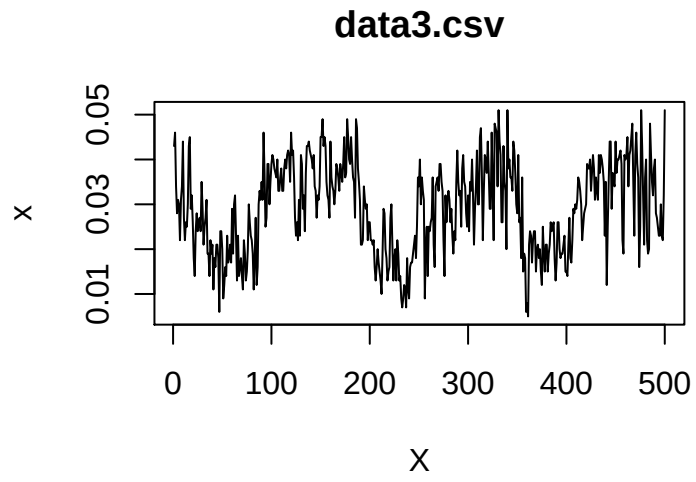
Midterm 2 - Report

Nicholas Lai

November 17, 2018

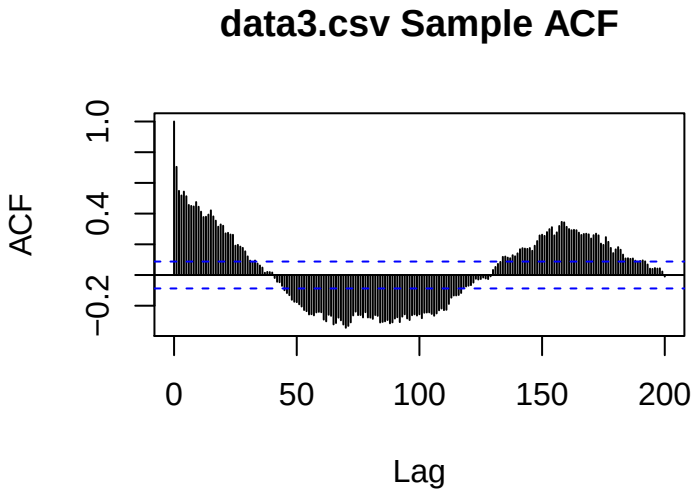
(A) Exploratory Data Analysis

This report deals with the analysis of the data in `data3.csv`. The raw data is shown below:



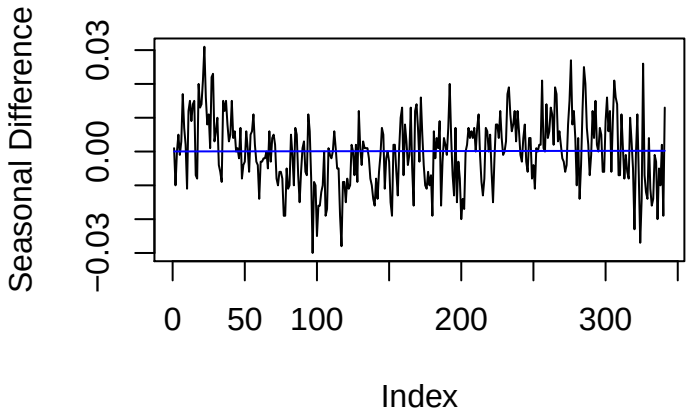
From the data, we see evidence of a seasonal trend and a non-zero mean. In order to analyze this data, we must deal with these issues.

To verify if our data is strongly seasonal, we can look at the `acf` of our data in order to see the correlation structure in our data.



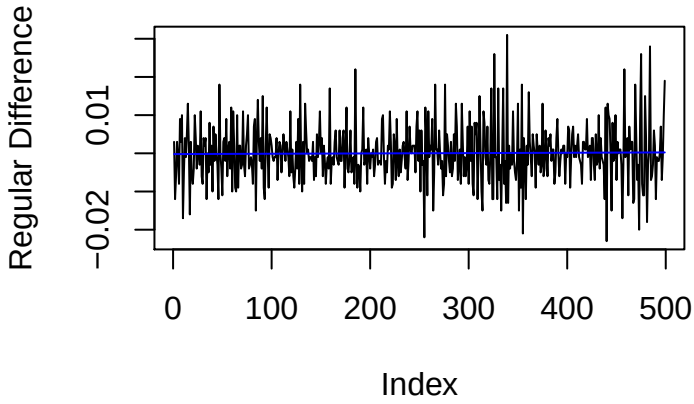
We clearly see a pattern in our `acf` for the raw time series, consistent with seasonality. A natural period to choose for seasonal differencing is the (positive) peak of the acf.

