ORIE 4630: Spring Term 2019
Homework #4
Due: Thursday, February 28, 2019

Students are required to work independently on homework. You should not give or receive help from other students. You should also not receive help from students or former students who took this course in previous years and who may have solutions to similar problems. The solutions you submit should be your own work and not copied from elsewhere.

Homework is due at the end of lecture (12:55pm) on the due date. You will usually have one week to do the assignments. Please don't wait until the homework is nearly due to start. Late homework is not accepted. Also, homework is not accepted by email. You can submit your assignment in lecture or in the drop box in Rhodes Hall.

Please print your name on the front of your homework so that it is legible.

Include your `R` code, output, graphs, and other work with your homework. This will allow the grader to find any errors you make and to give partial credit.

This assignment concerns daily prices and returns of certain stocks from Jan 4, 2006 to Aug 18, 2017. The log prices for the stocks under consideration are contained in a comma separated values (csv) file named `prices.csv`; the log gross returns are contained in a csv file named `returns.csv`. You should download these files from the course Blackboard site and put them into your `R` or `Rstudio` working directory. Each of the files has 35 columns. For each file, the first column shows the date (Date), and the next 30 columns are for the stocks that are the components of the Dow Jones Industrial average (DOW). The final four columns in each file are for the Dow Jones Industrial average index (DOW), the NASDAQ composite index (NASD), the NASDAQ 100 index (NASD100), and the S&P 500 index (SP500). There are 2928 days of log prices and 2927 days of log gross returns. The log prices and the log gross returns are calculated from the adjusted closing prices downloaded from Yahoo.

Start `R` or `Rstudio` and run the following lines:

```
1 Prices=read.csv("prices.csv")
2 Returns=read.csv("returns.csv")
3 Prices$Date=as.Date(Prices$Date, format="%m/%d/%Y")
4 Returns$Date=as.Date(Returns$Date, format="%m/%d/%Y")
```

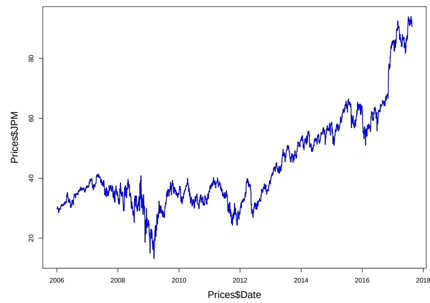
Line 1 reads the log prices into a data frame named `Prices`; line 2 reads the log gross returns into a data frame named `Returns`. If `prices.csv` is not in your working directory, then you need to give a complete path to that file in line 1; a similar comment applies for `returns.csv` line2. The function `as.date()` used in line 3 changes the class of `Prices$Date` from `factor` to `Date`; the same function is used similarly for `Returns$Date` in line 4.

To help determine whether a time series is stationary, it is useful to make a plot of the series in chronological order. The following lines produce time-series plots of the prices and returns for J. P. Morgan (JPM):

```
5 plot(Prices$Date , Prices$JPM, type="l", lwd=2, col="blue")
6 plot(Returns$Date, Returns$JPM, type="l", lwd=1, col="blue")
```

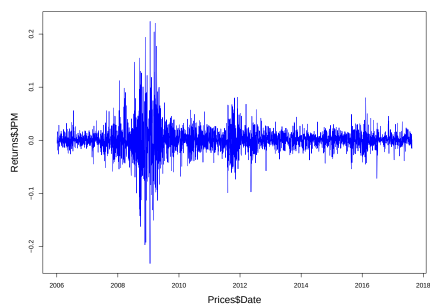
In the plots that result from lines 5 and 6, the dates from the object `Prices$Date` are plotted on the horizontal axis; in the plot that results from line 5, the log prices from the object `Prices$JPM` are plotted on the vertical axis; and in the plot that results from line 6, the log gross returns from the object `Returns$JPM` are plotted on the vertical axis. In both lines 5 and 6, the command `type="l"` causes points in the plots to be connected by lines in chronological order; the commands `lwd=2` and `lwd=1` determine the widths of the lines in the plots; and the command `col="blue"` causes the lines to be produced in the color blue.

Line 5 creates the plot:



The time-series plot of the log prices exhibits non-stationarity in two ways. First, the mean is not constant with time; and second, the variance also appears to be changing with time.

The plot produced by line 6 is



The time-series plot of the log gross returns suggests that the mean is constant over time; but the appearance of volatility clustering in 2008 and 2009 indicates that the variance is changing with time.

