Robert Smith

In-class Exercise: Fitting Dow Jones Data with ARIMA

October 15, 2013

1. Read in dowj.txt file. Plot the time series. Does the plot looks stationary? Plot ACF and PACF of the series. Test for the stationarity using Augmented Dickey-Fuller Unit-Root test.

```
Randomness.tests <- function(A, plott = FALSE) {

    library(tseries)

    L1 <- Box.test(A, lag = 15, type = "Ljung-Box")
    L2 <- Box.test(A, lag = 20, type = "Ljung-Box")
    L3 <- Box.test(A, lag = 25, type = "Ljung-Box")
    L4 <- Box.test(A^2, lag = 15, type = "Ljung-Box")
    L5 <- Box.test(A^2, lag = 20, type = "Ljung-Box")
    L6 <- wilcox.test(A)
    L7 <- jarque.bera.test(A)
    S1 <- sd(A)

    if (plott) {
        layout(matrix(c(rep(1, 8), 2, 3, 4, 8, 5, 6, 7, 8), 4, 4, byrow = T))
        plot(A, type = "l")
        acf(A)
        pacf(A)
        plot(density(A, bw = "SJ-ste"), main = "")
        acf(abs(A))
        acf(A^2)
        qqnorm(A)

        plot(c(-1, -1), xlim = c(0, 1), ylim = c(0, 1), ann = F, axes = F)

        text(0.5, 0.95, paste("Box-Ljung test"), cex = 1.3)
        text(0.5, 0.9, paste("H=15:p= ", round(L1$p.value, 4)), cex = 1)
        text(0.5, 0.85, paste("H=20:p= ", round(L2$p.value, 4)), cex = 1)
        text(0.5, 0.8, paste("H=25:p= ", round(L3$p.value, 4)), cex = 1)
```

```
        text(0.5, 0.7, paste("McLeod-Li test"), cex = 1.3)
        text(0.5, 0.65, paste("H=15:p= ", round(L4$p.value, 4)), cex = 1)
        text(0.5, 0.6, paste("H=20:p= ", round(L5$p.value, 4)), cex = 1)
        text(0.5, 0.5, paste("Wilcoxson test"), cex = 1.3)
        text(0.5, 0.45, paste("p= ", round(L6$p.value, 4)), cex = 1)
        text(0.5, 0.35, paste("Jaque-Bera test"), cex = 1.3)
        names(L7$p.value) <- ""
        text(0.5, 0.3, paste("p= ", round(L7$p.value, 4)), cex = 1)
        text(0.5, 0.2, paste("sample SD ", round(S1, 4)), cex = 1)

        layout(matrix(1, 1, 1))
    }

    return(t(t(c(BL15 = round(L1$p.value, 4), BL20 = round(L2$p.value, 4), BL25 = round(L3$p
        4), ML15 = round(L4$p.value, 4), ML20 = round(L5$p.value, 4), WX = round(L6$p.value,
        4), JB = round(L7$p.value, 4), SD = round(S1, 4)))))

}

D <- read.csv("http://gozips.uakron.edu/~nmimoto/pages/datasets/dowj.csv")
D1 <- ts(D, start = c(1, 1), freq = 1)

plot(D1, type = "o")
```

```
acf(D1)
```

```
pacf(D1)
```

```
adf.test(D1, alternative = "stationary")
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  D1
## Dickey-Fuller = -1.805, Lag order = 4, p-value = 0.6552
## alternative hypothesis: stationary
```

The as-is data does not appear to be stationary because the mean does not appear to be stationary over time. When we apply the acf() function we can see significant autocorrelation over approximately 20 lags. When the pacf() function is applied it tails off after lag-zero, thus indicating a potential AR or ARMA model.