The peak of the ACF corresponds to lag 159. Seasonal differencing with this parameter will eliminate the mean of the data.



After one seasonal differencing, we obtained a zero mean process (in blue, the regression of the data is shown to be roughly the x axis). On the surface, this process has lost its strong seasonality.

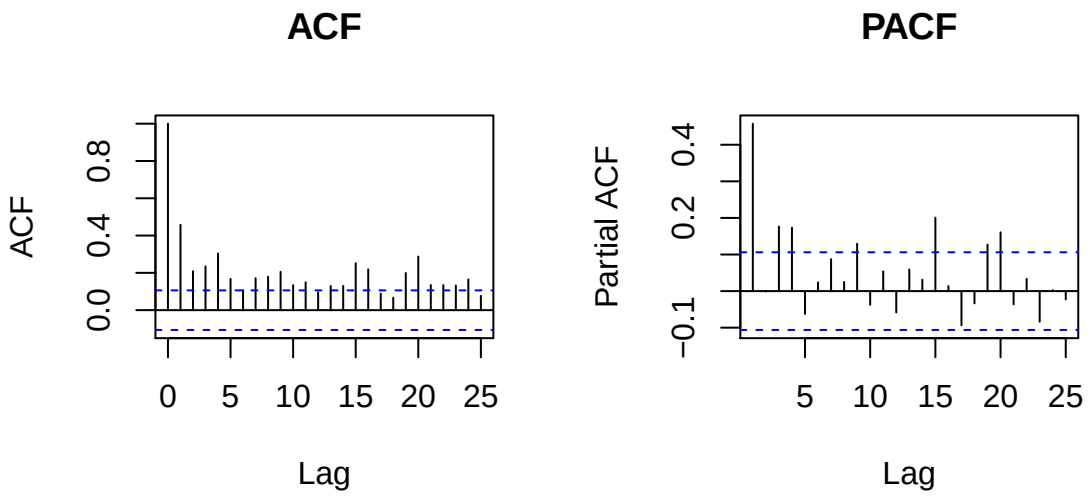
One interesting quirk of our data is that a regular difference of the data yields a seemingly stationary process, despite the data being seemingly seasonal.



(B) Identifying suitable ARIMA model

Note: The plots are in some insane order because of some quirk of R Markdown, so I apologise.

We can check the sample `acf` and `pacf` of our differenced series to try and determine a natural AR or MA parameter combination.



The plots don't reveal a natural model, so we may need to use criteria-based model selection.

In general, it is unlikely that the optimal MA and AR terms are more than the `auto.arima` defaults (and the plots give no reason to believe otherwise), so we can compute the optimal models by AIC and BIC:

```
## Series: dat3_main
## ARIMA(1,0,0)(0,1,0)[159]
##
## Coefficients:
##          ar1
##       0.4585
## s.e.  0.0481
##
## sigma^2 estimated as 8.284e-05:  log likelihood=1118.99
## AIC=-2233.98   AICc=-2233.94   BIC=-2226.32
```

An AR(1) model was selected by minimizing both AIC and BIC. The stepwise selection process and approximation in `auto.arima` was done to speed up the process. Other models I checked with seasonal differences performed similarly, so I am confident that despite these shortcuts we did not miss a good model.

We can perform diagnostics on this model:

The regression diagnostics show that this model, the most natural one from our seasonal differencing procedure, is not a perfect fit for the data. Our errors, by the Ljung-Box test, are uniformly below the threshold for non-zero autocorrelation for all lags, and our QQplot of the residuals shows long tails, which are evidence of non-normal error.

