Stat 510 Week 1 Homework Solutions

1. The data for this problem are the October levels of Lake Erie in the United States for 40 consecutive years. The dataset is named eriedata.dat and is linked in the Datasets folder of the course website. Download the file to your computer. We suggest that you create a Stat510 folder on your computer, if you haven't already done so. Then within that folder, create a Datasets folder. Download datasets to your Datasets folder.

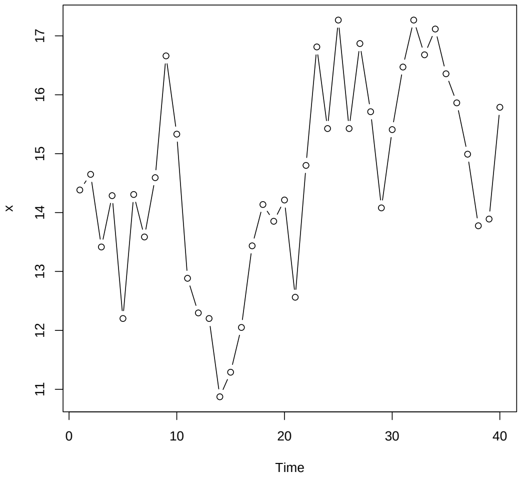
A. Start the R program. (You can use Minitab or SAS if you want, but I'm giving the instructions for R here. At the end of this document, there's a brief list of Minitab menu sequences for this assignment.)

Use **File>Change dir ..** (**Mac Users – Misc > Change Working Dir …**) as a menu sequence and select the folder in which you have stored the dataset for this problem. That will make that directory the "working" directory. Then, enter these three commands:

```
x = scan ("eriedata.dat")
x = ts (x)
plot (x, type="b")
```

The first command reads the data into a variable (object really) named x. The second command designates x as a time series object, a necessary step to make some of the time series commands work right. The third command creates a time series plot of x, the type = "b" part is optional – it puts symbols on the plot where data points fall.

As the answer to this part, copy and paste the time series plot of the data. (Right click the graph, copy it as a bitmap, and then switch to Word and paste. It can be resized there.



B. Refer to the plot made for part A. Briefly discuss features of the plot. See Lesson 1.1 for this week to see what we're worried about (trend, seasonality, outliers, etc.)

There's a tendency for values to stay on the same side of the mean for a while, though the data do not trend up or down over time. We do not see a repeating pattern to suggest seasonality as suspected because this is annual data. There are no issues with non-constant variance, outliers, cyclical behavior, or abrupt changes.

C. Use these two commands to get a plot of the autocorrelation function and a printed version of the values.

```
        acf1 = acf(x)
        acf1
```

Notes: The acf will include a lag 0 correlation which is simply the correlation between the variable and itself, so it must = 1. Ignore the lag 0 autocorrelation. The single command acf(x) all by itself will produce a plot of the ACF but won't give the numerical values.

As an answer to this part, first copy and paste the value of the acf given as a result of the second command. Then, discuss whether you think the pattern of the first few autocorrelations resembles the theoretical pattern of an AR(1) model.

Autocorrelations of series 'x', by lag

```
     0      1      2      3      4      5      6      7      8      9 
 1.000  0.698  0.541  0.363  0.212  0.174  0.117  0.182  0.197  0.161 
    10     11     12     13     14     15     16 
 0.028 -0.136 -0.191 -0.166 -0.133 -0.149 -0.160 
```

An AR(1) model is feasible. $r_2 = .541$ is slightly higher than $r_1^2 = .698^2$, but not substantially so. $r_3 = .363$ is close to $r_1^3 = .698^3$. Although there are some imperfections, the pattern isn't far off from the AR(1) pattern.

D. Use these four commands to do a regression for an AR(1) model:

```
       lag1x = lag (x, -1)
       y = cbind (x, lag1x)
       ar1model = lm (y[,1] ~ y[,2])
       summary (ar1model)
```

The first command creates the first lag of x. The second rather mysterious command creates a matrix named y in which the first column is x (original data) and the second is lag 1 of x. It turns out this is the only way to make the right link between x and lag 1 of x for the regression that's coming. The third command does a linear model (lm) relating x to lag 1of x (the two column of y). Results from the regression are stored in an object names ar1model. The final command gives a summary of the regression results.