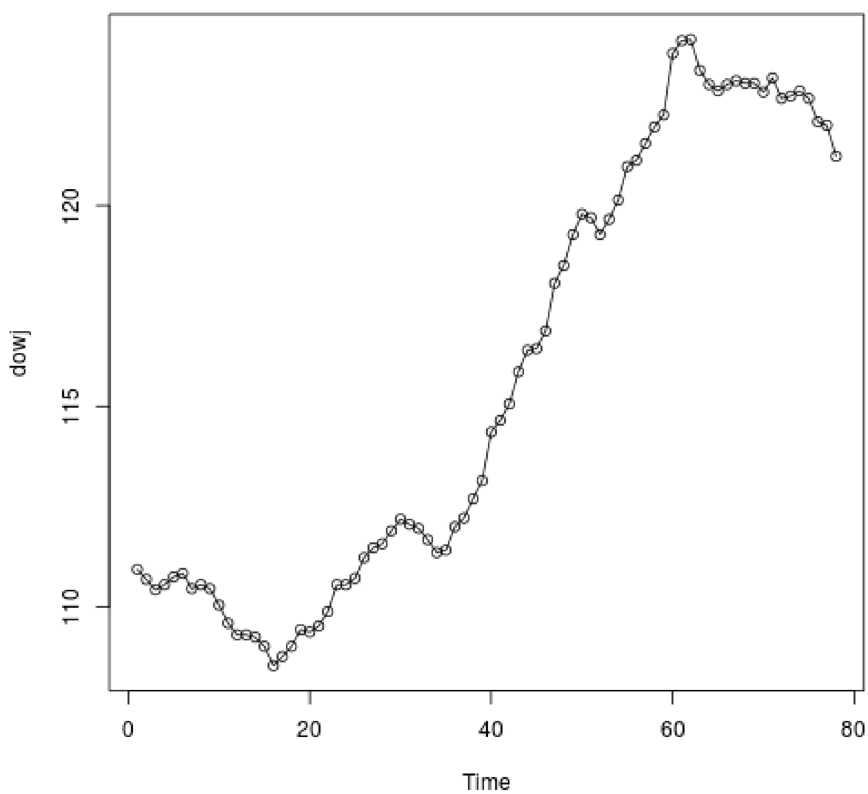

Figure 1: plot of chunk Q1

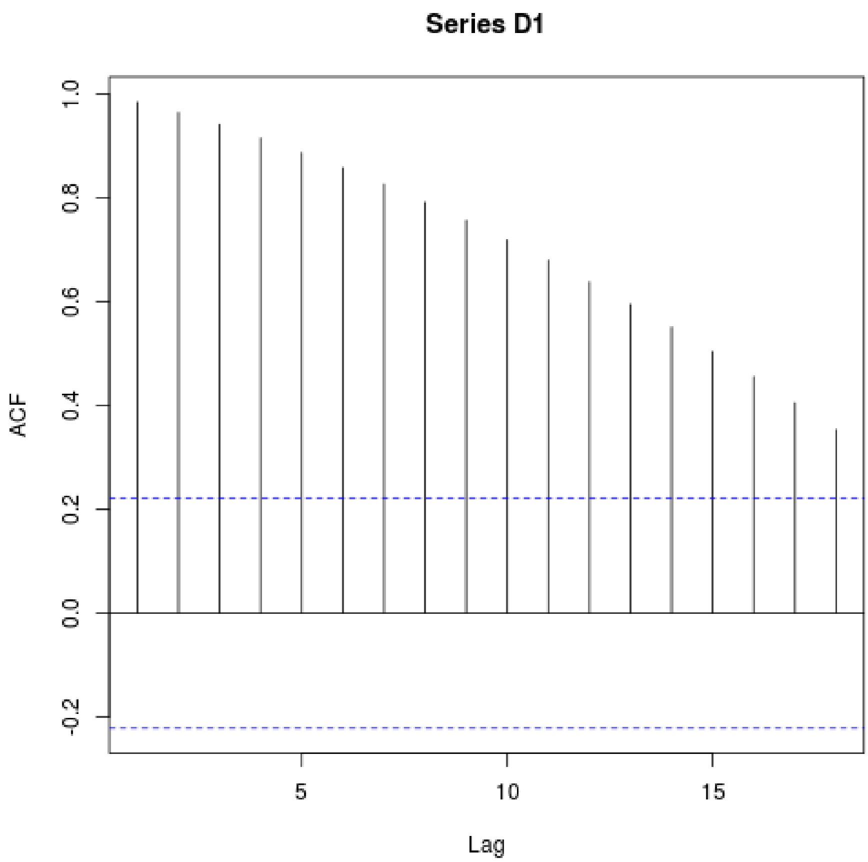

Figure 2: plot of chunk Q1

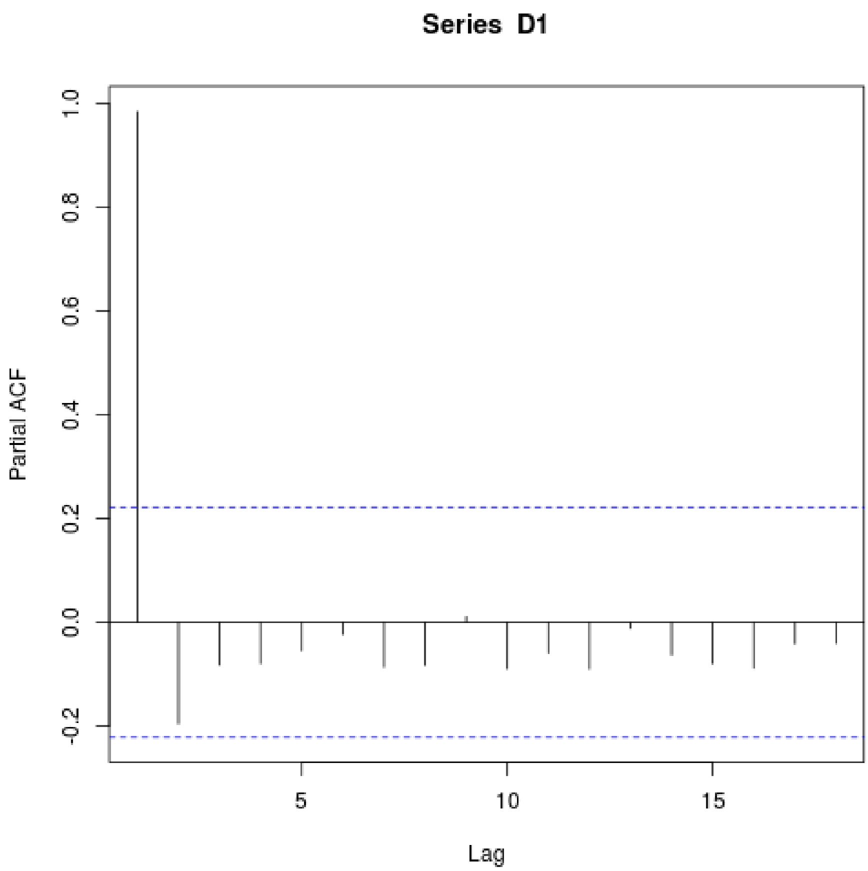

Figure 3: plot of chunk Q1

Based on the results of the Augmented Dickey-Fuller Unit-Root test for stationarity, p-value = 0.6552 and therefore we fail to reject H_0, and therefore cannot find that the time series has unit root and is stationary.

