Stat 153 Project

Nithin Raghavan

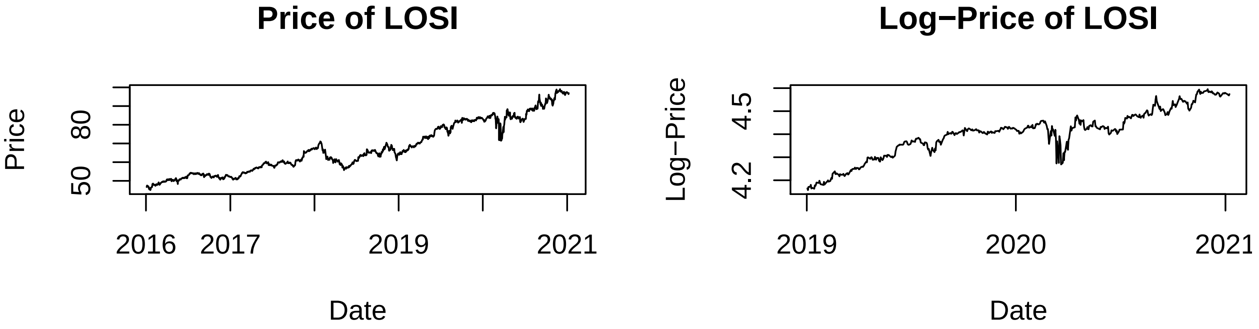
10 May 2021

Executive Summary

Despite the recent setback attributable to the onset of the COVID-19 pandemic, the stock price of Lots-of-stuff Incorporated has continued to rise at a substantial rate. According to my differencing model with ARMA(1,0) noise, the stock price will continue to rise at a linear rate for the next ten trading days, eventually attaining a value similar in magnitude to that of last November. If the price continues to rise at a similar rate as predicted by my model, the stock price will attain the highest value in its history.

Exploratory Data Analysis

The stock price for Lots-of-stuff Incorporated has seen an upwards, linear trend, as can be seen below in the left panel of Figure 1.


Price of LOSI

Log−Price of LOSI

Figure 1: Left: History of the stock prices of Lots-of-stuff Incorporated (LOSI) since the beginning of 2016. Right: The original time series data, truncated after 2019 and after a VST of $\log(x)$ is applied.

There does not appear to be any prominent seasonality, although it could be present in slight amounts. There is a one-off dip in early 2020 (due to the initial outbreak of COVID-19), and the company lost much value in the year 2018 (resulting in a net growth of zero during this year). The data is moderately heteroscedastic, so a variance-stabilizing transform (VST) of the initial time series such as $f(x) = \log(x)$ will be necessary. There was not much relative growth between the years 2016 and 2019, so a model attempting to use all the data to predict future values could end up under-predicting due to the impact of values before 2019. As such, we have chosen to ignore all the data before 2019 (and only use the dates in 2019-2021 for predictions), despite the dip due to COVID-19 occurring during this time. This is done in order to ensure that there will be enough data available to fit a parametric model. A plot of the modified data can be found in Figure 1 on the right panel. The trend of the graph appears linear, so first-order differencing could be another potential approach to modelling this series.

