UNIVERSITY OF BRITISH COLUMBIA
Department of Statistics
STAT 443: Time Series and Forecasting
Assignment 1 Solution: Analysis in the Time Domain

1. (10 marks) This question concerns a test to determine whether it seems likely that a time series is in fact a realization of a white noise process. As usual, we assume we have observed a series $x(1), \ldots, x(N)$ from a stochastic process $X(t)$, $t = 1, 2, \ldots$ Use the `rnorm` command in R to generate an i.i.d. series of size 200 from the $N(0, 2^2)$ distribution. Find the acf of your data and find out how to extract the values from R up to a given lag (you will find `help(acf)` useful).
   To simulate the vector use
   `x <- rnorm(200,0,2)`
   Plot should have no obvious pattern, and acf should have only occasional values (if any) outside $\pm 2/\sqrt{200} \approx \pm\ 0.141$.

   (a) One test for white noise is the *portmanteau lack–of–fit test.* This test has test statistic

   $$Q := N \sum_{k=1}^{M} r_k^2,$$

   where r_k is the sample autocorrelation at lag k, N is the number of terms in the series and M is an integer rather less than N, usually between 15 and 30. If the observations are from a white noise process, then approximately $Q \sim \chi^2_M$. Otherwise the value of Q is inflated. The choice of M is not straightforward, however. Perform this test for two different values of M for the data you generated, selecting the values of M at random between 15 and 30 inclusive. Quote the P–value of the test on each occasion, and comment on your results.
   Here the default number of lags may not be sufficient, so use
   `> rho <- acf(x, lag=50)$acf`
   to extract the acf values into a vector. In general we have
   `> Q <- 200*sum(rho[2:(M+1)]**2)`
   noting that the sum above starts with the second component of rho,

the first being r_0. *For example, choosing* $M = 15$, 25, *and 30 returned values of* Q *of 13.0481, 17.2237, and 20.8971 respectively. In turn, these have P-values 0.598, 0.873, and 0.891, obtained via, for example*

```
> pchisq(Q, M, ncp=0, lower.tail = F)
```

As would be expected here, the test statistics are far from being significant. (2 marks)

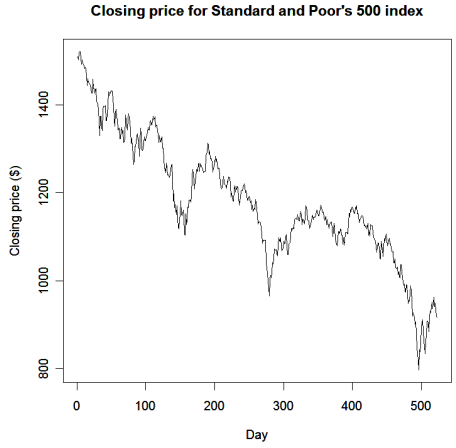
(b) The data file `SP500.txt` contains 523 consecutive closing values for Standard and Poor's 500 Index. Read the data set into R, and coerce the data into a time series object. Create a plot of the data, and of the acf of the series. Comment on what you observe.

The data are read into R via

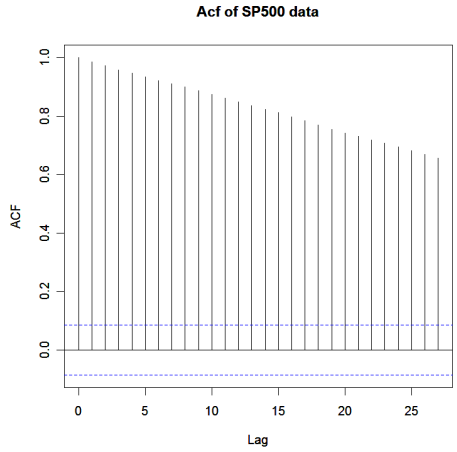
```
> SP500 <- read.table("SP500.txt", header=T)
> SP500 <- ts(SP500)
```

To plot the series, we use the following, or similar

```
> plot(SP500, main = "Closing price for Standard and Poor's
500 index", xlab = "Day", ylab = "Closing price ($)")
```