2. Take the difference of dowj data. Plot the time series. Does the plot looks stationary? Plot ACF and PACF of the series. What does ADF test say about stationarity?

```
D2 <- diff(D1)
plot(D2, type = "o")
```

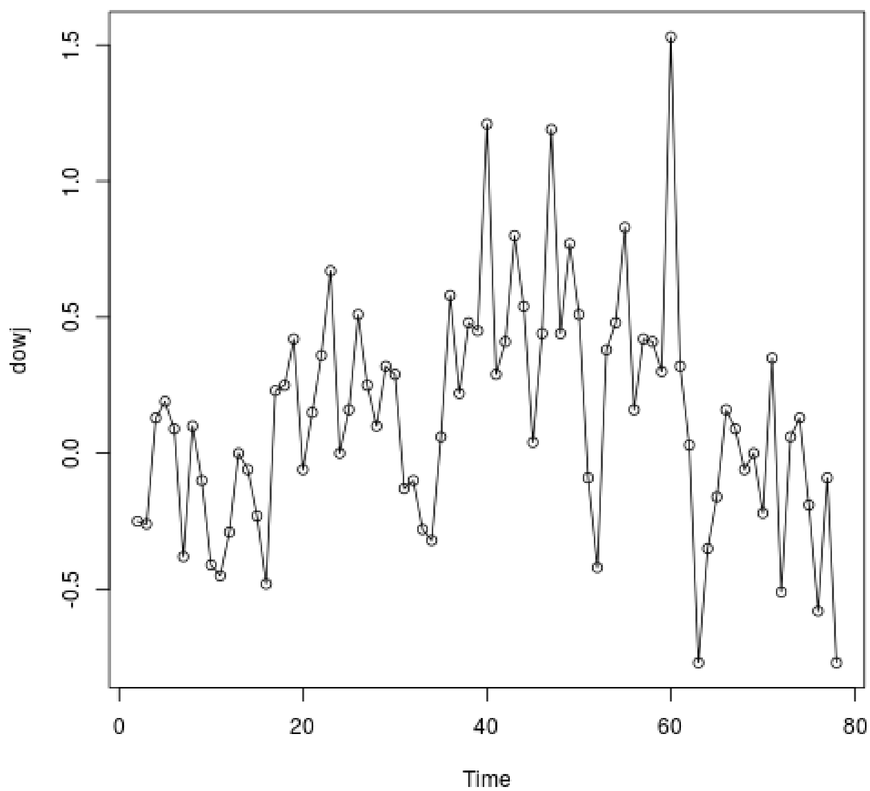

Figure 4: plot of chunk Q2

```
acf(D2)
```

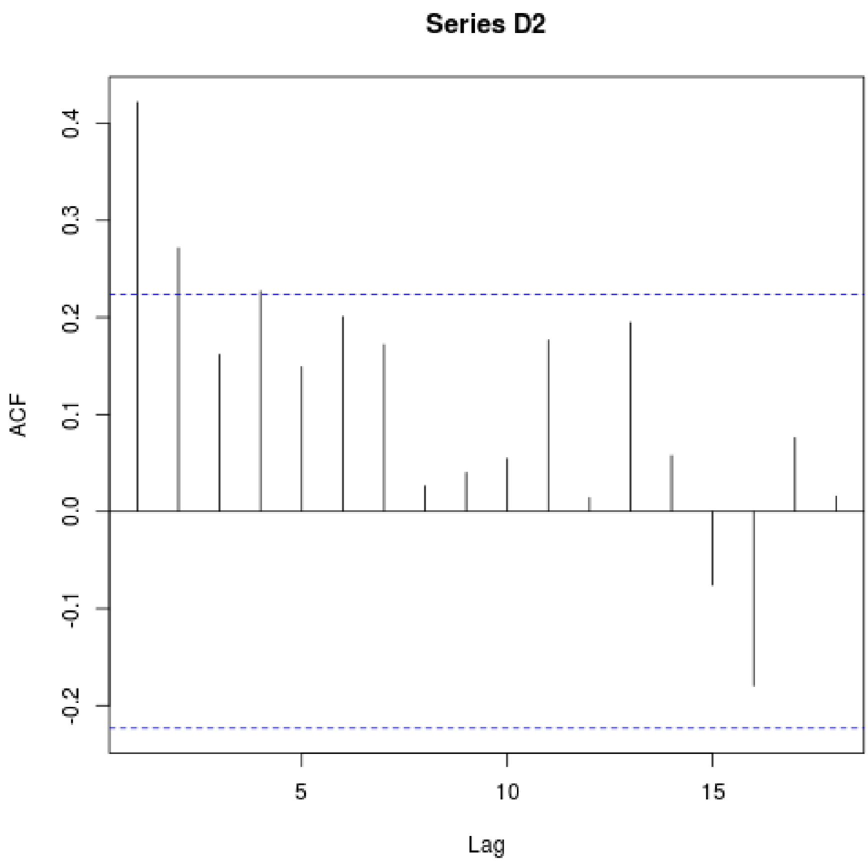

Figure 5: plot of chunk Q2

```r
pacf(D2)
```

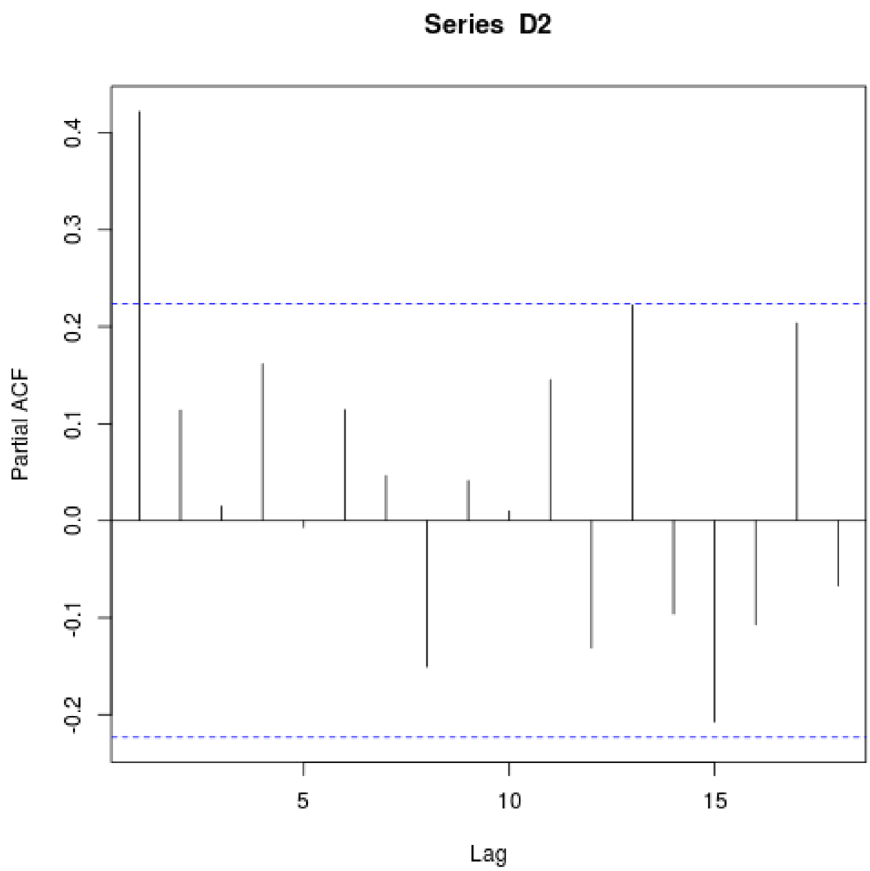

Figure 6: plot of chunk Q2

```r
adf.test(D2, alternative = "stationary")
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  D2
## Dickey-Fuller = -2.034, Lag order = 4, p-value = 0.5617
## alternative hypothesis: stationary
```

The plot appears more stationary than the raw data, but when the ADF test is applied the p-value of 0.5617 shows that the we still reject H_0 and cannot say the time-series is stationary.

3. Take additional difference of dowj data. Plot the time series. Does the plot looks stationary? Plot ACF and PACF of the series. What does ADF test say about stationarity?

```
D3 <- diff(D2)
plot(D3, type = "o")
```

