Stat 280 Exercise 5 - Veronica Bayani

Name: Veronica Bayani Student Number: 2009-00574

I. Use the hotel_occupancy data (Hotel Occupancy Rate in the Philippines, Jan 2000 – Dec 2009) in PhilMonthlyData.csv.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
## Warning: package 'readr' was built under R version 4.2.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fpp2)
```

```
## Warning: package 'fpp2' was built under R version 4.2.2
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
## -- Attaching packages ---------------------------------------------- fpp2 2.4 --
## v forecast  8.18     v expsmooth 2.3
## v fma       2.4
```

```
## Warning: package 'forecast' was built under R version 4.2.2
```

```
##
```

```
library(tinytex)
```

```
## Warning: package 'tinytex' was built under R version 4.2.2
```

```
library(forecast)

philmon <- read.csv("PhilMonthlyData.csv", stringsAsFactors = FALSE, na.strings = c("NA"))
```

Getting the Hotel Occupancy Rate in the Philippines

```
hotelocc <- ts(na.omit(philmon$hotel_occupancy),start=c(2000,1),frequency=12)
hotelocc
```

```
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 2000 49.15500 56.62250 51.56250 51.86500 52.77500 47.85750 50.82000 51.34500
## 2001 49.15500 56.62250 51.56250 51.86500 52.77500 47.85750 50.82000 51.34500
## 2002 51.59500 57.73250 53.09500 56.51000 56.41750 52.49000 54.53000 53.51250
## 2003 54.05821 64.03250 55.54250 50.84750 53.71750 52.58250 56.76750 59.74250
## 2004 63.68500 70.39750 68.06500 66.33250 64.20250 62.33250 65.07750 64.38750
## 2005 66.58750 71.44500 67.80000 71.28500 69.86500 67.29250 66.52000 66.00500
## 2006 72.96500 73.79750 68.38250 68.51500 70.97750 64.43500 67.22500 66.12000
## 2007 74.59250 77.24000 73.26250 70.20250 72.01250 69.46750 70.91000 68.91750
## 2008 71.37000 75.20000 69.26500 73.64500 71.40250 65.70250 65.32000 63.56250
## 2009 64.13390 68.89215 64.54464 64.83269 65.11408 61.90907 63.61740 63.26045
##           Sep      Oct      Nov      Dec
## 2000 49.66000 54.13750 54.24000 55.00000
## 2001 49.66000 54.13750 54.24000 55.00000
## 2002 54.78500 57.48250 59.05750 54.99500
## 2003 61.00250 65.70750 68.62250 63.82500
## 2004 65.76500 67.03250 68.81067 66.29829
## 2005 64.84750 67.34000 74.43750 70.08807
## 2006 66.96000 68.68500 75.53750 71.12639
## 2007 69.73500 68.91250 73.88250 67.51000
## 2008 63.20000 68.43000 69.23000 61.50000
## 2009 63.56350 66.02322 68.96274 65.02663
```

Split the data into training and test data set:

Training dataset = Jan 2000 – Dec 2007; and Test dataset = Jan 2008 – Dec 2009.

```
#train dataset
hotelocc_train <- window(hotelocc,start=2000,end=c(2007,12))
hotelocc_train
```