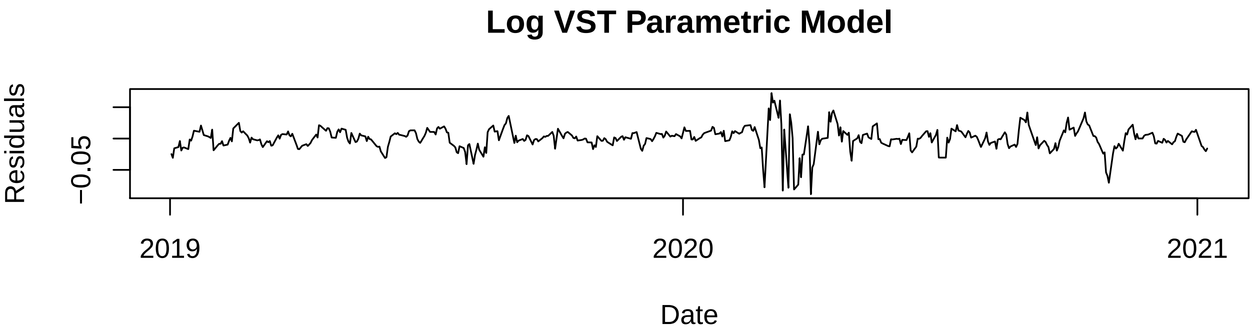


Figure 2: Residuals from a model fit using a parametric equation.

Models Considered

In order to pursue stationarity, two signal models have been chosen: a parametric model and a differencing model.

Parametric Signal Model

For the parametric model, we decided to model potential seasonality over each year (261 working days) and over each quarter (65.25 working days) using a cosine and sine term over each such frequency, despite the fact that this model does not appear to contain much signal. In addition, we use indicator variables over each month to model potentially different growth rates over each month, in addition to checking if the current timestamp occurs during the initial onset of the COVID-19 pandemic. Eq. 1 describes the formula used for the parametric model, where V_t refers to the variance-stabilized series series ($V_t = \log(Y_t)$, where Y_t is the original series), and $\mathbb{1}(x)$ is an indicator that returns 1 if x is true, and 0 else. The residuals of such a model are plotted in Figure 2. The linear trend is modelled by $\beta_1 t$. At a glance, this already appears less heteroscedastic than the residuals of the linear regression model above. There appears to be a large increase in variance in the beginning of 2020, which is larger than any variances previously seen. This is due to the onset of the pandemic brought on by COVID-19, which caused large decreases and other fluctuations in the stock price. Since this appears to be a one-time or rare event, which is essentially random with respect to the residuals, we can disregard the impact that this would have on the overall stability of the residuals. And without considering the impact of COVID-19 or the stochastic volatility of the stock itself, the residuals appear to be close to stable.

$$
\begin{aligned}
V_t = {} & \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{261}\right) + \beta_3 \cos\left(\frac{2\pi t}{261}\right) + \beta_4 \sin\left(\frac{2\pi t}{65.25}\right) + \beta_5 \cos\left(\frac{2\pi t}{65.25}\right) \\
& + \beta_6 t \sin\left(\frac{2\pi t}{261}\right) + \beta_7 t \cos\left(\frac{2\pi t}{261}\right) + \beta_8 t \sin\left(\frac{2\pi t}{65.25}\right) + \beta_9 t \cos\left(\frac{2\pi t}{65.25}\right) \\
& + \beta_{10} \mathbb{1}(\text{month} = \text{January}) + \cdots + \beta_{21} \mathbb{1}(\text{month} = \text{December}) \\
& + \beta_{22} t \mathbb{1}(\text{month} = \text{January}) + \cdots + \beta_{33} t \mathbb{1}(\text{month} = \text{December}) \\
& + \beta_{34} \mathbb{1}(\text{COVID-19}) \sin\left(\frac{2\pi t}{261}\right) + \beta_{35} \mathbb{1}(\text{COVID-19}) \cos\left(\frac{2\pi t}{261}\right) \\
& + \beta_{36} \mathbb{1}(\text{COVID-19}) \sin\left(\frac{2\pi t}{65.25}\right) + \beta_{37} \mathbb{1}(\text{COVID-19}) \cos\left(\frac{2\pi t}{65.25}\right)
\end{aligned}
\tag{Eq. 1}
$$

Parametric Signal Model with ARMA(1,0)

The ACF and PACF plots of the residuals of this parametric model can be seen in Figure 3. Note that as the number of lags increases, the ACF plot forms an oscillating pattern around the horizontal axis, which is indicative of an autoregressive (AR) structure with more than one parameter. The PACF shows a value of large magnitude at $h = 1$ on the horizontal axis, and values of reduced magnitude for $h > 1$. These two observations indicate that there is no moving average (MA) term, and could lead to two possible AR proposals.

