Module 7 Assignment

Umakanth Sai Balguri - (U01403700)



Screenshot from Canvas

# Preprocessing

### Loading Library files

rm(list=ls())  
library(rio)  
library(car)

## Loading required package: carData

library(moments)  
library(robustHD)

## Loading required package: ggplot2

## Loading required package: perry

## Loading required package: parallel

## Loading required package: robustbase

### Loading the file into R and

beer=import("6304 Module 7 Assignment Data.xlsx")

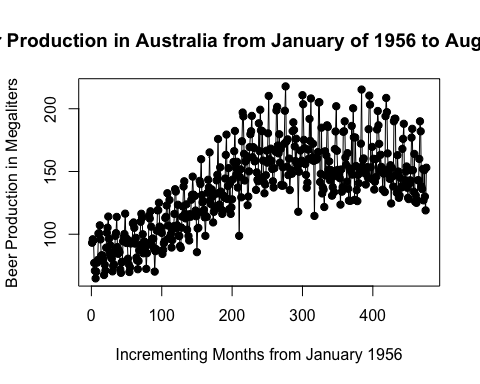
### Fixing names and Creating new variables

colnames(beer)=tolower(make.names(colnames(beer)))  
colnames(beer)[1]="index"  
colnames(beer)[2]="date"  
colnames(beer)[3]="production"  
beer$year=as.numeric(format(beer$date,'%Y'))  
beer$month=as.numeric(format(beer$date,'%m'))  
attach(beer)

# Analysis

### Response to Q1

plot(index, production,type="o",pch=19, main="Beer Production in Australia from January of 1956 to August 1995", xlab="Incrementing Months from January 1956",ylab="Beer Production in Megaliters")

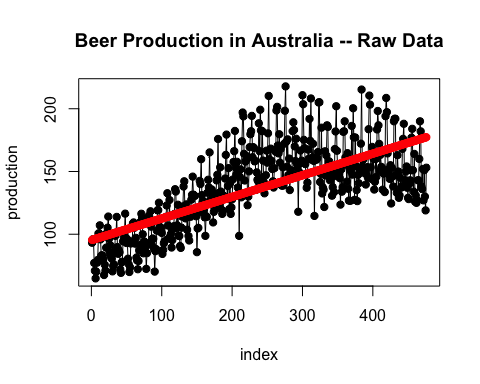


### Response to Q2

beer.out=lm(production~index,data=beer)  
summary(beer.out)

##   
## Call:  
## lm(formula = production ~ index, data = beer)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -58.136 -17.704 -2.887 14.167 74.945   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 95.31025 2.20744 43.18 <2e-16 \*\*\*  
## index 0.17227 0.00802 21.48 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 24.04 on 474 degrees of freedom  
## Multiple R-squared: 0.4933, Adjusted R-squared: 0.4922   
## F-statistic: 461.4 on 1 and 474 DF, p-value: < 2.2e-16

plot(index,production,type="o",pch=19,  
 main="Beer Production in Australia -- Raw Data")  
points(beer.out$fitted.values,type="o",pch=19,col="red")



confint(beer.out)

## 2.5 % 97.5 %  
## (Intercept) 90.9726784 99.6478258  
## index 0.1565061 0.1880233

cor(beer$production,beer.out$fitted.values)

## [1] 0.7023274

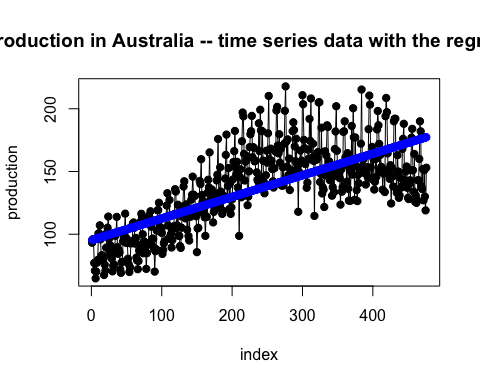
1. Based on the regression analysis, the pvalue is highly significant. It shows a positive secular trend. It indicates that for evey extra month, we expect Production of beer to rise by 0.1722 Megaliteres

Slope of the regression line is estimated to be 0.17227 from regression model analysis. And from confidence interval analysis,it is estimated to be between 0.1565061 (2.5% confindence) and 0.1880233 (97.5% confidence)

Co-relation coefficient is 0.7023274

### Response to Q3

plot(index,production,type="o",pch=19,  
 main="Beer Production in Australia -- time series data with the regression line")  
points(beer.out$fitted.values,type="o",pch=19,col="blue")