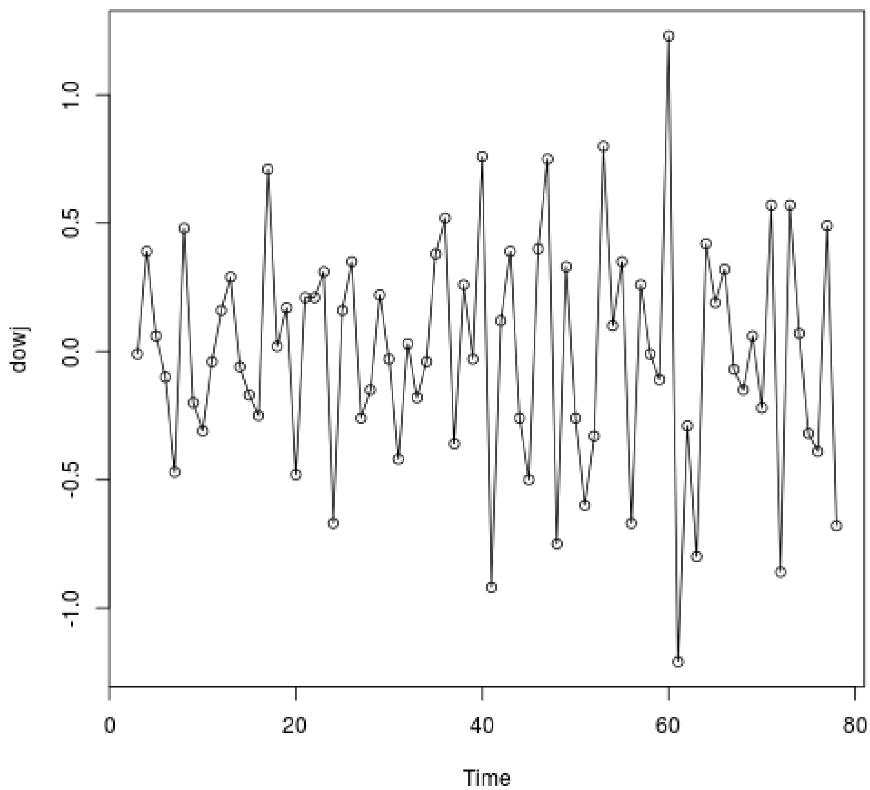

Figure 7: plot of chunk Q3

```
acf(D3)
```

```
pacf(D3)
```

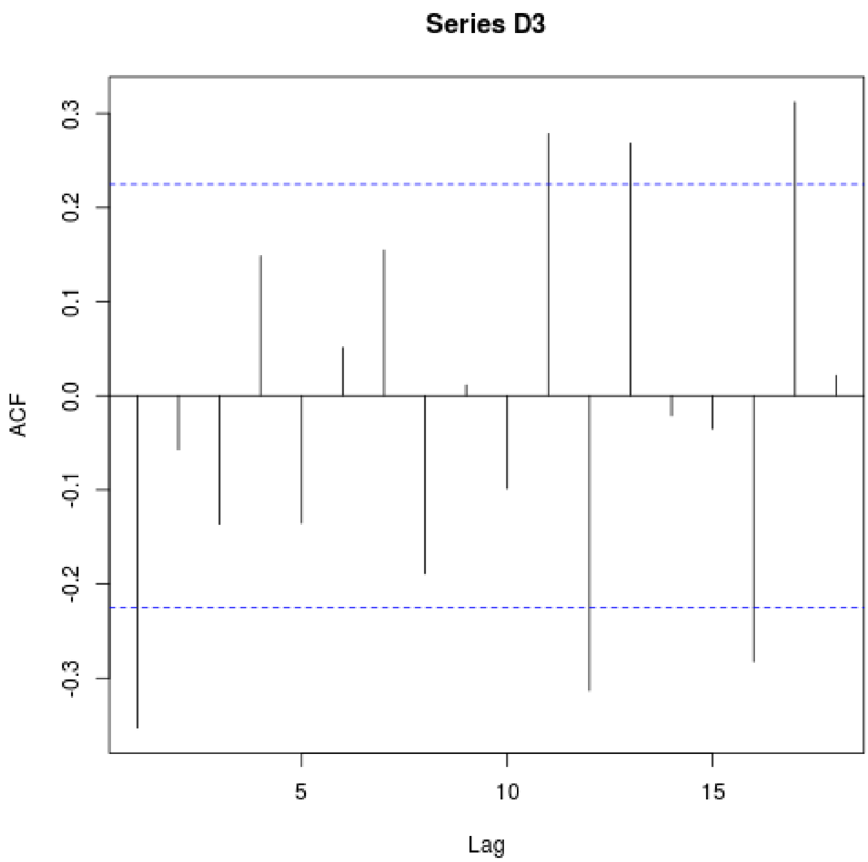

Figure 8: plot of chunk Q3

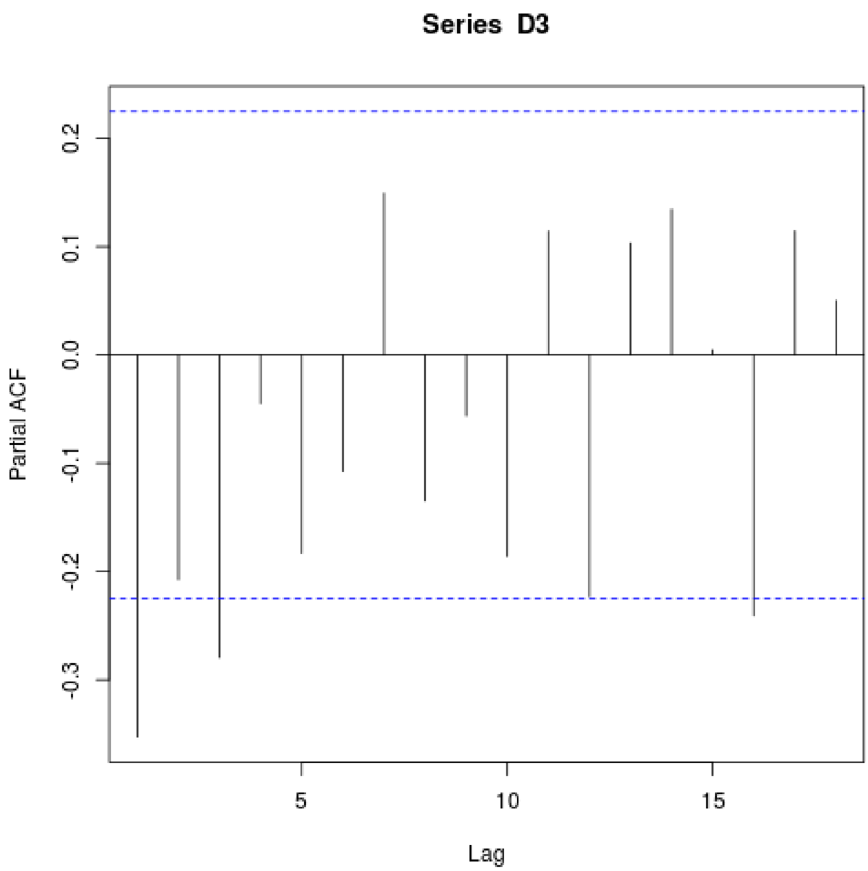

Figure 9: plot of chunk Q3

```r
adf.test(D3, alternative = "stationary")
```

```
## Warning: p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  D3
## Dickey-Fuller = -5.905, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

Based on the view of the data which has been further differenced the graph does not appear to be significantly different by eye, but when the ADF test was applied a p-value < 0.01 was found and therefore we fail to reject H_0 and find that it is possible that this time-series is stationary.

2b. Now based on what we saw in question 2, model the origianl dowj data with ARIMA(p, 1, q). Use auto.arima() in forecast package to choose p and q based on AICc. Diagnose the residual after the fit. Is the model fitting well? If not, manually search for better value of p and q.

