Stats326 - Assignment2

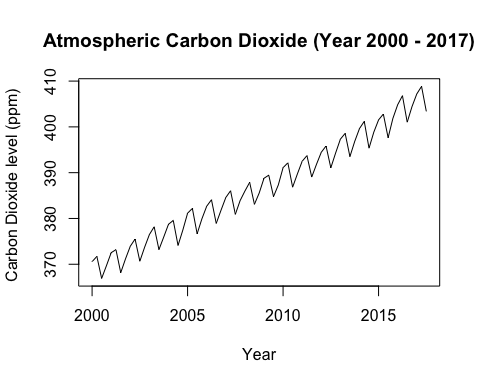
Cameron Todd (5611230)

7 Jan 2018

Question 1 - Looking at the plot below we can clearly see and increasing positive trend with strong seasonality. I think the variability in the data does not appear to be increasing over time and therefore I did not transform the response variable. Their could be ever so slight curvature in the plot and therefore would require perhaps a quadratic when we go to fit the model. Their appears to be no outliers of any sort or kinks in the linear trend, therefore a break is not required

A\_carbon.df = read.table("quarterly dioxide.txt", header=T)

series.ts = ts(A\_carbon.df$dioxide, frequency = 4, start=2000)  
plot.ts(series.ts,main="Atmospheric Carbon Dioxide (Year 2000 - 2017)", xlab="Year",ylab="Carbon Dioxide level (ppm) ")



Q2 - Fit Appropriate model after checking all of them. Seasonally Adjusted Moving Averages Model ended up being the best when comparing the RMSEP against all other potential models.

Time = 1:67  
Quarter = factor(c(rep(1:4,16),(1:3)))  
acd.ts = ts(A\_carbon.df$dioxide[1:67],start=2000, frequency=4)  
decomp.ma.acd = decompose(acd.ts)  
  
#SA MA Forecast with quadratic and lag  
acd\_sa\_ma.ts = acd.ts - decomp.ma.acd$seasonal  
acd.fit\_SA\_MA2 = lm(acd\_sa\_ma.ts[-1]~Time[-1]+ I(Time[-1]^2) + acd\_sa\_ma.ts[-67])  
  
#SA MA Forecast  
t68.ma.pred = acd.fit\_SA\_MA2$coefficients[1] + acd.fit\_SA\_MA2$coefficients[2]\*68 + acd.fit\_SA\_MA2$coefficients[3] \*(68^2) + acd.fit\_SA\_MA2$coefficients[4] \* acd\_sa\_ma.ts[67]  
cat("Q4 2016 prediction: ",t68.ma.pred,"\n")

t69.ma.pred = acd.fit\_SA\_MA2$coefficients[1] + acd.fit\_SA\_MA2$coefficients[2]\*69 + acd.fit\_SA\_MA2$coefficients[3] \*(69^2) + acd.fit\_SA\_MA2$coefficients[4] \* t68.ma.pred  
cat("Q1 2017 prediction: ",t69.ma.pred,"\n")

t70.ma.pred = acd.fit\_SA\_MA2$coefficients[1] + acd.fit\_SA\_MA2$coefficients[2]\*70 + acd.fit\_SA\_MA2$coefficients[3] \*(70^2) + acd.fit\_SA\_MA2$coefficients[4] \* t69.ma.pred  
cat("Q2 2017 prediction: ",t70.ma.pred,"\n")

t71.ma.pred = acd.fit\_SA\_MA2$coefficients[1] + acd.fit\_SA\_MA2$coefficients[2]\*71 + acd.fit\_SA\_MA2$coefficients[3] \*(71^2) + acd.fit\_SA\_MA2$coefficients[4] \* t70.ma.pred  
cat("Q3 2017 prediction: ",t71.ma.pred,"\n")

RMSEP.ma.sa = sqrt((1/4)\*sum((A\_carbon.df$dioxide[68]-t68.ma.pred)^2,(A\_carbon.df$dioxide[69]-t69.ma.pred)^2,(A\_carbon.df$dioxide[70]-t70.ma.pred)^2,(A\_carbon.df$dioxide[71]-t71.ma.pred)^2))  
cat("The RMSEP for this seasonally adjusted moving average model is",RMSEP.ma.sa,"\n")