As an answer, copy and paste the summarized regression results. Then, write the regression equation.

```
Residuals:
    Min       1Q      Median      3Q        Max
-2.25526 -0.80864  -0.04491  1.08912  2.06151

Coefficients:
             Estimate    Std. Error t value   Pr(>|t|)
(Intercept)   4.2878     1.7231   2.488     0.0175 *
y[, 2]        0.7078      0.1176   6.017  5.95e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.25 on 37 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.4946,     Adjusted R-squared: 0.4809
F-statistic:  36.2 on 1 and 37 DF,  p-value: 5.954e-07
```

Equation is $\hat{x}_t = 4.2878 + 0.7078 x_{t\text{-}1}$

E. Now we’ll look at a plot of residuals versus predicted values for the regression that we just did. The command is

```
plot(fitted(ar1model),residuals(ar1model))
```

Briefly discuss whether the plot looks the way it should for a good model. You don’t have to give the plot.

The plot of residuals versus predicted values looks more or less like a horizontal band with no pattern and relatively constant variance. This is a desirable plot.

F. Now we’ll determine the ACF of the residuals. The command is

```
acf(residuals(ar1model))
```

Briefly discuss whether the ACF of residuals looks like it should for a good model.

All autocorrelations are non-significant for lags ≥ 1. This is the way it should be for residuals from a model.

G. Use the regression equation found in part D to predict the level in the next year past the series. You’ll need the value of x at the end of the series to do this. One way to find that is to enter the command

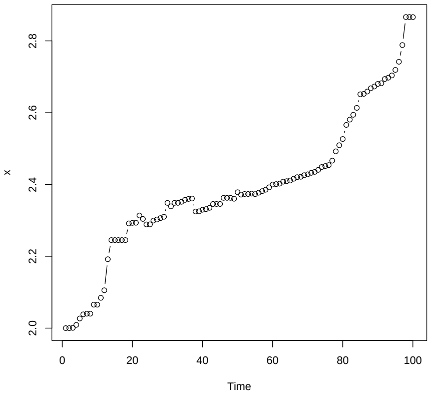
```
x
```

That will list the values of the series in order.

$\hat{x}_{41} = 4.2878 + 0.7078x_{40} = 4.2878 + 0.7078(15.787) = 15.462$

2. For this problem, use the dataset oildata.dat from the Dataset folder of the course website. Download the dataset to your computer. The data are a time series of a price index for oil in the United States for 100 consecutive months. For month t, this measure is $\log_{10}\left(\frac{x_t}{x_1}\right)$ where x = actual price.

A. Refer back to part A of problem 1 to see how to read the data and create a time series plot of the data. Briefly describe the noteworthy features. What's one obvious reason why the series is not a stationary series?



There's a strong upward trend, so the series mean is increasing. Thus the series is not stationary.

B. A first difference is defined as $x_t - x_{t-1}$. For series with a strong trend, first differences may (or may not) be stationary. An example of how to create a first difference in R, is the following:

```
diff1 = diff (x, 1)
```

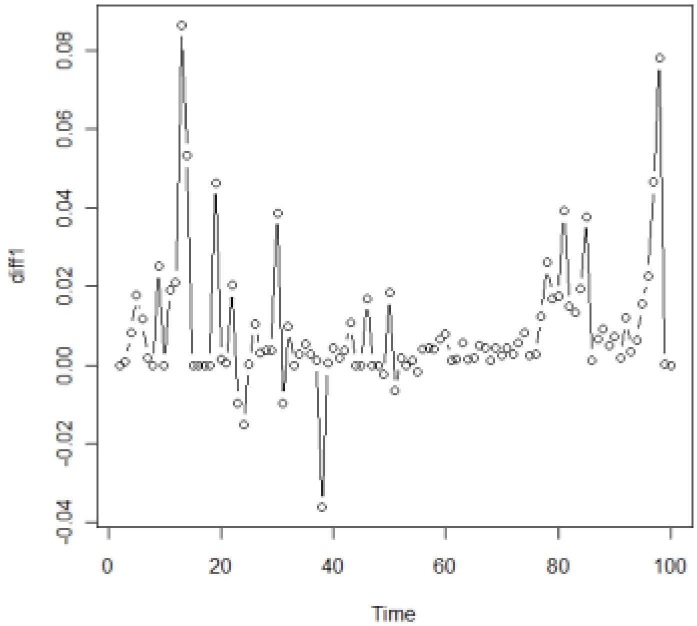
The diff1 name is arbitrary. You can call it anything you like. Also, the x within the parentheses is the name that you have already given the data. If you called the original data something else, then that's what goes there.