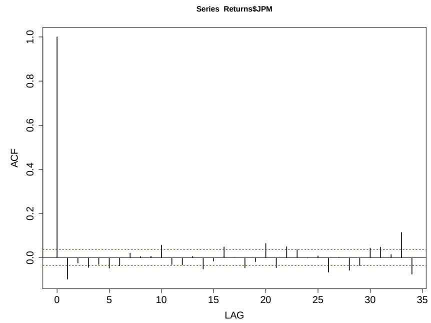
To help determine whether the returns are independent over time, it is useful to make a plot of the sample ACF (autocorrelation function). Recall that if the returns process $\ldots, r_{-1}, r_0, r_1, \ldots$ is stationary, then the autocorrelation of lag h is $\rho(h) = \mathrm{cor}(r_t, r_{t+h})$. Of course, $\rho(0) = 1$. If the returns are independent, or if the returns are a white noise process, then $\rho(h) = 0$ $(h = 1, 2, \ldots)$. Based on observed returns $r_1, \ldots, r_n$, the sample ACF function is $\hat{\rho}(0) = 1$, and

$$\hat{\rho}(h) = \sum_{i=1}^{n-h} (r_{i+h} - \bar{r})(r_i - \bar{r}) / \sum_{i=1}^{n} (r_i - \bar{r})^2, \qquad (h = 1, 2, \ldots).$$

Based on the returns for J. P. Morgan (JPM), the sample ACF is computed and plotted by running the following line:

```
7 acf(Returns$JPM)
```

Line 7 produces the plot:



The plot of the sample ACF indicates that the lag 1 autocorrelation, $\rho(1)$, is significant.

The Ljung-Box test is for testing the null hypothesis $H_0 : \rho(1) = \cdots = \rho(K) = 0$ against the alternative hypothesis H_A : at least one of $\rho(1), \ldots, \rho(K)$ is non-zero. The test statistic is

$$X^2 = n(n+2) \sum_{h=1}^{K} \frac{\{\hat{\rho}(h)\}^2}{n-h}.$$

For large n, the distribution of X^2 under the null hypothesis is approximately $\chi^2_{(K)}$. The Ljung-Box test for $K = 5$ can be performed in the case of the returns for J. P. Morgan (JPM) by running the following line:

```
8 Box.test(Returns$JPM, lag=5, type="Ljung-Box")
```

Note that in line 8, the command `lag=5` specifies that $K = 5$ for the test. Typically, the values $K = 5$, $K = 10$, and $K = 15$ are used.

Line 8 produces the output:

```
9          Box-Ljung test
10
11 data:  Returns$JPM
12 X-squared = 43.76, df = 5, p-value = 2.591e-08
```

Note that, in line 12, the value of the test statistic is $X^2 = 43.76$, and the $\chi^2_{(5)}$ distribution is used to obtain the p-value, since $K = 5$ in line 12. The p-value is very small, so the Ljung-Box test rejects the null hypothesis $H_0 : \rho(1) = \cdots = \rho(5) = 0$; similar results are obtained if `lag=10` or `lag=15` is used in line 12, so both of the null hypotheses $H_0 : \rho(1) = \cdots = \rho(10) = 0$ and $H_0 : \rho(1) = \cdots = \rho(15) = 0$ are rejected. Note that the conclusion from the Ljung-Box test is consistent with the pattern of the sample autocorrelations shown in the sample ACF plot.

If the sample ACF plot and the Ljung-Box test suggest that the returns show serial correlation, i.e., if the plot and test indicate non-zero auto-correlations, then it could be useful to model the distribution of the returns.

One simple model for the distribution of the returns is the AR(1) model. To fit the AR(1) model to the returns for J. P. Morgan (JPM) and to assess the goodness-of-fit of the model, run the following lines:

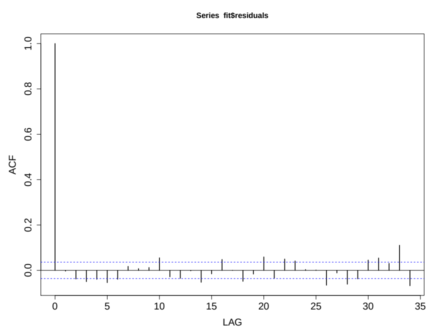
```
13 fit=arima(Returns$JPM, order=c(1,0,0))
14 fit
15 acf(fit$residuals)
16 Box.test(fit$residuals, lag=10, type="Ljung-Box", fitdf=1)
```

In line 13, maximum likelihood estimation is used to fit the AR(1) model, and the various quantities that result from the estimation process are stored in the object `fit`. The output that results from Line 13 produces a summary of the estimation process. Line 15 produces the sample ACF plot of the residuals, which are stored in the object `fit$residuals`. Line 16 performs the Ljung=Box test with `lag=10`, i.e., $K = 10$, on the residuals in the object `fit$residuals`. Since the AR(1) model is fit, one autoregressive parameter is estimated, so the command `fitdf=1` is used to reduce the degrees of freedom used for the χ^2-distribution in the Ljung-Box test from K to $K - 1$.