```
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 2000 49.15500 56.62250 51.56250 51.86500 52.77500 47.85750 50.82000 51.34500
## 2001 49.15500 56.62250 51.56250 51.86500 52.77500 47.85750 50.82000 51.34500
## 2002 51.59500 57.73250 53.09500 56.51000 56.41750 52.49000 54.53000 53.51250
## 2003 54.05821 64.03250 55.54250 50.84750 53.71750 52.58250 56.76750 59.74250
## 2004 63.68500 70.39750 68.06500 66.33250 64.20250 62.33250 65.07750 64.38750
## 2005 66.58750 71.44500 67.80000 71.28500 69.86500 67.29250 66.52000 66.00500
## 2006 72.96500 73.79750 68.38250 68.51500 70.97750 64.43500 67.22500 66.12000
## 2007 74.59250 77.24000 73.26250 70.20250 72.01250 69.46750 70.91000 68.91750
##           Sep      Oct      Nov      Dec
## 2000 49.66000 54.13750 54.24000 55.00000
## 2001 49.66000 54.13750 54.24000 55.00000
## 2002 54.78500 57.48250 59.05750 54.99500
```

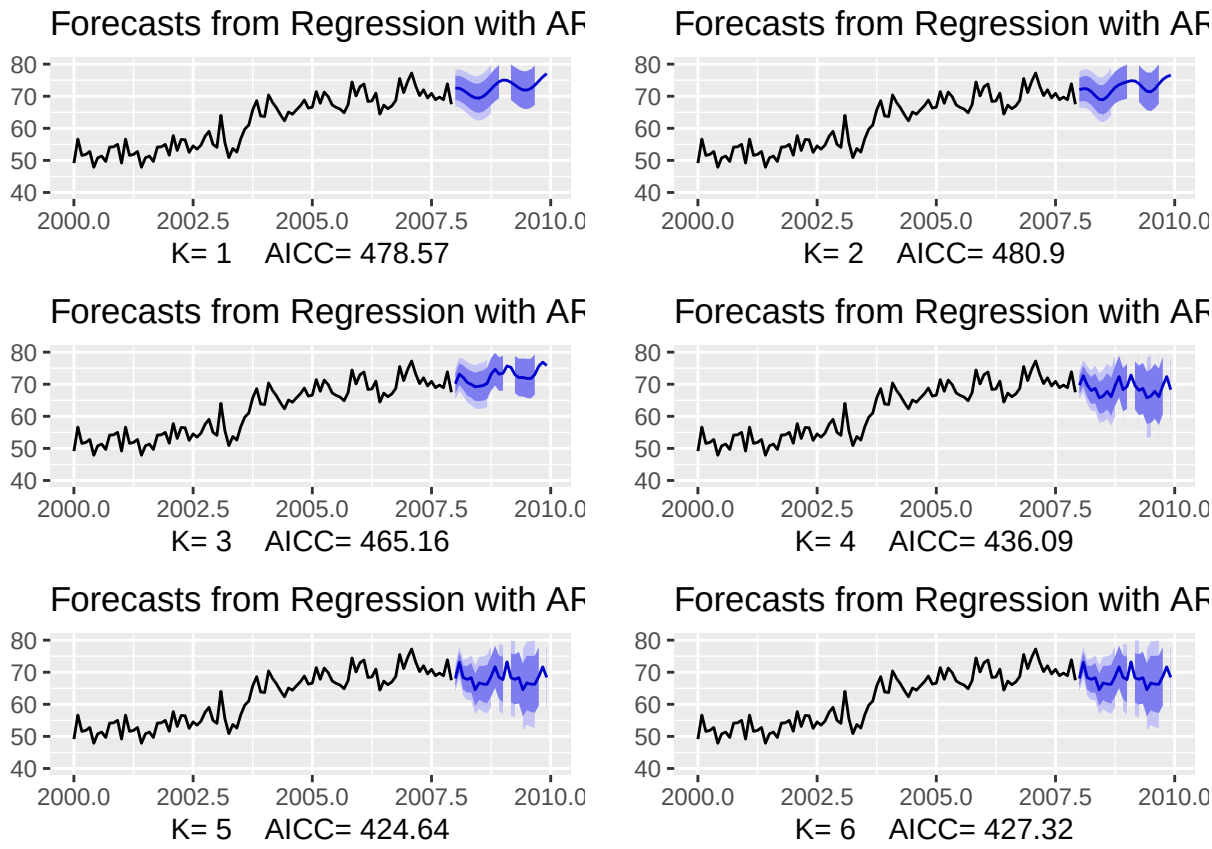
```
## 2003 61.00250 65.70750 68.62250 63.82500
## 2004 65.76500 67.03250 68.81067 66.29829
## 2005 64.84750 67.34000 74.43750 70.08807
## 2006 66.96000 68.68500 75.53750 71.12639
## 2007 69.73500 68.91250 73.88250 67.51000
```

```
#test dataset
hotelocc_test <- window(hotelocc,start=2008,end=c(2009,12))
hotelocc_test
```

```
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 2008 71.37000 75.20000 69.26500 73.64500 71.40250 65.70250 65.32000 63.56250
## 2009 64.13390 68.89215 64.54464 64.83269 65.11408 61.90907 63.61740 63.26045
##           Sep      Oct      Nov      Dec
## 2008 63.20000 68.43000 69.23000 61.50000
## 2009 63.56350 66.02322 68.96274 65.02663
```

1) [2pts] Using a dynamic harmonic regression with linear trend and ARIMA errors through auto.arima() function, show the best performing model for the training dataset and show the equations form of the model with estimated parameter values plugged in (for reference, see the fitted regression equation form in https://otexts.com/fpp2/regarima.html)

```
plots <- list()
for (i in seq(6)) {
  fit <- auto.arima(hotelocc_train, xreg = fourier(hotelocc_train, K = i),
    seasonal = FALSE)
  plots[[i]] <- autoplot(forecast(fit,
      xreg=fourier(hotelocc_train, K=i, h=24))) +
    xlab(paste("K=",i,"   AICC=",round(fit[["aicc"]],2))) +
    ylab("") + ylim(40,80)
}
gridExtra::grid.arrange(
  plots[[1]],plots[[2]],plots[[3]],
  plots[[4]],plots[[5]],plots[[6]], nrow=3)
```



The best performing model is the dynamic harmonic regression using K=5.

```
fit_fourier <- auto.arima(hotelocc_train, xreg = fourier(hotelocc_train, K = 5),
    seasonal = FALSE)
fit_fourier
```