```r
library(forecast)
```

```r
auto.arima(D1, d = 1, stepwise = FALSE, seasonal = FALSE, trace = TRUE)
```

```
##
##  ARIMA(0,1,0)                    : 95.78
##  ARIMA(0,1,0) with drift         : 89.48
##  ARIMA(0,1,1)                    : 83.58
##  ARIMA(0,1,1) with drift         : 80.34
##  ARIMA(0,1,2)                    : 79.96
##  ARIMA(0,1,2) with drift         : 78.31
##  ARIMA(0,1,3)                    : 81.95
##  ARIMA(0,1,3) with drift         : 80.55
##  ARIMA(0,1,4)                    : 80.31
##  ARIMA(0,1,4) with drift         : 79.84
##  ARIMA(0,1,5)                    : 82.67
##  ARIMA(0,1,5) with drift         : 82.25
##  ARIMA(1,1,0)                    : 76.54
##  ARIMA(1,1,0) with drift         : 75.71
##  ARIMA(1,1,1)                    : 75.71
##  ARIMA(1,1,1) with drift         : 76.38
##  ARIMA(1,1,2)                    : 76.58
##  ARIMA(1,1,2) with drift         : 77.75
##  ARIMA(1,1,3)                    : 78.42
```

```
##  ARIMA(1,1,3) with drift         : 79.7
##  ARIMA(1,1,4)                    : 79.8
##  ARIMA(1,1,4) with drift         : 81.05
##  ARIMA(2,1,0)                    : 76.98
##  ARIMA(2,1,0) with drift         : 76.88
##  ARIMA(2,1,1)                    : 79.21
##  ARIMA(2,1,1) with drift         : 78.89
##  ARIMA(2,1,2)                    : 78.66
##  ARIMA(2,1,2) with drift         : 81.02
##  ARIMA(2,1,3)                    : 79.36
##  ARIMA(2,1,3) with drift         : 80.73
##  ARIMA(3,1,0)                    : 78.92
##  ARIMA(3,1,0) with drift         : 79.09
##  ARIMA(3,1,1)                    : 78.66
##  ARIMA(3,1,1) with drift         : 79.93
##  ARIMA(3,1,2)                    : 80.61
##  ARIMA(3,1,2) with drift         : 81.95
##  ARIMA(4,1,0)                    : 78.26
##  ARIMA(4,1,0) with drift         : 79.11
##  ARIMA(4,1,1)                    : 80.56
##  ARIMA(4,1,1) with drift         : 81.39
##  ARIMA(5,1,0)                    : 80.61
##  ARIMA(5,1,0) with drift         : 81.51
```

```
## Series: D1
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1     ma1
##        0.851  -0.526
## s.e.  0.138   0.255
##
## sigma^2 estimated as 0.143:  log likelihood=-34.69
## AIC=75.38   AICc=75.71   BIC=82.41
```

```
ARIMA1 <- auto.arima(D1, d = 1, stepwise = FALSE, seasonal = FALSE)

adf.test(ARIMA1$residuals, alternative = "stationary")
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ARIMA1$residuals
## Dickey-Fuller = -3.961, Lag order = 4, p-value = 0.01591
## alternative hypothesis: stationary
```

```r
Randomness.tests(ARIMA1$residuals)
```

```
##                [,1]
## BL15         0.2869
## BL20         0.0877
## BL25         0.0943
## ML15         0.7007
## ML20         0.5222
## WX           0.5141
## JB.X-squared 0.0480
## SD           0.3771
```

```r
summary(ARIMA1)
```

```
## Series: D1
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1     ma1
##        0.851  -0.526
## s.e.  0.138   0.255
##
## sigma^2 estimated as 0.143:  log likelihood=-34.69
## AIC=75.38   AICc=75.71   BIC=82.41
##
## Training set error measures:
##                     ME   RMSE  MAE     MPE   MAPE   MASE     ACF1
## Training set 0.03566 0.3764 0.29 0.03178 0.2492 0.8488 -0.07309
```

Given the above code, I find that ARIMA(1,1,1) is the model with the lowest AIC & standard error. Using 'stepwise = FALSE' outputs the AIC statistic of each model tested, which is how I verified that this model was the best of the group.

2c. Using the model you came up in the previous question, give 5-day prediction of dowj value. Plot the data(black) and predictioin(red) on the same plot. The range of x-axis must be suitablly chosen.

```r
plot(forecast(ARIMA1, h = 5), fcol = 2)
```

2d. In part (2-b), your ARIMA parameter estimation gave standard errors for estimation. Can you trust that number? Why? How would you verify?

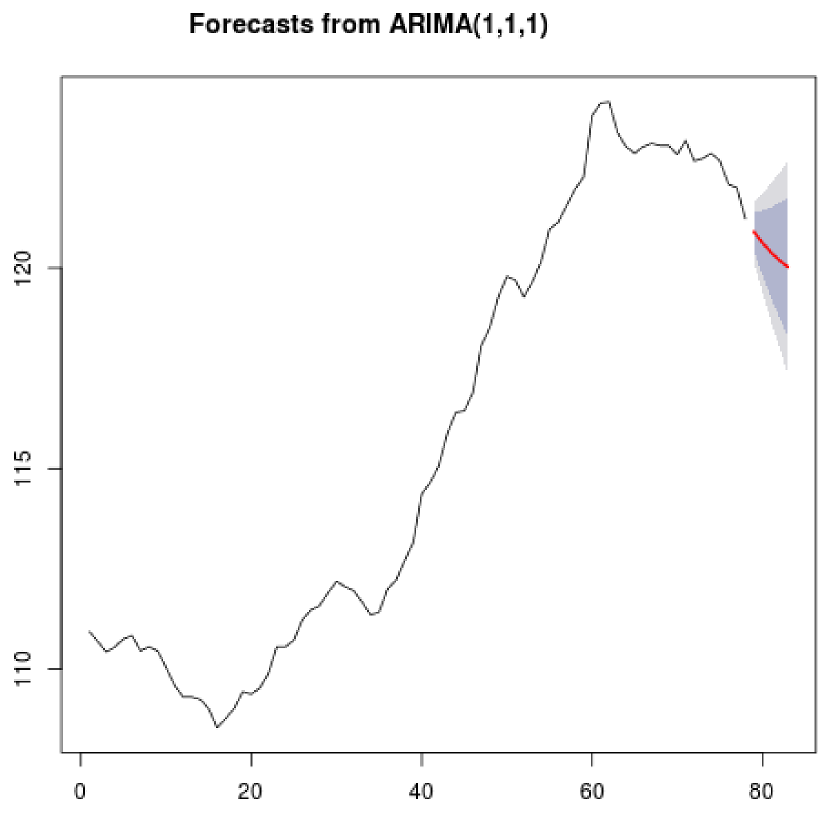

Figure 10: plot of chunk Q2C

I would verify the standard error through simulation of values from a known distribution. If the simulation comes up with similar values for the standard error and then track the ratio of predicted values vs. those that fall within the interval for the simulation.

3b. Now based on what we saw in question 3, model the origianl dowj data with ARIMA(p, 2, q). Use auto.arima() in forecast package to choose p and q based on AICc. Diagnose the residual after the fit. Is the model fitting well? If not, manually search for better value of p and q.