## Q4 2016 prediction: 404.6857

## Q1 2017 prediction: 405.1752

## Q2 2017 prediction: 405.7136

## Q3 2017 prediction: 406.2803

## The RMSEP for this seasonally adjusted moving average model is 2.359868

## (Intercept) Time[-1] I(Time[-1]^2) acd\_sa\_ma.ts[-67]   
## 1.623448e+02 2.013264e-01 4.514249e-04 5.604094e-01

The Model I finalised that had the highest predicting power using the RMSEP was:

Or more exact estimated model is:

ACDf= 162.3448 + 0.2013264Time + 0.0004514249Time2 + 0.5604094 \* yt-1 + ∊t, where yt-1 is the observation value at quarter before the time you will forecast.

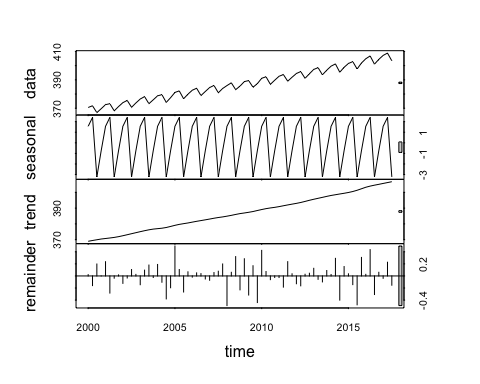
# Question 3 -

Re-running Seasonally adjusted Moving averages model - with all data

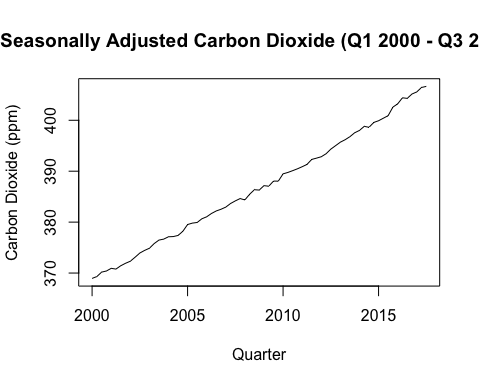
Time.sa = 1:71  
Quarter.full.sa = factor(c(rep(1:4,17),(1:3)))  
ACD.sa.ts = ts(A\_carbon.df$dioxide[1:71],start=2000, frequency=4)  
decomp.ma.acd = decompose(ACD.sa.ts)  
  
decomp.stl.acd = stl(ACD.sa.ts, s.window = "periodic")  
decomp.stl.acd$time.series[1:4,1]

## [1] 1.5596894 2.4042105 -3.2254920 -0.7384081

plot(decomp.stl.acd)



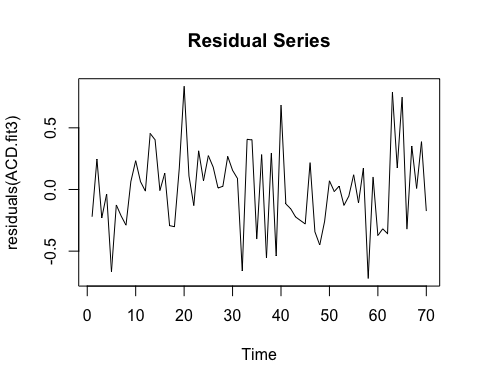
acd\_ma.sa.ts = ACD.sa.ts - decomp.ma.acd$seasonal  
  
plot(acd\_ma.sa.ts, main = "Seasonally Adjusted Carbon Dioxide (Q1 2000 - Q3 2017)" , xlab="Quarter", ylab="Carbon Dioxide (ppm)")



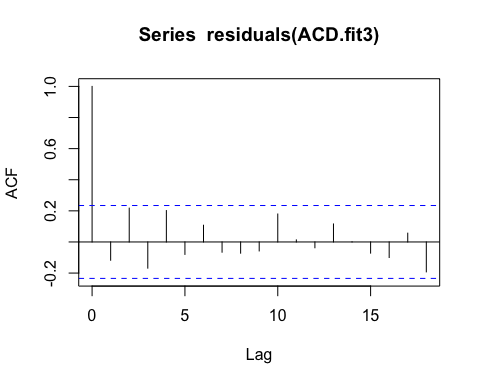
ACD.fit3 = lm(acd\_ma.sa.ts[-1]~Time.sa[-1]+ I(Time.sa[-1]^2) + acd\_ma.sa.ts[-71])  
summary(ACD.fit3)

