ANLY 500 Laboratory #2 – Predictive Statistics

*Evans Chapter 8 through and including Chapter 12*

*“Performance Lawn Equipment Case Study” from Evans,* ***Business Analytics***

[Introduction 1](#_Toc485808906)

[Chapter 8 2](#_Toc485808907)

[Part 1: Defects after Delivery 3](#_Toc485808908)

[Step 1 3](#_Toc485808909)

[Step 2 3](#_Toc485808910)

[Part 2: Defects after Delivery Continued 4](#_Toc485808911)

[Step 1 4](#_Toc485808912)

[Part 3: Employee Retention 7](#_Toc485808913)

[Step 1 7](#_Toc485808914)

[Part 4: Learning Associated with Adopting New Technology 10](#_Toc485808915)

[Step 1 10](#_Toc485808916)

[Chapter 9 13](#_Toc485808917)

[Part 1: Mower Sales 13](#_Toc485808918)

[Step 1 13](#_Toc485808919)

[Part 3: Production Costs 18](#_Toc485808920)

[Step 1 18](#_Toc485808921)

# Introduction

This laboratory follows the exercises in the book, specifically the Performance Lawn Equipment Case Study homework assigned exercises Chapters 8 through and including 12, except this laboratory requires that you use R to complete the exercises. That is, you should answer all questions in the textbook exercises and complete computations using R. Each laboratory in ANLY 500 will build on the laboratories you have completed before. So, you will want to continue to keep your work in the folder you set-up for Laboratory #1 so that you can refer back to previous laboratories if necessary. You will also continue to use the data files for ANLY 500 on Moodle.

# Chapter 8

In reviewing the PLE data, Elizabeth Burke noticed that defects received from suppliers have decreased (DefectsAfterDelivery.csv data file). Upon investigation, she learned that in 2010, PLE experienced some quality problems due to an increasing number of defects in materials received from suppliers. The company instituted an initiative in August 2011 to work with suppliers to reduce these defects, to more closely coordinate deliveries, and to improve materials quality through reengineering supplier production policies. Elizabeth noted that the program appeared to reverse an increasing trend in defects; she would like to predict what might have happened had the supplier initiative not been implemented and how the number of defects might further be reduced in the near future.

In meeting with PLE’s human resources director, Elizabeth also discovered a concern about the high rate of turnover in its field service staff. Senior managers have suggested that the department look closer at its recruiting policies, particularly to try to identify the characteristics of individuals that lead to greater retention. However, in a recent staff meeting, HR managers could not agree on these characteristics. Some argued that years of education and grade point averages were good predictors. Others argued that hiring more mature applicants would lead to greater retention. To study these factors, the staff agreed to conduct a statistical study to determine the effect that years of education, college grade point average, and age when hired have on retention. A sample of 40 field service engineers hired 10 years ago was selected to determine the influence of these variables on how long each individual stayed with the company. Data are compiled in the Employee Retention worksheet.

Finally, as part of its efforts to remain competitive, PLE tries to keep up with the latest in production technology. This is especially important in the highly competitive lawn-mower line, where competitors can gain a real advantage if they develop more cost-effective means of production. The lawn-mower division therefore spends a great deal of effort in testing new technology. When new production technology is introduced, firms often experience learning, resulting in a gradual decrease in the time required to produce successive units. Generally, the rate of improvement declines until the production time levels off. One example is the production of a new design for lawn-mower engines.

To determine the time required to produce these engines, PLE produced 50 units on its production line; test results are given on the worksheet Engines in the database. Because PLE is continually developing new technology, understanding the rate of learning can be useful in estimating future production costs without having to run extensive prototype trials, and Elizabeth would like a better handle on this. Use techniques of regression analysis to assist her in evaluating the data in these three worksheets and reaching useful conclusions. Summarize your work in a formal report with all appropriate results and analyses.