```r
auto.arima(D3, d = 2, stepwise = FALSE, seasonal = FALSE, trace = TRUE)
```

```
##
##  ARIMA(0,2,0)                    : 252.1
##  ARIMA(0,2,1)                    : 1e+20
##  ARIMA(0,2,2)                    : 1e+20
##  ARIMA(0,2,3)                    : 1e+20
##  ARIMA(0,2,4)                    : 97.1
##  ARIMA(0,2,5)                    : 98.47
##  ARIMA(1,2,0)                    : 199.2
##  ARIMA(1,2,1)                    : 1e+20
##  ARIMA(1,2,2)                    : 1e+20
##  ARIMA(1,2,3)                    : 96.03
##  ARIMA(1,2,4)                    : 1e+20
##  ARIMA(2,2,0)                    : 183.7
##  ARIMA(2,2,1)                    : 1e+20
##  ARIMA(2,2,2)                    : 1e+20
##  ARIMA(2,2,3)                    : 98.19
##  ARIMA(3,2,0)                    : 160.2
##  ARIMA(3,2,1)                    : 1e+20
##  ARIMA(3,2,2)                    : 1e+20
##  ARIMA(4,2,0)                    : 152.3
##  ARIMA(4,2,1)                    : 1e+20
##  ARIMA(5,2,0)                    : 151
```

```
## Series: D3
## ARIMA(1,2,3)
##
## Coefficients:
##          ar1     ma1    ma2     ma3
##        0.391  -2.953  2.923  -0.970
## s.e.   0.125   0.086  0.170   0.086
##
## sigma^2 estimated as 0.143:  log likelihood=-42.57
## AIC=95.14   AICc=96.03   BIC=106.7
```

```r
ARIMA3 <- auto.arima(D1, d = 2, stepwise = FALSE, seasonal = FALSE)

adf.test(ARIMA3$residuals, alternative = "stationary")
```

```
## Warning: p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  ARIMA3$residuals
## Dickey-Fuller = -4.163, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

```r
Randomness.tests(ARIMA3$residuals)
```

```
##                [,1]
## BL15         0.2781
## BL20         0.0687
## BL25         0.0655
## ML15         0.6047
## ML20         0.4360
## WX           0.8343
## JB.X-squared 0.2311
## SD           0.3786
```

```r
summary(ARIMA3)
```

```
## Series: D1
## ARIMA(1,2,1)
##
## Coefficients:
##         ar1     ma1
##       0.248  -0.839
## s.e.  0.145   0.082
##
## sigma^2 estimated as 0.145:  log likelihood=-34.85
## AIC=75.7   AICc=76.03   BIC=82.69
##
## Training set error measures:
##                     ME   RMSE    MAE       MPE   MAPE   MASE     ACF1
## Training set -0.006313 0.3762 0.2911 -0.002538 0.2498 0.8519 0.008372
```

```r
accuracy(ARIMA3)
```

```
##                          ME   RMSE    MAE       MPE   MAPE   MASE     ACF1
## Training set -0.006313 0.3762 0.2911 -0.002538 0.2498 0.8519 0.008372
```

With a MA $\sigma_e = 0.1447$ and AR $\sigma_e = 0.0819$ I see an improvement in the fit of the forecast. The improvement is light for AR, but quite significant for MA. Overall, I'm not sure if this is the best, but certain elements seem to be better.

3c. Using the model you came up in the previous question, give 5-day prediction of dowj value. Plot the data(black) and predictioin(red) on the same plot. The range of x-axis must be suitablly chosen.

```
plot(forecast(ARIMA3, h = 5), fcol = 2)
```