```
## Series: hotelocc_train
## Regression with ARIMA(1,1,0) errors
##
## Coefficients:
##           ar1   S1-12   C1-12    S2-12    C2-12    S3-12    C3-12    S4-12
##       -0.3945  1.0109  1.9307  -0.3682  -0.4003  -0.7175  -0.6604  -1.2439
## s.e.   0.0953  0.4079  0.4020   0.2290   0.2274   0.1867   0.1867   0.1873
##         C4-12    S5-12   C5-12
##       -1.2928  -0.7327  0.6512
## s.e.   0.1876   0.2126  0.2124
##
## sigma^2 = 4.309:  log likelihood = -198.42
## AIC=420.84   AICc=424.64   BIC=451.48
```

The equation for the dynamic harmonic regression is given as:

$$y_t = \beta_0 + \sum_{k=1}^{K} [\alpha_k s_k(t) + \gamma_k c_k(t)] + \epsilon_t$$

where

$$s_k(t) = sin(\frac{2\pi * kt}{m})$$

and

$$c_k(t) = cos(\frac{2\pi * kt}{m})$$

where m is the seasonal period, α_k and γ_k are regression coefficients and ϵ_t is modeled as a non-seasonal ARIMA process.

Substituting the values for the k=5 model into the equation for the dynamic harmonic regression with the AR term yields,

Line 1 of the equation

$$y_t = 1.0109s_1(t)+1.9307c_1(t)-0.3682s_2(t)-0.4003c_2(t)-0.7175s_3(t)-0.6604c_3(t)-1.2439s_4(t)-1.2928c_4(t)+...$$

Line 2 of the equation

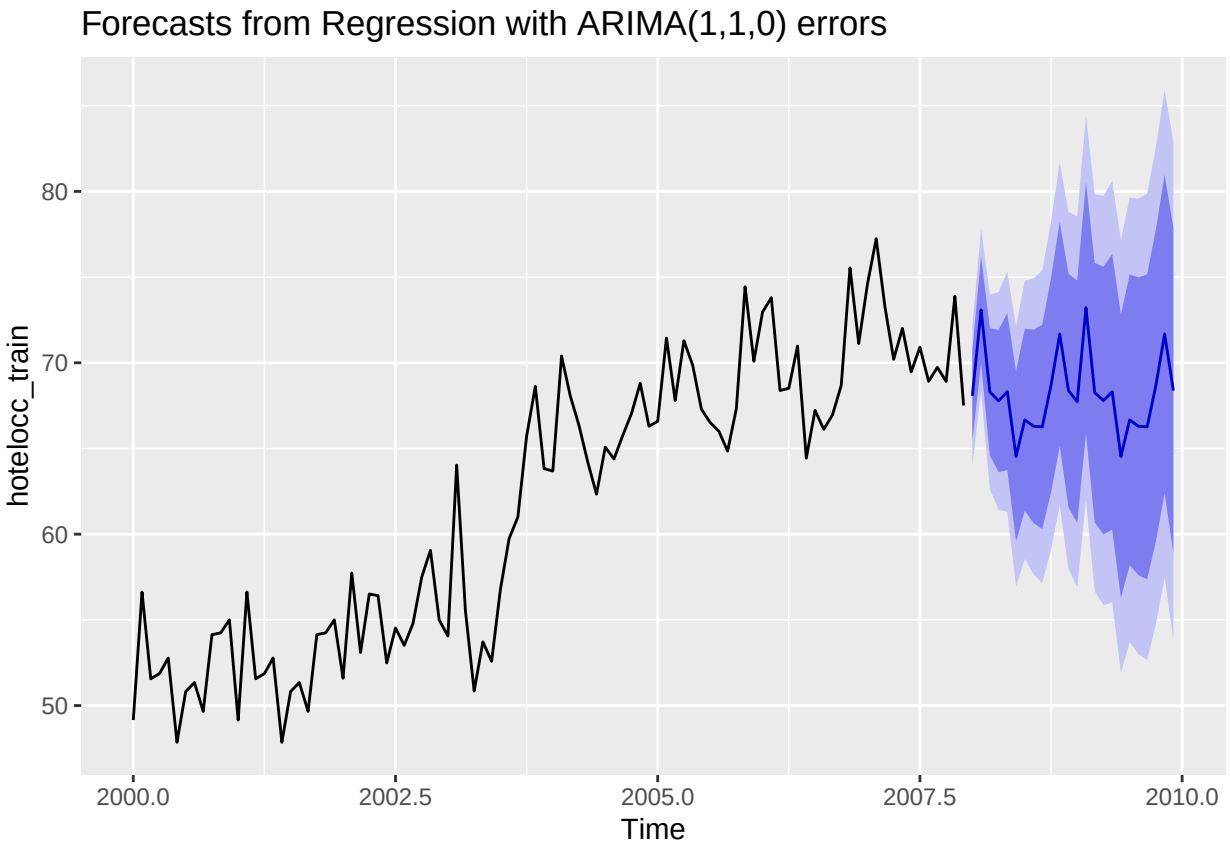
$$... - 0.7327s_5(t) + 0.6512c_5(t) - 0.3945\eta_{t-1} + \epsilon_t$$

and

$$\epsilon_t \sim NID(0, 4.309)$$

2) [1pt] Based on the model in (1), show a plot of the forecasted value of hotel_occupancy for the test data added into the plot of the full dataset. Analyze the plot in terms of the forecasting performance of the selected model in (1)

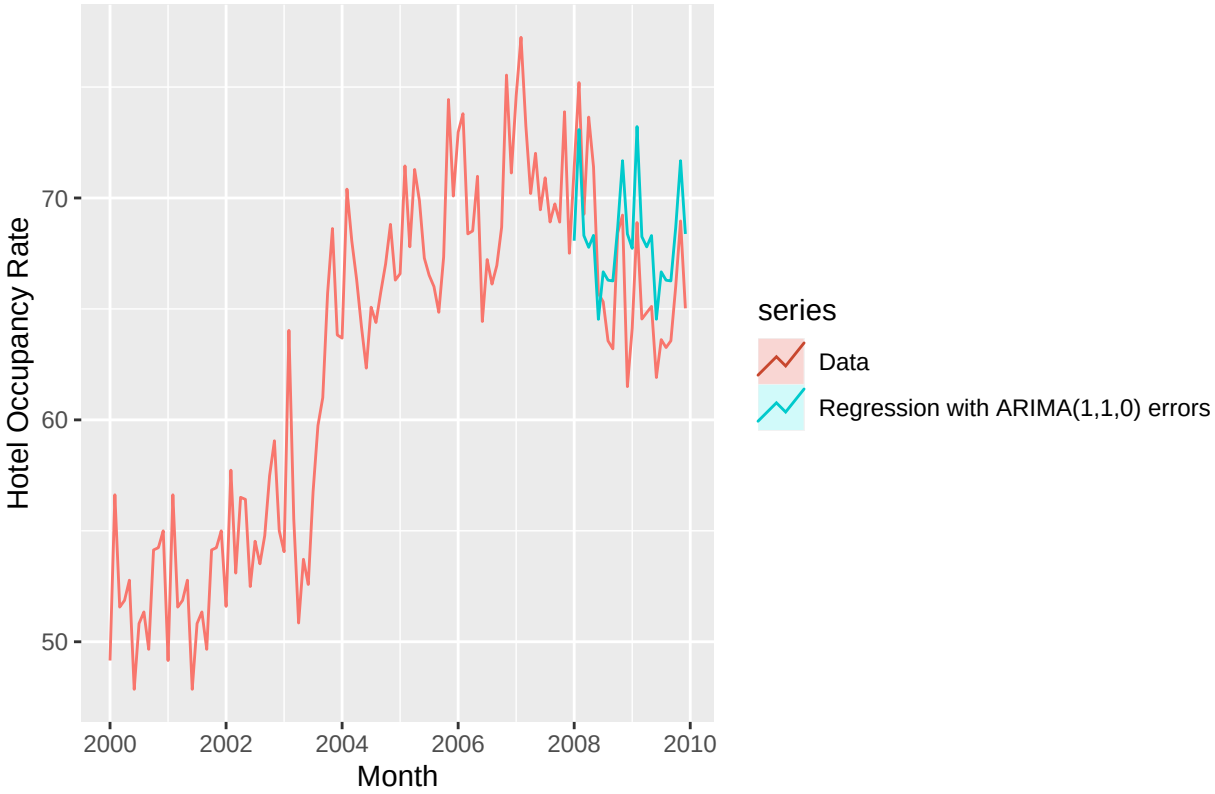
```
newharmonics2 <- fourier(hotelocc_train, K = 5,h=24)
fcx <- forecast(fit_fourier, xreg=newharmonics2)
autoplot(fcx)
```



Plotting the forecasts and the actual values together,

