Assignment 3

Isuri Willaddara Gamage

July 11, 2020

### Load packages

#Data management packages  
library(ggplot2)  
library(coefplot)  
library(dplyr)

library(imputeTS)

library(forecast)

## Exercise 3: Predicting household income with logistic regression

### Read the dataset of the 2010 American Community Survey (ACS) for New York state into R library

#Read the dataset into R library  
survey.data <- read.csv("acs\_ny.csv")

### Examine the dataset and get the statistical summary of the data

#examine dataset  
head(survey.data)

## Acres FamilyIncome FamilyType NumBedrooms NumChildren NumPeople NumRooms  
## 1 10-Jan 150 Married 4 1 3 9  
## 2 10-Jan 180 Female Head 3 2 4 6  
## 3 10-Jan 280 Female Head 4 0 2 8  
## 4 10-Jan 330 Female Head 2 1 2 4  
## 5 10-Jan 330 Male Head 3 1 2 5  
## 6 10-Jan 480 Male Head 0 3 4 1  
## NumUnits NumVehicles NumWorkers OwnRent YearBuilt HouseCosts  
## 1 Single detached 1 0 Mortgage 1950-1959 1800  
## 2 Single detached 2 0 Rented Before 1939 850  
## 3 Single detached 3 1 Mortgage 2000-2004 2600  
## 4 Single detached 1 0 Rented 1950-1959 1800  
## 5 Single attached 1 0 Mortgage Before 1939 860  
## 6 Single detached 0 0 Rented Before 1939 700  
## ElectricBill FoodStamp HeatingFuel Insurance Language  
## 1 90 No Gas 2500 English  
## 2 90 No Oil 0 English  
## 3 260 No Oil 6600 Other European  
## 4 140 No Oil 0 English  
## 5 150 No Gas 660 Spanish  
## 6 140 No Gas 0 English

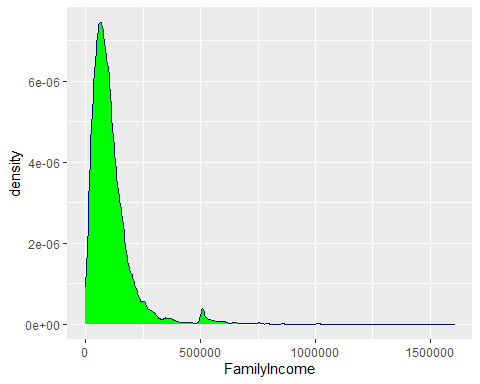
#summary stat of the dataset  
summary(survey.data)

## Acres FamilyIncome FamilyType NumBedrooms   
## Length:22745 Min. : 50 Length:22745 Min. :0.000   
## Class :character 1st Qu.: 52540 Class :character 1st Qu.:3.000   
## Mode :character Median : 87000 Mode :character Median :3.000   
## Mean : 110281 Mean :3.385   
## 3rd Qu.: 133800 3rd Qu.:4.000   
## Max. :1605000 Max. :8.000   
## NumChildren NumPeople NumRooms NumUnits   
## Min. : 0.0000 Min. : 2.00 Min. : 1.000 Length:22745   
## 1st Qu.: 0.0000 1st Qu.: 2.00 1st Qu.: 6.000 Class :character   
## Median : 0.0000 Median : 3.00 Median : 7.000 Mode :character   
## Mean : 0.9012 Mean : 3.39 Mean : 7.175   
## 3rd Qu.: 2.0000 3rd Qu.: 4.00 3rd Qu.: 8.000   
## Max. :12.0000 Max. :18.00 Max. :21.000   
## NumVehicles NumWorkers OwnRent YearBuilt   
## Min. :0.000 Min. :0.000 Length:22745 Length:22745   
## 1st Qu.:2.000 1st Qu.:1.000 Class :character Class :character   
## Median :2.000 Median :2.000 Mode :character Mode :character   
## Mean :2.113 Mean :1.745   
## 3rd Qu.:3.000 3rd Qu.:2.000   
## Max. :6.000 Max. :3.000   
## HouseCosts ElectricBill FoodStamp HeatingFuel   
## Min. : 4 Min. : 1 Length:22745 Length:22745   
## 1st Qu.: 650 1st Qu.:100 Class :character Class :character   
## Median :1200 Median :150 Mode :character Mode :character   
## Mean :1480 Mean :175   
## 3rd Qu.:2000 3rd Qu.:220   
## Max. :7090 Max. :580   
## Insurance Language   
## Min. : 0.0 Length:22745   
## 1st Qu.: 400.0 Class :character   
## Median : 720.0 Mode :character   
## Mean : 960.9   
## 3rd Qu.:1200.0   
## Max. :6600.0

