STAT 153 Midterm Project

Introduction

Time series data is very common in real life, such as stock prices. Analysis is often performed on the time series data to extract meaningful characteristics, and to predict future values based on previously observed values. To this end, this report aims to accomplish the following:

- Identify the characteristics of the time series data provided;
- Form an appropriate ARIMA model to predict the values in the future.

Seasonality and linear trend were considered, and various ARIMA models are fitted, among which an ARIMA($p = 3, d = 1, q = 3$) model performed the best.

Exploratory Data Analysis

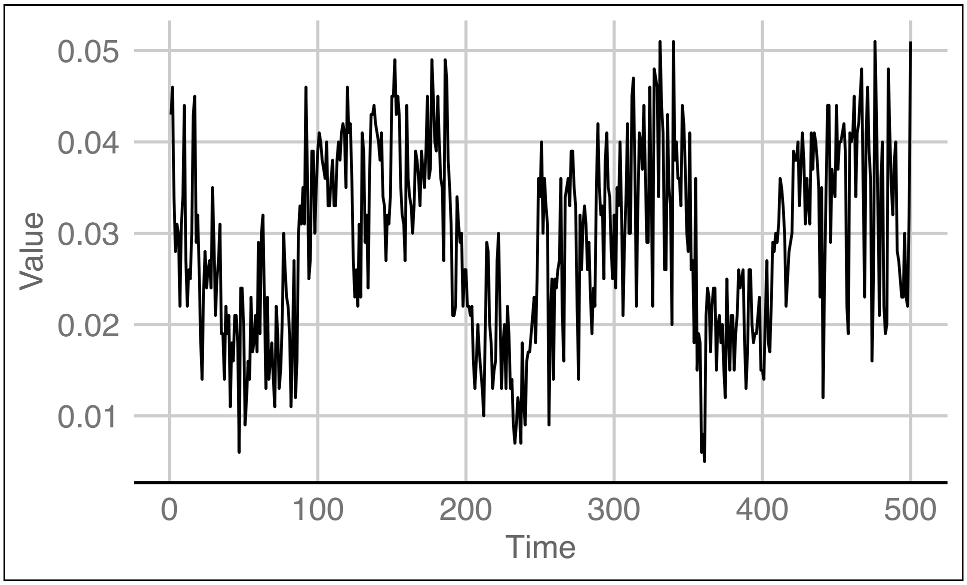

Figure 1: Original Time Series

To make this report easier to follow, we denote the original values in the time series dataset, data3, with $\{X_t\}$, where t is a discrete time point. Figure 1 shows that $\text{Var}(X_t)$ is pretty similar across different time point $t \geq 1$, so no power transformation is considered. There seems to be significant fluctuations present in the values of the data, so first order differencing is performed to remove this. We let $\{Y_t\}$ denote the differenced data, and $Y_t = (1 - B)X_t = X_t - X_{t-1}$. As shown in Figure 2, after this differencing, data becomes zero-meaned. Besides, there seems to be no trend or seasonal component present, so it is safe to conclude stationarity has been achieved at this point.

