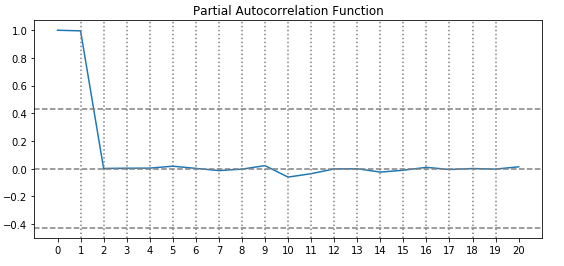


Figure - Partial Autocorrelation

What does this information tell us? First, it tells us that we definitely have non-stationary data. Second, it tells us that to further our analysis we are going to have to start incorporating differences (as well as, maybe, moving averages).

What follows next is an exploration of ARIMA with particular focus on hypertuning of p,d,q parameters. We will wrap up the ARIMAS exploration with a grid search of those same parameters in order to arrive at the best possible combination (as measured by Loss). We will also explore the metric of AIC - a basic accuracy score - in order to consider the parameter sets for the ARIMAS model. Ultimately, this all leads to the ability to decipher the past time events in order to understand/predict future events given time events that have occurred in a pattern over time that is not readily explainable outside of other, unprovided for variables.

We first run our ARIMA model with p, d, and q parameters all set to 0 as seen in Figure 6.

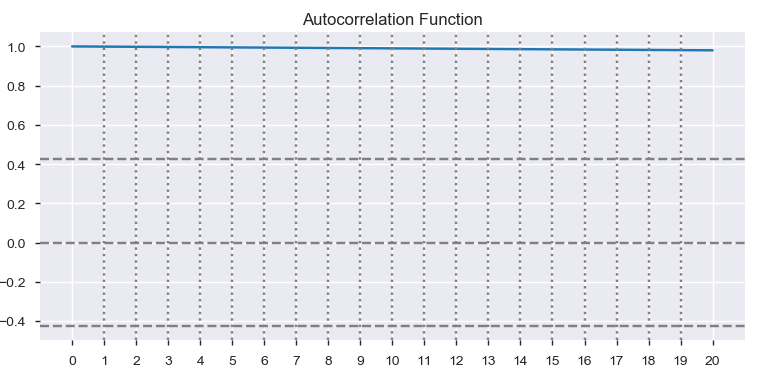


Figure - P=0, D=0, Q=0

Leveraging Rule #1, we observe that this series has positive correlations out to a high number of lags. For that reason, we will immediately re-run the model with 1 higher level of differencing, with results seen in Figure 7.

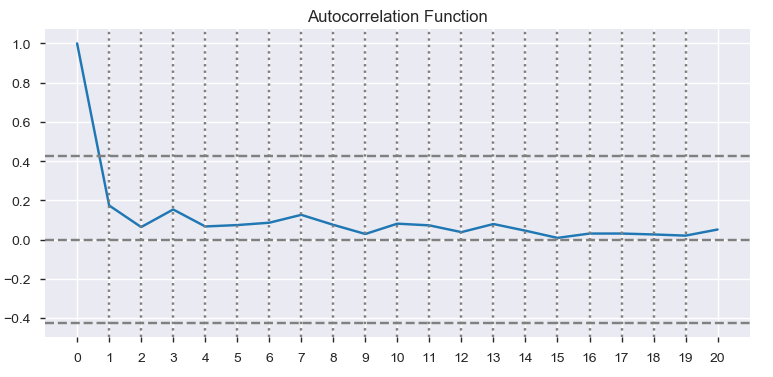


Figure - P=0, D=1, Q=0

Leveraging Rule #2, the lag-1 autocorrelation is not zero or negative, but the autocorrelations are all small and there doesn't seem to be much in the way of a discernable pattern. For purposes of really exploring this data, we will see what happens when a differencing of 2 is provided. We are doing this especially because we cannot see evidence of over differening after lag-1. These results are seen in Figure 8.

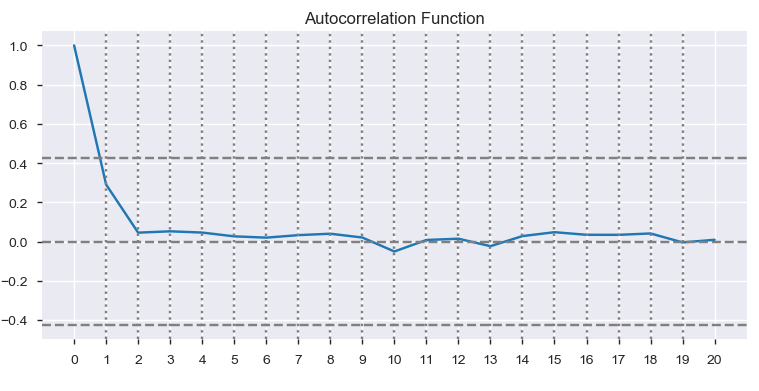


Figure - P=0,D=2,Q=0

These results are interesting. After lag-2, the autocorrelation seems to arrive at '0'. This would indicate that any additional differencing would severely run the risk of over differencing.