Other possible SARIMA models arise from alternative differencings. Consider the regular diffrence that yielded a stationary process above. With analagous logic as with the seasonal differencing case, we can select models.

```
## Series: dat3$x
## ARIMA(2,1,5)
##
## Coefficients:
##          ar1      ar2      ma1     ma2      ma3      ma4      ma5
##       0.8975  -0.8741  -1.3378  1.0284  -0.2135  -0.1003  -0.1449
```

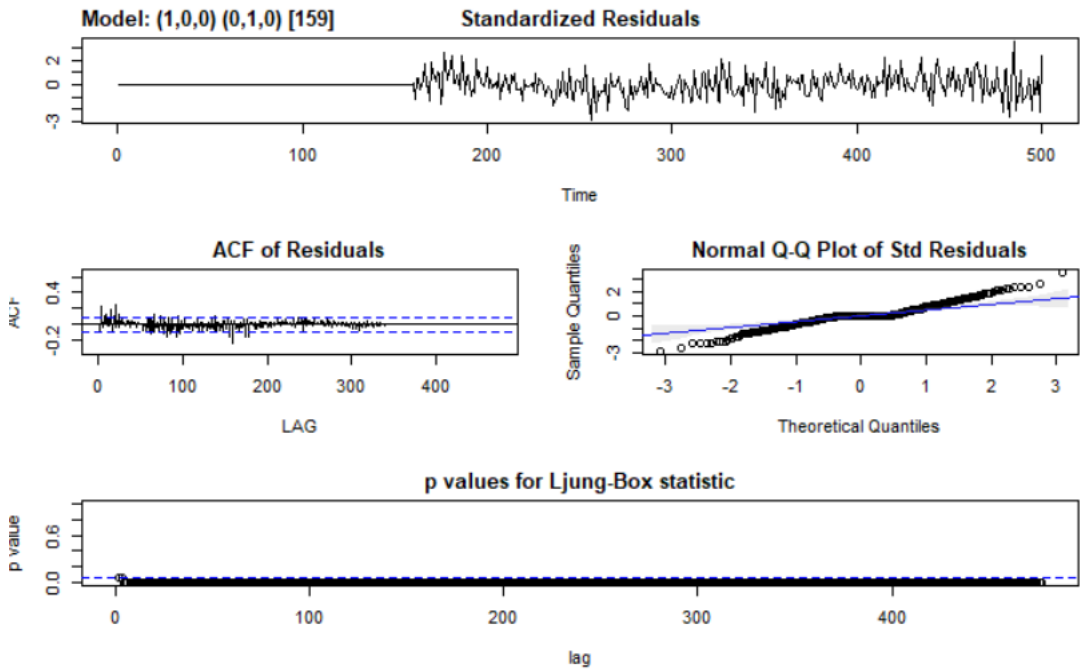


Figure 1:

```
## s.e.  0.0438   0.0458   0.0620  0.0883   0.0902   0.0822   0.0517
##
## sigma^2 estimated as 4.55e-05:  log likelihood=1789.19
## AIC=-3562.38   AICc=-3562.09   BIC=-3528.68
```

```
## Series: dat3$x
## ARIMA(0,1,2)
##
## Coefficients:
##           ma1      ma2
##       -0.4534  -0.2899
## s.e.   0.0414   0.0407
##
## sigma^2 estimated as 4.694e-05:  log likelihood=1779.28
## AIC=-3552.55   AICc=-3552.51   BIC=-3539.92
```

Both of these models, one preferred by BIC and one by AIC, are already huge improvements on the seasonal differencing selected model. Both models have QQplots that show normality of residuals as reasonable, and the Ljung-Box statistics for both models have some points above the threshold for rejection of uncorrelatedness, especially in the case of the ARIMA(2,1,5) model.

It is clear at this point that seasonal differencing is a lost cause, as the models above with seasonal differencing are worse by every diagnostic.

Now, choosing between our ARIMA(2,1,5) and our ARIMA(0,1,2) models can be done via cross-validation.

I calculate the sum of the squared error of models trained on 250-499 previous data points when fitted to the remaining values of the data, and use that as my cross-validation score.

```
## [1] 0.01453721
```