```
autoplot(hotelocc, series="Data") +
  autolayer(forecast(fcx, newdata = hotelocc_test), series="Regression with ARIMA(1,1,0) errors", PI=FA
  xlab("Month") + ylab("Hotel Occupancy Rate") +
  ggtitle("Figure 1. Hotel Occupancy Rate in the Philippines, Jan 2000 - Dec 2009")
```

Figure 1. Hotel Occupancy Rate in the Philippines, Jan 2000 – Dec 2009



The forecast using the Dynamic Harmonic Regression with ARIMA(1,1,0) errors seems to capture some seasonality from the original dataset but most of the forecasted values are higher than the actual values.

3) [1pt] Generate the accuracy measures of the selected model in (1) with respect to the testing dataset. Write a short analysis based on the accuracy measures.

Checking the accuracy for Dynamic Harmonic Regression with ARIMA(1,1,0) errors model,

```
accuracy(fcx, hotelocc_test)
```

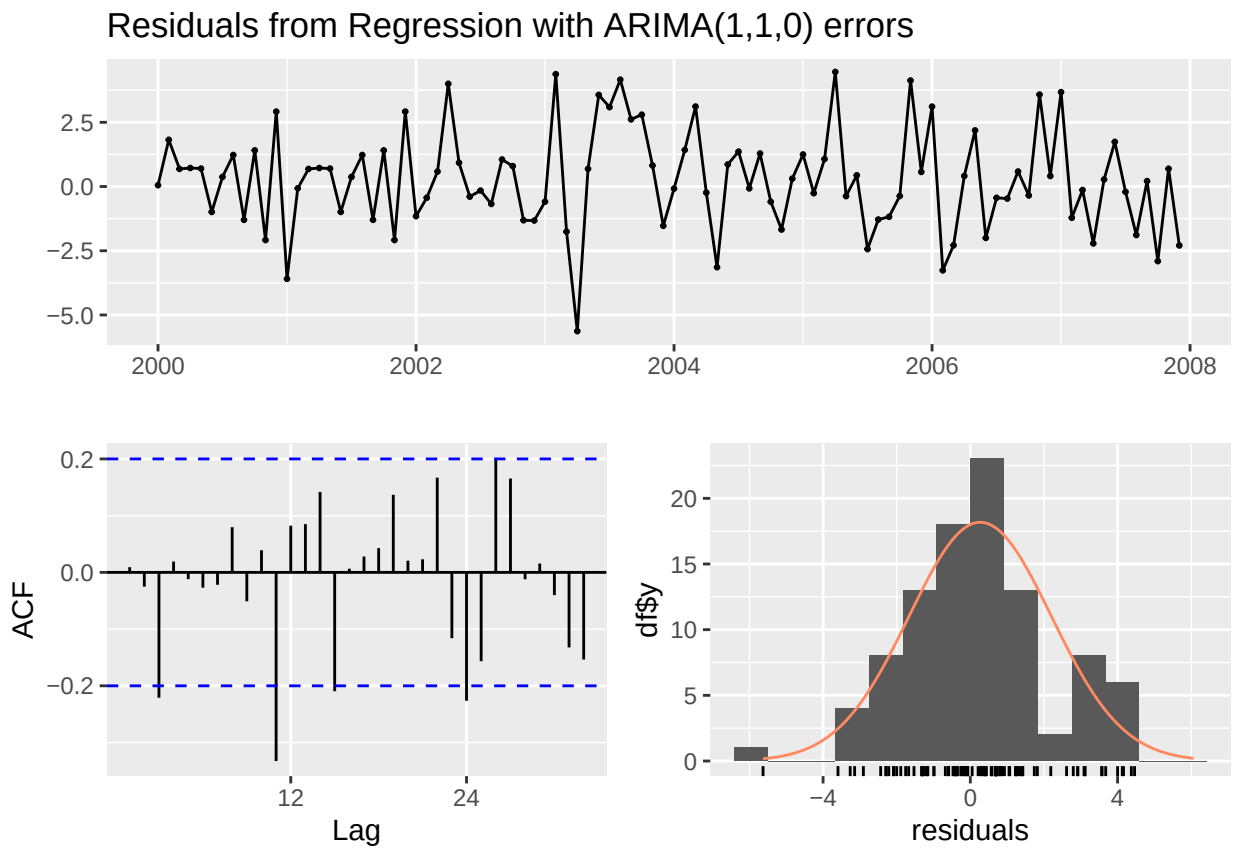
```
##                      ME     RMSE      MAE        MPE     MAPE      MASE
## Training set  0.2686672 1.941683 1.492589  0.3946599 2.439005 0.4600963
## Test set     -1.5897084 3.268773 2.964605 -2.5717442 4.479430 0.9138504
##                    ACF1 Theil's U
## Training set 0.009179168        NA
## Test set     0.695322031 0.9442845
```

The RMSE, MAE and MAPE of the Dynamic Harmonic Regression with ARIMA(1,1,0) errors model increased in the Test set as compared to the Train data set. This is expected since we saw in the forecast that, generally speaking, most of the forecast values are higher than the actual values.

4) [1pt] Check the residuals of the selected model in (1). Has the selected model in (1) complied with the properties that residuals should have for full extraction of the patterns from the time series? Any recommendations?

Testing for the presence of autocorrelation,

```
checkresiduals(fcx)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(1,1,0) errors
## Q* = 30.157, df = 18, p-value = 0.03594
##
## Model df: 1.   Total lags used: 19
```

Using a p value of 0.05, there is sufficient evidence to conclude that the residuals are not independently distributed and there is autocorrelation present in the data.

Testing for the normality of the residuals,

```
shapiro.test(residuals(fcx))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fcx)
## W = 0.98025, p-value = 0.157
```

Based on the results of the Shapiro-Wilk normality test, the residuals are normally distributed.

The Dynamic Harmonic Regression with ARIMA(1,1,0) errors model does not fully comply with the properties that residuals should have for full extraction of the patterns from the time series. There is sufficient evidence to conclude that autocorrelation is present based on the Ljung-Box test results even if the residuals are normally distributed.

To further improve the performance of the model, the autocorrelation issues must be addressed. This can be done by exploring the addition of other predictor variables in the model or via variable transformation.

II. Use the volpal data (Volume of Palay Production, Q1 1994 – Q4 2008) in PhilQuarterData.csv.

```
PhilQuarterlyData <- read.csv("PhilQuarterData.csv", stringsAsFactors = FALSE, na.strings = c("NA"))

#Quarterly Volume of Palay production data

PH_Quarter_palay <- ts(na.omit(PhilQuarterlyData$volpal), start=1994, frequency=4)
PH_Quarter_palay
```

```
##         Qtr1    Qtr2    Qtr3    Qtr4
## 1994 2288317 2090216 1876635 4282886
## 1995 2272045 2045286 1785510 4437808
## 1996 2523794 2427116 2116498 4216160
## 1997 2563757 2282704 1788141 4634361
## 1998 2220968 1338008 1284443 3711405
## 1999 2996188 2275865 2248702 4265870
## 2000 2856356 2586140 2412610 4534306
## 2001 2813930 2753901 2405187 4981852
## 2002 3058094 2614275 2020796 5577488
## 2003 3035561 2345717 2434696 5683910
## 2004 3434383 2604199 2871678 5586524
## 2005 3381885 2651140 2669137 5900843
## 2006 3615607 2923824 3006200 5781075
## 2007 3676704 3051176 3146959 6365355
## 2008 3748852 3371867 3467460 6227369
```

Split the data into training and test data set:

Training dataset = Q1 1994 – Q4 2005; and Test dataset = Q1 2006 – Q4 2008.