## Part 1: Defects after Delivery

## Step 1

As we found before, most of the work is getting the data into the proper format to analyze. This will still be true, even for performing regression analyses. So, the first thing to do to look at what the numbers of defects would have been if nothing had changed is to get a month number column along with that part of the DefectsAfterDelivery.csv data that is of interest. It looks like to me that the number of defects continues to rise through October 2011

To get the number of the month in a first column and the number of defects for that month in a second column I:

1. combine the years 2010 and 2011 into a vector, then remove the last two values because I’m not including November or December 2011.
2. create a month number vector for the 22 months of interest.
3. combine the month numbers and the number of defects into a matrix with two columns and check the first few rows.
4. Use the lm() function to do the regression and check the output.
5. Plot the data in a scatter plot, add the regression line and check it.

The question that you should answer at this point is why do we need to run an ANOVA on this data? You should say something in your lab report about whether you think it was necessary to do analysis of variance in this case or not.

## Step 2

In performing the regression analysis we always want to check and make sure that all the assumptions are true; linearity, normality of errors, equal variance or homoscedasticity, and independence. If we plot the residuals we see we are in trouble already. The plot of the residuals shown below has a pattern that to me looks cyclical. To create this plot I wrote another short R script. Note that to use the function xyplot() you will need to attach the lattice package again, i.e. library(lattice). The scrip and resulting plot are:

xyplot(resid(defects.old.rate.mod) ~ fitted(defects.old.rate.mod),

xlab="Fitted Values",

ylab="Residuals",

main="Residual Diagnostic Plot",

panel=function(x, y, ...)

{

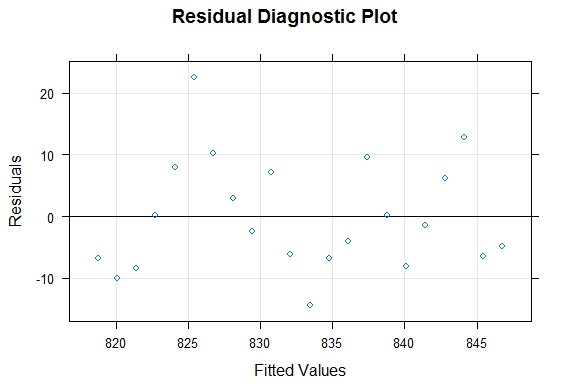
panel.grid(h=-1, v=-1)

panel.abline(h = 0)

panel.xyplot(x, y, ...)

}

)



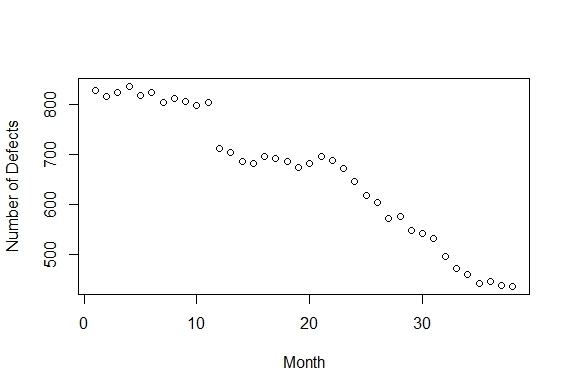
I believe you should see some periodicity in mower related data, more units sold in the spring, less in the fall and so on. So I think the best idea would be to use a method amenable to seasonal cycles. I will resume this analysis after the technique is presented in Chapter 9.

## Part 2: Defects after Delivery Continued

## Step 1

If we want to look at the trend in Defects after Delivery since the change was made, as usual, we need to put the data of interest in the proper format. Since we included the year 2010 and January through October 2011 in the first regression analysis, we’ll use November-December 2011 and the remaining years in this analysis.

You should use the same commands and logic as before for the data and create a scatter plot (Number of defects based on months). The result should be like:



Looking at the scatterplot of the data it appears that it was several months after the change was made before there was a real difference in the number of defects. This is where I see what looks like a “gap”. But, we’ll leave that alone for now.

Now, perform a linear regression for this data and draw the relative conclusions using following commands.