### There are 18 variables in the dataset including 8 categorical variables. Predictor variable of the dataset is family income. Create binary variable using threshold $150,000 to predict the income greater than or less than threshold. We can see some variables have possible outliers such as familyincome, NumChildren and NumRoom, where maximum value of these variables has a high deviation from mean value.

### Create a density plot of family income to see distribution.

#create a density plot of family income to see distribution  
p <- ggplot(survey.data, aes(x=FamilyIncome)) +   
 geom\_density(color="darkblue", fill="green")  
p



* According to density plot we can see most of the family income lies between $0 and $250,000 of New York State.
* Dependent variable of the model - FamilyIncome. Create binary variable for family income as 1 for income above $150,000 and 0 for income below $150,000

### Build the Model with all Variables

#creating binary variable for family income (1 for income above $150,000 and 0 for income below $150,000)  
survey.data$FamilyIncome <- ifelse(survey.data$FamilyIncome >= 150000, 1, 0)  
  
#model building using logistic regression  
model <- glm(formula = FamilyIncome~., data = survey.data, family = "binomial")  
summary(model)

##   
## Call:  
## glm(formula = FamilyIncome ~ ., family = "binomial", data = survey.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3124 -0.6095 -0.3793 -0.1266 3.2645   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.125e+01 1.788e+00 -6.292 3.13e-10 \*\*\*  
## Acres10+ 1.005e-01 1.162e-01 0.865 0.387088   
## AcresSub 1 2.141e-01 5.601e-02 3.822 0.000133 \*\*\*  
## FamilyTypeMale Head 3.427e-01 1.498e-01 2.288 0.022119 \*   
## FamilyTypeMarried 1.299e+00 8.951e-02 14.517 < 2e-16 \*\*\*  
## NumBedrooms 8.013e-02 2.238e-02 3.581 0.000342 \*\*\*  
## NumChildren 2.342e-02 2.619e-02 0.894 0.371174   
## NumPeople -1.296e-01 2.370e-02 -5.467 4.57e-08 \*\*\*  
## NumRooms 1.090e-01 9.802e-03 11.122 < 2e-16 \*\*\*  
## NumUnitsSingle attached 2.367e+00 4.576e-01 5.173 2.30e-07 \*\*\*  
## NumUnitsSingle detached 2.238e+00 4.532e-01 4.938 7.88e-07 \*\*\*  
## NumVehicles 2.003e-01 2.370e-02 8.453 < 2e-16 \*\*\*  
## NumWorkers 5.839e-01 3.121e-02 18.709 < 2e-16 \*\*\*  
## OwnRentOutright 1.416e+00 2.229e-01 6.354 2.10e-10 \*\*\*  
## OwnRentRented -2.633e-01 1.057e-01 -2.491 0.012753 \*   
## YearBuilt1940-1949 1.761e+00 1.687e+00 1.044 0.296677   
## YearBuilt1950-1959 1.908e+00 1.687e+00 1.131 0.257857   
## YearBuilt1960-1969 1.878e+00 1.687e+00 1.113 0.265630   
## YearBuilt1970-1979 1.788e+00 1.687e+00 1.060 0.289237   
## YearBuilt1980-1989 2.136e+00 1.687e+00 1.266 0.205652   
## YearBuilt1990-1999 2.087e+00 1.687e+00 1.237 0.216147   
## YearBuilt2000-2004 2.036e+00 1.688e+00 1.206 0.227838   
## YearBuilt2005 2.058e+00 1.696e+00 1.214 0.224938   
## YearBuilt2006 1.968e+00 1.698e+00 1.159 0.246439   
## YearBuilt2007 2.242e+00 1.700e+00 1.319 0.187243   
## YearBuilt2008 1.545e+00 1.714e+00 0.901 0.367374   
## YearBuilt2009 1.984e+00 1.717e+00 1.155 0.248019   
## YearBuilt2010 2.246e+00 1.731e+00 1.297 0.194460   
## YearBuiltBefore 1939 1.765e+00 1.686e+00 1.046 0.295355   
## HouseCosts 5.915e-04 1.987e-05 29.763 < 2e-16 \*\*\*  
## ElectricBill 1.165e-03 1.889e-04 6.169 6.89e-10 \*\*\*  
## FoodStampYes -5.517e-01 1.257e-01 -4.388 1.14e-05 \*\*\*  
## HeatingFuelElectricity 7.160e-01 3.712e-01 1.929 0.053741 .   
## HeatingFuelGas 8.921e-01 3.589e-01 2.486 0.012935 \*   
## HeatingFuelNone -7.277e-01 1.134e+00 -0.642 0.521167   
## HeatingFuelOil 9.072e-01 3.592e-01 2.526 0.011552 \*   
## HeatingFuelOther 8.087e-01 4.309e-01 1.877 0.060530 .   
## HeatingFuelSolar 7.563e-01 1.258e+00 0.601 0.547730   
## HeatingFuelWood 3.984e-02 3.767e-01 0.106 0.915755   
## Insurance 2.941e-04 2.146e-05 13.702 < 2e-16 \*\*\*  
## LanguageEnglish -1.873e-01 9.655e-02 -1.940 0.052364 .   
## LanguageOther -1.874e-01 1.846e-01 -1.016 0.309858   
## LanguageOther European -2.020e-01 1.096e-01 -1.844 0.065168 .   
## LanguageSpanish -4.175e-01 1.162e-01 -3.593 0.000327 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 22808 on 22744 degrees of freedom  
## Residual deviance: 17226 on 22701 degrees of freedom  
## AIC: 17314  
##   
## Number of Fisher Scoring iterations: 7

