Assignment 2

Ong Jia Hui JONG119@e.ntu.edu.sg
G1903467L

1. Problem

This data is for monthly anti-diabetic drug sales in Australia from 1992 to 2008. Total monthly scripts for pharmaceutical products falling under ATC code A10, as recorded by the Australian Health Insurance Commission. Please build a good model to predict the drug sales.

1.1 Dataset

The drug.txt data file contains 204 rows of time series data as shown in listing 1.

```
> data = read.delim("./data/drug.txt", header=TRUE, sep = ",")
> head(data)
        date    value
1 1991-07-01 3.526591
2 1991-08-01 3.180891
3 1991-09-01 3.252221
4 1991-10-01 3.611003
5 1991-11-01 3.565869
6 1991-12-01 4.306371
> nrow(data)
[1] 204
```

Listing 1: Read data contents

2. Examine Dataset

This section describes the inital steps taken to find out the appropriate SARIMA model for the time series data described in Section 1.1.

2.1 Original Time Plot

The original time plot in Figure 1 appears to have a upward trending component as well as a seasonal pattern.

2.2 Seasonal Decomposition

Seasonal decomposition can be performed on the data using *SLT()*, which further indicates that there is a clear trend and seasonal pattern.

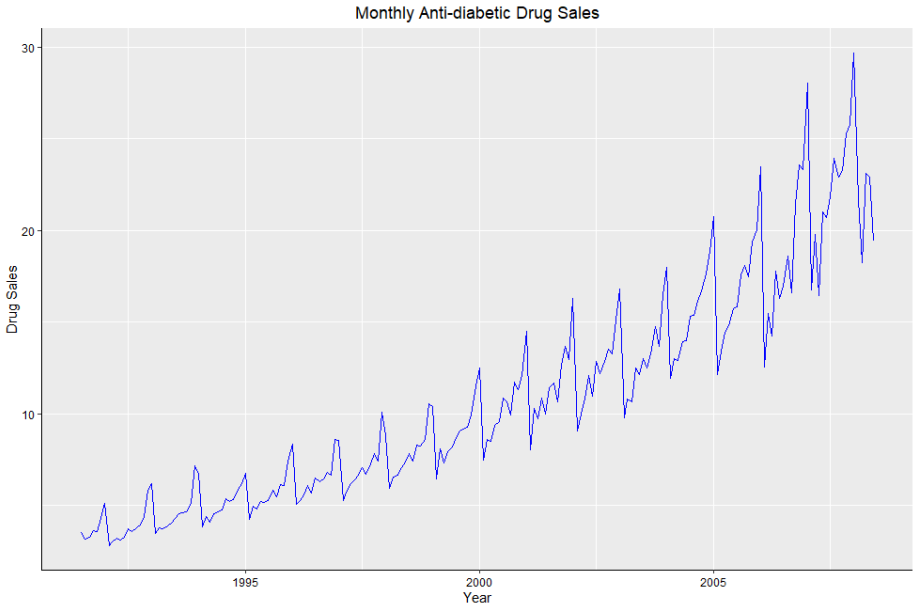


Figure 1: Time Series Plot of Original Data

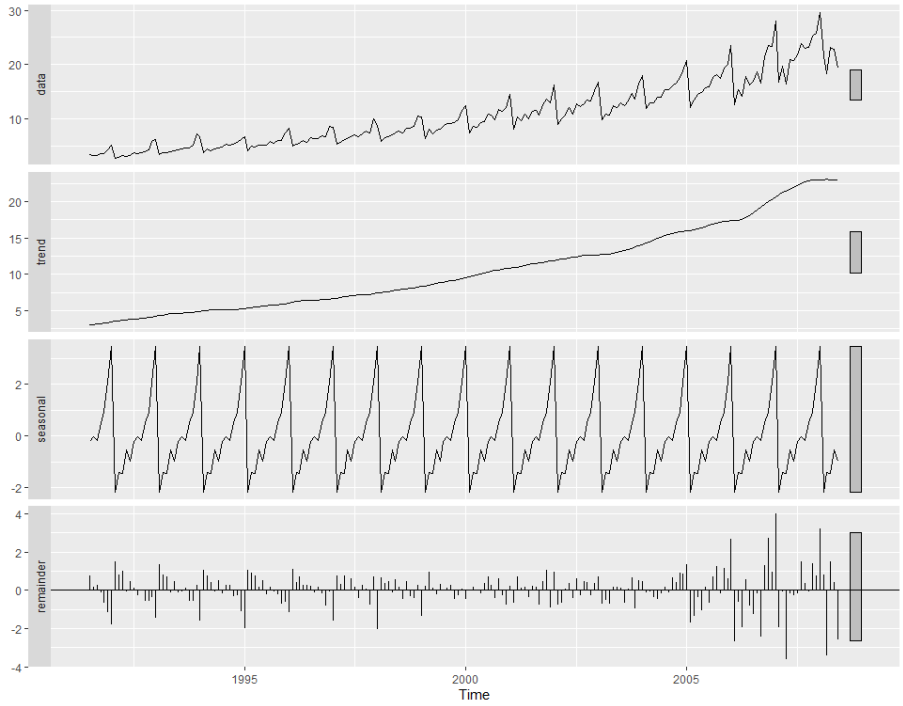


Figure 2: Seasonal Decomposition

2.3 ACF and PACF of Original Data

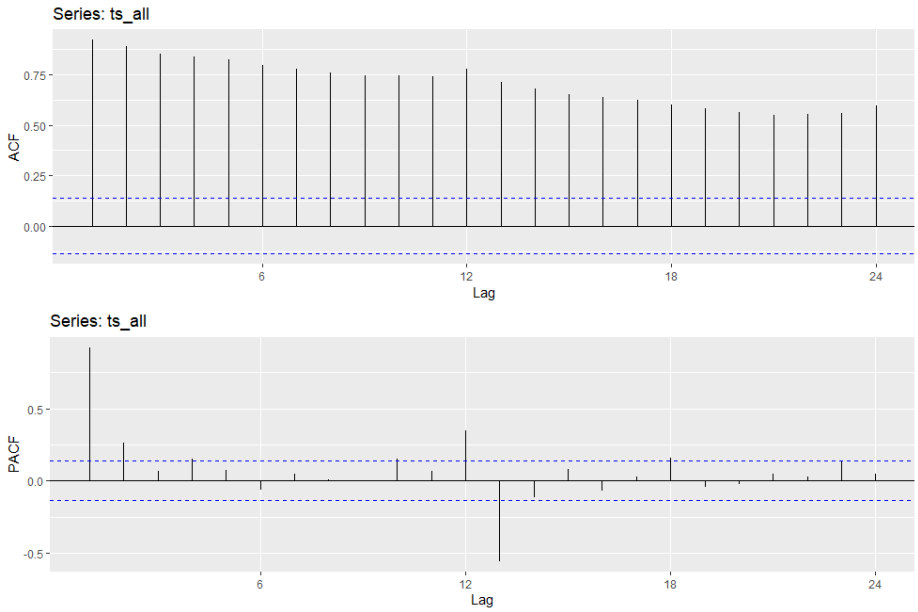