The output from line 14 is

```
17 Call:
18 arima(x = Returns$JPM, order = c(1, 0, 0))
19
20 Coefficients:
21           ar1  intercept
22       -0.0968      4e-04
23 s.e.   0.0184      4e-04
24 sigma^2 estimated as 0.0006674:  log likelihood = 6548.02,  aic = -13090.04
```

The output from line 15 is



The plot of the sample ACF indicates that significant autocorrelations persist, so this plot provides evidence that the AR(1) model is inadequate.

The output from line 16 is

```
25 Box-Ljung test
26
27 data:  fit$residuals
28 X-squared = 39.653, df = 9, p-value = 8.781e-06
```

In line 28, the Ljung-Box test with $K = 10$ based on the residuals gives a value of the test statistic $X^2 = 39.653$. The $\chi^2_{(9)}$-distribution is used to compute the p-value since one autoregressive parameter ϕ was estimated to compute the residuals, `fitdf=1`, and $K-1 = 9$. The p-value in line 28 is very small, so the null hypothesis $H_0 : \rho(1) = \cdots = \rho(10) = 0$ is rejected. Thus, the test suggests that the AR(1) model disturbances, the ϵ_ts, which are estimated by the residuals in the object `fit$residuals`, do not follow a white noise process. This conclusion of non-zero autocorrelations is also supported by the sample ACF plot of the residuals from line 15.

The value of the information criterion AIC$= -2 * \log$ likelihood $+ 2 * p$, where p is the number of fitted parameters, is automatically given in the summary of the estimation process; see line 23. Recall that BIC$= -2 * \log$ likelihood $+ \log(n) * p$. To output the values of the AIC and BIC information criteria contained in the object `fit` defined in line 13, run the following lines:

```
29 AIC(fit)
30 BIC(fit)
```

Recall that the AR(1) model is $(r_t - \mu) = \phi(r_{t-1} - \mu) + \epsilon_t$, which can be written in the form $r_t = \mu(1-\phi) + \phi r_{t-1} + \epsilon_t = c + \phi r_{t-1} + \epsilon_t$, where $c = \mu(1-\phi)$. In line 21, the autoregressive parameter ϕ is referred to as `ar1`, and the constant c is referred to as `intercept`. Thus, from line 22, the maximum likelihood estimate of ϕ is -0.0968 and the maximum likelihood estimate of c is 0.0004. Standard errors of the maximum likelihood estimates are given in line 23. Note that the constant c is not significantly different from 0, which implies that the mean $\mu = c/(1-\phi)$ is also not significantly different from 0.

The approach to model building is straightforward, and is based on the assumption that the model disturbances, the ϵ_ts, are a white noise process. A model is fit to the returns, and then the residuals are computed. The sample ACF based on the residuals is plotted, and the Ljung-Box test on the residuals is performed. If the sample autocorrelation function or the Ljung-Box test indicate significant autocorrelations, then the model is judged to be inadequate, and another model is sought.

In the case of the returns for J. P. Morgan (JPM), the AR(1) model is judged to be inadequate, and a higher-order autoregressive model, AR(p), might be entertained. To fit the AR(6) model ($p = 6$) to the returns for J. P. Morgan (JPM) and to assess the goodness-of-fit of the model, run the following lines:

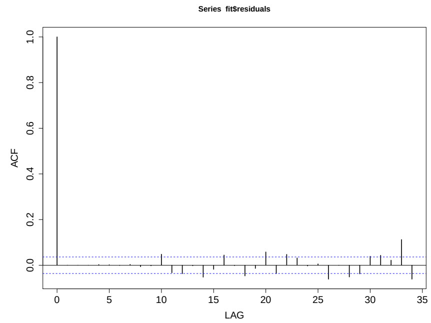
```
31 fit=arima(Returns$JPM, order=c(6,0,0))
32 fit
33 acf(fit$residuals)
34 Box.test(fit$residuals, lag=10, type="Ljung-Box", fitdf=6)
```

Note that the command `lag=10` is used in line 34 for the Ljung-Box test, so the test is for $K = 10$. Since a model with $p = 6$ autoregressive parameters is fit in line 31, the command `fitdf` $=$ `6` is used in line 31 for the Ljung-Box test. Thus, the degrees of freedom for the χ^2-distribution in the Ljung-Box test is $K - p = 10 - 6 = 4$.