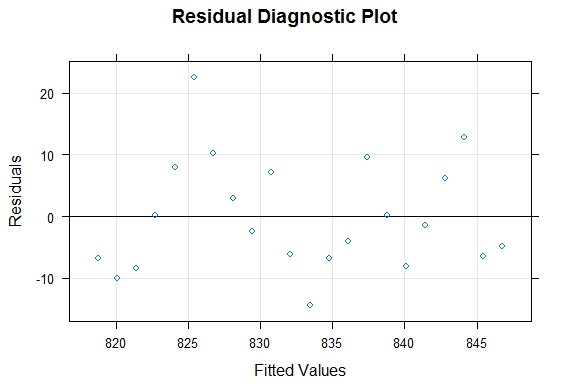
> defects.new.rate.mod <- lm(defects.new.rate[,2] ~ defects.new.rate[,1])

> summary(defects.new.rate.mod)

Here we see that by omitting the months prior to the change we find that a linear model explains just over 95% of the variance in the data. This is a very good result. Plot the graph and regression line together.

What I have done is a “piecewise” solution based on the information in the case study. You should comment in your lab report on what you believe is the most correct way to approach this problem.

I used the same R script as before to plot the residuals for this analysis. And, I see the same cyclical nature in the residuals as before. In fact, if you look for it you can see the cyclical nature of the data in the scatter plot. The plot of the residuals is:



For our analysis let’s complete a normal probability plot of the residuals to look at normality.

After writing the code, you should see a graph like this:



which looks pretty good. Evans uses the Durbin-Watson test to evaluate serial dependence. This test can be done in R with the dwtest() function in the lmtest package. So, if you haven’t already installed and attached the lmtest package you’ll need to do that. Output from the Durbin-Watson test can be interpreted using the following table:



Then, perform the Durbin-Watson test as follows:

> dwtest(defects.new.rate.mod)

Durbin-Watson test

data: defects.new.rate.mod

DW = 0.42875, p-value = 1.29e-10

alternative hypothesis: true autocorrelation is greater than 0

Do you see any correlation?

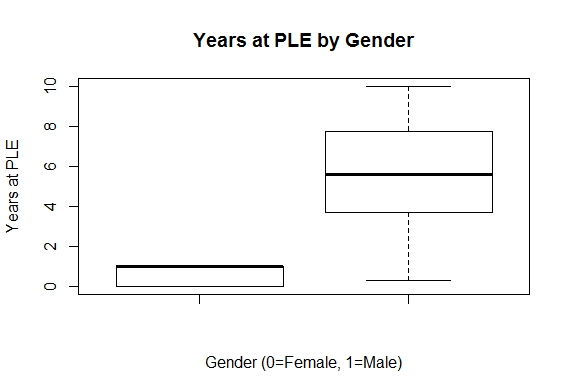
## Part 3: Employee Retention

## Step 1

This is a pretty straight forward look at the variables PLE has available with regard to employee retention, i.e. gender, number of years at PLE, number of years of education, college GPA, college graduate, employee age, and if the employee is local or not. We can do several things to analyze these data. We could look at correlations. And we could use ANOVA to assess any relationships between variables.

We can convert categorical variables, e.g. gender, to numeric variables in order to conduct an analysis using them in R relatively easily as follows:

1. convert to numeric value, i.e. female = 0, male = 1
2. Create a plot to explore how gender is related to the number of years at PLE and view it. It should look like following plot.



1. look at the mean and standard deviation of the categorical variable
2. create a matrix of categorical variables to look at how they might be correlated

> cat.employee.retention.variables <- matrix(c(dGender, dCollege.Grad, dLocal), ncol = 3)

> colnames(cat.employee.retention.variables) <- c("Gender", "College.Grad", "Local")

> str(cat.employee.retention.variables)

num [1:40, 1:3] 0 1 1 0 0 1 1 1 0 1 ...

- attr(\*, "dimnames")=List of 2

..$ : NULL

..$ : chr [1:3] "Gender" "College.Grad" "Local"

> head(cat.employee.retention.variables)

Gender College.Grad Local

[1,] 0 1 1

[2,] 1 1 1

[3,] 1 1 0

[4,] 0 1 1

[5,] 0 1 1

[6,] 1 1 1

*Write your correlation code here*

Compare male employees at PLE and college graduates.