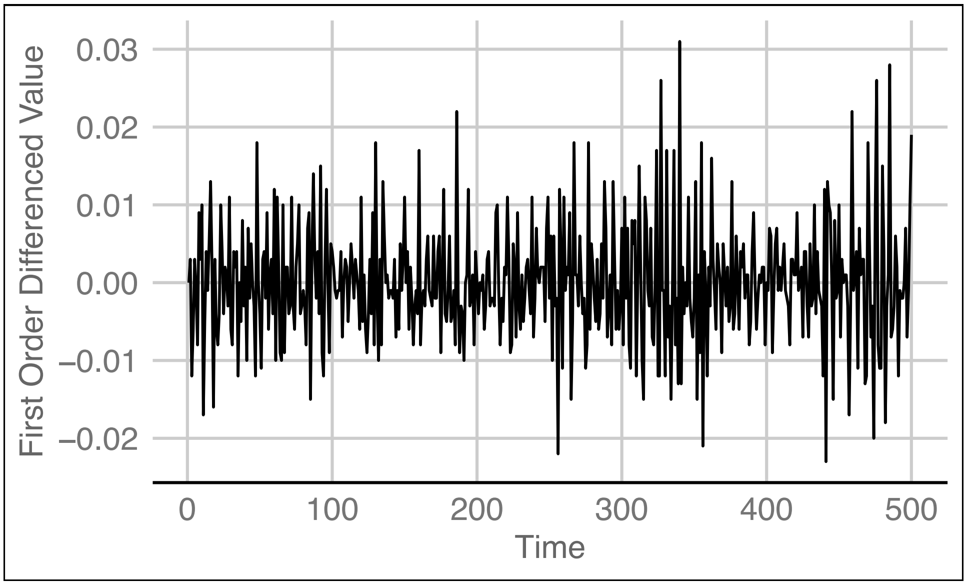

Figure 2: First Order Differenced Log Time Series

Identifying Suitable ARIMA Model

Both ACF and Partial ACF plots were used to form initial guesses of the possible model. Figure 3 shows that the individual autocorrelations seem to be weak, since the only sample autocorrelation r_k's that exceed the 95% confidence band by a lot are at lags 1 and 2. This suggests $q = 2$ for the ARIMA model. To compare model performances, an MA($q = 3$) component is also considered. This is because r_3 is also outside the 95% confidence band but only by a little, and in order to keep the model simple, $q > 5$ will not be considered unless there is seasonal component or significant ACF values after lag 5, which seem to be absent from this dataset.

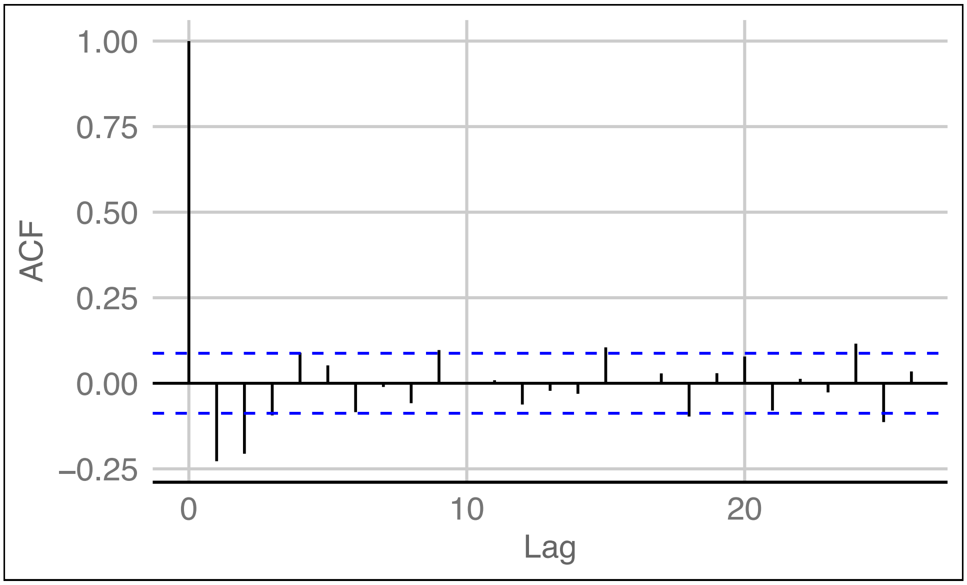

Figure 3: Correlogram of First Order Differenced Time Series

As for the Partial the ACF plot, Figure 4 suggests a similar pattern as ACF plot. Only values at the first three lags are outside the 95% confidence band by a lot. This observation leads to an initial guess of $p = 3$ for the ARIMA model. An AR($p = 2$) component is also considered for model comparison.