Due to the sharp cutoff in terms of magnitude of the PACF values, we propose $p = 1$ and $q = 0$ as a potential fit. Figure 7 illustrates the Ljung-Box plot for the entire dataset (April 2020 - Dec 2020), and shows that the Ljung-Box p-values are quite low until $h = 250$, and then jump up. This occurs because of the high variance caused by the onset of the COVID-19 pandemic, which ends at approximately this lag. Thus, when considering all the data after the high variance from COVID-19 dies down, the Ljung-Box p-values for the residuals of the parametric model show a very good fit.

Parametric Signal Model with ARMA(2,0)

The oscillation of the ACF values could indicate that there is an autoregression with two AR parameters, with the second being small and negative, so we propose $p = 2$ and $q = 0$ as a potential fit. Figure 7 illustrates the Ljung-Box plot for the entire dataset (April 2020 - Dec 2020), and shows that the Ljung-Box p-values are quite low until $h = 250$, and then jump up for the same reasons as before. When considering all the data after the high variance from COVID-19 dies down to the end (after $h = 250$), the Ljung-Box p-values for the residuals of the parametric model show a very good fit.

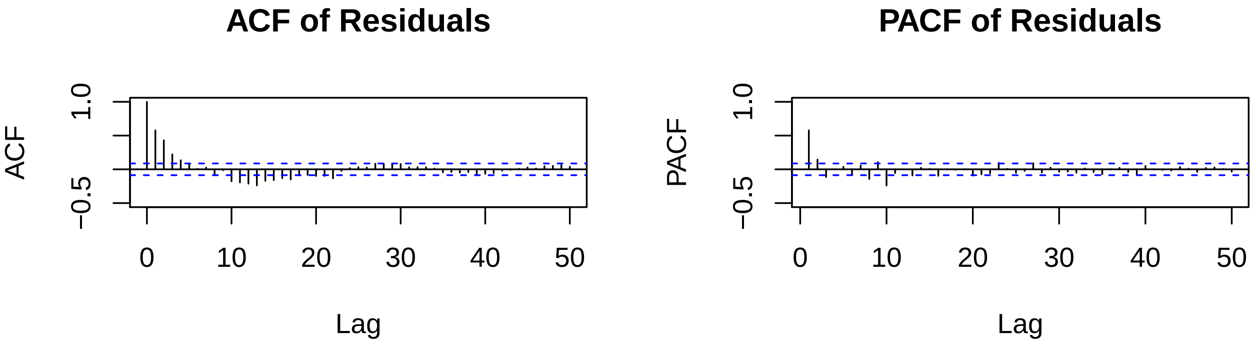


Figure 3: Autocorrelation function (ACF) and partial autocorrelation function (PACF) for the residuals of the parametric model.

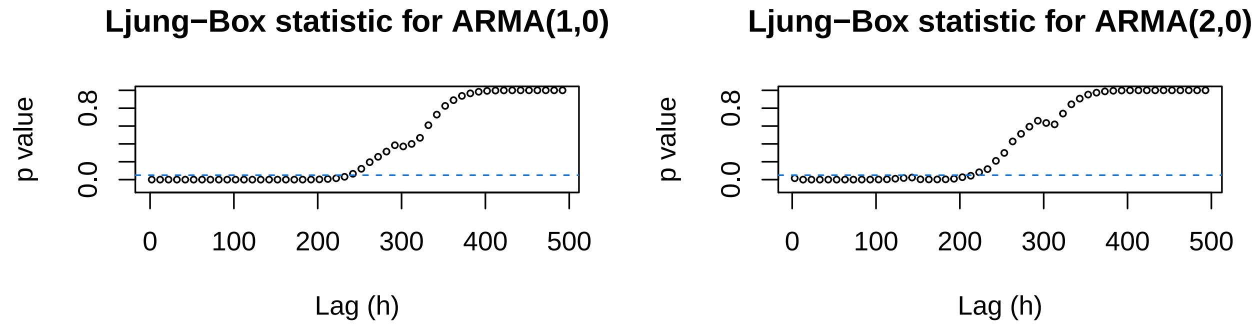


Figure 4: The parametric residual p-values for the Ljung-Box statistic for the whole dataset (Jan 2019 - Dec 2020) for ARMA(1,0) (left) and ARMA(2,0) (right).