```
#train dataset
palay_train <- window(PH_Quarter_palay,start=1994,end=c(2005,4))
palay_train
```

```
##         Qtr1    Qtr2    Qtr3    Qtr4
## 1994 2288317 2090216 1876635 4282886
## 1995 2272045 2045286 1785510 4437808
## 1996 2523794 2427116 2116498 4216160
## 1997 2563757 2282704 1788141 4634361
## 1998 2220968 1338008 1284443 3711405
## 1999 2996188 2275865 2248702 4265870
## 2000 2856356 2586140 2412610 4534306
## 2001 2813930 2753901 2405187 4981852
## 2002 3058094 2614275 2020796 5577488
```

```
## 2003 3035561 2345717 2434696 5683910
## 2004 3434383 2604199 2871678 5586524
## 2005 3381885 2651140 2669137 5900843
```

```
#test dataset
palay_test <- window(PH_Quarter_palay,start=2006,end=c(2008,4))
palay_test
```

```
##         Qtr1    Qtr2    Qtr3    Qtr4
## 2006 3615607 2923824 3006200 5781075
## 2007 3676704 3051176 3146959 6365355
## 2008 3748852 3371867 3467460 6227369
```

1) [2pts] Using a dynamic harmonic regression with linear trend and ARIMA errors through auto.arima() function, show the best performing model for the training dataset and show the equations form of the model with estimated parameter values plugged in (for reference, see the fitted regression equation form in https://otexts.com/fpp2/regarima.html)

Since k must not be greater than period/2, we will try k=1 and k=2

Using k=1,

```
fit_fourier_palay1 <- auto.arima(palay_train, xreg = fourier(palay_train, K = 1),
    seasonal = FALSE)
fit_fourier_palay1
```

```
## Series: palay_train
## Regression with ARIMA(1,1,0) errors
##
## Coefficients:
##           ar1      S1-4        C1-4
##       -0.9301  336071.1  1212936.84
## s.e.   0.0517   48735.8    48734.76
##
## sigma^2 = 2.18e+11:  log likelihood = -679.67
## AIC=1367.35   AICc=1368.3   BIC=1374.75
```

Using k=2,

```
fit_fourier_palay2 <- auto.arima(palay_train, xreg = fourier(palay_train, K = 2),
    seasonal = FALSE)
fit_fourier_palay2
```