Create first differences for this series and the plot the first differences. For example, if you called the first differences diff1, the command could be

```
plot (diff1, type='b')
```

As the answer to this part, give the plot and briefly describe any noteworthy features of the plot.

There is no trend in the differenced series. There are two or three extreme spikes indicating large changes at some points in time, though there is not an abrupt change to the level of the series. There are no issues with non-constant variance, or cyclical behavior.



C. Determine the ACF of the first differences. See part C of problem 1 for guidance.

As an answer to this part, copy and paste the value of the acf given as a result of the second command. Then, discuss whether you think the pattern of the first few autocorrelations resembles the theoretical pattern of an AR(1) model.

Autocorrelations of series ‘diff1’, by lag

```
   0     1         2       3         4       5       6       7        8        9       10
 1.000  0.317  0.082  0.061  0.059  0.028  0.080  0.059  0.027 -0.007 -0.048

   11        12      13      14      15        16      17       18      19
-0.042  -0.080  0.169  0.084  0.033  0.126  0.194  0.031  -0.037
```

For lags past 2, the autocorrelations are a bit higher than r_1^k**, but an AR(1) isn’t a bad guess for this series.**

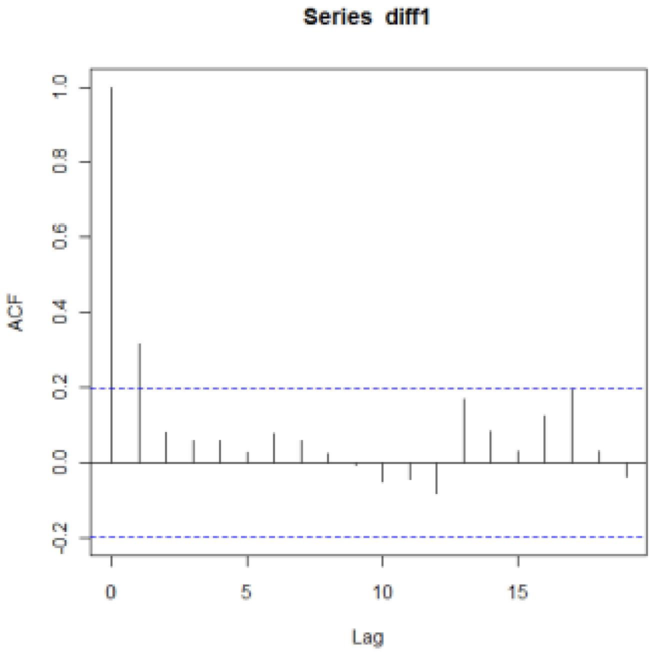
D. A “moving average model of order 1” (abbreviated MA(1)) is defined as

$x_t = \mu + w_t + \theta_1 w_{t-1}$ where w_t are independently distributed with mean 0 and variance σ_w^2. That is, the value of x at time t = mean + random error at this time + random error from last time. The theoretical ACF of an MA1 is simple – the first lag autocorrelation is non-zero while all other autocorrelations are 0.

Refer back to the ACF found in part C. Explain why a MA(1) might work as a model for the first differences