Figure 3: ACF and PACF Plots of Original Data

The ACF plot in Figure 3 also shows that the ACF dies down slowly, which indicate that the data is non-stationary.

3. Data Transform

3.1 BoxCox Transform

Box-Cox transformation is to stabilize the variance of the data. Both a zero lambda value and a non-zero lambda value were experimented in this report. For the non-zero lambda, the value 0.1313326 was obtained using *BoxCox.lambda()*. Using a zero lambda will mean a pure logarithmic function applied to the dataset.

3.2 Train-Test Split

The data of 204 observations are split into 80:20, with the training set to be used for fitting the model and the testing set to be used for calculating the accuracy of the model's forecasts. As a result of the split, the training set contains 163 observations, while the test set contains the remaining 41 observations.

3.3 Trend and Seasonal Differencing

As the data appears to have a trending component, we apply one time differencing to remove the trending component. $\mathbf{Z_t} = \mathbf{X_t} - \mathbf{X_{t-1}}$ using R function *d*iff(y, differences=1). After applying one time differencing, seasonal differencing was also applied by using the same R

function, but with another parameter, lag = 12. Figure 4 shows the effect of differencing, both the trending and seasonal pattern were reduced in the bottom plot. Hence, we have determined the d and D values as 1 to be used for the SARIMA models.