```
## Series: palay_train
## Regression with ARIMA(0,1,1) errors
##
## Coefficients:
##           ma1     drift       S1-4       C1-4       C2-4
##       -0.7899  25308.41  339118.58  1216330.4  538776.04
## s.e.   0.1275  11004.80   60635.45    60631.5   42517.32
##
## sigma^2 = 1.209e+11:  log likelihood = -664.21
## AIC=1340.42   AICc=1342.52   BIC=1351.52
```

K=2 has the lowest AICc between the 2 models.

The equation for the dynamic harmonic regression is given as:

$$y_t = \beta_0 + \sum_{k=1}^{K}[\alpha_k s_k(t) + \gamma_k c_k(t)] + \epsilon_t$$

where

$$s_k(t) = sin(\frac{2\pi * kt}{m})$$

and

$$c_k(t) = cos(\frac{2\pi * kt}{m})$$

where m is the seasonal period, α_k and γ_k are regression coefficients and ϵ_t is modeled as a non-seasonal ARIMA process.

Substituting the values for the k=2 model into the equation for the dynamic harmonic regression with the MA and drift term yields,

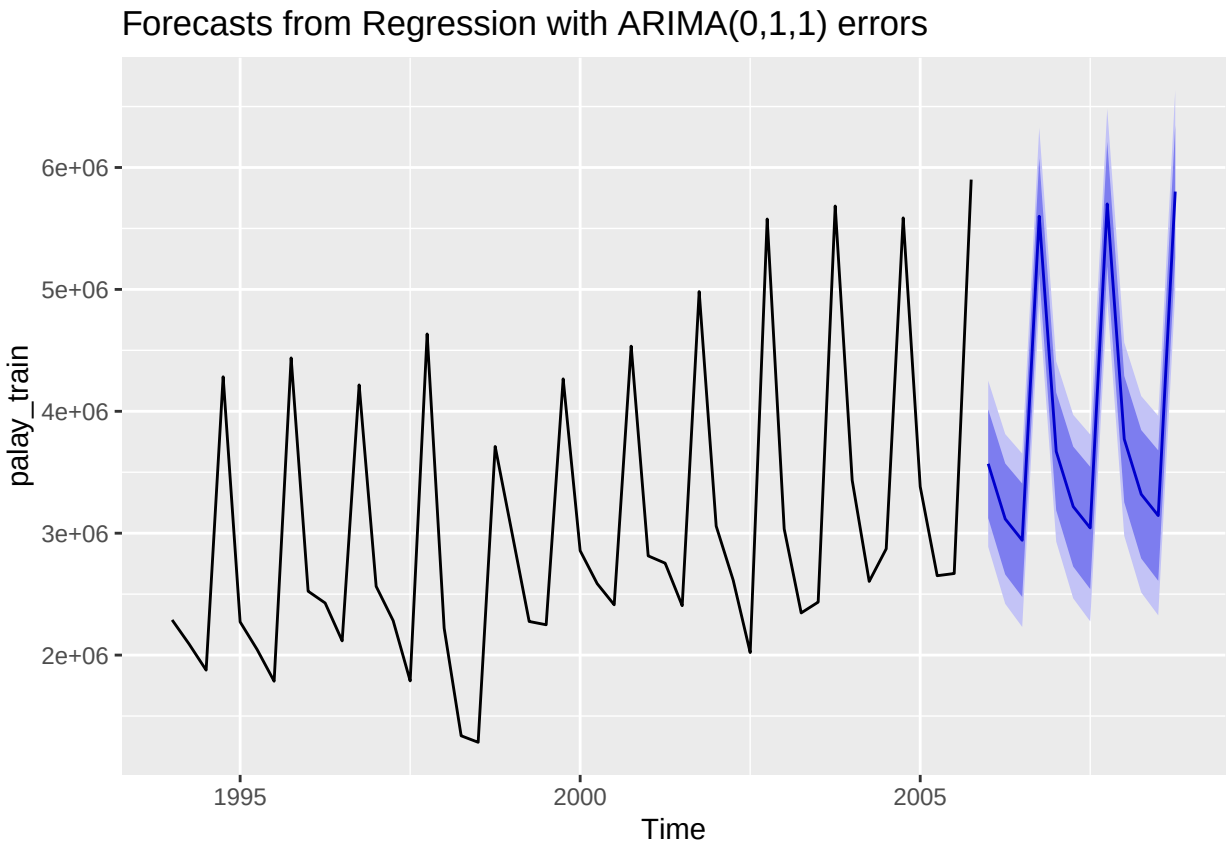
$$y_t = 339118.58s_1(t) + 1216330.4c_1(t) + 538776.04c_2(t) - 0.7899\epsilon_{t-1} + 25308.41 + \epsilon_t$$

and

$$\epsilon_t \sim NID(0, 1.209e + 11)$$

2) [1pt] Based on the model in (1), show a plot of the forecasted value of volpal for the test data added into the plot of the full dataset. Analyze the plot in terms of the forecasting performance of the selected model in (1)