Line 32 produces the output

```
35 Call:
36 arima(x = Returns$JPM, order = c(6, 0, 0))
37
38 Coefficients:
39           ar1      ar2      ar3      ar4      ar5      ar6  intercept
40       -0.1091  -0.0456  -0.0591  -0.0489  -0.0631  -0.0522      4e-04
41 s.e.   0.0185   0.0185   0.0185   0.0185   0.0185   0.0184      3e-04
42
43 sigma^2 estimated as 0.0006599:  log likelihood = 6564.58,  aic = -13113.15
```

Line 33 produces the sample ACF plot



Line 34 produces the Ljung-Box output

```
44          Box-Ljung test
45
46 data:  fit$residuals
47 X-squared = 6.9317, df = 4, p-value = 0.1395
```

The plot of the sample ACF suggests that there are no non-zero autocorrelations, at least for lags less than 10; the p-value in line 47 from the Ljung-Box test with $K = 10$ is insignificant. The AR(6) model might be judged to be adequate; however, there is some evidence that non-zero autocorrelations might exist at lags 10 and greater. Consequently, alternative models might be sought.

Another class of models the might be considered are the moving average models. To fit the MA(1) model to the returns for J. P. Morgan (JPM) and to assess the appropriateness of the model, run the following lines:

```
48 fit=arima(Returns$JPM, order=c(0,0,1))
49 fit
50 acf(fit$residuals)
51 Box.test(fit$residuals, lag=10, type="Ljung-Box", fitdf=1)
```

Note that the command `fitdf=1` is used in line 34 because the residuals in the object `fit$residuals` are produced by fitting a model with 1 moving average parameter.

A higher-order autoregressive model, MA(q), might be entertained. To fit the MA(6) model ($q = 6$) to the returns for J. P. Morgan (JPM) and to assess the goodness-of-fit of the model, run the following lines:

```
52 fit=arima(Returns$JPM, order=c(0,0,6))
53 fit
54 acf(fit$residuals)
55 Box.test(fit$residuals, lag=10, type="Ljung-Box", fitdf=6)
```

Note that the command `fitdf=6` is used in line 55 because the residuals in the object `fit$residuals` are produced by fitting a model with 6 moving average parameters.

It turns out that the moving average models do not fit the returns for J. P. Morgan (JPM) especially well. A more complicated class of models are the autoregressive-moving average models, ARMA(p,q). To fit the ARMA(2,1) model to the returns for J. P. Morgan (JPM) and to assess the appropriateness of the model, run the following lines:

```
56 fit=arima(Returns$JPM, order=c(2,0,1))
57 fit
58 acf(fit$residuals)
59 Box.test(fit$residuals, lag=10, type="Ljung-Box", fitdf=3)
```

Note that the command `fitdf=3` is used in line 59 because the residuals in the object `fit$residuals` are produced by fitting a model with 2 autoregressive parameters and 1 moving average parameter.

The process of model selection can be automated by using the function `auto.arima()` in the package `forecast`. To install and load the package `forecast`, run the following lines.

```
60 install.packages("forecast")
61 library(forecast")
```

To identify the "best fitting" AR(p) model for $p \leq 20$ by using the AIC information criterion in the case of the returns from J. P. Morgan (JPM), run the following line:

```
62 auto.arima(Returns$JPM, max.p=20, max.q=0, d=0, ic="aic")
```

To identify the "best fitting" MA(q) model for $q \leq 20$ by using the AIC information criterion in the case of the returns from J. P. Morgan (JPM), run the following line:

```
63 auto.arima(Returns$JPM, max.p=0, max.q=20, d=0, ic="aic")
```

To identify the "best fitting" ARMA(p,q) model for $p \leq 20$ and $q \leq 20$ by using the AIC information criterion in the case of the returns from J. P. Morgan (JPM), run the following line:

```
64 auto.arima(Returns$JPM, max.p=20, max.q=20, d=0, ic="aic")
```

To identify the "best fitting" ARIMA(p,d,q) model by using the AIC information criterion in the case of the returns from J. P. Morgan (JPM), run the following line:

```
65 auto.arima(Returns$JPM, ic="aic")
```

To use the BIC information criterion instead of using the AIC information criterion to select the model, replace the command `ic="aic"` by `ic="bic"` in lines 62 through 65.