As for the other rules not yet covered (for parameter d), in using two orders of differencing (thus far) it is apparent that the original series has a time varying trend (Rule #4). Simply put, the model does not seem to be stationary or constant in trend. Rules 5 and Rules 3 do not necessarily apply: this model has at least 1 order of differencing so it does not require a constant. With respect to Rule 3, we will see how the potential of std results may fair with MA terms.

We next attempt adjusting the autoregressive parameter (p) leaving the other parameters are 0. We will attempt p=2 and the results are seen in Figure 9.

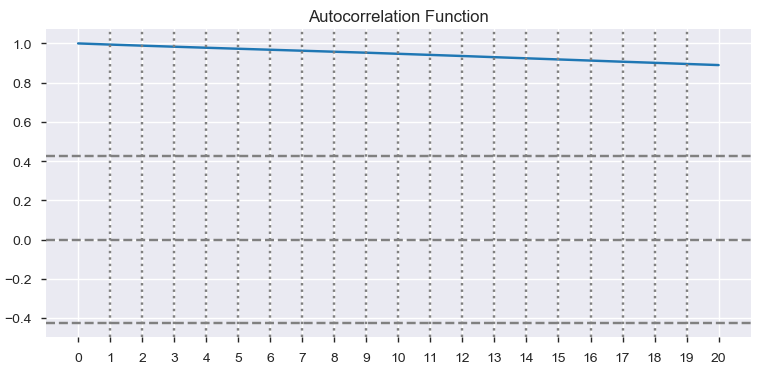


Figure - P=2, D=0, Q=0

This autocorrelation has the same pattern at the original autocorrelation function. We can infer from this that there is consistency in applied models. We can also infer - understandably - that whether the parameter is 1 or 2 the autocorrelation is occurring between the current time series observations and prior time series observations. Thus, the primary parameter of consideration really is d (differences). We will run the model again with p=2, d=2, and q=0 and these results can be seen in Figure 10.

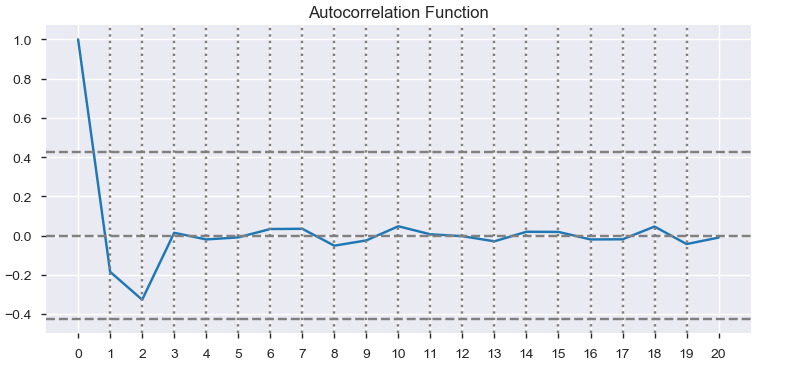


Figure - P=2, D=2, Q=0

Of course, combining the autoregression with the differences creates a problem. A problem that violates rule #7. The series is now over differenced. This is true whether p is 1 or 2. We will next run the model with all parameters set to 2 and the results are seen in Figure 11.

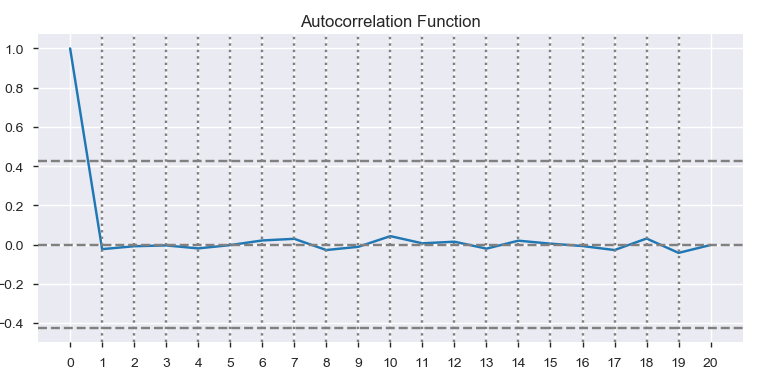


Figure - P=2, D=2, Q=2

Here we set AR and MA terms to equal each other. What is interesting is that, given prior variations of the model run, it is clear the terms cancel each other out. That is, it wouldn't matter if they were set to "2" or "1" in this case in terms of the overall model correlation. As a result, we will revert both parameters back to 0.