* First build the model using dependent binary variable “FamilyIncome” and all other variables as independent variables using Generalized Linear Models function glm().
* The first model (full model) is,  
  formula = FamilyIncome ~ ., family = “binomial”, data = survey.data
* From the P-values of for the regression coefficients, we can see that there are some variables that may not make a significant contribution to the equation (P > 0.05), such as NumChildren, YearBuilt and Language. Build the second model called “survey.model” without non-significant variables.

### Build New Model with Significant Variables

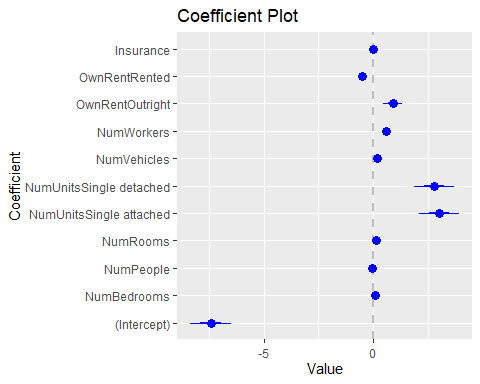
#select significant varibles to create new model  
survey.model <- glm(formula = FamilyIncome~NumBedrooms+NumPeople+NumRooms+NumUnits+NumVehicles  
 +NumWorkers+NumWorkers+OwnRent+Insurance, data = survey.data, family = "binomial")  
summary(survey.model)

