Stats 326: Assignment 3

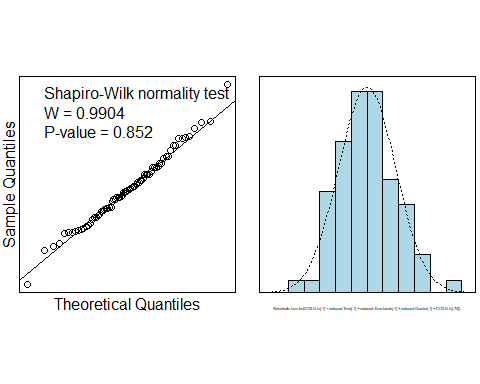
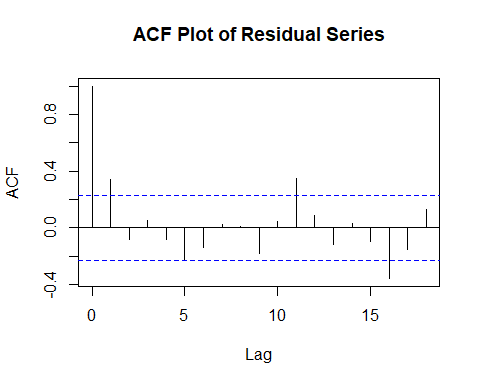
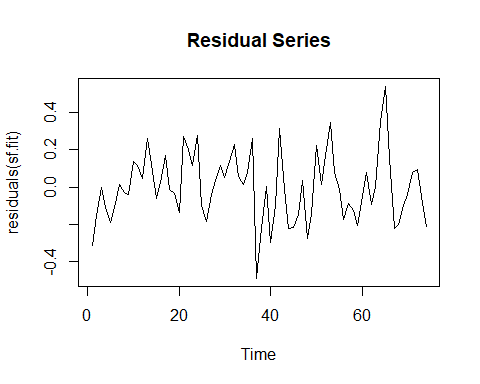
Hasnain Cheena

16/04/2020

## Question 1

sf.fit = lm(CO2.fit.ts[-1] ~ reduced.Time[-1] + reduced.Time.break[-1 ]+ reduced.Quarter[-1] + CO2.fit.ts[-75])  
  
#model summary  
summary(sf.fit)

##   
## Call:  
## lm(formula = CO2.fit.ts[-1] ~ reduced.Time[-1] + reduced.Time.break[-1] +   
## reduced.Quarter[-1] + CO2.fit.ts[-75])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.48990 -0.12209 -0.00581 0.11306 0.53995   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 108.82449 29.39775 3.702 0.000435 \*\*\*  
## reduced.Time[-1] 0.14548 0.03860 3.769 0.000349 \*\*\*  
## reduced.Time.break[-1] 0.04182 0.01196 3.496 0.000843 \*\*\*  
## reduced.Quarter[-1]2 0.43876 0.07593 5.778 2.14e-07 \*\*\*  
## reduced.Quarter[-1]3 1.14763 0.07477 15.348 < 2e-16 \*\*\*  
## reduced.Quarter[-1]4 0.41510 0.06543 6.344 2.22e-08 \*\*\*  
## CO2.fit.ts[-75] 0.70187 0.08043 8.727 1.18e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1889 on 67 degrees of freedom  
## Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997   
## F-statistic: 4.282e+04 on 6 and 67 DF, p-value: < 2.2e-16



#forecast 2018 Q4  
t76.sf.pred = sf.fit$coefficients[1] + (sf.fit$coefficients[2] \* 76) + (sf.fit$coefficients[3] \* 26) +   
 (sf.fit$coefficients[6]\*1) + (sf.fit$coefficients[7]\*CO2.fit.ts[75])  
  
#forecast 2019 Q1  
t77.sf.pred = sf.fit$coefficients[1] + (sf.fit$coefficients[2] \* 77) + (sf.fit$coefficients[3] \* 27) +   
 (sf.fit$coefficients[7]\*t76.sf.pred)  
  
#forecast 2019 Q2  
t78.sf.pred = sf.fit$coefficients[1] + (sf.fit$coefficients[2] \* 78) + (sf.fit$coefficients[3] \* 28) +   
 (sf.fit$coefficients[4]\*1) + (sf.fit$coefficients[7]\*t77.sf.pred)  
  