Let’s look at the numeric variables.

First, let’s put the variables we want to analyze in a matrix as follows:

> num.employee.retention.variables <- matrix(c(EmployeeRetention$YearsPLE, EmployeeRetention$YrsEducation, EmployeeRetention$College.GPA, EmployeeRetention$Age), ncol = 4)

> colnames(num.employee.retention.variables) <- c("YearsPLE", "YrsEducation", "College.GPA", "Age")

> str(num.employee.retention.variables)

num [1:40, 1:4] 10 10 10 10 9.6 8.5 8.4 8.4 8.2 7.9 ...

- attr(\*, "dimnames")=List of 2

..$ : NULL

..$ : chr [1:4] "YearsPLE" "YrsEducation" "College.GPA" "Age"

Then, see how/if these variables are correlated.

Is college GPA correlated with years of education? How about age? Is it correlated with years of education?

The book uses 0.7 as a threshold for stating variables are correlated. You should note that other references use 0.5 as a threshold. The only time two variables are truly not correlated is if the correlation coefficient is in fact 0.0. However, we are conducting the correlation in part to assess any multicollinearity. Hint: Since the correlation is low we can state that multicollinearity is not an issue.

We can just do a linear regression of these variables to determine each independent variable’s effect on the dependent variable.

Start the linear regression as follows (remembering that the dependent variable or “y” is YearsPLE):

> EmployeeRetention.mod <- lm(EmployeeRetention$YearsPLE ~ EmployeeRetention$YrsEducation + EmployeeRetention$College.GPA + EmployeeRetention$Age)

> summary(EmployeeRetention.mod)

Call:

lm(formula = EmployeeRetention$YearsPLE ~ EmployeeRetention$YrsEducation +

EmployeeRetention$College.GPA + EmployeeRetention$Age)

Residuals:

Min 1Q Median 3Q Max

-5.3299 -1.6122 -0.2433 1.8893 4.6312

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.73711 4.50415 -0.608 0.5472

EmployeeRetention$YrsEducation -0.06705 0.35516 -0.189 0.8513

EmployeeRetention$College.GPA 0.67998 1.18355 0.575 0.5692

EmployeeRetention$Age 0.29154 0.13504 2.159 0.0376 \*

---

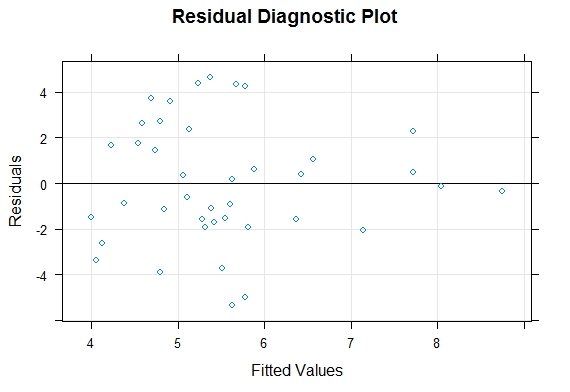
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.726 on 36 degrees of freedom

Multiple R-squared: 0.1502, Adjusted R-squared: 0.07939

F-statistic: 2.121 on 3 and 36 DF, p-value: 0.1146

Which shows that Age is the only statistically significant variable when it comes to employee retention at PLE. We can plot and view the residuals to verify using the R script from before (but changing the name of the linear model to EmployeeRetention.mod):



This may be alright, but shows that the residuals do not really have an equal variance across all fitted values. But moreover, based on R-squared this really isn’t a good model for the data. If we eliminate the variables that are not significant and run it again we get:

> EmployeeRetention.mod2 <- lm(EmployeeRetention$YearsPLE ~ EmployeeRetention$Age)

> summary(EmployeeRetention.mod2)

Call:

lm(formula = EmployeeRetention$YearsPLE ~ EmployeeRetention$Age)

Residuals:

Min 1Q Median 3Q Max

-4.9927 -1.6430 -0.1952 2.0328 4.8078

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.0149 3.0425 -0.662 0.5118

EmployeeRetention$Age 0.3003 0.1198 2.506 0.0166 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.666 on 38 degrees of freedom

Multiple R-squared: 0.1419, Adjusted R-squared: 0.1193

F-statistic: 6.282 on 1 and 38 DF, p-value: 0.01659

Which still shows that Age is statistically significant but explains very little of the variation in the data.