### Response to Q4

durbin.out=durbinWatsonTest(beer.out)  
durbin.out

## lag Autocorrelation D-W Statistic p-value  
## 1 0.6647934 0.6682376 0  
## Alternative hypothesis: rho != 0

From Durbin-Watson test, we could see P-value is ‘0’ means null hypothesis is rejected. (there exists auto-correlation = 0.6647934)

### Response to Q5

indices=data.frame(month=1:12,average=0,index=0)  
for(i in 1:12) {  
 count=0  
 for(j in 1:nrow(beer)) {  
 if(i==beer$month[j]) {  
 indices$average[i]=indices$average[i]+beer$production[j]  
 count=count+1  
 }  
 }  
 indices$average[i]=indices$average[i]/count  
 indices$index[i]=indices$average[i]/mean(beer$production)}  
  
  
for(i in 1:12){  
 for(j in 1:nrow(beer)){  
 if(i==beer$month[j]){  
 beer$deseason.production[j]=beer$production[j]/indices$index[i]  
 }  
 }  
}

### Response to Q6

des.prod1.out=lm(deseason.production~index,data=beer)  
summary(des.prod1.out)

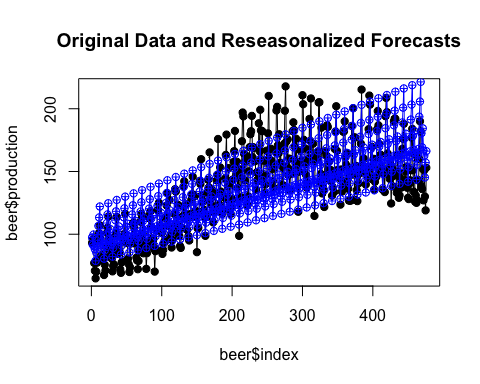
##   
## Call:  
## lm(formula = deseason.production ~ index, data = beer)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -43.076 -12.943 -2.126 13.175 51.737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 95.069346 1.617548 58.77 <2e-16 \*\*\*  
## index 0.173275 0.005877 29.48 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.62 on 474 degrees of freedom  
## Multiple R-squared: 0.6472, Adjusted R-squared: 0.6464   
## F-statistic: 869.4 on 1 and 474 DF, p-value: < 2.2e-16

des.prod2.out=lm(deseason.production~poly(index,2),data=beer)  
summary(des.prod2.out)

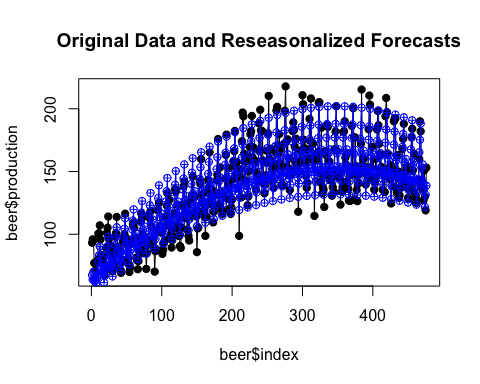
##   
## Call:  
## lm(formula = deseason.production ~ poly(index, 2), data = beer)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -35.069 -8.714 -0.766 7.904 40.011   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 136.3954 0.5327 256.06 <2e-16 \*\*\*  
## poly(index, 2)1 519.4625 11.6216 44.70 <2e-16 \*\*\*  
## poly(index, 2)2 -288.5060 11.6216 -24.82 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.62 on 473 degrees of freedom  
## Multiple R-squared: 0.8468, Adjusted R-squared: 0.8461   
## F-statistic: 1307 on 2 and 473 DF, p-value: < 2.2e-16

### Response to Q7

beer$deseason.forecast=des.prod1.out$fitted.values  
for(i in 1:12){  
 for(j in 1:nrow(beer)){  
 if(i==beer$month[j]){  
 beer$reseason.forecast[j]=beer$deseason.forecast[j]\*  
 indices$index[i]  
 }  
 }  
}  
beer$deseason.forecast2=des.prod2.out$fitted.values  
for(i in 1:12){  
 for(j in 1:nrow(beer)){  
 if(i==beer$month[j]){  
 beer$reseason.forecast2[j]=beer$deseason.forecast2[j]\*  
 indices$index[i]  
 }  
 }  
}  
plot(beer$index,beer$production,type="o",pch=19,  
 main="Original Data and Reseasonalized Forecasts")  
points(beer$index,beer$reseason.forecast,  
 type="o",pch=10,col="blue")



plot(beer$index,beer$production,type="o",pch=19,  
 main="Original Data and Reseasonalized Forecasts")  
points(beer$index,beer$reseason.forecast2,  
 type="o",pch=10,col="blue")

 From a visual review, second model, i.e second order polynomial (x2) appears to have the better fit to the original beer production data as it closely follows the trend of original values.