##   
## Call:  
## glm(formula = FamilyIncome ~ NumBedrooms + NumPeople + NumRooms +   
## NumUnits + NumVehicles + NumWorkers + NumWorkers + OwnRent +   
## Insurance, family = "binomial", data = survey.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0658 -0.6541 -0.4819 -0.2219 3.3371   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.422e+00 4.584e-01 -16.193 < 2e-16 \*\*\*  
## NumBedrooms 1.090e-01 2.032e-02 5.366 8.05e-08 \*\*\*  
## NumPeople -3.688e-02 1.422e-02 -2.594 0.00947 \*\*   
## NumRooms 1.195e-01 9.063e-03 13.182 < 2e-16 \*\*\*  
## NumUnitsSingle attached 3.003e+00 4.552e-01 6.597 4.19e-11 \*\*\*  
## NumUnitsSingle detached 2.771e+00 4.514e-01 6.140 8.27e-10 \*\*\*  
## NumVehicles 1.833e-01 2.050e-02 8.943 < 2e-16 \*\*\*  
## NumWorkers 5.701e-01 2.717e-02 20.984 < 2e-16 \*\*\*  
## OwnRentOutright 9.056e-01 2.172e-01 4.169 3.06e-05 \*\*\*  
## OwnRentRented -4.999e-01 1.019e-01 -4.903 9.42e-07 \*\*\*  
## Insurance 5.941e-04 1.946e-05 30.529 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 22808 on 22744 degrees of freedom  
## Residual deviance: 19229 on 22734 degrees of freedom  
## AIC: 19251  
##   
## Number of Fisher Scoring iterations: 7

* The second model (reduced model) is  
  FamilyIncome ~ NumBedrooms + NumPeople + NumRooms + NumUnits + NumVehicles + NumWorkers + NumWorkers + OwnRent + Insurance
* Each regression coefficient in the new model is statistically significant (P < 0.05). The coefficient tells us how much that variables contribute to the log odds.
* Coefficient of the variable “NumBedrooms”  
  Each one-unit change in number of bedrooms will increase the log odds of family income by 1.090e-01, and its p-value indicates that it is significant in determining the income.
* Coefficient of the variable “OwnRent”  
  Home ownership outright has an increase relationship with the log odds of family income and rented has a decrease relationship with the log odds of family income.
* Coefficient of the variable “NumPeople”  
  Each one-unit change in number of people will decrease the log odds of family income by 3.688e-02, and its p-value indicates that it is somewhat significant in determining the income.
* Coefficient of other variable Number of rooms, number of units, number of vehicles, number of workers and Insurance have positive relationships with family income.
* The difference between Null deviance and Residual deviance tells us that the model is a good fit. Greater the difference better the model. Null deviance is the value when only interception in the equation with no variables and Residual deviance is the value when take all the variables into account. It makes sense to consider the model good if that difference is big enough.
* The AIC provides a method for assessing the quality of the model through comparison of related models. Lower the AIC better the model. According to AIC the full model (AIC: 17314) with non-significant variables is better than the reduced model (AIC: 19251). But I do not choose first model since it has considerable number of non-significant variables.

### Create a Coefficient Plot

#create model for family income greater than or equal $150,000  
survey.model2 <- glm(FamilyIncome == 1~NumBedrooms+NumPeople+NumRooms+NumUnits+NumVehicles  
 +NumWorkers+NumWorkers+OwnRent+Insurance, data = survey.data, family = "binomial")  
  
#Create a coefficient plot for logistic regression on family income greater than $150,000   
coefplot(survey.model2,col.pts="red", intercept=TRUE)



* In the coefficient plot, we can see which coefficients are significantly different from zero and positive and negative impact for the dependent variable. According to coefficient plot of the model we can see home ownership (outright and rented), number of workers and number of units (single detached and single attached) have significant different from zero.

## Exercise 4: Predicting patient count for 12 months for a Dental Clinic

### Read the dataset of the 2010 American Community Survey (ACS) for New York state into R library

#Read the dataset into R library  
dental <- read.csv("BestSmileDental.csv")

### Examine the dataset and get the statistical summary of the data

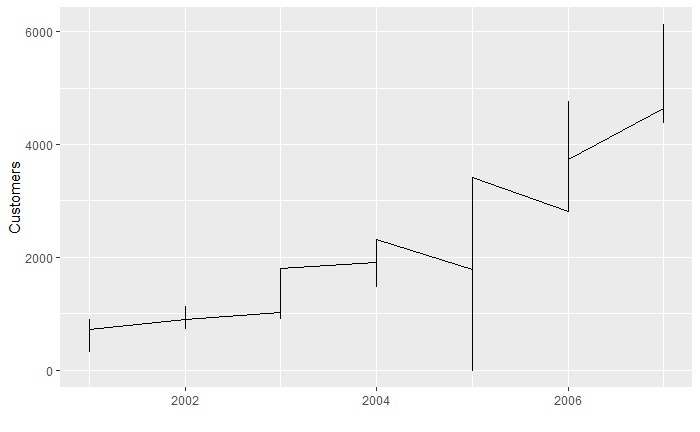
#examine dataset  
head(dental)