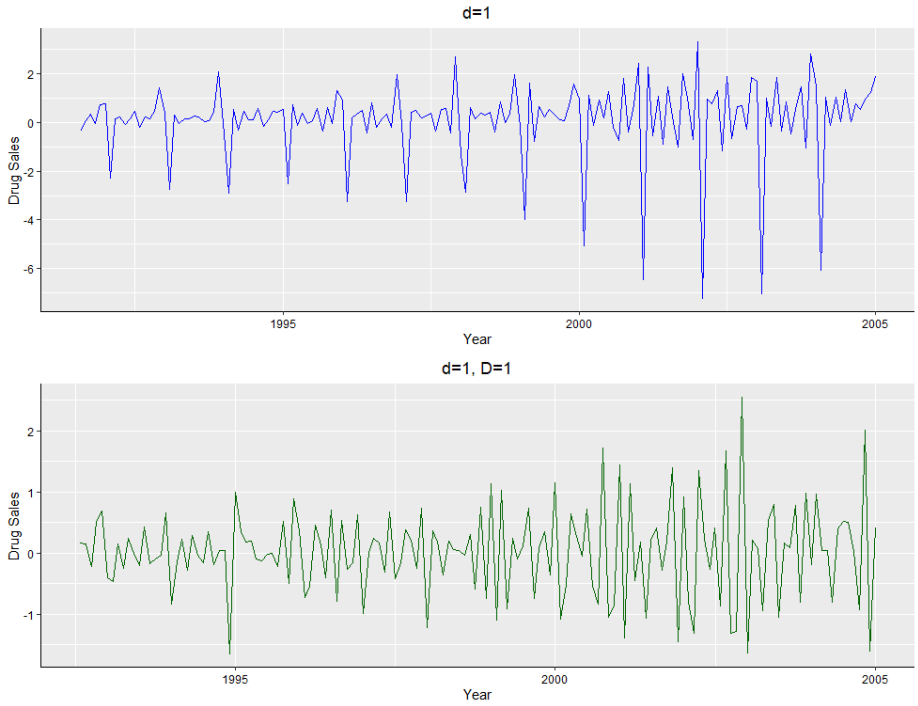


Figure 4: Time Plots after Trend and Trend & Seasonal Differencing

4. SARIMA Models

4.1 ACF and PACF Plots

Post-processing of data, the normalized set of train data was plotted to view their ACF and PACF as shown in Figure 5.

From the ACF and PACF plots, we can determine the following p, q, P and Q values as 5, 4, 0, 2 respectively. Along with d = 1 and D = 1, we can try to fit both ARMA, MA and AR models using the parameters.

4.2 Model Fitting

```
# Try ARMA(p,q) model
fit.arimaARMA = arima(ts_train_t, order=c(5,1,4),
                      seasonal=list(order=c(0,1,2), period=global.freq))
checkresiduals(fit.arimaARMA)
tsdiag(fit.arimaARMA)
summary(fit.arimaARMA)
pred.arimaARMA <- InvBoxCox(forecast(fit.arimaARMA, h = length(ts_test)+20)$
    mean, lambda = lambda)
```