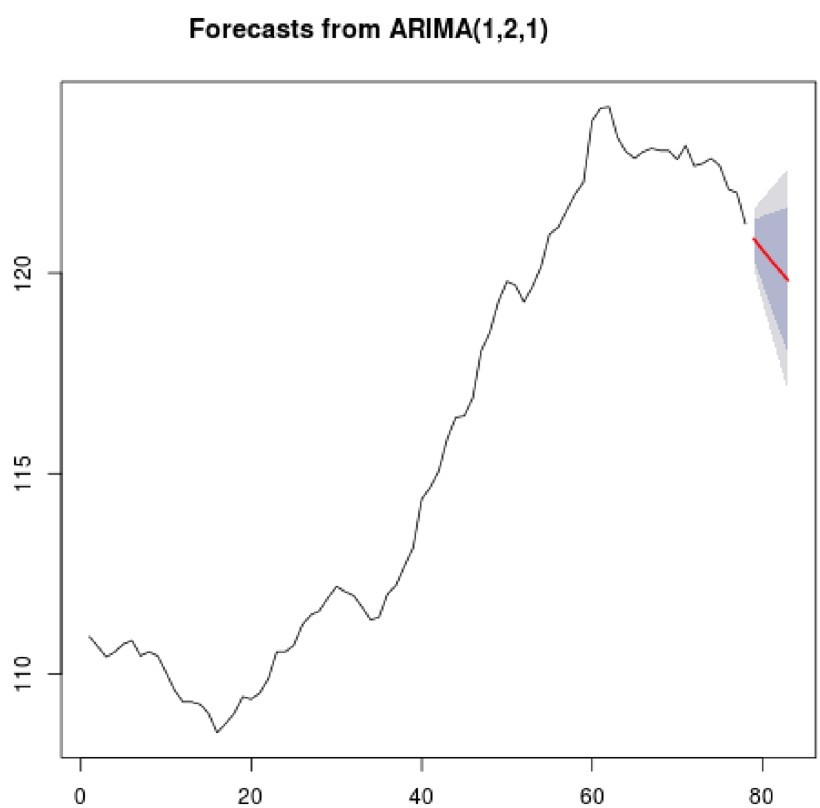

Figure 11: plot of chunk Q3C

4. (optional) Can you come up with some other way of fitting the dowj model?

```
ARIMA.MDL <- vector("list", length = 4)

bestAIC <- NULL

for (p in 0:5) {
    for (d in 0:5) {
        for (q in 0:5) {
            temp <- Arima(D1, order = c(p, d, q))
            if (is.null(bestAIC))
                bestAIC <- temp else {
                if (bestAIC$aic > temp$aic) {
                  bestAIC <- temp
                }
            }
        }
    }
}

## Warning: possible convergence problem: optim gave code = 1

## Error: non-stationary AR part from CSS
```

```
summary(bestAIC)

## Series: D1
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1     ma1
##        0.851  -0.526
## s.e.  0.138   0.255
##
## sigma^2 estimated as 0.143:  log likelihood=-34.69
## AIC=75.38   AICc=75.71   BIC=82.41
##
## Training set error measures:
##                    ME   RMSE  MAE     MPE   MAPE   MASE     ACF1
## Training set 0.03566 0.3764 0.29 0.03178 0.2492 0.8488 -0.07309
```

```
plot(forecast(bestAIC, 5), fcol = 2)
```

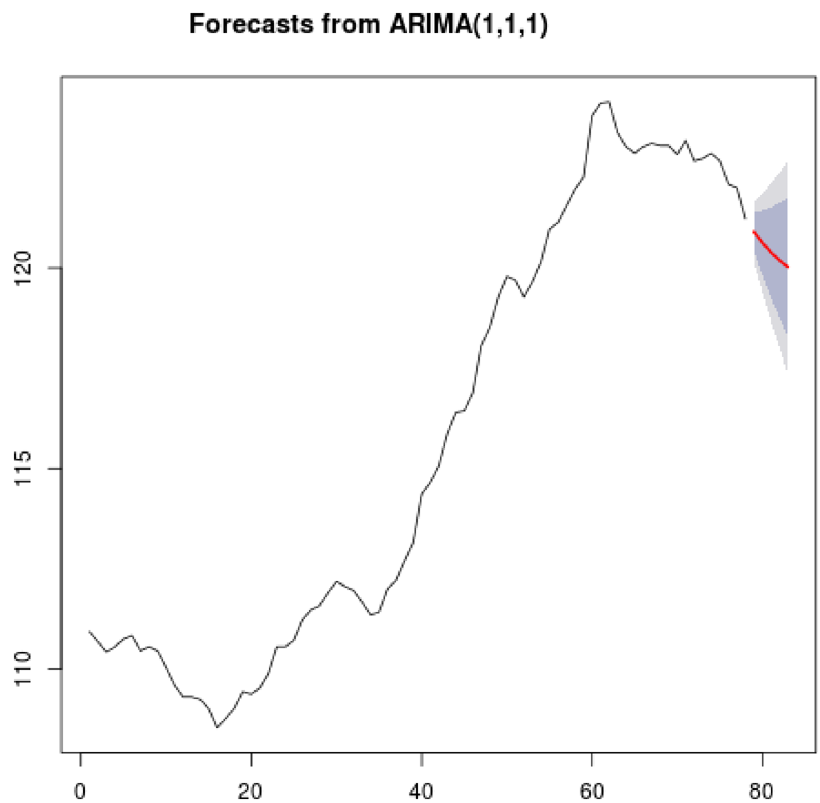

Figure 12: plot of chunk Q4

```r
adf.test(bestAIC$residuals, alternative = "stationary")
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  bestAIC$residuals
## Dickey-Fuller = -3.961, Lag order = 4, p-value = 0.01591
## alternative hypothesis: stationary
```

```r
Randomness.tests(bestAIC$residuals)
```

```
##                [,1]
## BL15         0.2869
## BL20         0.0877
## BL25         0.0943
## ML15         0.7007
## ML20         0.5222
## WX           0.5141
## JB.X-squared 0.0480
## SD           0.3771
```

5. Which model do you like better (2-b), (3-b) or 4? Why?

The exhaustive search I used in 4 found that ARIMA(1,1,1) is the best fit, which makes me agree with 2-b.