Assignment 7

Alka Santosh Naik

2023-10-17

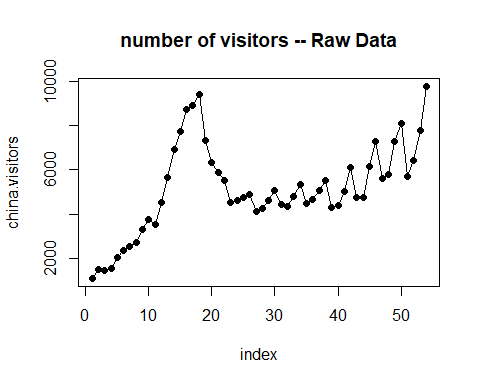
#Alka Santosh Naik  
rm(list=ls())  
library("rio")  
library(car)

## Loading required package: carData

library(readxl)  
visi=import("6304 Module 7 Assignment Data.xlsx")  
visi$index=seq(1:nrow(visi))  
colnames(visi)=tolower(make.names(colnames(visi)))  
attach(visi)

Analysis: Preprocessing the data.

#1  
plot(index,china.visitors,type="o",pch=19,  
 main="number of visitors -- Raw Data")



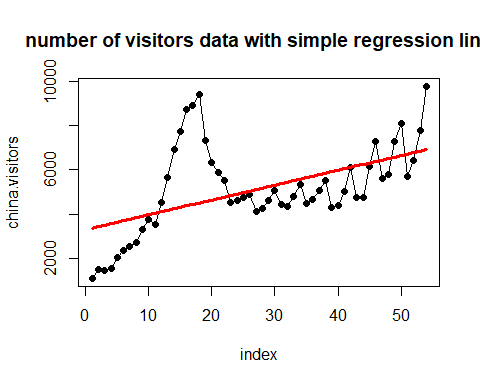
Analysis1: Plot for number of visitors to the United States from the People's Republic of China.

#2  
base.out=lm(china.visitors~index,data=visi)  
summary(base.out)

##   
## Call:  
## lm(formula = china.visitors ~ index, data = visi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2273.0 -1071.4 -436.2 739.3 4929.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3294.93 477.58 6.899 7.17e-09 \*\*\*  
## index 67.07 15.11 4.439 4.73e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1730 on 52 degrees of freedom  
## Multiple R-squared: 0.2748, Adjusted R-squared: 0.2609   
## F-statistic: 19.71 on 1 and 52 DF, p-value: 4.728e-05

Analysis2: Here the R-square value is 0.2748 implying a weak fit.

#3  
plot(index,china.visitors,type="o",pch=19,main="number of visitors data with simple regression line")  
points(base.out$fitted.values,type="l",lwd=3,col="red")



Analysis3: Here we see the simple regression line doesn’t fit the pronounced cyclical pattern in the data.

#4  
durbin.out=car::durbinWatsonTest(base.out)  
durbin.out

## lag Autocorrelation D-W Statistic p-value  
## 1 0.8219683 0.2701599 0  
## Alternative hypothesis: rho != 0

Analysis4: From the Durbin-Watson test we can see the D-W Statistic is 0.2701599 which is approaching 0 indicating positive serial auto correlation. The p-value is 0 there for we reject the null hypothesis and accept the alternating hypothesis that there is correlation.

#5  
indices=data.frame(month=1:4,average=0,index=0)  
for(i in 1:4) {  
 count=0  
 for(j in 1:nrow(visi)) {  
 if(i==visi$quarter[j]) {  
 indices$average[i]=indices$average[i]+visi$china.visitors[j]  
 count=count+1  
 }  
 }  
 indices$average[i]=indices$average[i]/count  
 indices$index[i]=indices$average[i]/mean(visi$china.visitors)  
}  
  
#deseasonalize  
for(i in 1:4){  
 for(j in 1:nrow(visi)){  
 if(i==visi$quarter[j]){  
 visi$deseason.visitor[j]=visi$china.visitors[j]/indices$index[i]  
 }  
 }  
}

Analysis5: The data is deseasonalized and stored in a column in the original data frame.

#6  
desreg.out=lm(deseason.visitor~index,data=visi)  
summary(desreg.out)

##   
## Call:  
## lm(formula = deseason.visitor ~ index, data = visi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2339.6 -892.8 -595.9 457.4 4830.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3358.33 459.19 7.314 1.57e-09 \*\*\*  
## index 64.77 14.53 4.458 4.43e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1664 on 52 degrees of freedom  
## Multiple R-squared: 0.2765, Adjusted R-squared: 0.2626   
## F-statistic: 19.88 on 1 and 52 DF, p-value: 4.434e-05

desreg.out2=lm(deseason.visitor~index+I(index^2),data=visi)  
summary(desreg.out2)

##   
## Call:  
## lm(formula = deseason.visitor ~ index + I(index^2), data = visi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1549.0 -1140.6 -670.9 675.2 4640.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2474.697 693.171 3.570 0.000789 \*\*\*  
## index 159.441 58.146 2.742 0.008400 \*\*   
## I(index^2) -1.721 1.025 -1.680 0.099142 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1635 on 51 degrees of freedom  
## Multiple R-squared: 0.3145, Adjusted R-squared: 0.2876   
## F-statistic: 11.7 on 2 and 51 DF, p-value: 6.59e-05

desreg.out3=lm(deseason.visitor~index+I(index^2)+I(index^3),data=visi)  
summary(desreg.out3)

##   
## Call:  
## lm(formula = deseason.visitor ~ index + I(index^2) + I(index^3),   
## data = visi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1477.9 -880.3 -128.4 389.7 3254.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -996.37824 650.74976 -1.531 0.132   
## index 883.74782 101.52939 8.704 1.38e-11 \*\*\*  
## I(index^2) -34.34425 4.26835 -8.046 1.41e-10 \*\*\*  
## I(index^3) 0.39543 0.05104 7.747 4.10e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1113 on 50 degrees of freedom  
## Multiple R-squared: 0.6884, Adjusted R-squared: 0.6697   
## F-statistic: 36.83 on 3 and 50 DF, p-value: 1.045e-12

Analysis6: Here we see as the R-square value increasing as we include second order and third order polynomial models, this indicates we are more likely to capture the longer-term secular fluctuations in the deseasonalized data.

#7  
indices$index2=indices$index^2  
indices$index3=indices$index^3  
  
  
visi$deseason.forecast=desreg.out$fitted.values  
visi$deseason.forecast2=desreg.out$fitted.values  
visi$deseason.forecast3=desreg.out$fitted.values  
  
for(i in 1:4){  
 for(j in 1:nrow(visi)){  
 if(i==visi$quarter[j]){  
 visi$reseason.forecast[j]=visi$deseason.forecast[j]\*indices$index[i]  
 visi$reseason.forecast2[j]=visi$deseason.forecast2[j]\*indices$index2[i]  
 visi$reseason.forecast3[j]=visi$deseason.forecast3[j]\*indices$index3[i]  
 }  
 }  
}  
  
plot(index,china.visitors,type="o",pch=19,main="number of visitors data with forecast’s")  
points(visi$index,visi$reseason.forecast,type="o",pch=19,col="red")  
points(visi$index,visi$reseason.forecast2,type="o",pch=19,col="green")  
points(visi$index,visi$reseason.forecast3,type="o",pch=19,col="blue")



Analysis8: The third order polynomial model gives the highest R-squared value 0.6884 as it capture the longer-term secular fluctuations in the deseasonalized data.

The p-value indicates high significance for all independent variable index, index^2 and index^3.

From the graph in 7th analysis we can see the blue line is capturing the data better then red and green lines.