##   
## Call:  
## lm(formula = acd\_ma.sa.ts[-1] ~ Time.sa[-1] + I(Time.sa[-1]^2) +   
## acd\_ma.sa.ts[-71])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.72099 -0.24587 0.01019 0.20833 0.83587   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.502e+02 3.647e+01 4.118 0.000109 \*\*\*  
## Time.sa[-1] 1.822e-01 4.541e-02 4.014 0.000155 \*\*\*  
## I(Time.sa[-1]^2) 4.884e-04 1.569e-04 3.112 0.002743 \*\*   
## acd\_ma.sa.ts[-71] 5.936e-01 9.903e-02 5.994 9.47e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3452 on 66 degrees of freedom  
## Multiple R-squared: 0.999, Adjusted R-squared: 0.999   
## F-statistic: 2.275e+04 on 3 and 66 DF, p-value: < 2.2e-16

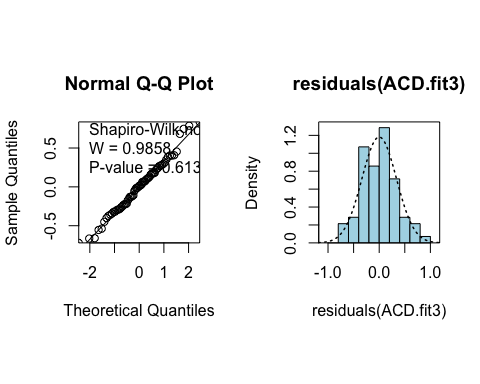
plot.ts(residuals(ACD.fit3), main = "Residual Series")



acf(residuals(ACD.fit3))



normcheck(residuals(ACD.fit3), shapiro.wilk = T)



#SA MA Forecast  
  
t72.ma.pred = ACD.fit3$coefficients[1] + ACD.fit3$coefficients[2]\*72 + ACD.fit3$coefficients[3] \*(72^2) + ACD.fit3$coefficients[4] \* acd\_ma.sa.ts[71]  
  
t73.ma.pred = ACD.fit3$coefficients[1] + ACD.fit3$coefficients[2]\*73 + ACD.fit3$coefficients[3] \*(73^2) + ACD.fit3$coefficients[4] \* t72.ma.pred  
  
t74.ma.pred = ACD.fit3$coefficients[1] + ACD.fit3$coefficients[2]\*74 + ACD.fit3$coefficients[3] \*(74^2) + ACD.fit3$coefficients[4] \* t73.ma.pred  
  
t75.ma.pred = ACD.fit3$coefficients[1] + ACD.fit3$coefficients[2]\*75 + ACD.fit3$coefficients[3] \*(75^2) + ACD.fit3$coefficients[4] \* t74.ma.pred

Technical Notes

#TO DO PREDICTION INTERVALS   
pred72.high = t72.ma.pred + 1.96 \* 0.3452  
pred72.low = t72.ma.pred - 1.96 \* 0.3452  
pred73.high = t73.ma.pred + 1.96 \* 0.3452  
pred73.low = t73.ma.pred - 1.96 \* 0.3452  
pred74.high = t74.ma.pred + 1.96 \* 0.3452  
pred74.low = t74.ma.pred - 1.96 \* 0.3452  
pred75.high = t75.ma.pred + 1.96 \* 0.3452  
pred75.low = t75.ma.pred - 1.96 \* 0.3452  
  
cat("Prediction point Q4 2017= ",t72.ma.pred,"Lower limit t72 = ",pred72.low, " Upper limit t72 = ",pred72.high, "\n")

cat("Prediction point Q1 2018= ",t73.ma.pred,"Lower limit t73 = ",pred73.low, " Upper limit t73 = ",pred73.high, "\n")

cat("Prediction point Q2 2018= ",t74.ma.pred,"Lower limit t74 = ",pred74.low, " Upper limit t74 = ",pred74.high, "\n")

cat("Prediction point Q3 2018= ",t75.ma.pred,"Lower limit t75 = ",pred75.low, " Upper limit t75 = ",pred75.high, "\n")