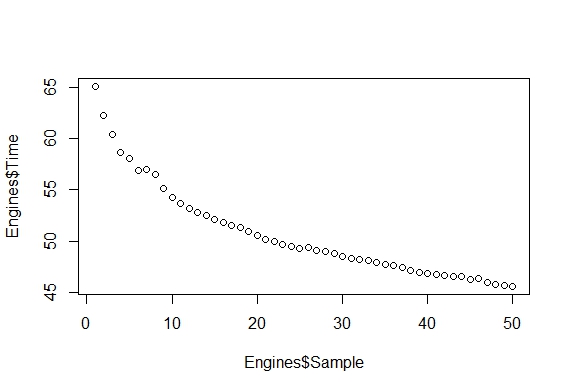
## Part 4: Learning Associated with Adopting New Technology

## Step 1

Just to clean things up a bit and make it easier to work with the variables, after I imported the Engines.csv data I changed the column names to just “Sample” and “Time” as follows:

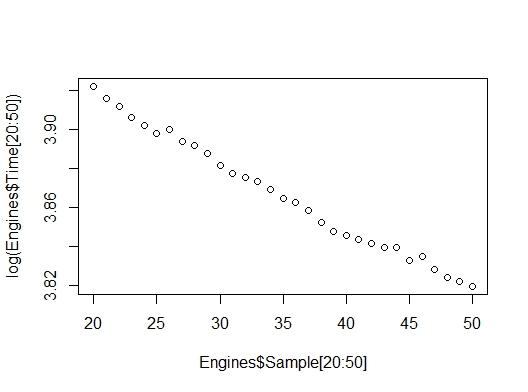
> colnames(Engines) <- c("Sample", "Time")

Then, just to see what’s going on create scatter plot of the data as follows:



It doesn’t look like a linear model will fit these data very well. It really looks like a logarithmic decay. Also, when I played around with the data a bit I found a better fit by just looking at the last 30 data points rather than the entire series. See what you think below:

> plot(Engines$Sample[20:50], log(Engines$Time[20:50]))



which is nearly linear. So, let’s try that and see what we get. To transform the data we’ll use the exponential as follows:

> m <- lm(log(Engines$Time[20:50]) ~ Engines$Sample[20:50])

> summary(m)

Call:

lm(formula = log(Engines$Time[20:50]) ~ Engines$Sample[20:50])

Residuals:

Min 1Q Median 3Q Max

-0.0052441 -0.0016683 -0.0004296 0.0020182 0.0052949

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.9837148 0.0019472 2045.84 <2e-16 \*\*\*

Engines$Sample[20:50] -0.0033450 0.0000539 -62.05 <2e-16 \*\*\*

---

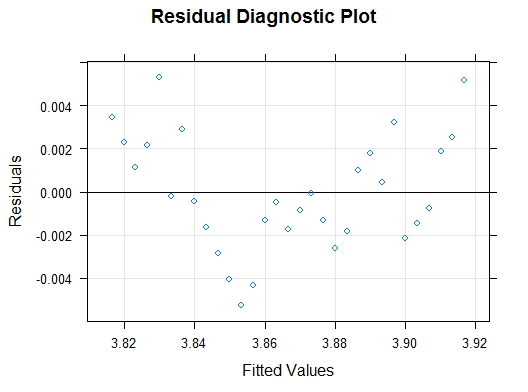
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002684 on 29 degrees of freedom

Multiple R-squared: 0.9925, Adjusted R-squared: 0.9923

F-statistic: 3851 on 1 and 29 DF, p-value: < 2.2e-16

We find that both the intercept and the slope are statistically significant. And, from R-squared it looks like the model explains nearly all the variation in the data. Then last, running the R script again to plot the residuals we get:



which is not too bad.