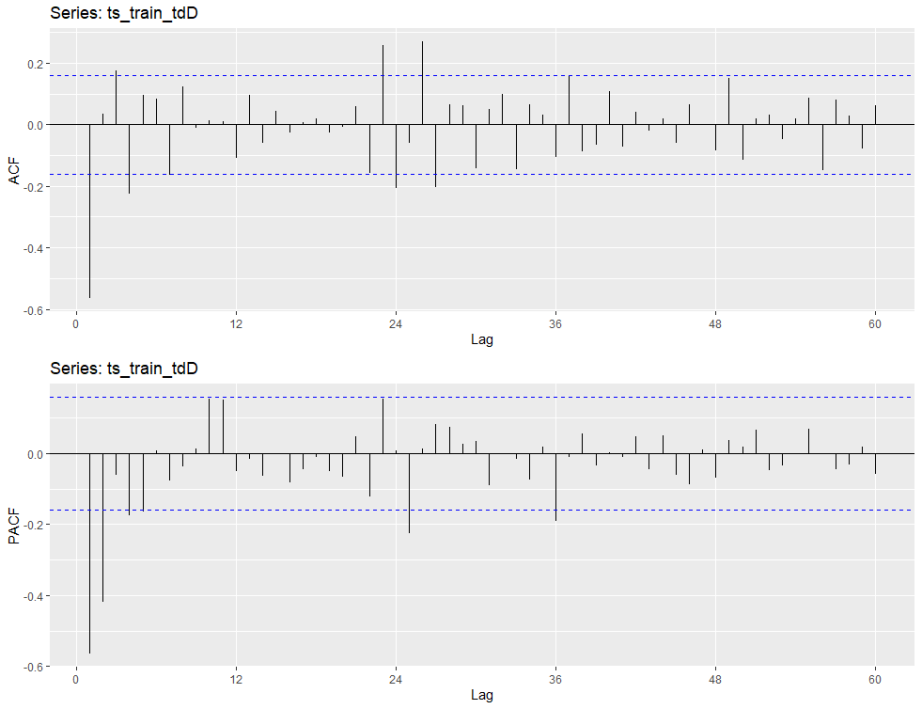


Figure 5: ACF and PACF Plots of transformed data

```
accuracy(pred.arimaARMA, ts_all)

# Try MA(q) model
fit.arimaMA = arima(ts_train_t, order=c(0,1,4),
                    seasonal=list(order=c(0,1,2), period=global.freq))
checkresiduals(fit.arimaMA)
tsdiag(fit.arimaMA)
summary(fit.arimaMA)
pred.arimaMA <- InvBoxCox(forecast(fit.arimaMA, h = length(ts_test)+20)$mean,
    lambda = lambda)
accuracy(pred.arimaMA, ts_all)

# Try AR(p) model
fit.arimaAR = arima(ts_train_t, order=c(5,1,0),
                    seasonal=list(order=c(0,1,0), period=global.freq))
checkresiduals(fit.arimaAR)
tsdiag(fit.arimaAR)
summary(fit.arimaAR)
pred.arimaAR <- InvBoxCox(forecast(fit.arimaAR, h = length(ts_test)+20)$mean,
    lambda = lambda)
accuracy(pred.arimaAR, ts_all)
```

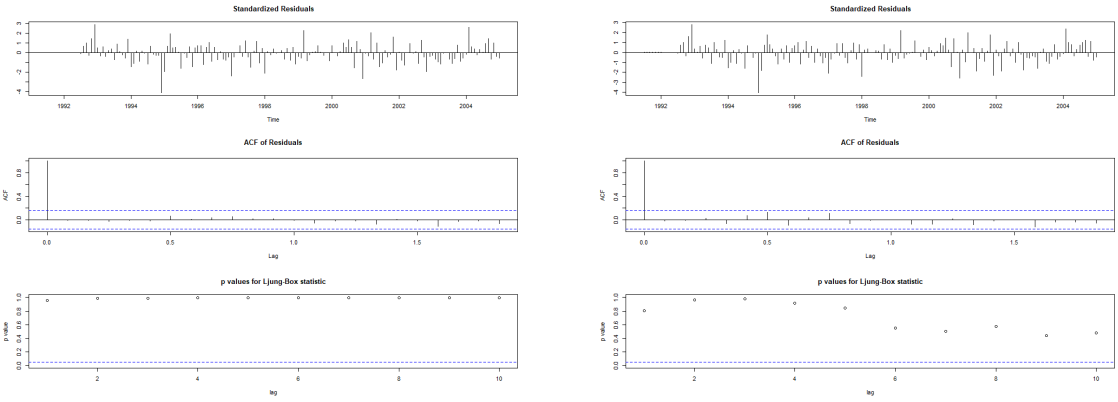
Listing 2: R Codes to fit SARIMA model

4.3 Diagnostic Checking

A *tsdiag()* diagnostic check as well were done on the three models. The ARMA and MA models are shown to be adequate fits as they have the following characteristics:

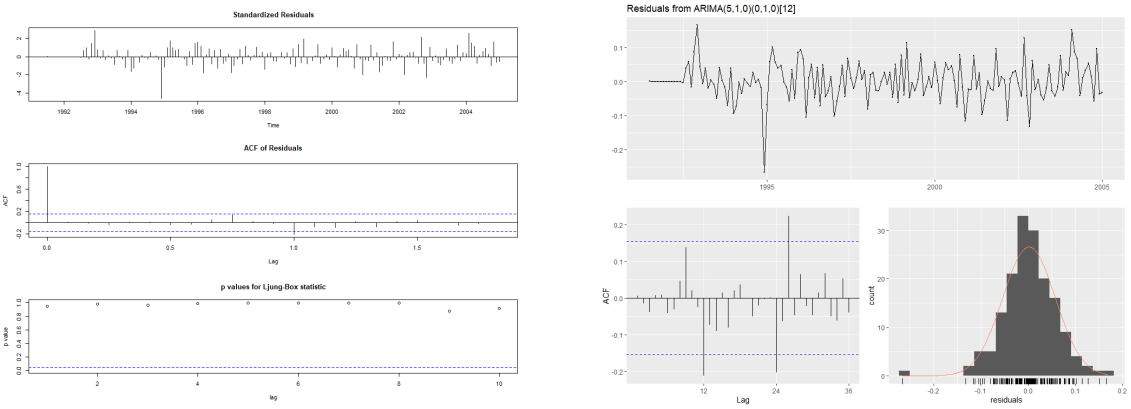
- The residuals looks to be random, which means it resembles white noise.
- The ACF of the residuals cuts off after lag 0.
- The p-values of Ljung-Box statistics are all above 0.05, therefore significant.

However, for the ACF diagnostic plot, the ACF of the residuals did not cut off after lag 0. Its residual plot using *checkresiduals()* also indicates that the AR model may not be a good fit.