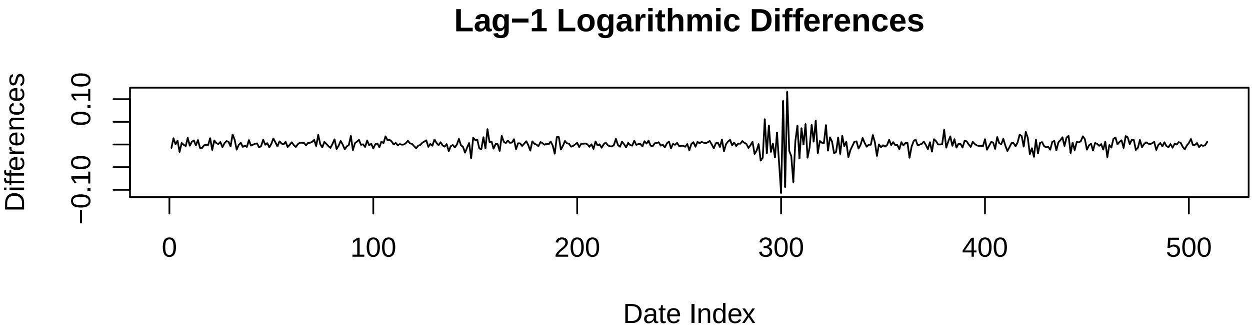


Figure 5: Residuals from the first difference of the logarithm of the original time series, similar to the percent change transform (if the changes are small).

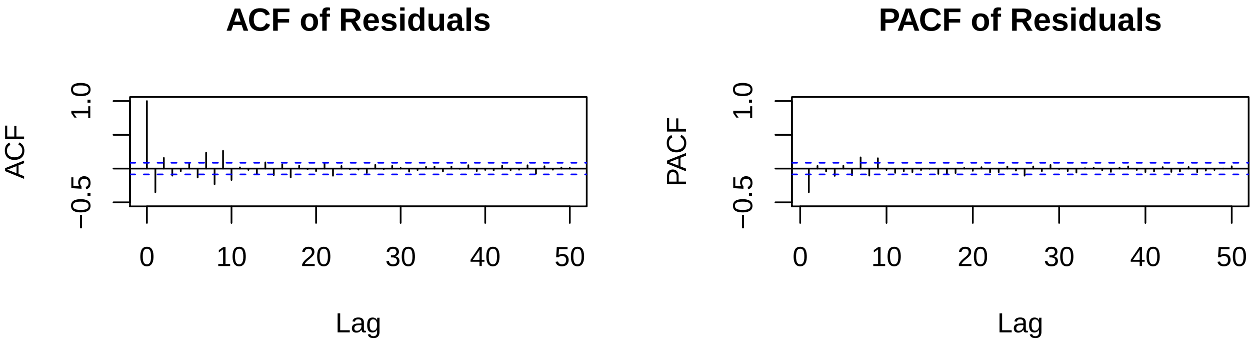


Figure 6: Autocorrelation function (ACF) and partial autocorrelation function (PACF) for the residuals of the differenced model.