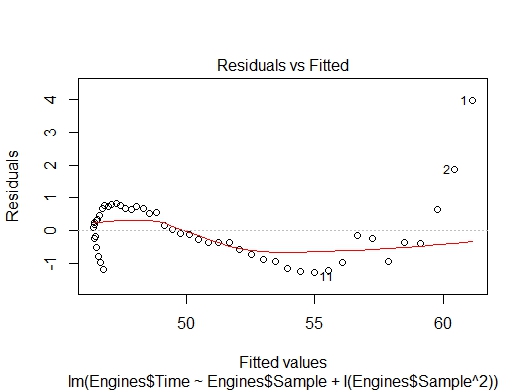
So, if we use a 2nd order fit:

> m <- lm(Engines$Time ~ Engines$Sample + I(Engines$Sample^2))

> summary(m)

Does it look promising? What about the residuals?

> plot(m, 1)



which really looks bad. I’d go with the logarithmic decay model. The residuals look the best. If you do this makes the model for learning associated with adopting this new technology

Production Time (min) = 3.99 – 0.0034\*log(Unit)

# Chapter 9

An important part of planning manufacturing capacity is having a good forecast of sales. Elizabeth Burke is interested in forecasting sales of mowers and tractors in each marketing region as well as industry sales to assess future changes in market share. She also wants to forecast future increases in production costs. Develop forecasting models for these data and prepare a report of your results with appropriate charts and output from R.

## Part 1: Mower Sales

Plot data for mower sales in each marketing region as well as industry sales. Then, establish a method for use in predicting mower sales.

## Step 1

Import and plot the mower sales. As usual, the hardest part will be getting the data in the proper format. First, I’ll get rid of the “NA” column name for North America, just to get that out of the way.

> colnames(MowerUnitSales) <- c("Date", "NorthA", "SouthA", "Europe", "Pacific", "China", "World")

> str(MowerUnitSales)

'data.frame': 60 obs. of 7 variables:

$ Date : Factor w/ 60 levels "April 2010","April 2011",..: 21 16 36 1 41 31 26 6 56 51 ...

$ NorthA : int 6000 7950 8100 9050 9900 10200 8730 8140 6480 5990 ...

$ SouthA : int 200 220 250 280 310 300 280 250 230 220 ...

$ Europe : int 720 990 1320 1650 1590 1620 1590 1560 1590 1320 ...

$ Pacific: int 100 120 110 120 130 120 140 130 130 120 ...

$ China : int 0 0 0 0 0 0 0 0 0 0 ...

$ World : int 7020 9280 9780 11100 11930 12240 10740 10080 8430 7650 ...

Now plot the mower sales. The only tricky part is getting tick marks and labels on the x-axis and a grid if you want one. First, I set up a data object for the number of tick marks I wanted every 6 months, then one for the labels.

> c = seq(1, length(MowerUnitSales$Date), 6)

> ticks=c("Jan 2010", "Jun 2010", "Jan 2011", "Jun 2011", "Jan 2012", "Jun 2012", "Jan 2013", "Jun 2013", "Jan 2014", "Jun 2014")

Then, I created the plot in separate pieces; 1) plot the data, 2) add the tick marks and labels, 3) add the y-axis grid, 4) add a custom grid to match my tick marks and labels for every 6 months of data:

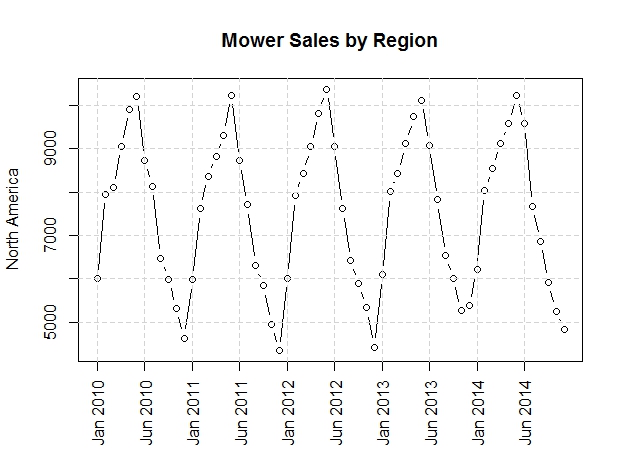
> plot(df$NorthA, type = "b", xlab = "", xaxt="n", ylab = "North America", main = "Mower Sales by Region")

> axis(1, at=c, labels=ticks, las=2)

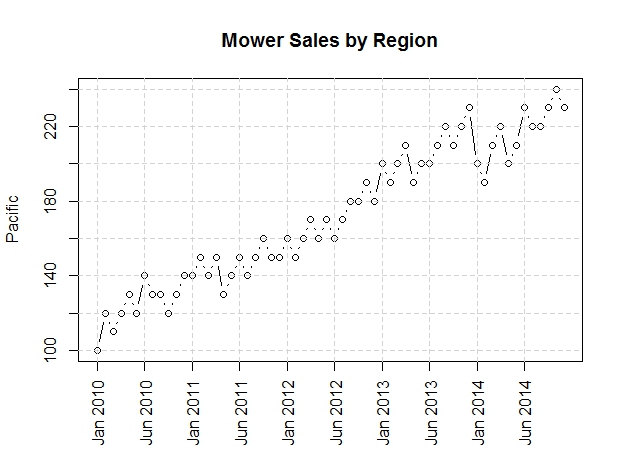
> grid(NA, NULL, lty = 2)

> abline(v=seq(1, length(Date), 6), lty=2, col="lightgray")

And, here’s the plot:



You can change the colors and everything in R. For this lab you might want to just do the basic plots to get done fastest. Each would be similar to this one. In fact, you can use this same process to plot the tractor sales too. Now plot for the mower sales for the Pacific region. Your result should be like this:



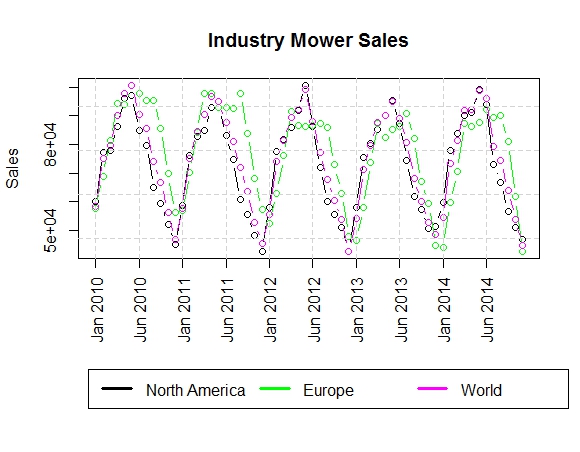
You may plot the industry sales data for multiple regions on the same plot. In fact, you could do that for both PLE’s sales data and the industry sales data. There are a couple ways to do this. The way we’ve used before is the par() function, i.e. par(new=TRUE), so that we can add to the existing plot. A second way to do the same thing is just to use the lines() function to add the second “line” to the same plot. You may plot the North America, Europe and World regions together because they show seasonality without any trend.

Using the technique of inserting the command par(new=TRUE) each time you want to add another curve to a plot the North America, Europe and World data commands are:

par(oma=c(…)) sets the outer margin area. Here I’ve set the bottom of the plot to have 5 line0s.

par(xpd=NA) tells R to put the next thing in the outer margin area, in this case the legend. I’ve used the parameter horiz=TRUE to make the legend print along the center (x=0) bottom (y=1). You can play with the colors, the size of the lines and font, etc. It is all very flexible.