For the acf

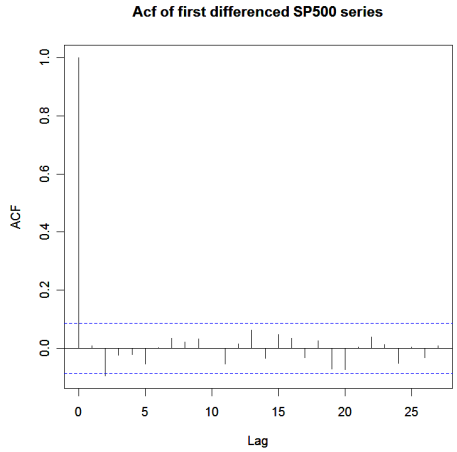


Series has apparent downward trend, indicative of the resulting acf. (2 marks)

(c) Use a portmanteau lack–of–fit test with $M = 25$ to decide whether the series appears to be a realization from a white noise process.
With $M = 25$ in the portmanteau test we have
`> rhoSP <- acf(SP500, lag=40)$acf`
`> 523*sum(rhoSP[2:26]**2)`
and the test statistic is $Q = 9201.848$. Comparing with χ^2_{25} this has a miniscule P-value and so rejects the null hypothesis that the series is white noise. (2 marks)

(d) Apply an operator to the SP500 series that might reasonably be expected to remove the non-stationary component. Plot the new series and its acf, and comment.
To remove the trend it seems reasonable to take first differences:
`> diffSP500 <- diff(SP500, lag=1)`
`> plot(diffSP500, main="First differenced SP500 data", ylab="Differences")`
The resulting series has no obvious trend or seasonal component.
`> acf(diffSP500)`

> `acf(diffSP500, main="Acf of first differenced SP500 series")`



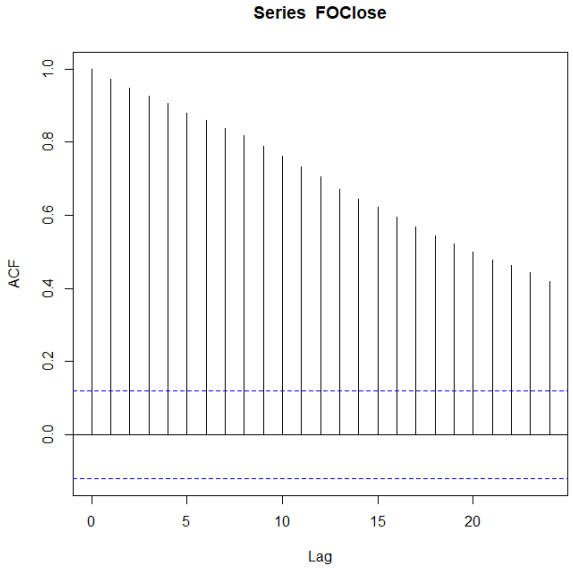
The resulting acf looks consistent with a stationary series, possibly even white noise. Only the value at lag 2 is (just) significantly different from zero, of the first 25 values. (2 marks)

(e) Use a portmanteau lack–of–fit test with $M = 25$ to decide whether the series you created in (d) appears to be a realization from a white noise process.
With $M = 25$ in the portmanteau test we have
> `rhodiffSP <- acf(diffSP500, lag=40)$acf`
> `522*sum(rhodiffSP[2:26]**2)`
and the test statistic is $Q = 22.9012$. Comparing with χ^2_{25} this has a P-value of 0.5833 and so we cannot reject the null hypothesis that the differenced series is white noise. (2 marks)

2. (10 marks) The variable `Close` in the data file `FeedOneCloseJan15Feb16` gives closing price (in Yen) of Feed One Co. Ltd., as listed on the Tokyo Stock Exchange, 5th January 2015 to 29th January 2016.