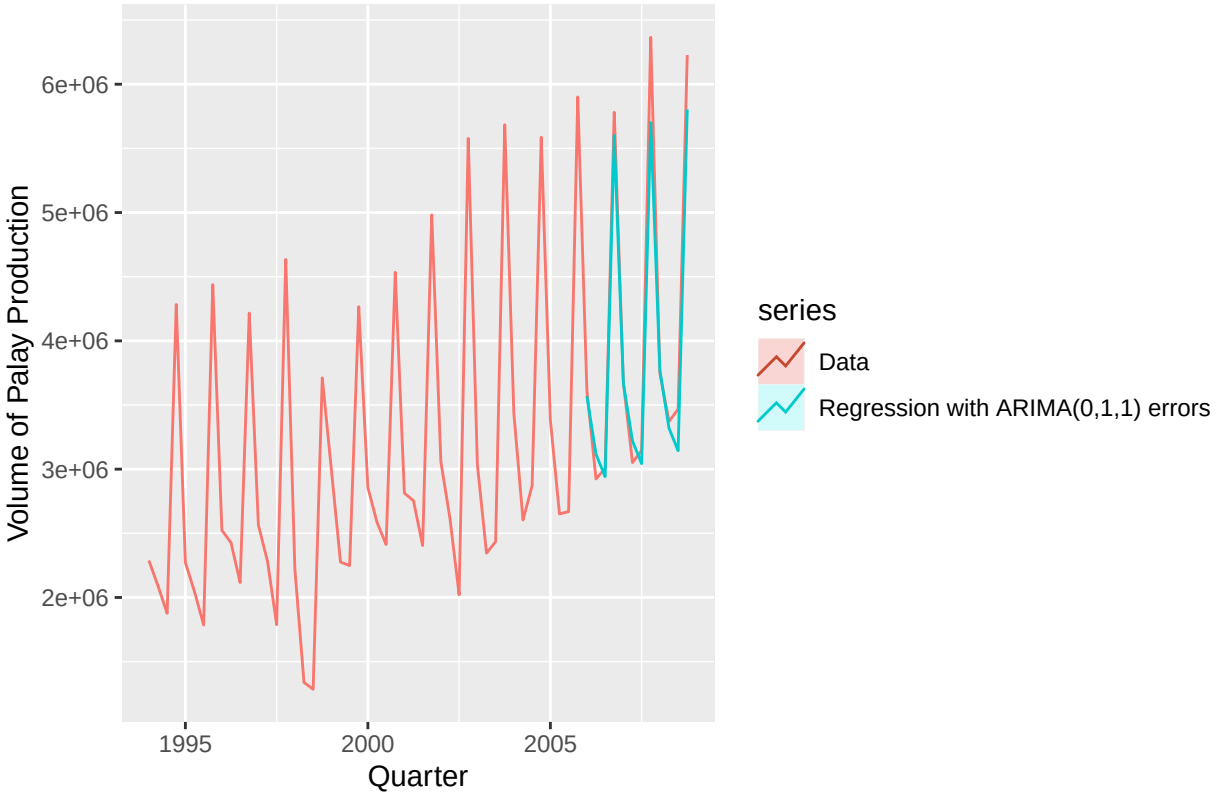
```
newharmonics3 <- fourier(palay_train, K = 2,h=12)
fcx2 <- forecast(fit_fourier_palay2, xreg=newharmonics3)
autoplot(fcx2)
```



Plotting the forecasts and the actual values together,

```
autoplot(PH_Quarter_palay, series="Data") +
  autolayer(forecast(fcx2, newdata = palay_test), series="Regression with ARIMA(0,1,1) errors", PI=FALSE
  xlab("Quarter") + ylab("Volume of Palay Production") +
  ggtitle("Figure 2. Volume of Palay Production, Q1 1994 - Q4 2008")
```

Figure 2. Volume of Palay Production, Q1 1994 – Q4 2008



Based on the plot of the actual and the forecast values, the Dynamic Harmonic Regression with ARIMA(0,1,1) errors model provides a good forecast for the Volume of Palay Production. The seasonality and trend are captured well in the forecast and the actual and forecast values are close to each other.

3) [1pt] Generate the accuracy measures of the selected model in (1) with respect to the testing dataset. Write a short analysis based on the accuracy measures.

Checking the accuracy for the Dynamic Harmonic Regression with ARIMA(0,1,1) errors model

