3.R

harshitaggarwal

2022-04-14

setwd("~/Documents/GWU/Forecasting/Assignment 3")  
library(greybox)

## Warning: package 'greybox' was built under R version 4.1.2

## Package "greybox", v1.0.5 loaded.

library(forecast)

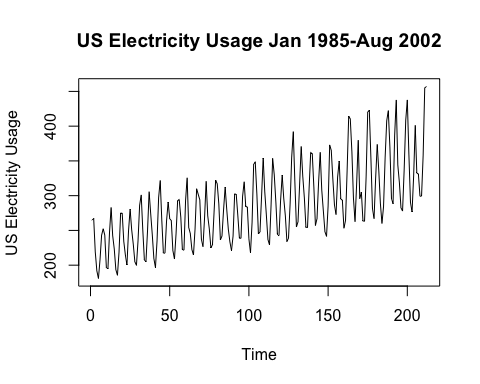
## Warning: package 'forecast' was built under R version 4.1.2

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

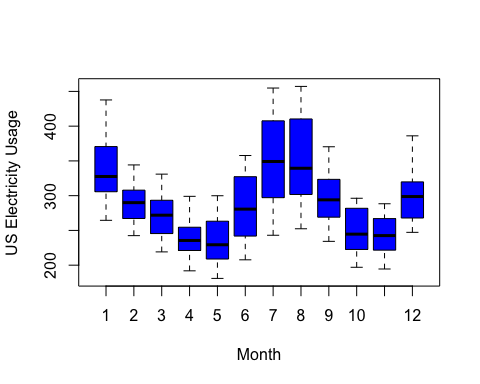
##   
## Attaching package: 'forecast'

## The following object is masked from 'package:greybox':  
##   
## forecast

data <- read.table(file="US\_Electricity\_Usage.txt", header=TRUE)  
  
attach(data)  
  
ts.plot(USAGE,ylab="US Electricity Usage",main="US Electricity Usage Jan 1985-Aug 2002")



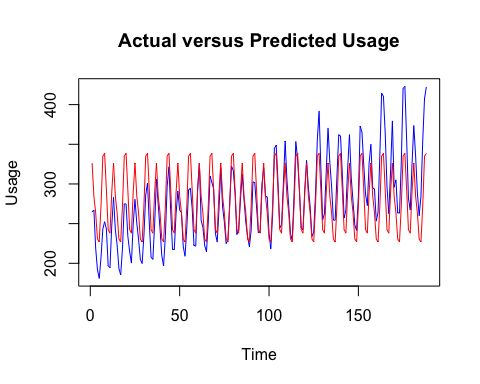
boxplot(USAGE/1~MONTH,xlab="Month",ylab="US Electricity Usage",col="blue")



n\_USAGE=USAGE[1:188]  
n\_MONTH=MONTH[1:188]  
  
fit<-lm(n\_USAGE~as.factor(n\_MONTH))  
summary(fit)

##   
## Call:  
## lm(formula = n\_USAGE ~ as.factor(n\_MONTH))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -91.894 -26.456 0.734 24.578 85.166   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 326.360 8.963 36.411 < 2e-16 \*\*\*  
## as.factor(n\_MONTH)2 -41.560 12.676 -3.279 0.00126 \*\*   
## as.factor(n\_MONTH)3 -62.678 12.676 -4.945 1.77e-06 \*\*\*  
## as.factor(n\_MONTH)4 -95.683 12.676 -7.548 2.28e-12 \*\*\*  
## as.factor(n\_MONTH)5 -99.414 12.676 -7.843 4.08e-13 \*\*\*  
## as.factor(n\_MONTH)6 -53.839 12.676 -4.247 3.50e-05 \*\*\*  
## as.factor(n\_MONTH)7 8.574 12.676 0.676 0.49968   
## as.factor(n\_MONTH)8 12.186 12.676 0.961 0.33768   
## as.factor(n\_MONTH)9 -33.975 12.885 -2.637 0.00912 \*\*   
## as.factor(n\_MONTH)10 -83.747 12.885 -6.499 8.02e-10 \*\*\*  
## as.factor(n\_MONTH)11 -88.391 12.885 -6.860 1.13e-10 \*\*\*  
## as.factor(n\_MONTH)12 -36.821 12.885 -2.858 0.00478 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 35.85 on 176 degrees of freedom  
## Multiple R-squared: 0.5512, Adjusted R-squared: 0.5231   
## F-statistic: 19.65 on 11 and 176 DF, p-value: < 2.2e-16