## Year Month Customers  
## 1 2001 1 416  
## 2 2001 2 329  
## 3 2001 3 750  
## 4 2001 4 904  
## 5 2001 5 794  
## 6 2001 6 485

#summary stat of the dataset  
summary(dental)

## Year Month Customers   
## Min. :2001 Min. : 1.00 Length:84   
## 1st Qu.:2002 1st Qu.: 3.75 Class :character   
## Median :2004 Median : 6.50 Mode :character   
## Mean :2004 Mean : 6.50   
## 3rd Qu.:2006 3rd Qu.: 9.25   
## Max. :2007 Max. :12.00

ggplot(dental, aes(Year, Customers)) + geom\_line() + ylab("Customers") +  
 xlab("")



The graph shows the irregular pattern of the data set because of the outliers, invalid or missing values. We need to clean the dataset before forecasting.

## Data cleaning

ImputeTS library (Imputation by using the Kalman filter) is used to clean the values of Customers variable, which has some missing and invalid values.

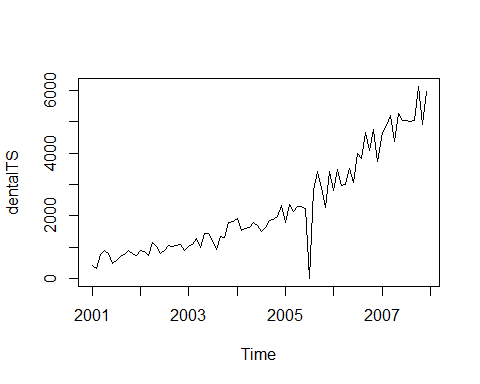
#convert non numeric to NA value  
dental$Customers <- as.integer(dental$Customers)  
  
#convert negative numbers to NA value  
dental$Customers[dental$Customers < 0] <- NA  
  
#imputation for missing/invalid data using ImputeTS  
new.dental <- dental %>% na.kalman

## Imputed values list

|  |  |  |
| --- | --- | --- |
| Year | Month | Customers (Imputed Values) |
| 2001 | 8 | 692.5879 |
| 2002 | 9 | 1015.7212 |
| 2002 | 10 | 1040.7576 |
| 2002 | 11 | 1066.4332 |
| 2004 | 2 | 1543.3404 |
| 2005 | 3 | 2140.4198 |
| 2005 | 7 | 0.0000 |
| 2006 | 8 | 3838.5003 |
| 2006 | 10 | 4083.1495 |
| 2007 | 6 | 5030.0054 |

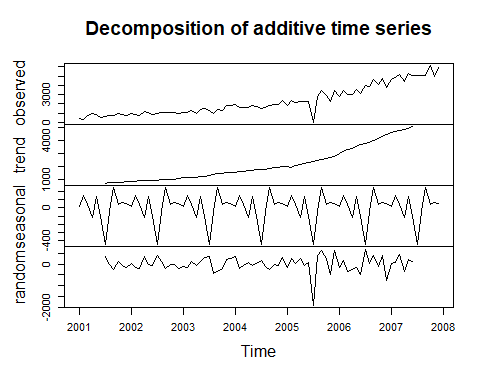
## Build Time Series

# build a time series from data  
dentalTS <- ts(new.dental$Customers, start = c(2001,1), frequency = 12)  
plot(dentalTS)



The “new.dental” is the univariate data which we are converting to time series. start gives the starting time of the data, in this case, its Jan 2001. As it is a monthly data so ‘frequency=12’. We can infer from the graph itself that the data points follow an overall upward trend with some outliers in terms of sudden lower values.