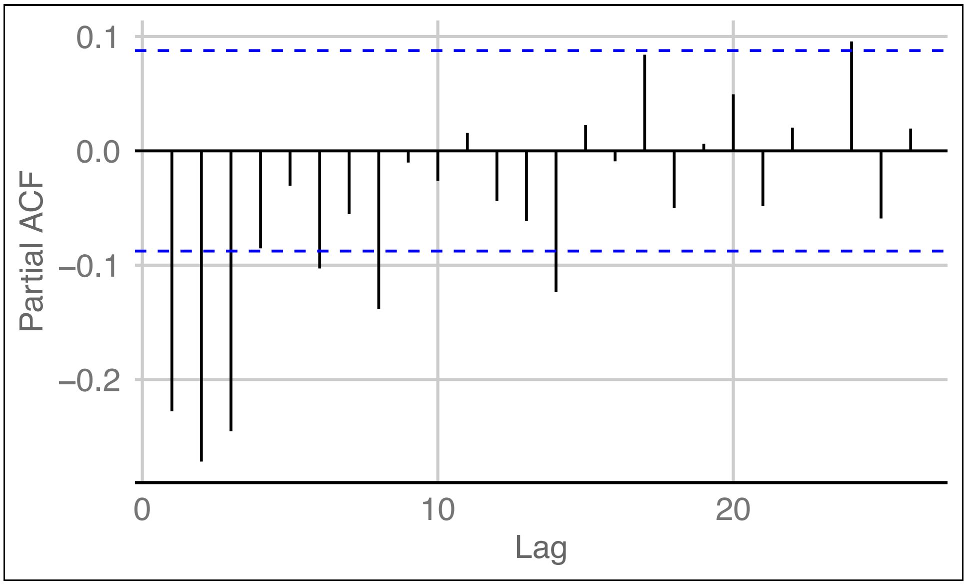

Figure 4: Partial ACF Plot of First Order Differenced Time Series

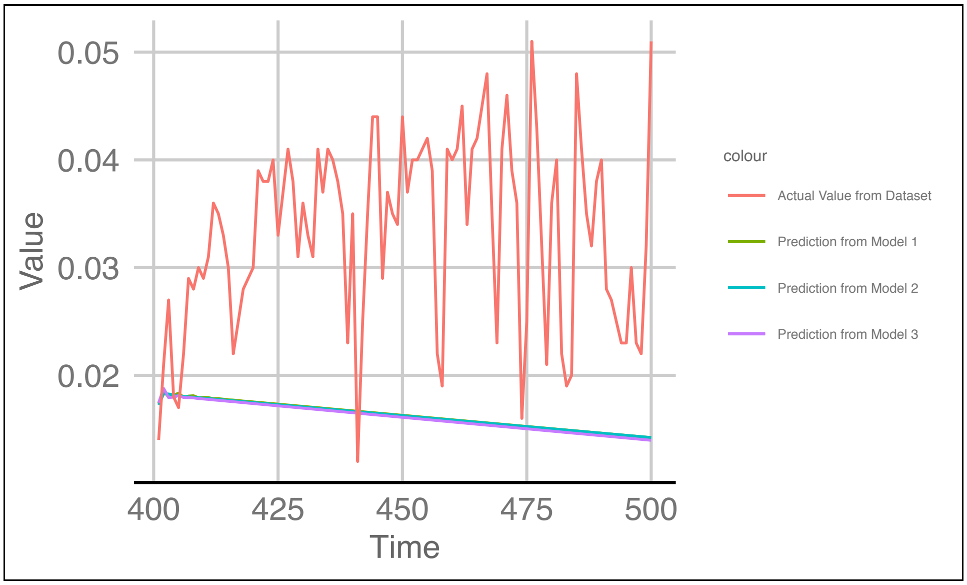

Figure 5: Actual vs. Predicted Time Series Data

Performance testing is done in order to determine the better model of the three proposed so far. Model 1 is an ARIMA($p = 3, d = 1, q = 3$) model, model 2 is an ARIMA($p = 3, d = 1, q = 2$) model, and model 3 is an ARIMA($p = 2, d = 1, q = 3$). An 8-2 train-test split is adopted, which means that the first 80% of the data is used as the training set, and the last 20% test set.

As Figure 5 shows, none of the predictions captures the real trend of the data. However, because there seems to be no clear seasonal component in the data itself, it is safe to conclude that these simple models are the best guesses that can be provided without further information on the strcture of this time series dataset. Nonetheless, cross-validation is performed in order to pick the best model.

Cross-validation is done with the same train-test split. Starting from x_{401}, which is the first datapoint in the test set, $x_{401}, \ldots, x_{500}$ are divided into groups of size 10 sequentially. All the data up to a certain test group is used to train the model, and is validated on the test group selected. Then average sum of squared errors are calculated from this process, and results are listed in the table below. The average sum of squared errors for model 1 is $7.60 \cdot 10^{-4}$, for model 2 is $7.68 \cdot 10^{-4}$, and for model 3 is $7.69 \cdot 10^{-4}$. Clearly, model 1 has the smallest average sum of squared errors, so cross-validation suggests that model 1 performs the best.

	Average Sum of Squared Errors
model1	0.000760
model2	0.000768
model3	0.000769

Looking at the model selection criteria listed in the table below, AIC and AICc values suggest model 1 is the best model, same as cross-validation. At the same time, BIC suggests model 3 is the best model. However, because the differences among these numbers are very small, these models are comparable in terms of all three model selection criteria. This table, combined with the CV errors from above, show that model 1 has the best overall performance.

