ECOM90024
Forecasting in Economics and Business
Tutorial 8 Solutions

---

1. Consider the following AR(4) process,

$$Y_t = \mu + \delta t + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} + \varepsilon_t$$

a.) Given the above process, derive a regression model that will allow you to test whether the process possesses a unit root. In doing so, make sure to derive expressions of the regression coefficients in terms of the parameters of the above process.

The easiest way to approach this is to work backwards from the result that we desire. From the lecture notes, we know that we want a regression model of the form

$$\Delta Y_t = \mu + \delta t + \rho Y_{t-1} + \sum_{j=2}^{4} \beta_j \Delta Y_{t-j+1} + \varepsilon_t$$

Where $\rho = \phi_1 + \phi_2 + \phi_3 + \phi_4 - 1$

Therefore, as a first step, we will need to add and subtract $\phi_2 Y_{t-1}$, $\phi_3 Y_{t-1}$ and $\phi_4 Y_{t-1}$ to the right-hand side:

$$Y_t = \mu + \delta t + (\phi_1 + \phi_2 + \phi_3 + \phi_4)Y_{t-1} - (\phi_2 + \phi_3 + \phi_4)Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} + \varepsilon_t$$

The next step is that we will need to form the term $\beta_1 \Delta Y_{t-1} = \beta_1(Y_{t-1} - Y_{t-2})$. To do this, we will need to add and subtract $\phi_3 Y_{t-2}$ and $\phi_4 Y_{t-2}$ to the right-hand side:

$$Y_t = \mu + \delta t + (\phi_1 + \phi_2 + \phi_3 + \phi_4)Y_{t-1} - (\phi_2 + \phi_3 + \phi_4)(Y_{t-1} - Y_{t-2}) - \phi_3 Y_{t-2} - \phi_4 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} + \varepsilon_t$$

Then, we can proceed to form the terms $\beta_2 \Delta Y_{t-2} = \beta_2(Y_{t-2} - Y_{t-3})$ and $\beta_2 \Delta Y_{t-2} = \beta_3(Y_{t-3} - Y_{t-4})$. To do this, we will need to add and subtract $\phi_4 Y_{t-3}$ to the right-hand side:

$$Y_t = \mu + \delta t + (\phi_1 + \phi_2 + \phi_3 + \phi_4)Y_{t-1} - (\phi_2 + \phi_3 + \phi_4)(Y_{t-1} - Y_{t-2}) - (\phi_3 + \phi_4)(Y_{t-2} - Y_{t-3}) - \phi_4(Y_{t-3} - Y_{t-4}) + \varepsilon_t$$

Then, as a final step, we subtract Y_{t-1} from both sides of the equation to obtain the desired regression equation:

$$\Delta Y_t = \mu + \delta t + \rho Y_{t-1} + \sum_{j=2}^{4} \beta_j \Delta Y_{t-j+1} + \varepsilon_t$$

Where,

$$\rho = \phi_1 + \phi_2 + \phi_3 + \phi_4 - 1$$

$$\beta_1 = -(\phi_2 + \phi_3 + \phi_4)$$

$$\beta_2 = -(\phi_3 + \phi_4)$$

$$\beta_3 = -(\phi_4)$$

b.) Having derived the regression equation, explain how it can be used to test for the presence of a unit root. Make sure to write down the elements of the test (i.e., null and alternative hypotheses, test statistic and decision rule).

Writing our process in lag operator form,

$$(1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3 - \phi_4 L^4)Y_t = \mu + \delta t + \varepsilon_t$$

It is clear that the process will possess a unit root if $z = 1$ is a solution to the polynomial equation,

$$(1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3 - \phi_3 z^4) = 0$$

Thus, if we set $z = 1$, this will imply that,

$$\phi_1 + \phi_2 + \phi_3 + \phi_4 = 1$$

Looking at the regression equation that we have just derived, since

$$\rho = \phi_1 + \phi_2 + \phi_3 + \phi_4 - 1$$

We can therefore test for the presence of a unit root by performing the following hypothesis test,

$$H_0: \rho = 0$$

$$H_A: \rho \neq 0$$

The Augmented Dickey Fuller test statistic will be given by

$$ADF = \frac{\hat{\rho}}{\hat{\sigma}_{\hat{\rho}}}$$

If the test statistic exceeds the critical value, or equivalently, the associated p-value is smaller than the chosen significance level, we would reject the null hypothesis and conclude that the process does not contain a unit root.