```
accuracy(forecast(fcx2,h=12), palay_test)
```

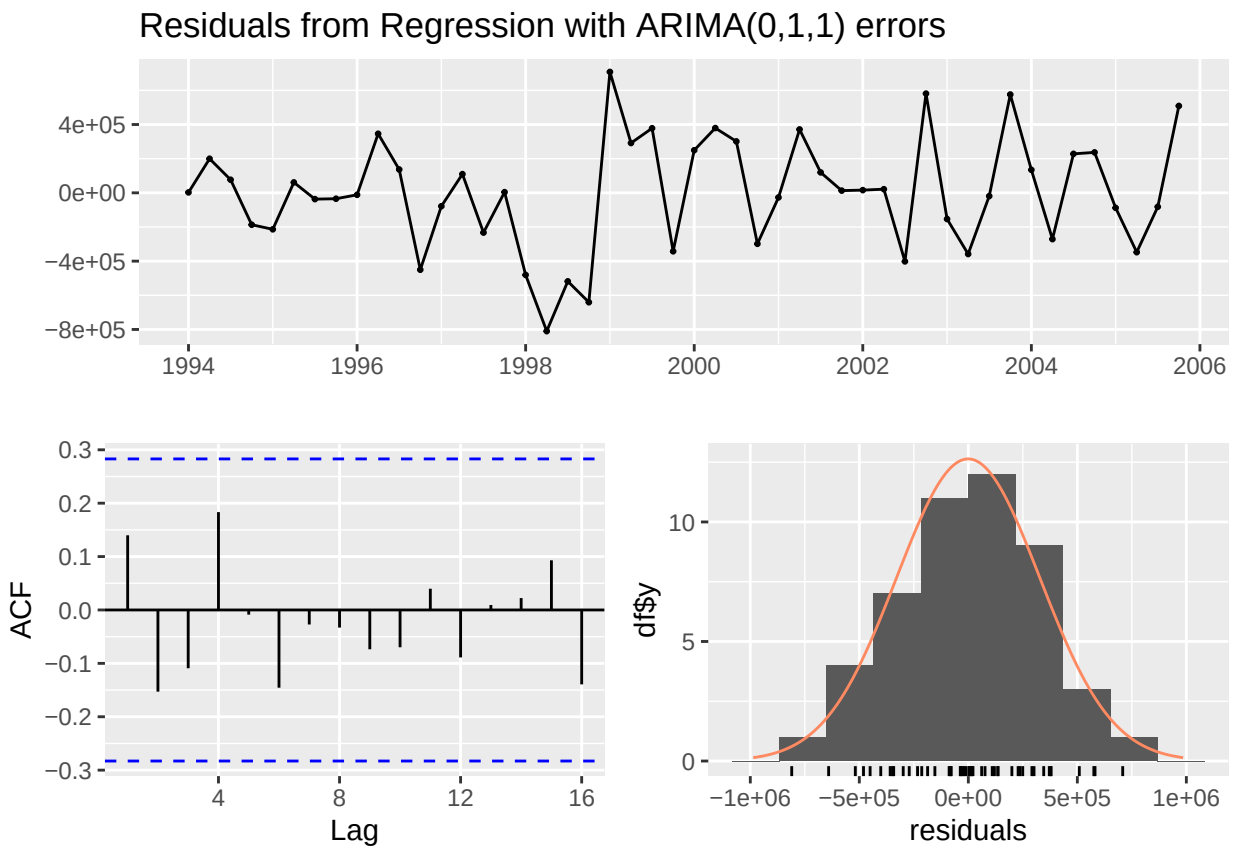
```
##                      ME     RMSE    MAE       MPE     MAPE      MASE
## Training set   -690.1792 325212.9 252814 -1.511737 9.712809 0.8002086
## Test set     123627.0792 265230.8 187434  2.122844 4.236595 0.5932674
##                    ACF1 Theil's U
## Training set 0.13985248        NA
## Test set     0.05970835 0.1648527
```

The RMSE, MAE and MAPE of the Dynamic Harmonic Regression with ARIMA(0,1,1) errors model decreased in the test set versus the training set. This is reasonable since, as seen on the plot, the forecast values are very close to the actual values.

4) [1pt] Check the residuals of the selected model in (1). Has the selected model in (1) complied with the properties that residuals should have for full extraction of the patterns from the time series? Any recommendations?

Testing for the presence of autocorrelation,

```
checkresiduals(fcx2)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,1,1) errors
## Q* = 6.0099, df = 7, p-value = 0.5386
##
## Model df: 1.   Total lags used: 8
```

Using a p value of 0.05, there is sufficient evidence to conclude that the residuals are independently distributed and there is no autocorrelation present in the data.

Testing for the normality of the residuals,

```
shapiro.test(residuals(fcx2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fcx2)
## W = 0.99188, p-value = 0.9828
```

Based on the results of the Shapiro-Wilk normality test, the residuals are normally distributed.

The Dynamic Harmonic Regression with ARIMA(0,1,1) errors model fully complies with the properties that residuals should have for full extraction of the patterns from the time series. There is enough evidence to conclude that autocorrelation is not present and the residuals are normally distributed. This is validated by the plot of the forecast and the actual values where it is seen that the Dynamic Harmonic Regression with ARIMA(0,1,1) errors model does an excellent job in forecasting for the volume of palay production.