After going through various iterations of the ARIMA model parameters p,d,q, it appears visually that the most appropriate p, d, q, settings are either 0, 1, 0 or 0, 2, 0. It is possible a value of d=2 might be slightly over differenced and d=1 might be more appropriate.

In order to assess this conclusion, we will now perform a 'brute force' analysis using Loss scores (while also pulling AIC scores). The RSS and AIC values for each run are seen in Figure 12.

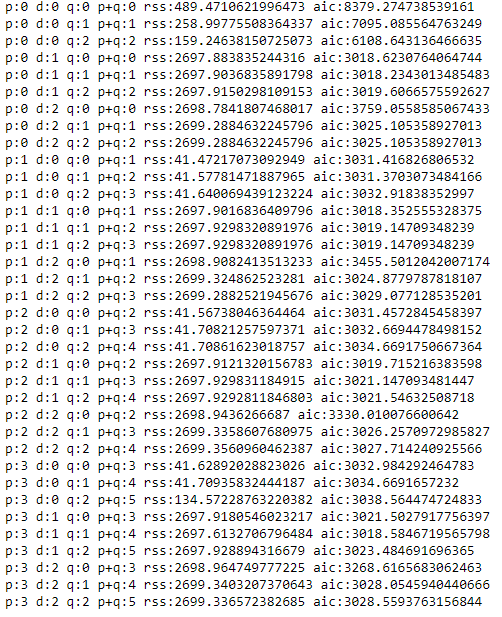


Figure - Brute Force Results

Using RSS values as the indicator of the best model, the optimal parameters determined by the brute-force method are seen in Figure 13.

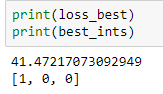


Figure - Optimal Parameters

These results are very interesting. We are looking at the loss function as the determinant for the best combination of p,d,q parameter set for the ARIMA model. The loss function used is Residual Sum of Squares. Running the grid search around these parameters (p,d,q) we arrive at a set of 1,0,0. This is of course very different from the parameter combinations we arrived at when manually tuning the model. The team appreciates that the parameters do in fact interact, and with those interactions the grid search is useful (where human intuition becomes limited).

However, the grid search model was not constructed with 'rules' to prevent overfitting of the model with the exception of putting into the function rule 8 (if (p > 1 and q <= 1) or (q > 1 and p <= 1) or (p <= 1 and q <= 1)). This all results in the necessity to utilize the results of both the grid search (to maximize accuracy) while also considering the manual fine-tuning (to ensure that the model is not overfitted - as considered through the autocorrelation visualizations against period lags).

With this combination, the team would not set the parameters of p,d,q at 1,0,0 because rules 1 and 2 would be immediately violated (i.e., the data clearly remains 'unstationary'). The team questioned the sole use of RSS in the brute-force method as sole determiner of best model. The team decided to also look at the AIC (Akaike Information Criteria) value in combination with intuition.

With the AIC values produced by the brute-force attempt in consideration, the team concludes that the correct parameter set would be 0,1,1. With that parameter set, all rules are met and the AIC score is minimized. Simply, the team focused on setting the difference parameter to its minimum where the rules could be met in order to maximize the accuracy. If the team only focused on RSS values, no difference parameter (d) would have been included.

The AIC value for our optimal parameter set of p=0, d=1, and q=1 is 3018.2343013485483.

**6 Conclusion**

The team found that a combination of intuition and brute-force parameter tuning produced the best results. Each method by itself produced sub-optimal parameter values. Our intuition based approached immediately demonstrated some order of differencing would be necessary, however, we failed to determine autoregressive or moving average parameters would be necessary. In contrast, the brute-force method identified an autoregressive parameter of 1 was necessary but failed to determine differencing was needed. Combining our intuition with the brute-force approach allowed us to determine looking at additional metrics might be prudent. By also looking at AIC values produced from the brute-force method, the team determined p, d, q values of 0, 1, 1 produced the optimal results while still adhering to rules.

**Appendix I Code**