2. Using the R statistical environment, simulate 1000 observations from an AR(4) (with no mean or deterministic trend) using a set of coefficients generated from the generateAR() function from the DREGAR package (see R-Week5.pdf)

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \phi_4 Y_{t-4} + \varepsilon_t$$

a.) Install and load the urca package. The function ur.df() performs an Augmented Dickey Fuller Test using the following syntax:

```
ur.df(y, type = c("none", "drift", "trend"), lags = k)
```

Where `y` is the time series data, `type` is the form of the ADF regression where

`"none" =` $\Delta Y_t = \rho Y_{t-1} + \sum_{j=2}^{k} \beta_j \Delta Y_{t-j+1} + \varepsilon_t$

"drift" = $\Delta Y_t = \mu + \rho Y_{t-1} + \sum_{j=2}^{4} \beta_j \Delta Y_{t-j+1} + \varepsilon_t$

"trend" = $\Delta Y_t = \mu + \delta t + \rho Y_{t-1} + \sum_{j=2}^{4} \beta_j \Delta Y_{t-j+1} + \varepsilon_t$

and `lags` is the number of lags to include in the ADF regression. Use the ur.df() function to perform an ADF test on your simulated data. Make sure to explain the lag that you've chosen to specify in the function. Do your results conform to your expectations?

The test statistic from the ADF test will be greater than the critical value and hence we would reject the null hypothesis and conclude that the data is stationary.

We should choose the max lag length of the differences in the testing equation to be equal to 3.

See R Code.

b.) Using the data that you have simulated, construct a new time series Z_t that is defined as:

$$Z_t = \sum_{s=0}^{t} Y_s$$

(Hint: you can use the cumsum() function).

Generate a plot of Z_t and describe what you see. Use the ur.df() function to perform an ADF test on Z_t. You can use the same lag specification used in part (a). Do your results conform to your expectations?

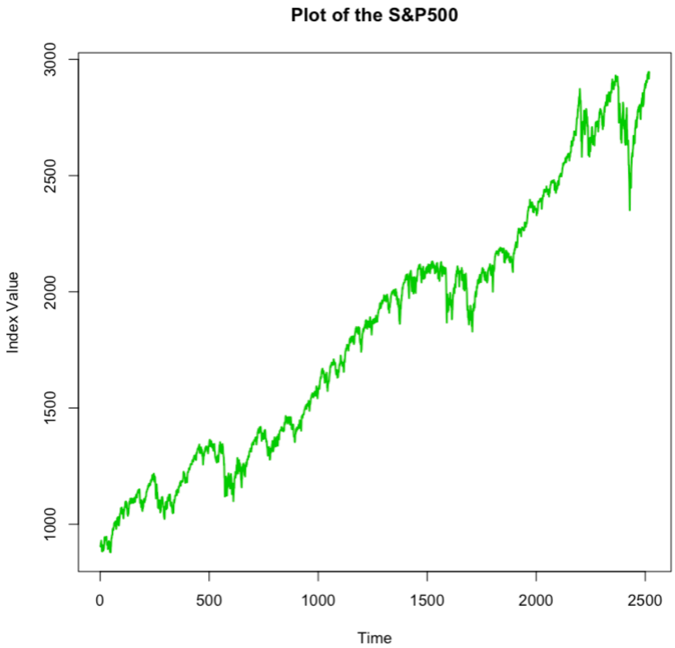
The plot should look like a random walk.

The test statistic from the ADF test will be less than the critical value and hence we would fail to reject the null hypothesis and conclude that the data has a unit root.

See R Code.