(i) Diagnostic Check on ARMA model (ii) Diagnostic Check on MA model

Figure 6: Diagnostic Checking on ARMA and MA model



(i) Diagnostic Check on AR model (ii) Residual Check on Ar model

Figure 7: Diagnostic and Residual Checks on AR model

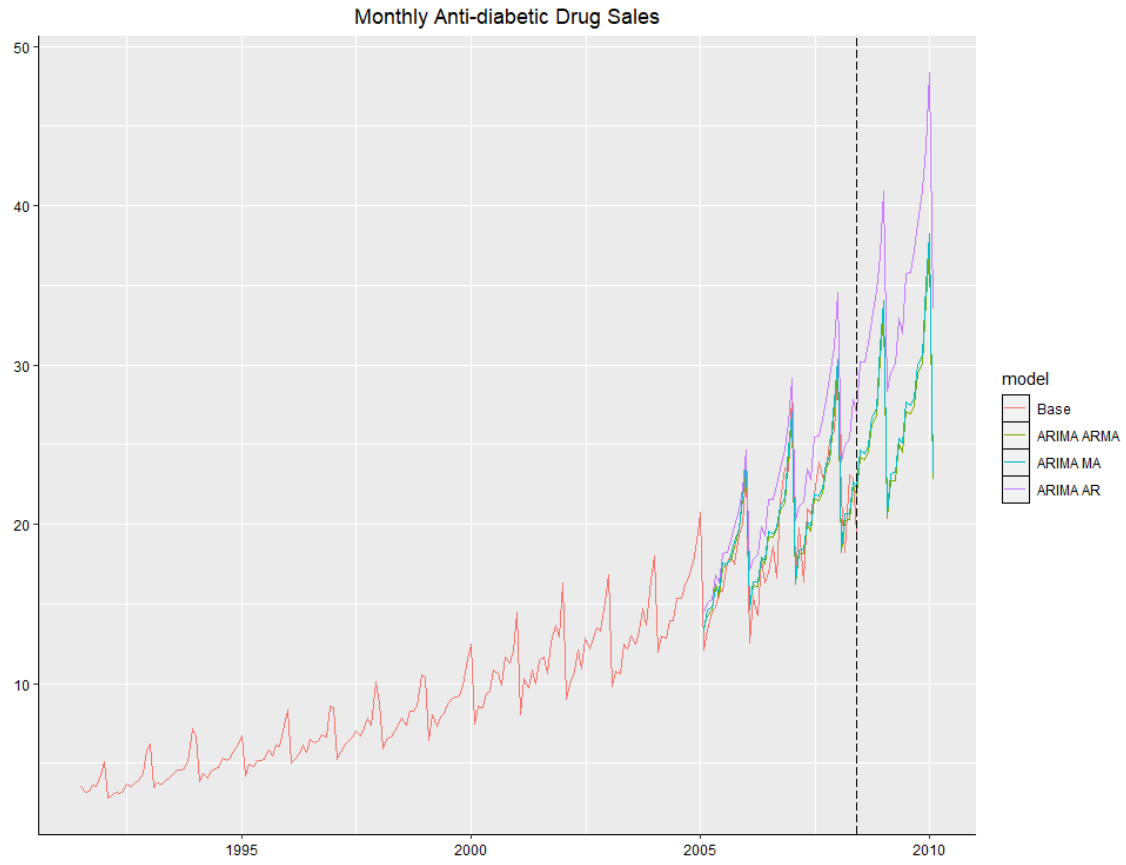


Figure 8: Prediction Plots for ARIMA Models

4.4 Forecasting and Results

Predictions were done for the test set duration + 20 future observations (total: 61 prediction). Figure 8 shows the plot of the predictions by the three models. The accuracy scores of each models are tabulated in Table 1.

Table 1: Accuracy of fitted SARIMA models

ARIMA	*AIC*	*Set*	*RMSE*	*MAE*	*MPE*	*MAPE*
ARMA	-441.25	Training	0.04691566	0.03377543	-0.1488159	1.669852
ARMA	-441.25	Test	1.500039	1.217112	-2.009964	6.565177
MA	-441.89	Training	0.04889734	0.03598504	-0.1005136	1.784139
MA	-441.89	Test	1.545632	1.257072	-3.231983	6.814599
AR	-416.29	Training	0.001626564	0.05547478	0.04048461	0.07318799
AR	-416.29	Test	3.400101	2.956103	-15.71631	15.71631

Table 1 revealed that the model with the lowest AIC was found to be the MA seasonal ARIMA model, followed by the ARMA model, then the AR model. However, ARMA model has slightly lower RMSE and MAE than the MA model in terms of prediction accuracy.

5. Alternative Models

5.1 Holt Winter Models

The Holt-Winters' Trend and Seasonality Model was also used to fit the model as the data is deemed to have both trend and seasonal. Similarly Box-Cox transformation was performed on the training set before modelling of the Holt-Winters' model with additive seasonal components and Holt-Winters' model with multiplicative seasonal components.