# The intercept reflect the usage in the month og January which is used as a reference month  
# Each coefficient associated with the months is the expected change on usage during that month on average  
# The coefficient of July is 8.574 with a p-value of 0.49968, due to a high p-value we cannot say that the average usage in July different than the average usage during January  
# The coefficient of September is -33.975 with a p-value of 0.00912, as it is below 0.05 we can say that usage in September is 33.975 units below usage in January on average  
  
plot.ts(n\_USAGE, main="Actual versus Predicted Usage",ylab="Usage", col="blue")  
lines(predict(fit),col="red")



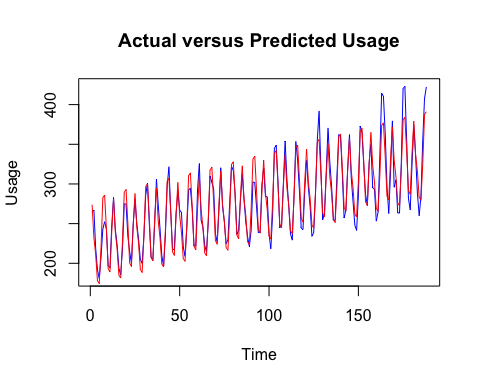
pred=predict(fit, data.frame(n\_MONTH=MONTH[189:212]), interval="prediction")  
  
mape=mean(abs(USAGE[189:212]-pred[,1])/USAGE[189:212])  
mape

## [1] 0.198047

time<-seq(1, length(n\_USAGE))  
fit2<-lm(n\_USAGE~time+as.factor(n\_MONTH))  
summary(fit2)

##   
## Call:  
## lm(formula = n\_USAGE ~ time + as.factor(n\_MONTH))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -39.374 -9.245 1.018 7.070 41.012   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 273.2567 4.0979 66.682 < 2e-16 \*\*\*  
## time 0.5836 0.0198 29.474 < 2e-16 \*\*\*  
## as.factor(n\_MONTH)2 -42.1435 5.2052 -8.096 9.25e-14 \*\*\*  
## as.factor(n\_MONTH)3 -63.8452 5.2054 -12.265 < 2e-16 \*\*\*  
## as.factor(n\_MONTH)4 -97.4338 5.2055 -18.717 < 2e-16 \*\*\*  
## as.factor(n\_MONTH)5 -101.7480 5.2058 -19.545 < 2e-16 \*\*\*  
## as.factor(n\_MONTH)6 -56.7565 5.2061 -10.902 < 2e-16 \*\*\*  
## as.factor(n\_MONTH)7 5.0724 5.2066 0.974 0.331   
## as.factor(n\_MONTH)8 8.1014 5.2070 1.556 0.122   
## as.factor(n\_MONTH)9 -35.1418 5.2914 -6.641 3.78e-10 \*\*\*  
## as.factor(n\_MONTH)10 -85.4980 5.2916 -16.157 < 2e-16 \*\*\*  
## as.factor(n\_MONTH)11 -90.7249 5.2918 -17.144 < 2e-16 \*\*\*  
## as.factor(n\_MONTH)12 -39.7391 5.2922 -7.509 2.92e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.72 on 175 degrees of freedom  
## Multiple R-squared: 0.9247, Adjusted R-squared: 0.9196   
## F-statistic: 179.2 on 12 and 175 DF, p-value: < 2.2e-16

# The coefficient of July is 5.0724 with a p-value of 0.331, due to a high p-value we cannot say that the average usage in July different than the usage during January on average beyond trend (controlling for trend)  
# The coefficient of September is -35.1418 with a p-value of 3.78e-10, as it is below 0.05 we can say that usage in September is 35.1418 units below usage in January on average beyond trend (controlling for trend)  
  
plot.ts(n\_USAGE, main="Actual versus Predicted Usage",ylab="Usage", col="blue")  
lines(predict(fit2),col="red")



pred=predict(fit2, data.frame(n\_MONTH=MONTH[189:212], time=c(189:212)), interval="prediction")  
  
mape=mean(abs(USAGE[189:212]-pred[,1])/USAGE[189:212])  
mape

## [1] 0.04725319

# MAPE with only seasonal dummy variables : 19.805%  
# MAPE with seasonal dummy variables and trend : 4.72%  
# If we see the plots the predicted usage of the model with dummy seasonal variables and trend variable is a much better fit  
# This can be seen from the MAPE also.  
  
# Yes there is on the analysis there is both trend and seasonality in the series  
# As we can see from fit2 the trend term (time) is has a p-value < 2e-16, this means we can reject the null hypothesis that there is no trend in the series  
# We can see from the p-value of dummy seasonal variables that out of 11, 9 of them have p-values lower than 0.05, so we can reject the null hypothesis that the usage is the same throughout the year beyond trend  
  
# If we use December as reference month, the coefficient for January will be  
# (273.2567 - 0) - (273.2567- 39.7391) = 39.7391  
  
# The coefficient for September will be:  
# (273.2567-35.1418) - (273.2567-39.7391) = 4.5973