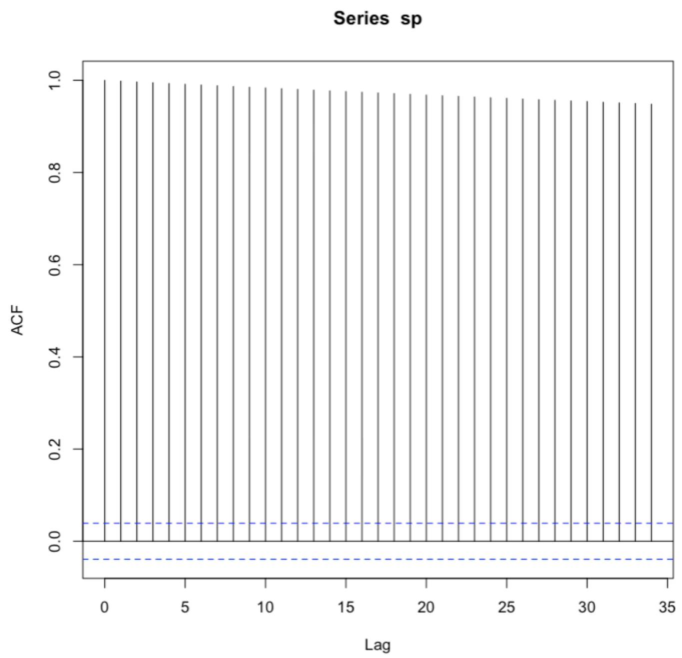
3.) Download the data contained in the file sp500.csv into your R environment.

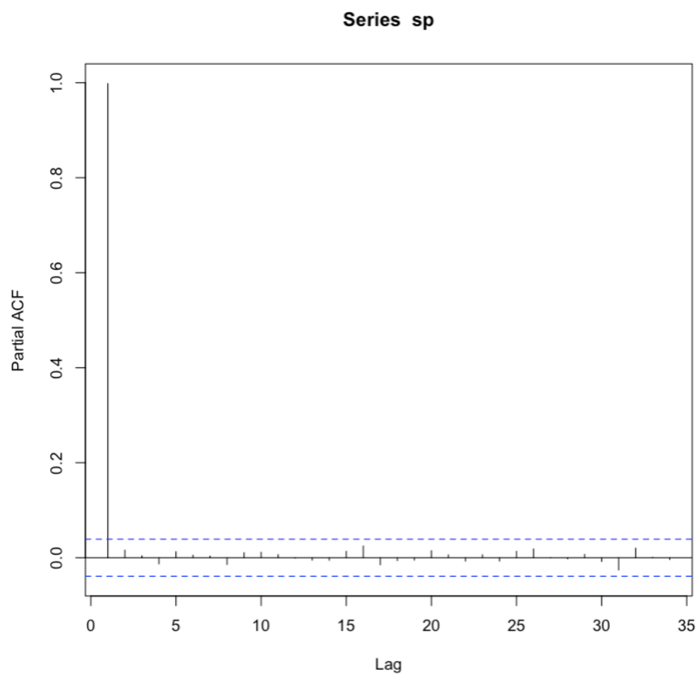
a.) Generate a plot of the data and briefly describe the primary visual characteristics of the data?



From the plot we can see that the data is clearly trending upwards. As it is financial data, it would not be reasonable to impose any sort of deterministic trend. Rather, financial theory tells us that it would be more appropriate to describe this data would be a random walk with a positive drift. See: https://en.wikipedia.org/wiki/Random_walk_hypothesis

b.) Generate sample ACF and PACF plots. What are they telling us about the dependence structure of the data?