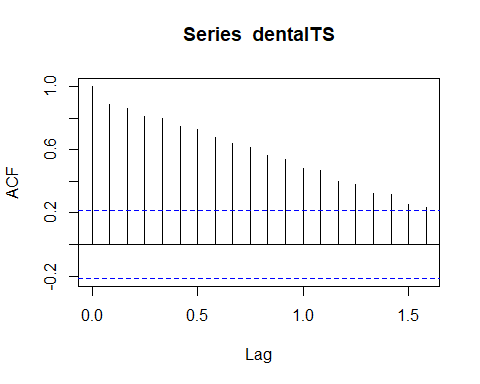
#find the components of the time series  
components.ts = decompose(dentalTS)  
plot(components.ts)



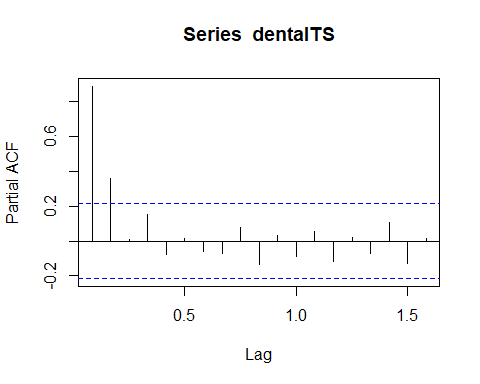
Here we get 4 components:

* Observed - the actual data plot
* Trend - the overall upward movement of the data points
* Seasonal - yearly pattern of the data points
* Random - unexplainable part of the data Observing these 4 graphs closely, we can find out if the data satisfies all the assumptions of ARIMA modeling, mainly, stationarity and seasonality.

# Assess the time series using ACF and PACF  
acf(dentalTS)



pacf(dentalTS)

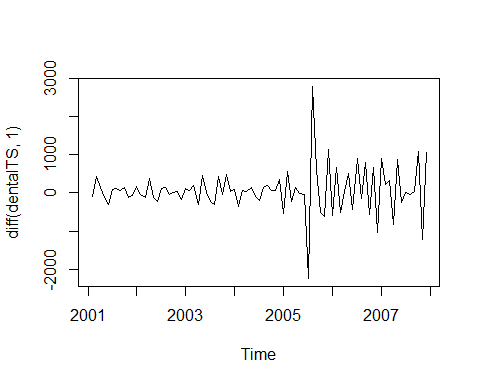


* The autocorrelation function (acf()) gives the autocorrelation at all possible lags. As we can infer from the graph above, the autocorrelation continues to decrease as the lag increases, confirming that there is no linear association between observations separated by larger lags.

# use diffing for data transformation. R can find optimal diffing  
ndiffs(x = dentalTS)

## [1] 1

# plot to see the effect of diffing  
plot(diff(dentalTS, 1))



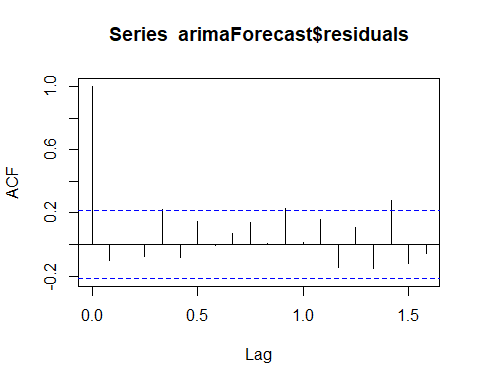
* ndiffs() function is used to fix the drifts of the time series we can see in the graph. Optimal number of diffins for the time series is 1. This creates more reliable time series for the forecasting.

## Build ARIMA time series model for forecasting

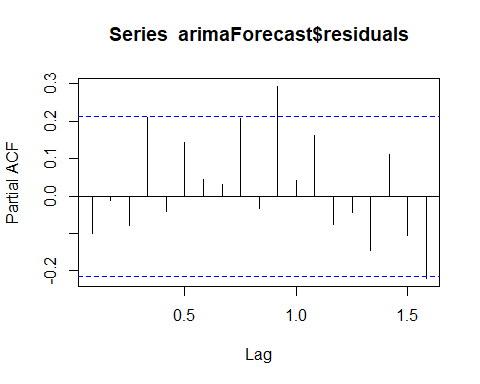
# fit the ARIMA model  
arimaForecast <- auto.arima(x=dentalTS)  
  