And, here is the figure (you should provide the code for generating this plot) :



As you can see the data are very different for each of these regions. Which data looks seasonal? Which one has trends? So, you’ll need to decide what method to use to analyze the data by region; e.g. seasonal with no trend = Holt-Winter no trend option, seasonal with trend = Holt-Winter’s additive model, etc.

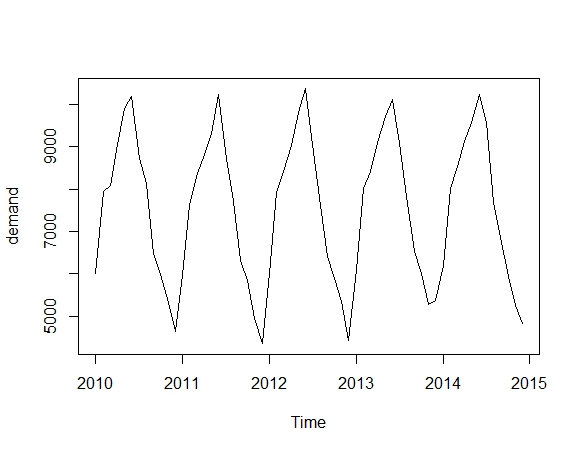
You should use the link: <http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html> to study using R for analyzing and making predictions using time-series data. It is absolutely the best reference I have found. For this lab, seasonal data with no trend, I used the following:

First I set up a data object with the mower sales data from the North America region and plotted it:

> demand <- ts(MowerUnitSales[,3], start = c(2010, 1), frequency=12)

> plot(demand)

Here’s the plot:



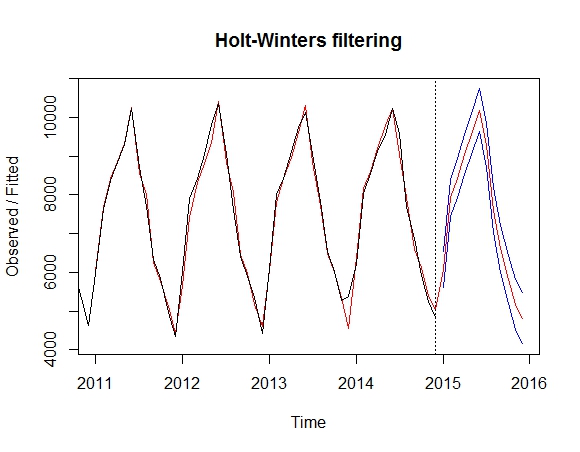
which looks like the correct data. To use Holt Winters in R takes you just:

> hw <- HoltWinters(demand)

Then, you can use this to make the forecast using:

> forecast <- predict(hw, n.ahead = 12, prediction.interval = T, level = 0.95)

where the number of months ahead is 12 and the significance level is .05. Plot Holt Winters based on forecast. Provide your code and, the plot looks like:



where you have the prediction and the upper and lower bounds. Pretty simple – right? The thing to keep in mind is to use the seasonal parameter with the HoltWinters() function and set to additive or multiplicative as required. The website I’ve given you talks about moving average and exponential smoothing.

Also, when you look through the webpage I’ve given you the link for above you’ll see that a more sophisticated approach is to decompose the data into its estimated trend, its estimated seasonal component and its estimated irregular (noise) component then work on each of these. That is beyond the scope of this coursework for me to require. However, should you be so inclined? There is also extra credit available.

## Part 3: Production Costs

Plot data for mower and tractor production costs. Then, establish a method for use in predicting future costs. Note that the file for this is UnitProductionCosts.csv

## Step 1

You should already be able to do this. You might want to spend a few minutes to see if you can before reading on. It really only requires developing a linear model for the data using the lm() function in R and then printing out the summary. This gives you the intercept and slope of the regression line and hence the equation to use for forecasting future production costs. Always remember, this is based (obviously based) on historical data and care should be taken in making predictions of future whatever using historical data…

What I did was plot up the data to see what I had. Then, I set-up a separate data object for the number of the months, 1 through and including 60. Then I used lm() and summary() as follows for tractor production costs. Provide the code to generate following plot.