Point predictions and prediction intervals can be obtained by using the function `forecast()`. To produce the point predictions along with 50%, 80%, and 95% prediction intervals for the next 5 returns from J. P. Morgan (JPM) based on the ARIMA(1,0,1) model, run the following lines:

```
66 fit=arima(Returns$JPM, order=c(1,0,1))
67 fc=forecast(fit, h=5, level=c(50,80,95))
68 fc
```

The ARIMA(1,0,1) model is fit in line 66, and the point predictions and the prediction intervals are are computed and stored in the object `fc` in line 67. The output of the point predictions and the prediction intervals results from line 68.

Plots of the predictions and prediction intervals can also be obtained. To obtain such a plot for prediction based on the ARIMA(1,1,1) model in the case of predicting the next 400 prices J. P. Morgan (JPM) , run the following lines:

```
69 fit=arima(Prices$JPM, order=c(1,1,1))
70 fc=forecast(fit, h=400, level=c(75,90,99))
71 plot(fc,shadecols=c("gray","orange","red"))
```

The ARIMA(1,1,1) is fit to the prices in line 69. The point predictions for the next 400 prices from J. P. Morgan (JPM) are computed in line 70. Also computed in line 65 are the 75%, 90%, and 99% prediction intervals. A plot of the data, the point predictions, and the prediction intervals is produced by the function `plot()` in line 71.

Questions:

1. [**10 points**] Run lines 1 to 4. Adapt lines 5 and 6 to produce time series plots for the prices and returns from United Technologies (UTX). Submit your plots. Do the plots show evidence of non-stationarity? Justify your answer briefly.

2. [**5 points**] By adapting line 7, produce a plot of the sample ACF for the returns from United Technologies (UTX). Submit your plot. Does the plot show evidence of serial autocorrelation? Justify your answer briefly.

3. [**5 points**] By adapting line 8, conduct the Ljung-Box test for the returns from United Technologies (UTX). Use $K = 12$. Submit your output. Does the test show evidence of serial autocorrelation? Justify your answer briefly.

4. [**15 points**] By adapting lines 13 to 16, fit the AR(1) model to the returns from United Technologies (UTX). Use $K = 7$ for the Ljung-Box test. Submit your output. Do the plot and the test show evidence of serial autocorrelation? Justify your answer briefly. Do the plot and test suggest that the AR(1) model is appropriate for the returns?

5. [**15 points**] By adapting lines 31 to 34, fit the AR(4) model to the returns from United Technologies (UTX). Use $K = 15$ for the Ljung-Box test. Submit your output. Do the plot and the test show evidence of serial autocorrelation? Justify your answer briefly. Do the plot and test suggest that the AR(4) model is appropriate for the returns?

6. [**15 points**] By adapting lines 52 to 55, fit the MA(5) model to the returns from United Technologies (UTX). Use $K = 13$ for the Ljung-Box test. Submit your output. Do the plot and the test show evidence of serial autocorrelation? Justify your answer briefly. Do the plot and test suggest that the MA(5) model is appropriate for the returns?

7. [**5 points**] Based on the AIC information criterion, would the AR(4) model or the MA(5) model for the returns from United Technologies (UTX) be preferred? Justify your answer.

8. [**10 points**] Run lines 60 and 61 to install the package `forecast`. By adapting line 64, find the best fitting ARMA(p,q) model for the returns from United Technologies (UTX) as identified by the `auto.arima()` function using the AIC criterion. Submit your output. Based on the AIC information criterion, would you prefer the model determined in this question or would you prefer the MA(5) model?

9. [**10 points**] By adapting lines 66 to 68, fit an ARIMA(2,1,2) model to the prices from United Technologies (UTX), and find point predictions along with 75% and 90% prediction intervals for the next 7 prices. Submit your output. What is the point prediction of the price for period $t = 2932$? What is the 75% prediction interval for the price at period $t = 2934$.

10. [**10 points**] By adapting lines 69 to 71, plot the point predictions of the prices from Proctor & Gamble (PG) along with the 50%, 75%, and 95% prediction intervals for the next 500 periods based on the ARIMA(2,1,2) model. Submit your plot.