The sample ACF and PACF also indicate that the data is being generated by a random walk. The sample autocorrelations are decaying very slowly (i.e., at a linear rate) while the first partial autocorrelation is very close to 1.

c.) Perform an ADF test using the ur.df() function in which `type = "drift"`. Set the lag length using the method of Ng & Perron discussed in class. In the output, the relevant test statistic is the first number that is reported. Compare this test statistic with the critical value labelled tau2. Do your results conform with your results in parts (a) and (b)?

```
Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max 
-118.160   -6.707    0.501    8.528  108.948 

Coefficients:
                 Estimate   Std. Error t value Pr(>|t|)    
(Intercept)   1.040433294  1.133331550   0.918 0.358693    
z.lag.1      -0.000002724  0.000592233  -0.005 0.996331    
z.diff.lag1  -0.033877840  0.020149514  -1.681 0.092827 .  
z.diff.lag2  -0.006205559  0.020160684  -0.308 0.758257    
z.diff.lag3  -0.000263526  0.020168602  -0.013 0.989576    
z.diff.lag4  -0.031869346  0.020165883  -1.580 0.114153    
z.diff.lag5  -0.055893619  0.020168123  -2.771 0.005624 ** 
z.diff.lag6  -0.007004686  0.020202261  -0.347 0.728825    
z.diff.lag7   0.014351853  0.020182665   0.711 0.477091    
z.diff.lag8  -0.044661528  0.020183454  -2.213 0.027004 *  
z.diff.lag9  -0.029474942  0.020187297  -1.460 0.144397    
z.diff.lag10  0.000477693  0.020183409   0.024 0.981120    
z.diff.lag11 -0.000133108  0.020155762  -0.007 0.994731    
z.diff.lag12  0.002068957  0.020105864   0.103 0.918048    
z.diff.lag13 -0.008036884  0.020101949  -0.400 0.689334    
z.diff.lag14 -0.068987971  0.020102091  -3.432 0.000609 ***
z.diff.lag15 -0.042848586  0.020157195  -2.126 0.033626 *  
z.diff.lag16  0.031553788  0.020164075   1.565 0.117746    
z.diff.lag17  0.047944489  0.020164409   2.378 0.017498 *  
z.diff.lag18  0.001644291  0.020171595   0.082 0.935039    
z.diff.lag19 -0.031079947  0.020160639  -1.542 0.123295    
z.diff.lag20 -0.009450534  0.020174652  -0.468 0.639514    
z.diff.lag21  0.020996948  0.020136447   1.043 0.297174    
z.diff.lag22 -0.010910369  0.020123288  -0.542 0.587746    
z.diff.lag23  0.016449047  0.020122406   0.817 0.413751    
z.diff.lag24 -0.007908351  0.020140683  -0.393 0.694608    
z.diff.lag25 -0.036863077  0.020133486  -1.831 0.067231 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.49 on 2465 degrees of freedom
Multiple R-squared:  0.02089,	Adjusted R-squared:  0.01057 
F-statistic: 2.023 on 26 and 2465 DF,  p-value: 0.001654


Value of test-statistic is: -0.0046 4.5389 

Critical values for test statistics: 
      1pct  5pct 10pct
tau2 -3.43 -2.86 -2.57
phi1  6.43  4.59  3.78
```

From the above estimation results we can conclude the following:

1. Coefficient in front of the 25th lagged difference has a t-value greater than |1.6|, so we don't reduce the lag length any further.