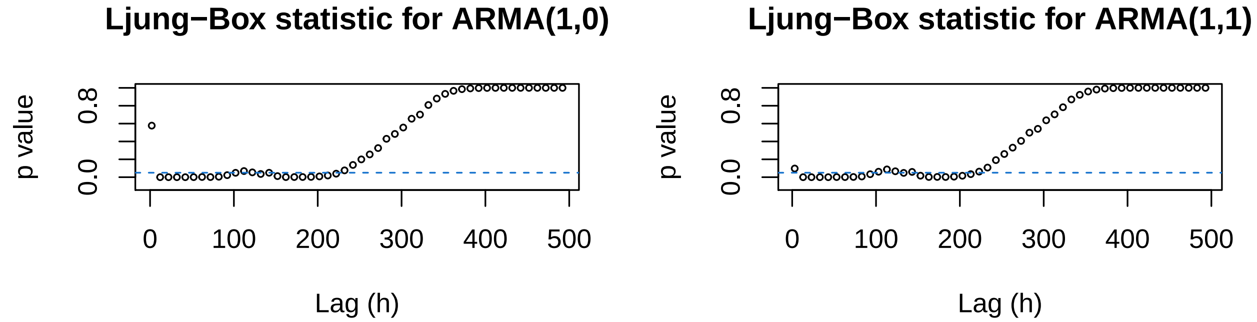


Figure 7: The differenced residual p-values for the Ljung-Box statistic for the whole dataset (Jan 2019 - Dec 2020) for ARMA(1,0) (left) and ARMA(1,1) (right).

First-Differencing

We use a first difference using the logarithm VST, which is similar to the percent change transform assuming the change between Y_t and Y_{t-1} is small:

$$\nabla \log(Y_t) = \log(Y_t) - \log(Y_{t-1}) = \log\left(\frac{Y_t}{Y_{t-1}}\right) \approx \frac{Y_t - Y_{t-1}}{Y_{t-1}}$$

Since the trend is linear, this first difference is expected to remove it (since there is not much signal in the model, we ignore seasonality). Figure 5 illustrates the residuals of such a model, which appear quite stable.

First-Differencing with ARMA(1,0)

The ACF and PACF plots of the residuals of this differenced model can be seen in Figure 6. Note that as the number of lags increases, the ACF plot exponentially decreases around the horizontal axis, which is indicative of an AR structure. The PACF shows a value of somewhat large magnitude at $h = 1$ on the horizontal axis, and values of reduced magnitude for $h > 1$. These two observations indicate that there could possibly be one moving average term, and could lead to two possible proposals.

Due to the sharp cutoff in terms of magnitude of the PACF values, there could potentially be no moving average term, and so we propose $p = 1$ and $q = 0$ as a potential fit. Figure 7 illustrates the Ljung-Box plot for the entire dataset (April 2020 - Dec 2020), and shows that the Ljung-Box p-values are quite low until $h = 250$, and then jump up for the same reasons as before. When considering all the data after the high variance from COVID-19 dies down to the end (after $h = 250$), the Ljung-Box p-values for the residuals of the differenced model show a very good fit.

First-Differencing with ARMA(1,1)

If the moving average term is small, then it could result in fast exponential decay and show a similar plot, so we propose $p = 1$ and $q = 1$ as a potential fit. Figure 7 illustrates the Ljung-Box plot for the entire dataset (April 2020 - Dec 2020), and shows that the Ljung-Box p-values are quite low until $h = 250$, and then jump up for the same reasons as before. When considering all the data after the high variance from COVID-19 dies down to the end (after $h = 250$), the Ljung-Box p-values for the residuals of the differenced model show a very good fit.

Model Comparison and Selection

The four aforementioned models (parametric and differencing plus two different ARMA models each) are compared through time series cross validation. The nonoverlapping testing sets roll through the last 91 trading days in the data in ten-day segments, from 17 August 2020 to 23 December 2020 (there are thus 91 forecasted points over these ten-day windows). The training sets consist of all data that occur before the appropriate testing set. The models' forecasting performances are compared using the root-mean-square prediction error (RMSPE), and the model with the lowest RMSPE will be chosen as the model for making a forecast for the next 10 trading days.

Table 1: Cross-validated out-of-sample root mean squared prediction error for the four models under consideration.