	Model1	Model2	Model3	Best Model
AIC	-8.96	-8.95	-8.96	Model1
AICc	-8.96	-8.95	-8.95	Model1
BIC	-9.90	-9.90	-9.91	Model3

A look at the normal Q-Q plot of standardized residuals from ARIMA($p = 3, d = 1, q = 3$) in Figure 6 suggests that the residuals roughly follow a normal distribution, which shows little evidence of departure from the assumption of white noise.

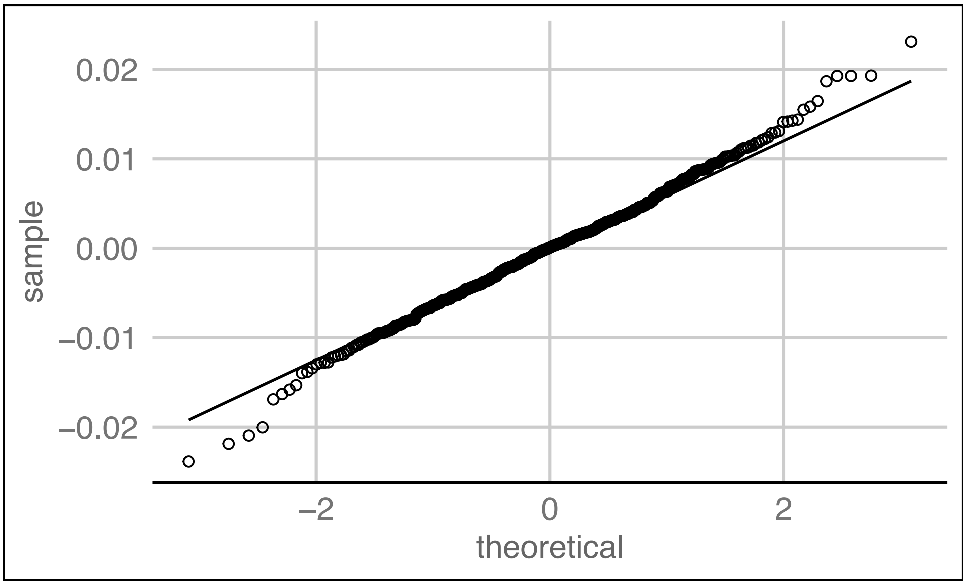

Figure 6: Normal Q-Q Plots of Standardized Residuals from ARIMA(3,1,3) Model

In order to examine the model fit of ARIMA($p = 3, d = 1, q = 3$), Ljung-Box test statistics are computed for lags 1 to 20. The large p-values for the Ljung-Box statistics from Figure 7 suggest that the groupwise residual autocorrelations are weak, indicating the failure of rejecting the null hypothesis that data $x_1, \ldots, x_{500}$ are generated from an invertible and causal ARIMA($p = 3, d = 1, q = 3$) process.