#forecast 2019 Q3  
t79.sf.pred = sf.fit$coefficients[1] + (sf.fit$coefficients[2] \* 79) + (sf.fit$coefficients[3] \* 29) +   
 (sf.fit$coefficients[5]\*1) + (sf.fit$coefficients[7]\*t78.sf.pred)  
  
#results  
results.sf.df = data.frame(Time=c("2018.4", "2019.1", "2019.2", "2019.3"),  
 Predictions=c(t76.sf.pred,t77.sf.pred,t78.sf.pred,t79.sf.pred))  
results.sf.df

## Time Predictions  
## 1 2018.4 406.0347  
## 2 2019.1 406.1401  
## 3 2019.2 406.8401  
## 4 2019.3 408.2276

#RMSEP  
sf.RMSEP = sqrt(1/4\*sum((results.sf.df$Predictions-CO2.pred.ts)^2))  
sf.RMSEP

## [1] 0.2384888

The Seasonal Factor model included a Time variable, break-in trend variable, seasonal factor and lagged response variable. The Residual Series showed reasonably constant scatter about 0. There is a slight upward trend in the early part of the Residual Series. The plot of the autocorrelation function shows signficant lag at 1, 11 and 16. The residuals appeared to follow a normal distribution (Shapiro-Wilk P-value = 0.852, min=-0.49 and max=0.54).

The RMSEP indicates that, on average, the prediction error is 0.24 ppm.

## Question 2

#model fit  
FH.fit = lm(CO2.fit.ts[-1] ~ reduced.Time[-1] + reduced.Time.break[-1 ]+ c1[-1] + s1[-1] + c2[-1] + CO2.fit.ts[-75])  
#model summary   
summary(FH.fit)

##   
## Call:  
## lm(formula = CO2.fit.ts[-1] ~ reduced.Time[-1] + reduced.Time.break[-1] +   
## c1[-1] + s1[-1] + c2[-1] + CO2.fit.ts[-75])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.48990 -0.12209 -0.00581 0.11306 0.53995   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 109.32486 29.38088 3.721 0.000408 \*\*\*  
## reduced.Time[-1] 0.14548 0.03860 3.769 0.000349 \*\*\*  
## reduced.Time.break[-1] 0.04182 0.01196 3.496 0.000843 \*\*\*  
## c1[-1] -0.01183 0.04370 -0.271 0.787404   
## s1[-1] -0.57381 0.03739 -15.348 < 2e-16 \*\*\*  
## c2[-1] -0.07344 0.02232 -3.290 0.001597 \*\*   
## CO2.fit.ts[-75] 0.70187 0.08043 8.727 1.18e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1889 on 67 degrees of freedom  
## Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997   
## F-statistic: 4.282e+04 on 6 and 67 DF, p-value: < 2.2e-16

#forecast 2018 Q4  
t76.FH.pred = FH.fit$coefficients[1] + (FH.fit$coefficients[2] \* 76) + (FH.fit$coefficients[3] \* 26) +   
(FH.fit$coefficients[4] \*cos(2\*pi\*76\*(1/4))) +   
(FH.fit$coefficients[5] \*sin(2\*pi\*76\*(1/4))) +  
(FH.fit$coefficients[6] \*cos(2\*pi\*76\*(2/4))) +  
(FH.fit$coefficients[7]\*CO2.fit.ts[75])  
  
#forecast 2019 Q1  
t77.FH.pred = FH.fit$coefficients[1] + (FH.fit$coefficients[2] \* 77) + (FH.fit$coefficients[3] \* 27) +   
(FH.fit$coefficients[4] \*cos(2\*pi\*77\*(1/4))) +   
(FH.fit$coefficients[5] \*sin(2\*pi\*77\*(1/4))) +  
(FH.fit$coefficients[6] \*cos(2\*pi\*77\*(2/4))) +  
(FH.fit$coefficients[7]\*t76.FH.pred)  
  