	RMSPE
Parametric Model + ARMA(1,0)	18.933561
Parametric Model + ARMA(2,0)	18.906501
Lag-1 Differencing + ARMA(1,0)	1.913419

	RMSPE
Lag-1 Differencing + ARMA(1,1)	1.919008

Table 1 shows that the Lag-1 Differencing model with ARMA(1,0) has the lowest cross-validated forecast error as measured by RMSPE. The forecasts made by the parametric model, though more accurate near the end of the given data, performed more poorly than the differencing models when trained only on the first four-fifths of the data and hence resulted in a worse RMSPE score. Thus, the differencing model with ARMA(1,0) is the chosen forecasting model.

Results

To make a forecast for the next 10 trading days, we use lag-1 differencing on the data recorded after 2019 after a VST of $\log(x)$ is applied. This is similar to the percent change transform. There is an additive noise term X_t, which represents a stationary process defined by ARMA(1,0). Eq. 2 describes the relevant model of the equation, where V_t is the variance-stabilized time series ($V_t = \log(Y_t)$), t refers to the most recent recorded time, μ is the mean of the difference, and W_t is white noise with variance σ_W^2.

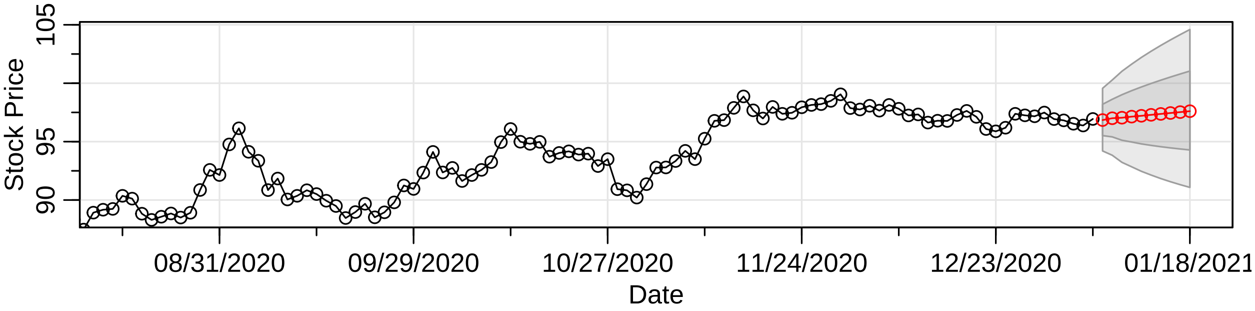
$$\begin{aligned} \hat{V}_t &= V_{t-1} + \mathrm{E}(\nabla V_t) + X_t + \mu \\ X_t &= \phi X_{t-1} + W_t \end{aligned} \qquad \text{(Eq. 2)}$$

Since the trend in the original time series was linear, $\mathrm{E}(\nabla V_t)$ can be estimated as just the average difference: $\widehat{\nabla V_t} = \frac{1}{t-2} \sum_{k=2}^{t-1} \nabla V_k$. This gives us a mechanism for predicting future values (any forecasts made using this model must be exponentiated to reverse the VST). ϕ and μ will be estimated in the next section.

Estimation of model parameters

Estimates of the model parameters are given in Table 2 in Appendix 1. It appears that the mean of the stationary process is relatively close to zero, and the white noise has low variance.

Prediction



Since a differencing method is used to predict, the next ten forecasts are located in a relatively straight line that represents an increasing, linear trend. The model predicts that the price will continue to rise and eventually overcome the small dip in price since November of last year. If the trend continues, then the price of the stock will eventually reach an all-time high. There will be no permanent dip in price during this time.

Appendix 1 - Table of Parameter Estimates

Table 2: Estimates of the forecasting model parameters for the lag-1 differencing model, with their standard errors (SE).

Parameter	Estimate	SE	Coefficient Description
μ	0.0008	0.0005	Mean
ϕ	-0.3505	0.0415	AR coefficient
σ_W^2	0.0001939		Variance of White Noise