# review the ARIMA model  
arimaForecast

## Series:   
## ARIMA(0,1,1) with drift   
##   
## Coefficients:  
## ma1 drift  
## -0.7291 61.9169  
## s.e. 0.0589 14.0171  
##   
## sigma^2 estimated as 212819: log likelihood=-626.27  
## AIC=1258.54 AICc=1258.84 BIC=1265.79

# check the ACF and PACF of the ARIMA model residuals  
acf(arimaForecast$residuals)



pacf(arimaForecast$residuals)



# check the coefficients  
coef(arimaForecast)

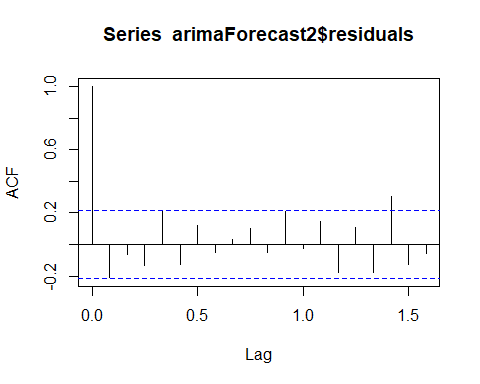
## ma1 drift   
## -0.7290858 61.9168539

The best fitting model R was found out is ARIMA(0,1,1). Let’s build 2nd ARIMA model with the p,d,q values that gives the best fitting time series model for the data set.

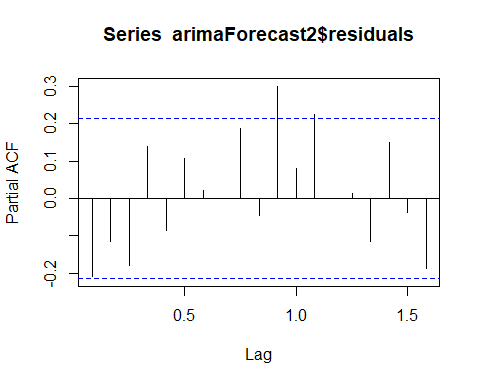
# try second ARIMA models - this time provide p,d,q values  
arimaForecast2 <- arima(dentalTS, order=c(0,1,1))  
  
# review the second ARIMA model  
arimaForecast2

##   
## Call:  
## arima(x = dentalTS, order = c(0, 1, 1))  
##   
## Coefficients:  
## ma1  
## -0.5956  
## s.e. 0.0688  
##   
## sigma^2 estimated as 240167: log likelihood = -632.14, aic = 1268.28

# check the ACF and PACF of the ARIMA model residuals  
acf(arimaForecast2$residuals)



pacf(arimaForecast2$residuals)



coef(arimaForecast2)

## ma1   
## -0.595624

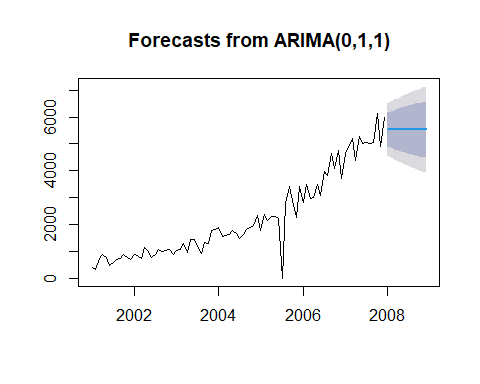
After examine different P,d,q values, I choose the values suggested by R in the previous model(0,1,1) by comparing AIC values of the residuals. From the two ARIMA models I will choose model 2 (“arimaForecast2” with p,d,q values), which has higher log-likelihood and lower AIC value than model 1 (arimaForecast).