#forecast 2019 Q2  
t78.FH.pred = FH.fit$coefficients[1] + (FH.fit$coefficients[2] \* 78) + (FH.fit$coefficients[3] \* 28) +   
(FH.fit$coefficients[4] \*cos(2\*pi\*78\*(1/4))) +   
(FH.fit$coefficients[5] \*sin(2\*pi\*78\*(1/4))) +  
(FH.fit$coefficients[6] \*cos(2\*pi\*78\*(2/4))) +  
(FH.fit$coefficients[7]\*t77.FH.pred)  
  
#forecast 2019 Q3  
t79.FH.pred = FH.fit$coefficients[1] + (FH.fit$coefficients[2] \* 79) + (FH.fit$coefficients[3] \* 29) +   
(FH.fit$coefficients[4] \*cos(2\*pi\*79\*(1/4))) +   
(FH.fit$coefficients[5] \*sin(2\*pi\*79\*(1/4))) +  
(FH.fit$coefficients[6] \*cos(2\*pi\*79\*(2/4))) +  
(FH.fit$coefficients[7]\*t78.FH.pred)  
  
#results  
results.FH.df = data.frame(Time=c("2018.4", "2019.1", "2019.2", "2019.3"),  
 Predictions=c(t76.FH.pred,t77.FH.pred,t78.FH.pred,t79.FH.pred))  
results.FH.df

## Time Predictions  
## 1 2018.4 406.0347  
## 2 2019.1 406.1401  
## 3 2019.2 406.8401  
## 4 2019.3 408.2276

#RMSEP  
FH.RMSEP = sqrt(1/4\*sum((results.FH.df$Predictions-CO2.pred.ts)^2))  
FH.RMSEP

## [1] 0.2384888

The Full Harmonic model produced the same results as the Seasonal Factor, which is expected. The Full Harmonic model had the smallest RMSEP of all the Harmonic models (0.238 ppm).

The Full Harmonic model included a Time variable, break-in trend variable, 3 harmonics (c1 with a P-value 0.79 being non-significant) and a lagged response variable. The Residual Series showed reasonably constant scatter about 0. There is a slight upward trend in the early part of the Residual Series. The plot of the autocorrelation function shows signficant lag at 1, 11 and 16. The residuals appeared to follow a normal distribution (Shapiro-Wilk P-value = 0.852, min=-0.49 and max=0.54).

A reduced harmonic model created by removing non-signficant pairs of harmonics is the same as the Full Harmonic model (as there are no non-significant pairs). Another reduced harmonic model created by removing non-significant harmonics resulted in removing c1 and produced an RMSEP of 0.243 ppm. Therefore, it as rejected. Furthermore, a cosine model was not appropriate as the observations did not follow a smooth harmonic curve for each year.

## Question 3

The Seasonal Factor model included a Time variable, break-in trend variable, seasonal factor and lagged response variable to account for the autocorrelation.

The Residual Series showed reasonably constant scatter about 0. There is a slight upward trend in the beginning of the Residual Series. However, as this is in the early part of the series it is not of concern. The plot of the autocorrelation function shows significant autocorrelation at lag1, lag11and lag16. However, this is not a concern as lag 11 and lag 16 are unusual lags to be significant in quarterly data. Furthermore, even though the plot shows significant positive autocorrelation at lag 1, our model has a lagged response variable to account for autocorrelation at lag 1 and itis only slightly significant; hence not a worry. The residuals appeared to follow a normal distribution (Shapiro-Wilk P-value = 0.852, min=-0.49 and max=0.54). Therefore, model assumptions are satisified.

We have strong evidence against the hypothesis that the coefficient associated time variable is 0 (p-value 0.0003). Further, we have strong evidence against the hypothesis that the coefficient associated with the break-in trend time variable is 0 (p-value 0.0008). Additionally, we have very strong evidence against the hypothesis of no autocorrelation (p-value 0).

The F-statistic provides extremely strong evidence against the hypothesis that none of the variables are related to the CO2 concentration (p-value 0). The Multiple R^2 is almost 1 (0.9997) indicating that nearly all the variation in the CO2 concentration is explained by the model.

The Residual Standard Error is 0.19 ppm so the prediction intervals should be reasonably narrow. The model predictions can be relied upon as the assumptions are satisfied. The RMSEP for the predictions from 2018 Q4 to 2019 Q3 was 0.238 ppm which was smaller than the Reduced Harmonic model (0.244 ppm) and the same as the Full Harmonic model.