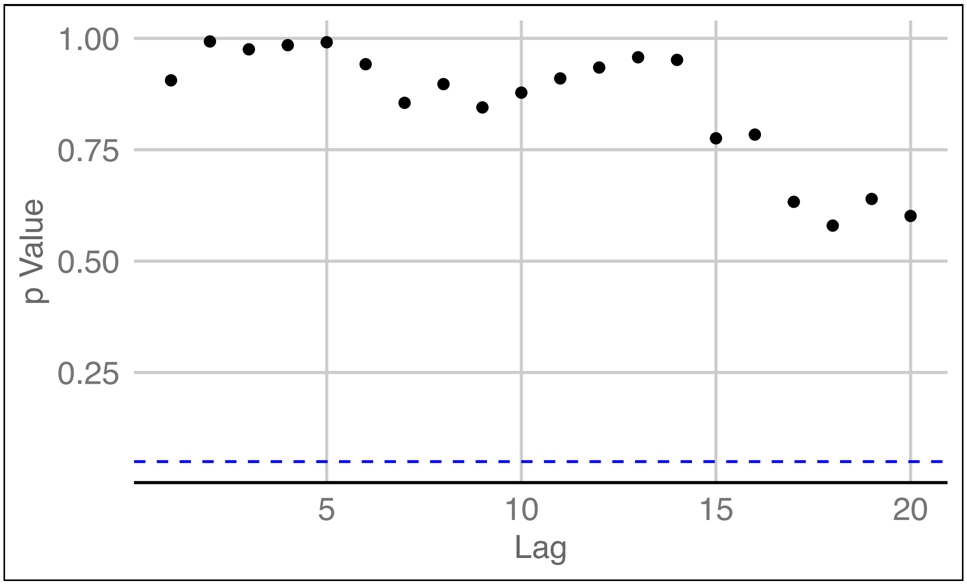

Figure 7: p Values for Ljung-Box Statistic

Model Fitting and Forecasting

Since an ARIMA($p = 3, d = 1, q = 3$) model is fitted, $\phi_1, \phi_2, \phi_3, \theta_1, \theta_2, \theta_3$, and μ are the coefficients that need to be estimated in $\phi(B)Y_t - \mu = \phi(B)(1 - B)X_t - \mu = \theta(B)W_t$ where $\{W_t\}$ is white noise.

It can be seen from the t table below that all coefficients are significantly different from 0 but the mean μ, which is understandable because the first order differenced data Y_t looks zero-mean and stationary.

	Estimate	SE	t.value	p.value
ar1	1.02	0.15	6.69	0.00
ar2	-1.10	0.11	-10.35	0.00
ar3	0.39	0.06	6.00	0.00
ma1	-1.46	0.15	-9.79	0.00
ma2	1.28	0.16	8.22	0.00
ma3	-0.66	0.10	-6.50	0.00
constant	-0.00	0.00	-0.03	0.97

Predicted values for $x_{501}, \ldots, x_{510}$ are listed in the table below and are plotted in Figure 8, where the 95% confidencen interval for the predicted values are also included.

	$\hat{X}_{501}$	$\hat{X}_{502}$	$\hat{X}_{503}$	$\hat{X}_{504}$	$\hat{X}_{505}$	$\hat{X}_{506}$	$\hat{X}_{507}$	$\hat{X}_{508}$	$\hat{X}_{509}$	$\hat{X}_{510}$
Value	0.0406	0.0336	0.0326	0.0352	0.0362	0.0340	0.0317	0.0321	0.0343	0.0350

Discussion and Summary

While an ARIMA($p = 3, d = 1, q = 3$) model works the best among all the initial guesses deduced from the ACF and PACF plots, there is no guarantee that this is the true model underlying this dataset. The ARIMA models at hand have all failed to capture the fluctuations across time, and other models should be considered for better prediction results.

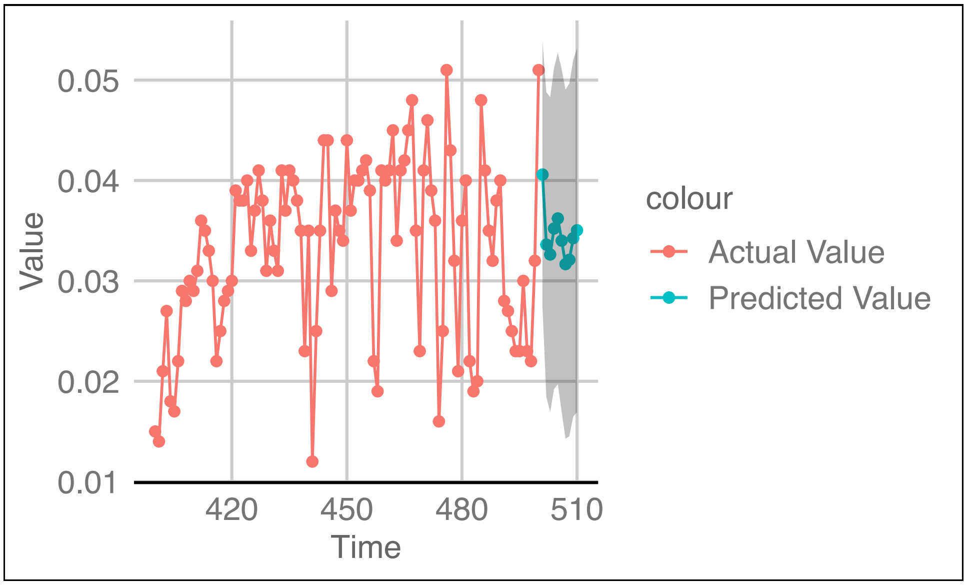

Figure 8: Predicted Values from ARIMA(3,1,3) model with 95% Confidence Interval