### Forecast the Customer count for Year 2008 using best ARIMA model

#predict next five months using the second ARIMA model  
nextForecast <- forecast (arimaForecast2, h=12)  
# review the predictions from the second ARIMA model  
nextForecast

## Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  
## Jan 2008 5532.141 4904.093 6160.189 4571.625 6492.657  
## Feb 2008 5532.141 4854.687 6209.595 4496.065 6568.217  
## Mar 2008 5532.141 4808.648 6255.635 4425.653 6638.629  
## Apr 2008 5532.141 4765.367 6298.915 4359.461 6704.821  
## May 2008 5532.141 4724.402 6339.880 4296.811 6767.471  
## Jun 2008 5532.141 4685.417 6378.865 4237.189 6827.094  
## Jul 2008 5532.141 4648.150 6416.133 4180.193 6884.089  
## Aug 2008 5532.141 4612.391 6451.891 4125.505 6938.778  
## Sep 2008 5532.141 4577.971 6486.311 4072.864 6991.418  
## Oct 2008 5532.141 4544.751 6519.531 4022.058 7042.224  
## Nov 2008 5532.141 4512.612 6551.670 3972.906 7091.376  
## Dec 2008 5532.141 4481.456 6582.826 3925.258 7139.025

# plot the predictions from the second ARIMA model  
plot(nextForecast)



## Build Holt-Winters time series model for forecasting

# forecast using Holt-Winters Exponential Smoothing  
holtForecast.mean <- HoltWinters(dentalTS, gamma=FALSE)  
  
# check alpha and beta suggested by R  
holtForecast.mean

## Holt-Winters exponential smoothing with trend and without seasonal component.  
##   
## Call:  
## HoltWinters(x = dentalTS, gamma = FALSE)  
##   
## Smoothing parameters:  
## alpha: 0.09266244  
## beta : 0.8665188  
## gamma: FALSE  
##   
## Coefficients:  
## [,1]  
## a 5519.25732  
## b 68.58858

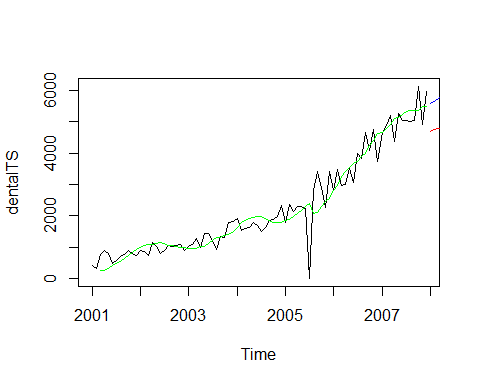
* R choose the alpha 0.09 and beta 0.87 for the time series.

### Forecast the Customer count for Year 2008 using Holt-Winters model

holtForecast.predict <- predict(holtForecast.mean, n.ahead = 12, prediction.interval = TRUE)  
holtForecast.predict

## fit upr lwr  
## Jan 2008 5587.846 6488.938 4686.754  
## Feb 2008 5656.434 6570.904 4741.965  
## Mar 2008 5725.023 6667.536 4782.510  
## Apr 2008 5793.612 6782.886 4804.337  
## May 2008 5862.200 6919.424 4804.976  
## Jun 2008 5930.789 7077.949 4783.629  
## Jul 2008 5999.377 7257.907 4740.848  
## Aug 2008 6067.966 7457.912 4678.020  
## Sep 2008 6136.555 7676.233 4596.876  
## Oct 2008 6205.143 7911.116 4499.170  
## Nov 2008 6273.732 8160.957 4386.506  
## Dec 2008 6342.320 8424.365 4260.275

# Plot data  
plot.ts(dentalTS)  
lines(holtForecast.mean$fitted[,1], col="green")  
lines(holtForecast.predict[,1], col="blue")  
lines(holtForecast.predict[,2], col="red")  
lines(holtForecast.predict[,3], col="red")



When comparing forecasting plots of best ARIMA model and Holt-Winters model, I can see Holt-Winters suggests better prediction than ARIMA model.

## forecast values of the customer/patient count for the 12 months of 2008 using Holt-Winters.

|  |  |
| --- | --- |
| Jan 2008 | 5587.846 |
| Feb 2008 | 5656.434 |
| Mar 2008 | 5725.023 |
| Apr 2008 | 5793.612 |
| May 2008 | 5862.200 |
| Jun 2008 | 5930.789 |
| Jul 2008 | 5999.377 |
| Aug 2008 | 6067.966 |
| Sep 2008 | 6136.555 |
| Oct 2008 | 6205.143 |
| Nov 2008 | 6273.732 |
| Dec 2008 | 6342.320 |