(a) Plot the time series and the acf of the series, and comment on what you observe. Does the closing price time series appear to be stationary?
A plot of the raw data shows variation with no obvious pattern or trend. The acf however, is not consistent with a series that is

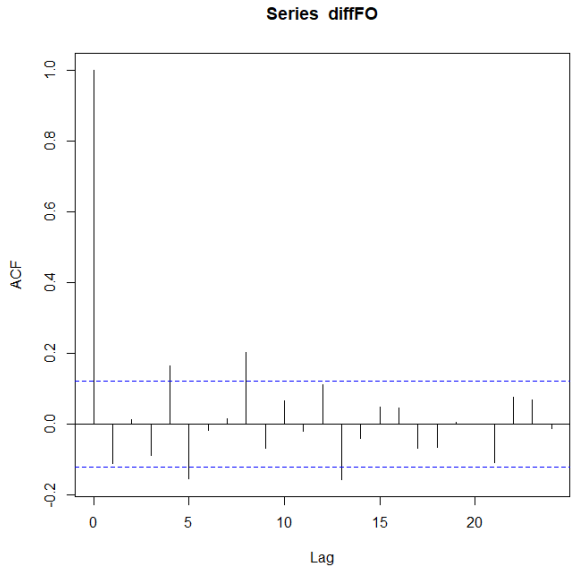
stationary as it decays very slowly.



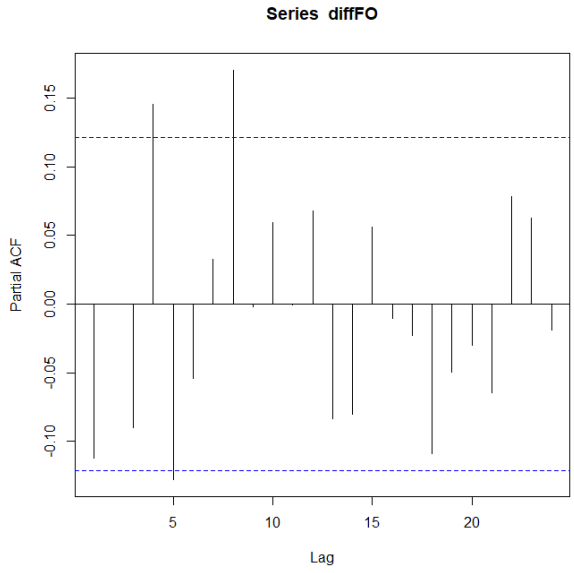
(1 mark)

(b) Apply the first difference operator to the time series. Plot the new series, the acf, and pacf of the new series, and comment on their patterns. Does the new series appear to be stationary?
The natural step is to apply ∇ *to the series. This gives a new series, "diffFO" say, which shows variation without apparent trend (some evidence of nonconstancy of variance due to higher volatility near the middle of the series) with acf that shows some spikes at*

low-ish lags but otherwise appears to decay quite slowly.



The pacf has more pronounced behaviour, apparently cutting off after the first lag.



There is good evidence to suggest the first differenced series is stationary. (1 mark)

(c) Fit various models from the ARMA(p, q) family to the new series. You should only consider models for which $p + q \leq 6$, and you should state which three models you would consider to be the best options and clarify why you chose those three models. Write out your three possible models in full, providing all the parameter estimates.

Note 1: Fit your models directly to the series created in (b), as `arima(x, order = c(p, 1, q))` gives a different model to `arima(diff(x), order = c(p, 0, q))`. Clarify this for yourself and identify how the two differ.

Note 2: For models with an AR component, R's use of the term "intercept" is not intuitive.

The plots suggest perhaps a white noise model could be fitted to the data. In fitting such a model however the diagnostics indicate low P–values for the Box–Ljung tests and spikes in the acf of the residuals. Moreover the AIC value of 1440.92 is not one of the lowest.

Trial and error indicates no clear winners, but of models that fit well in terms of model diagnostics and have the lowest AIC values, the following are about the best:

$ARMA(3, 3)$:

$$\begin{aligned}\nabla Y(t) &= X(t) \\ &= -1.100\,(X(t-1) - 0.0566) - 1.103\,(X(t-2) - 0.0566) \\ &\quad -0.856\,(X(t-3) - 0.0566) + Z(t) + 1.16Z(t-1) \\ &\quad +1.018Z(t-2) + 0.701Z(t-3)\end{aligned}$$

where $Z(t)$ has variance estimated at 13.05. This model has AIC of 1433.12, and diagnostics look good in that there are no spikes in the residual acf and the P-values of the Ljung–Box tests are all high.