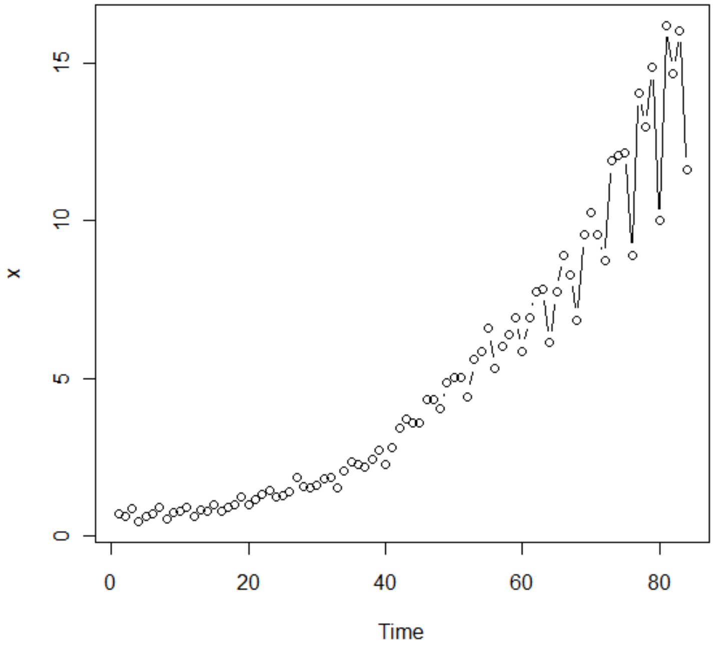
Note: This brings up one of the "fun" parts of time series - more than one model might work for a series.

The ACF plot is below. Note that the only significant autocorrelation is at the first lag so a MA(1) is a possibility.



3. Use the dataset jj.dat from the Datasets folder. The series gives measure of quarterly profits of the Johnson & Johnson Corporation for 84 consecutive quarters.

A. Create a time series plot of the data. Give the plot as the answer to this part.



B. Refer to the plot in part A. Discuss noteworthy features of the plot. See Lesson 1.1 for this week to review what we're looking for.

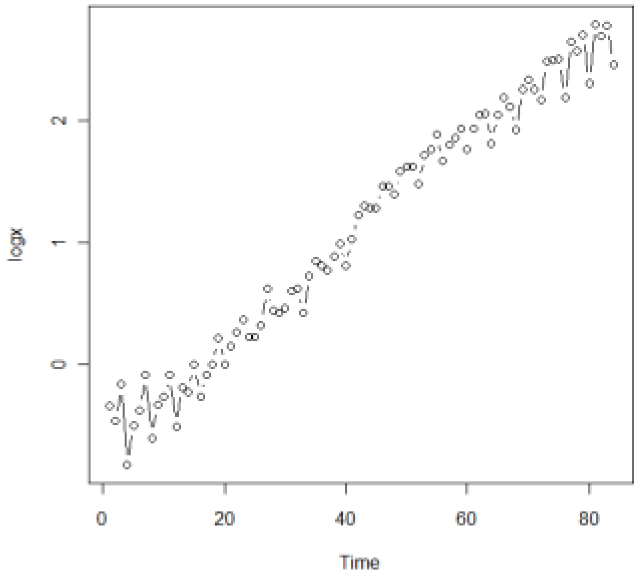
There's a curved upward trend. There's seasonality – for example, the 4th quarter of every year is the low for the year. The variation is increasing over time. No single values stand out as outliers and the level of the series does not change abruptly.

C. Calculate a series that is the logarithm of the original series, and then plot the logarithm series. For example, if you had called the original series x, then this R sequence would do the job:

```
logx = log(x)
plot (logx, type ="b")
```

Give the plot of the logarithm data as the answer to this part and then discuss any differences between this plot and the plot for parts A and B. (For instance, are there changes in trend, variance issues, etc.)

There's a linear trend and seasonality. The variance now seems to be relatively constant.



D. What regression model would you use to model the series in part C? Briefly describe the predictor variables in the model. See the second example of Lesson 1.1 for this week for guidance.

The predictors could be t = time (ranging from 1 to 84) and four indicators, one for each quarter (say, Q1= 1 if quarter is 1 and = 0 if not quarter 1; Q2= 1 if quarter is 2 and = 0 if not quarter 2; Q3= 1 if quarter is 3 and = 0 if not quarter 3; Q4= 1 if quarter is 4 and = 0 if not quarter 4).

If we leave out the intercept, the model is
$y_t = \beta t + \alpha_1 Q_1 + \alpha_2 Q_2 + \alpha_3 Q_3 + \alpha_4 Q_4 + w_t$ **, where** $y_t = \log(x_t)$ **and** w_t **is the usual error term. If we include an intercept** β_0**, we should drop one of the quarter indicators to avoid multicollinearity.**

MINITAB USERS
For a time series plot, Stat>Time Series > Time Series Plot
For an ACF, Stat>Time Series >Autocorrelation
To lag a variable, Stat>Time Series >Lag
To determine a first difference, Stat>Time Series >Differences