2. The value of the first reported test statistic is -0.0046

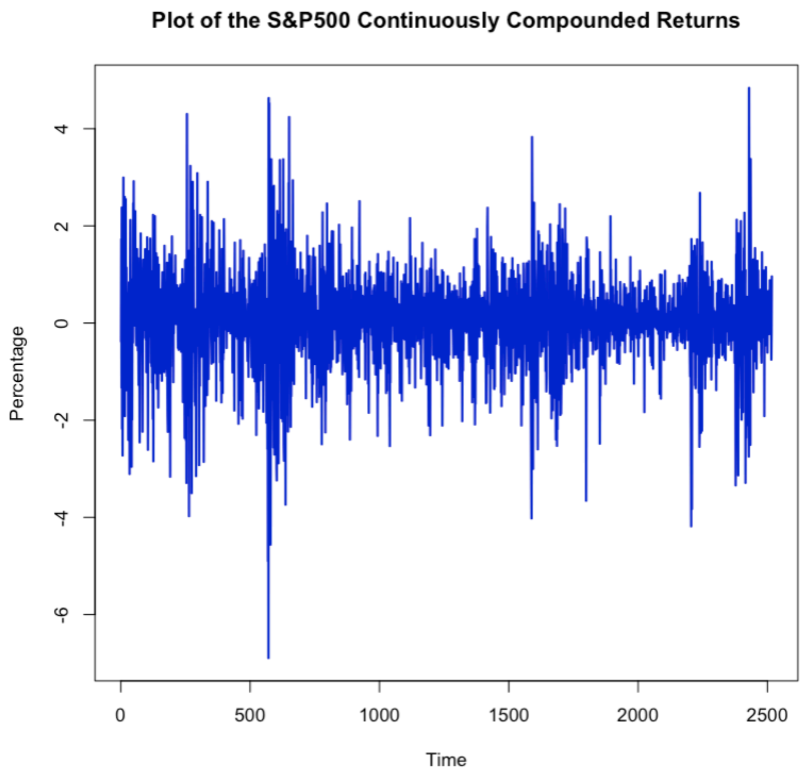
3. The value of the critical value at the 5% level of significance is tau2 = -2.86

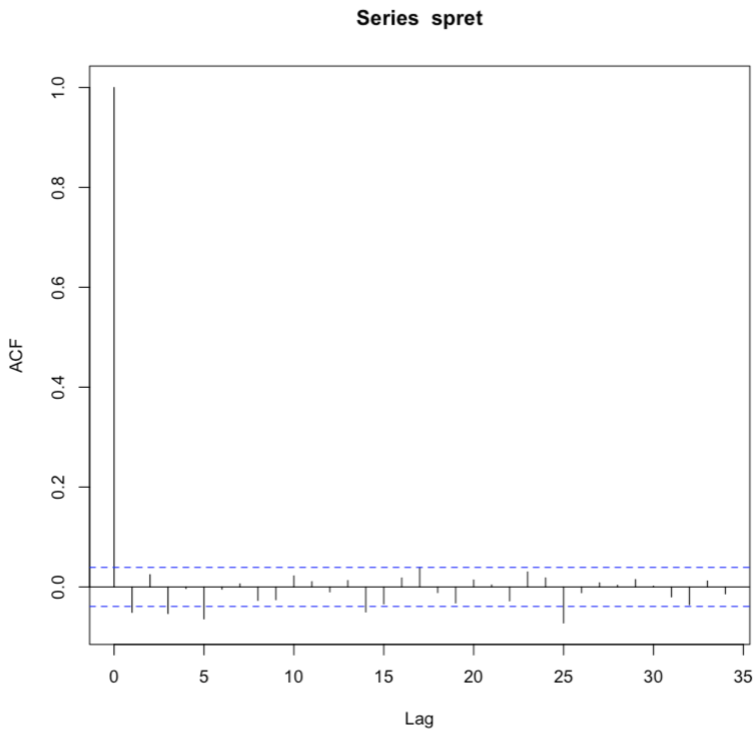
Therefore, we fail to reject the null hypothesis that there is a unit root. Thus, our test results are consistent with what we observe in the plot and the sample ACF and PACF.

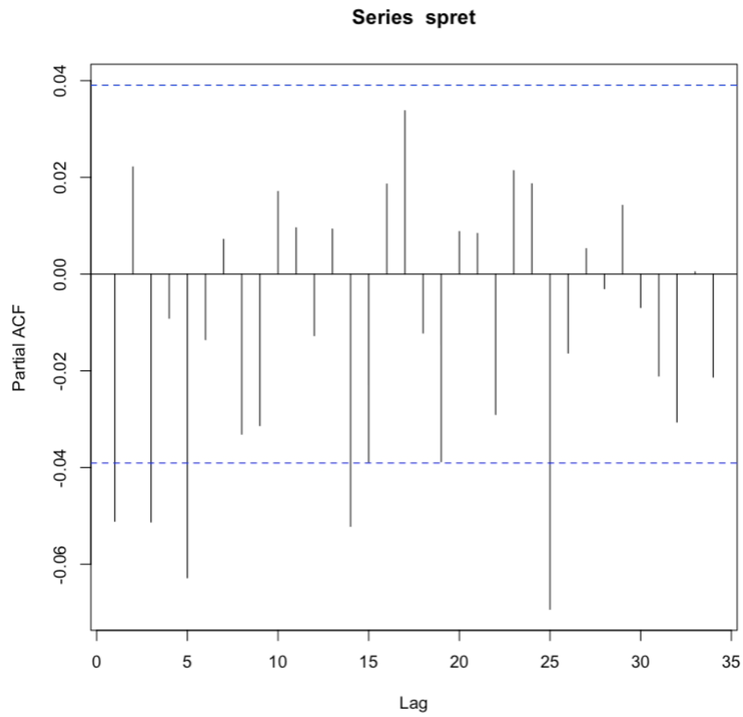
d.) Compute the continuously compounded return on the S&P 500 index using the following formula:

$$r_t = 100 \times (\ln(P_t) - \ln(P_{t-1}))$$

Plot the returns and repeat part (c) using the continuously compounded returns that you've just generated and report your results.







As we can see from the plot, the returns are clearly mean reverting and there are no clear trends or drifts. Moreover, the sample ACF and PACF are indicating to us that the data is stationary. Hence, we should perform the ADF test in which we set `type = "none"`.

```
Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
      Min        1Q    Median        3Q       Max
-0.066824 -0.003336  0.000832  0.005454  0.046797

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
z.lag.1      -1.25168    0.11582 -10.808 < 0.0000000000000002 ***
z.diff.lag1   0.21225    0.11344   1.871              0.06145 .
z.diff.lag2   0.23569    0.11089   2.125              0.03365 *
z.diff.lag3   0.19285    0.10819   1.783              0.07478 .
z.diff.lag4   0.17843    0.10552   1.691              0.09096 .
z.diff.lag5   0.11967    0.10270   1.165              0.24404
z.diff.lag6   0.10834    0.09971   1.087              0.27734
z.diff.lag7   0.11245    0.09682   1.161              0.24560
z.diff.lag8   0.08463    0.09390   0.901              0.36748
z.diff.lag9   0.05644    0.09066   0.623              0.53361
z.diff.lag10  0.07619    0.08717   0.874              0.38215
z.diff.lag11  0.08400    0.08372   1.003              0.31578
z.diff.lag12  0.08410    0.08044   1.046              0.29589
z.diff.lag13  0.09307    0.07696   1.209              0.22663
z.diff.lag14  0.04284    0.07335   0.584              0.55921
z.diff.lag15  0.01015    0.06947   0.146              0.88382
z.diff.lag16  0.03102    0.06523   0.476              0.63446
z.diff.lag17  0.06812    0.06091   1.118              0.26355
z.diff.lag18  0.05865    0.05649   1.038              0.29930
z.diff.lag19  0.02535    0.05164   0.491              0.62347
z.diff.lag20  0.03605    0.04639   0.777              0.43717
z.diff.lag21  0.04640    0.04111   1.129              0.25912
z.diff.lag22  0.01871    0.03519   0.532              0.59489
z.diff.lag23  0.04604    0.02877   1.600              0.10967
z.diff.lag24  0.06521    0.01986   3.283              0.00104 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009448 on 2467 degrees of freedom
Multiple R-squared:  0.5305,    Adjusted R-squared:  0.5257
F-statistic: 111.5 on 25 and 2467 DF,  p-value: < 0.00000000000000022


Value of test-statistic is: -10.8075

Critical values for test statistics:
      1pct  5pct 10pct
tau1 -2.58 -1.95 -1.62
```

From the above estimation results we can conclude the following:

1. Coefficient in front of the 24th lagged difference has a t-value greater than |1.6|, so we don't reduce the lag length any further.

2. The value of the first reported test statistic is -10.8075

3. The value of the critical value at the 5% level of significance is tau2 = -2.58

Therefore, we reject the null hypothesis that there is a unit root. Thus, our test results are consistent with what we observe in the plot and the sample ACF and PACF.