## Prediction point Q4 2017= 407.2015 Lower limit t72 = 406.5249 Upper limit t72 = 407.8781

## Prediction point Q1 2018= 407.7812 Lower limit t73 = 407.1046 Upper limit t73 = 408.4578

## Prediction point Q2 2018= 408.3793 Lower limit t74 = 407.7028 Upper limit t74 = 409.0559

## Prediction point Q3 2018= 408.9894 Lower limit t75 = 408.3128 Upper limit t75 = 409.666

Technical notes for question 3:

The plot of the quarterly measurements of atmospheric carbon dioxide in parts per million data shows an increasing trend from 370.56 ppm to 403.38 ppm from Q1 2000 to 3Q 2017, with slight curvature, with a reasonable constant seasonal component.

Model building steps (Your question was not clear if you wanted this answered in Q3 or Q2 or none at all, so I put them here like our lecture examples) The first (fit1 model) was a simple moving average seasonally adjusted model. After reviewing the residual plot, it was clear their was some curvature and we needed to fit a quadratic. The quadratic was clearly significant after reviewing the summery. Fit2, we plotted the acf and that also made it clear their was significant positive autocorrelation for lag(1) and slight positive autocorrelation for lag(2) and slight negative autocorrelation for lag(18).

Final fit (ACD.fit3) – A lagged response variable was fitted into the model. The plot of autocorrelation function shows that lag(1)(2) and (18) was no longer significant, therefore highlitting that fitting a lagged response variable was significant. We then plotted for our assumptions check of the normality or symmetry of the data distribution looked fine. There is no evidence against the underlying errors having come from a normal distribution (P-value = 0.613)

The significane of time, looking at the p-value of 0.00015 for Time and 0.0027 for the quadratic term and the lagged response p-value ≈ 0, means all the variables are significant in the model as all p-value < 0.05.

The adjusted R-squared also improved (just) after correcting for autocorrelation by adding the lagged response from 0.9984 to 0.9988.

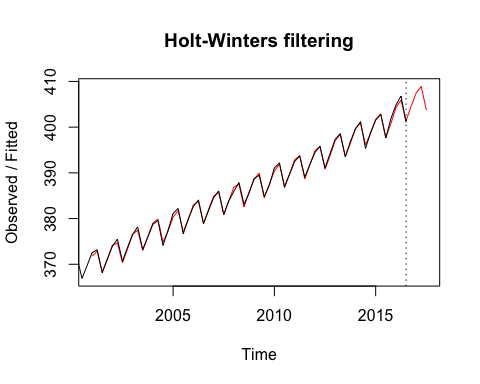
The F-statistics provides extremely strong evidence against the hypothesis that none of the variables are related to the seasonally adjusted atmospheric carbon dioxide (p-value ≈ 0). The multiple R2 of 0.9988 indicating that 99% of the variation in the seasonally adjusted atmospheric carbon dioxide levels is explained by the model. Prediction will be reliable….?

Question 4 - Holt Winters Model Additive Model

ACD.fit.hw = HoltWinters(acd.ts)  
  
ACD.fit.hw.pred = predict(ACD.fit.hw, n.ahead = 4)  
ACD.fit.hw.pred

## Qtr1 Qtr2 Qtr3 Qtr4  
## 2016 404.5566  
## 2017 407.4691 408.9130 403.7358

plot(ACD.fit.hw,ACD.fit.hw.pred)



actual = A\_carbon.df$dioxide[68:71]  
  
RMSEP\_ACD.hw.pred = sqrt(1/4 \* sum((actual-ACD.fit.hw.pred)^2))  
cat("The RMSEP for the Holt winters model is: ",RMSEP\_ACD.hw.pred)

## The RMSEP for the Holt winters model is: 0.2419712

As you can see there is a better model, the Holt-Winters additive model produces a far better model then the Seasonally adjusted moving averages model. We chose the additive model because our data shows that the seasonal component is reasonably constant through time, if the seasonal component was not constant we would use the multiplicative model. The better model is defined as having the strongest predictive power by predicting 4 quarters that we already know and calculate the RMSEP, the RMSEP for the Holt winters is significantly smaller of 0.2419712 Comments: justify answer for this being better = RMSEP is smaller