$AR(5)$:

$$\begin{aligned}\nabla Y(t) &= X(t) \\ &= -0.0796\,(X(t-1) - 0.059) - 0.0167\,(X(t-2) - 0.059) \\ &\quad -0.0759\,(X(t-3) - 0.059) + 0.134\,(X(t-4) - 0.059) \\ &\quad -0.129\,(X(t-5) - 0.059) + Z(t)\end{aligned}$$

where $Z(t)$ has variance estimated at 13.45. This model has AIC of 1438.6. Again, P-values for L-B test high, no spikes in acf of residuals.
$AR(6)$:

$$\begin{aligned}\nabla Y(t) &= X(t) \\ &= -0.0872\,(X(t-1)-0.0578) - 0.0890\,(X(t-2)-0.0578) \\ &\quad -0.0801\,(X(t-3)-0.0578) - 0.133\,(X(t-4)-0.0578) \\ &\quad -0.133\,(X(t-5)-0.0578) - 0.0552\,(X(t-6)-0.0578) + Z(t)\end{aligned}$$

where $Z(t)$ has variance estimated at 13.41. This model has AIC of 1439.8. Again, P-values for L-B test high, no spikes in acf of residuals.
Other possibilities include $ARMA(3,2)$ (AIC = 1441.11, diagnostics good) and $AR(4)$ (AIC of = 1441.03, diagnostics OK). (3 marks)

(d) An alternative, and modern, way of model selection is to split the data into a *training set* and a *test set.* The idea is that we use the training set to determine fit and explore competing models, and then assess how well the models perform when fitting values in the test set. A popular criterion to adopt in assessing a model's performance on the test set is mean squared error.
The above can be applied to time series in R. Suppose the time series is `x` and we wish to use the final `m` values as the test set. Then

```
train <- 1:(length(x)-m)
trainx <- x[train]
testx <- x[-train]
# sets up training set and test set
model <- arima(trainx, order = c(p, 0, q))
foremodel = predict(model, m)
#fits model to training set, uses model to predict test
set
error <- sum((testx - foremodel$pred)^2)
# computes squared errors over test set
```

Using the differences in January 2016 closing prices as the test set for your models and the remaining data as the training set,

compare the three models you selected in (c) for performance over the test set as above.
For the ARMA$(3,3)$ model the error SS is 181.401, for the AR(5) it is 193.772, and for the AR(6) it is 192.213. So ARMA $(3,3)$ is the clear winner. By comparison, error SS for ARMA$(3,2)$ and AR(4) are 193.884 and 189.729 respectively. So the model with the best AIC also wins the test set comparison. (3 marks)

(e) Use your winning model from (d), when fitted to the entire series, to forecast the next two values in the differenced series. Hence forecast the next two values in the closing price series.
We can use
`predict(ARMA33, n.ahead=2)`
This gives

$$\begin{aligned} \hat{x}\,(262,1) &= -1.8511, \\ \hat{x}\,(262,2) &= 3.6129 \end{aligned}$$

with estimated s.e.'s of 0.98178 and 3.6256 respectively. Now since $\hat{x}\,(262,1)$ *estimates* $y\,(264) - y\,(263)$ *we have*

$$\begin{aligned} \hat{y}\,(263,1) &= \hat{x}\,(262,1) + y\,(263) \\ &= -1.8511 + 126 \\ &= 124.15. \end{aligned}$$

In a similar fashion,

$$\begin{aligned} \hat{y}\,(263,2) &= \hat{x}\,(262,2) + y\,(264) \\ &= 3.6129 + 124.\,15 \\ &= 127.\,76. \end{aligned}$$

So the forecasts for the next two trading days would be 124.15 Yen and 127.76 Yen respectively. (2 marks)

BD