The predictions for 2018 Q4 to 2019 Q3 are: 2018 Q4: 406.03 ppm 2019 Q1: 406.14 ppm 2019 Q2: 406.84 ppm 2019 Q3: 408.23 ppm

## Question 4

#model fit  
sf.full.fit = lm(CO2.full.ts[-1] ~ Time[-1] + Time.break[-1]+ Quarter[-1] + CO2.full.ts[-79])  
summary(sf.full.fit)

##   
## Call:  
## lm(formula = CO2.full.ts[-1] ~ Time[-1] + Time.break[-1] + Quarter[-1] +   
## CO2.full.ts[-79])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.50285 -0.12867 -0.00066 0.12102 0.53787   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 108.36491 28.78665 3.764 0.000341 \*\*\*  
## Time[-1] 0.14500 0.03785 3.831 0.000273 \*\*\*  
## Time.break[-1] 0.04082 0.01114 3.665 0.000474 \*\*\*  
## Quarter[-1]2 0.46150 0.07438 6.205 3.25e-08 \*\*\*  
## Quarter[-1]3 1.16834 0.07242 16.132 < 2e-16 \*\*\*  
## Quarter[-1]4 0.41786 0.06380 6.549 7.79e-09 \*\*\*  
## CO2.full.ts[-79] 0.70309 0.07876 8.927 3.20e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1888 on 71 degrees of freedom  
## Multiple R-squared: 0.9998, Adjusted R-squared: 0.9997   
## F-statistic: 5.127e+04 on 6 and 71 DF, p-value: < 2.2e-16

#FORECASTING  
  
#forecast 2019 Q4  
t80.sf.pred = sf.full.fit$coefficients[1] + (sf.full.fit$coefficients[2] \* 80) + (sf.full.fit$coefficients[3] \* 30) +   
 (sf.full.fit$coefficients[6]\*1) + (sf.full.fit$coefficients[7]\*CO2.full.ts[79])  
  
#forecast 2020 Q1  
t81.sf.pred = sf.full.fit$coefficients[1] + (sf.full.fit$coefficients[2] \* 81) + (sf.full.fit$coefficients[3] \* 31) +   
 (sf.full.fit$coefficients[7]\*t80.sf.pred)  
  
#forecast 2020 Q2  
t82.sf.pred = sf.full.fit$coefficients[1] + (sf.full.fit$coefficients[2] \* 82) + (sf.full.fit$coefficients[3] \* 32) +   
 (sf.full.fit$coefficients[4]\*1) + (sf.full.fit$coefficients[7]\*t81.sf.pred)  
  
#forecast 2020 Q3  
t83.sf.pred = sf.full.fit$coefficients[1] + (sf.full.fit$coefficients[2] \* 83) + (sf.full.fit$coefficients[3] \* 33) +   
 (sf.full.fit$coefficients[5]\*1) + (sf.full.fit$coefficients[7]\*t82.sf.pred)  
  
#results  
results.sf.df = data.frame(Time=c("2019.4", "2020.1", "2020.2", "2020.3"),  
 Predictions=c(t80.sf.pred,t81.sf.pred,t82.sf.pred,t83.sf.pred))  
results.sf.df

## Time Predictions  
## 1 2019.4 408.6447  
## 2 2020.1 408.6901  
## 3 2020.2 409.3694  
## 4 2020.3 410.7396

## [1] 0.740096

## CO2.full.ts[-79]   
## 0.9047168

## CO2.full.ts[-79]   
## 0.9758918

The summary shows the seasonal factor model parameter estimates for the full timeframe model () are similar to those of the reduced timeframe model (). As the model assumptions were satifised our predictions should be reasonably reliable.

The prediction intervals are between 0.74 and 0.98 ppm.

## Question 5

The best predicting model is the seasonal-trend-lowess seasonally adjusted model as it had the lowest RMSEP (0.2 ppm). As the model assumptions for the seasonally adjusted model were satisified, we should be able to rely on any predictions.

The prediction intervals for the Seasonally Adjusted and Seasonal Factors models are the same (both are 0.74 to 0.98 ppm).