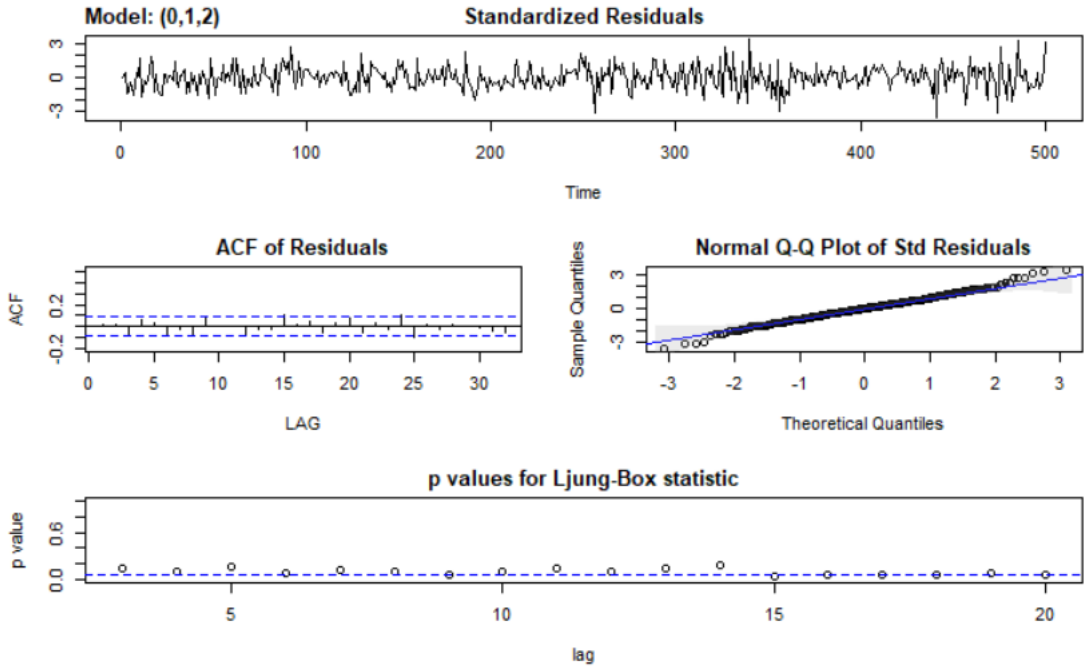


Figure 2:

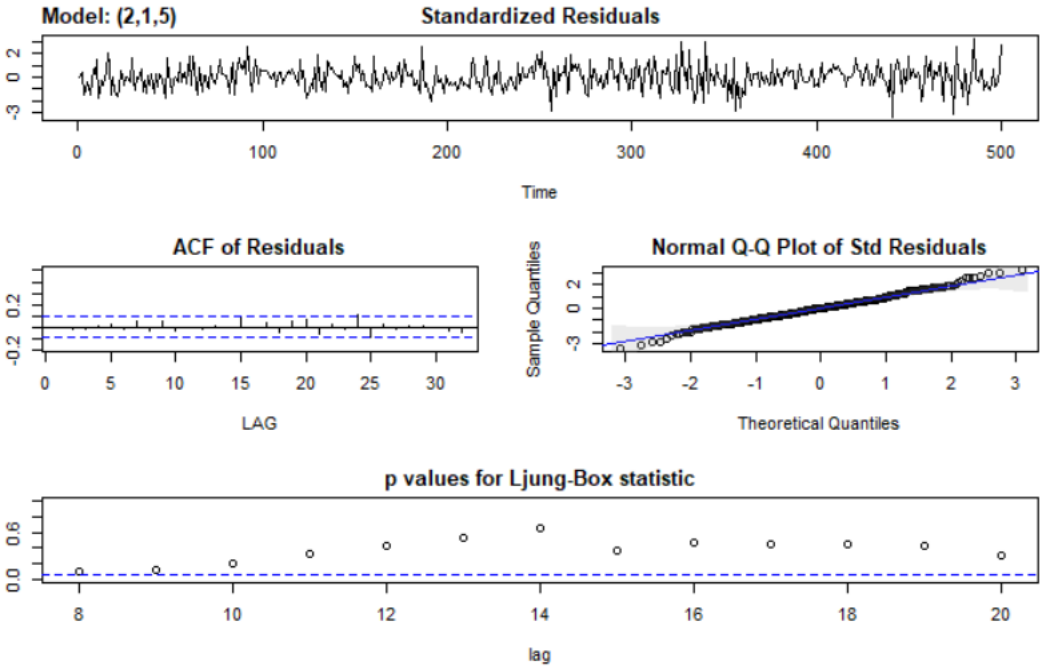


Figure 3:

```
## [1] 0.01453924
```

The cross-validation score for the ARIMA(2,1,5) model is smaller than the ARIMA(0,1,2) model, so we select the model for our forecast.

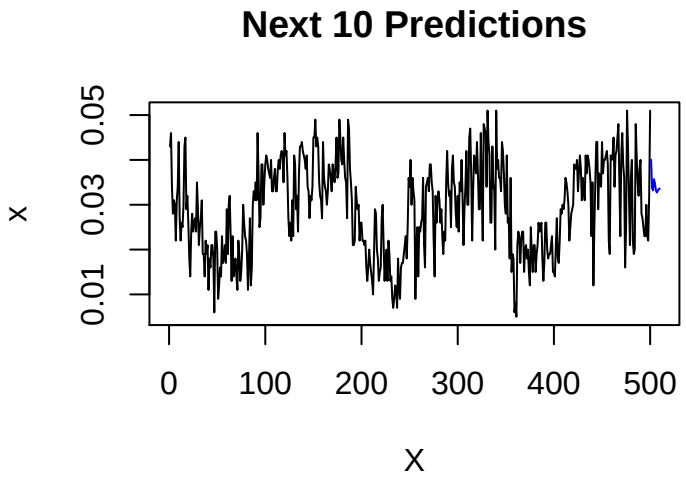
(C) Model Fitting and Forecasting

Estimation of Parameters:

Using Conditional Sum of Squares to fit the data,

```
## Series: dat3$x
## ARIMA(2,1,5)
##
## Coefficients:
##          ar1      ar2      ma1     ma2      ma3     ma4      ma5
##       0.5444  -0.4153  -0.9868  0.3904  -0.1236  0.0547  -0.1298
## s.e.  0.2345   0.2226   0.2333  0.2700   0.1243  0.0823   0.0528
##
## sigma^2 estimated as 4.624e-05:  part log likelihood=1784.91
```

Forecast: With all the parameters of our model, we can simply plug in our data into the model.




```
## Time Series:
## Start = 501
## End = 510
## Frequency = 1
##  [1] 0.04003692 0.03371628 0.03319125 0.03571217 0.03474855 0.03317689
##  [7] 0.03272149 0.03312635 0.03353590 0.03359072
```

Appendix

Code:

```r
knitr::opts_chunk$set(echo = FALSE, cache = TRUE, fig.pos = '!h')
library(aTSA)
library(astsa)
library(forecast)

dat3 <- read.csv("data3.csv")
plot(dat3, type = 'l', main = "data3.csv")

acf3 <- acf(dat3$x, lag.max = 200, main = 'data3.csv Sample ACF')

dat3_test <- diff(dat3$x, lag = 159, differences = 1)
dat3_main <- ts(dat3$x, frequency = 159)
plot(dat3_test, type = 'l', ylab = 'Seasonal Difference')
lines(lm(dat3_test~I(1:length(dat3_test)))$fitted, col = "blue")

dat3_test <- diff(dat3$x,differences = 1)
plot(dat3_test, type = 'l', ylab = "Regular Difference")
lines(lm(dat3_test~I(1:length(dat3_test)))$fitted, col = "blue")

acf(diff(dat3$x,differences = 1, lag = 159), main = 'ACF')
pacf(diff(dat3$x,differences = 1, lag = 159), main = "PACF")

auto.arima(dat3_main, D=1,
           allowmean = FALSE, ic = 'aic')
auto.arima(dat3_main, D=1,
           allowmean = FALSE, ic = 'bic')

invisible(sarima(dat3$x, 1,0,0,0,1,0,159))

auto.arima(dat3$x, approximation = FALSE, stepwise = FALSE,max.order = 14, d=1,
           allowmean = FALSE, ic = 'aic')
auto.arima(dat3$x, approximation = FALSE, stepwise = FALSE,max.order = 14, d=1,
           allowmean = FALSE, ic = 'bic')
sarima(dat3$x, 2,1,5,0,0,0)
sarima(dat3$x, 0,1,2,0,0,0)

fore1 <- function(x,h){forecast(Arima(x, order = c(2,1,5)), h = h)}
fore2 <- function(x,h){forecast(Arima(x, order = c(0,1,2)), h = h)}
e1 <- tsCV(dat3$x, fore1, h=1)
e2 <- tsCV(dat3$x, fore2, h=1)

sum(e1[250:500]^2, na.rm = TRUE)
sum(e2[250:500]^2, na.rm = TRUE)

model3 <- Arima(dat3$x, order = c(2,1,5), method = 'CSS')
model3

predictions = predict(model3, n.ahead=10)
plot(dat3, xlim = c(0,510), type = 'l', main = "Next 10 Predictions")
lines(predictions$pred, col='blue')
predictions$pred
```