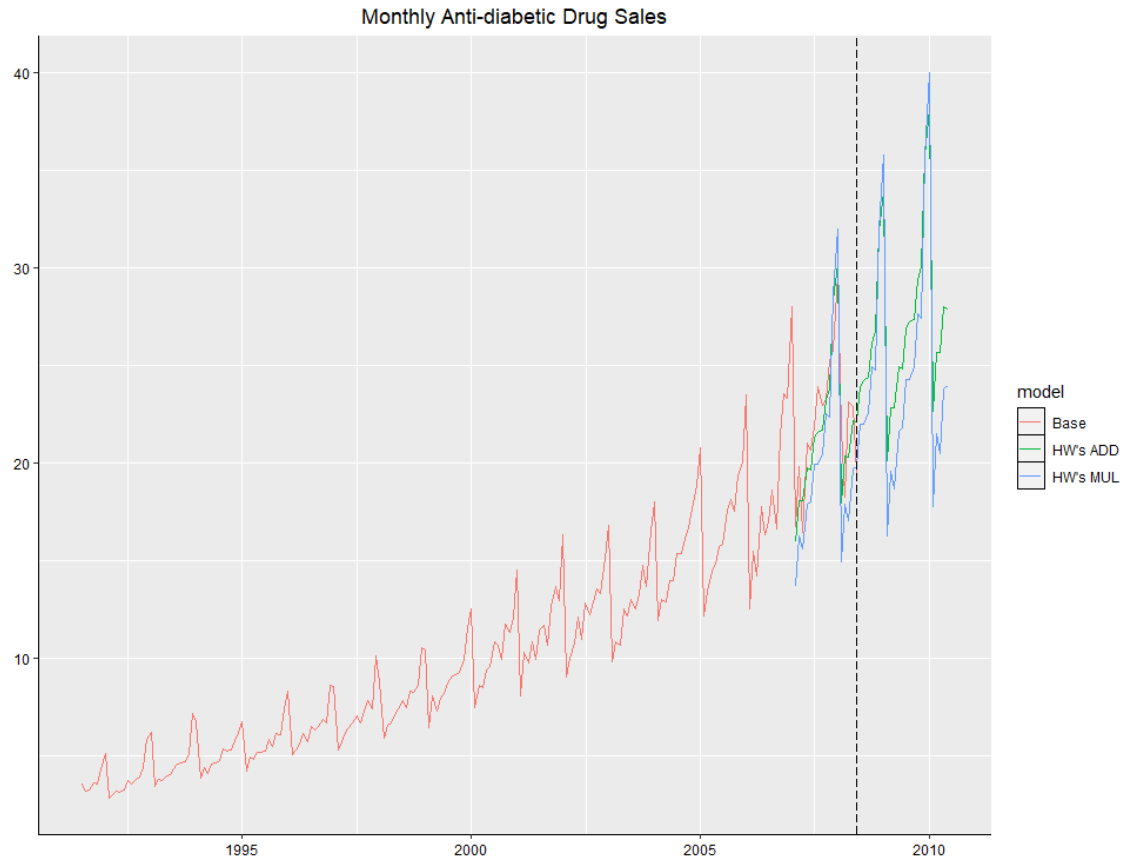


Figure 9: Prediction Plots for Holt-Winters' Models

5.2 Linear Regression Model

As the original timeplot in Figure 1 shows that the trend seems to be linear, a regression model is fitted on the dataset. The R function used is *tslm()*, which is used to fit linear

models to time series including trend and seasonality components. The appropriate lambda is passed into the function, which will perform Box-Cox transform on the data itself.

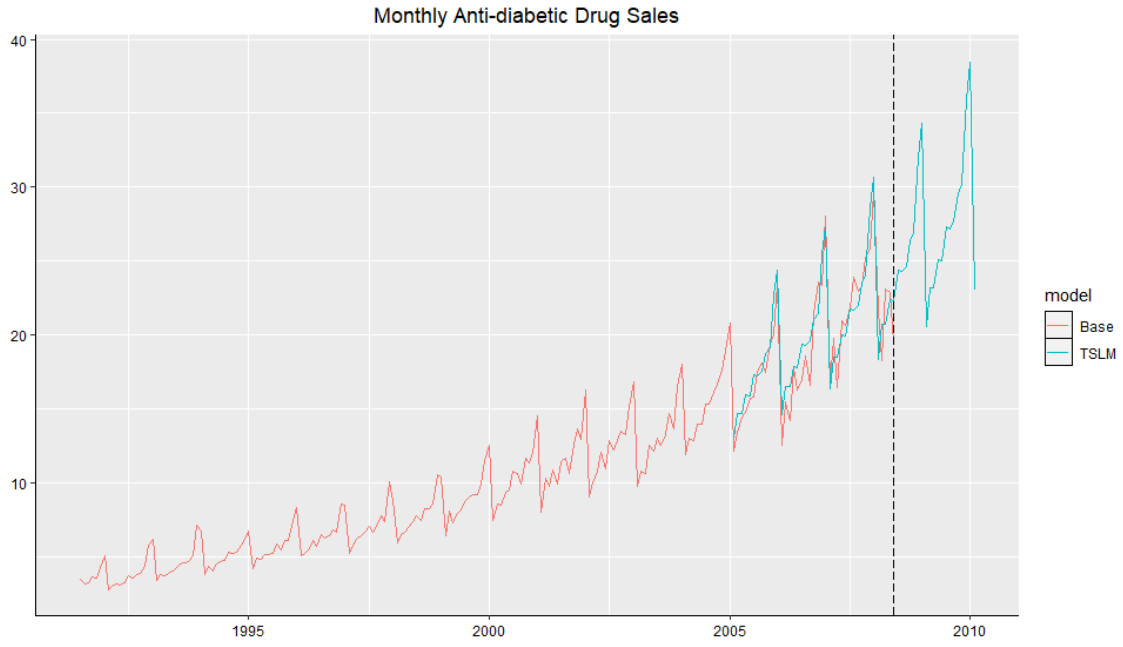


Figure 10: Prediction Plots for ARIMA Models

5.3 Alternative Model Forecast Accuracy Results

Table 2: Accuracy of fitted Alternative models

HoltWinter	*Set*	*RMSE*	*MAE*	*MPE*	*MAPE*
HW's Add	Test	1.857903	1.593702	1.979081	7.473986
HW's Mul	Test	3.258563	2.777955	10.14675	12.61482
TSLM	Test	1.633469	1.346868	-3.174507	7.197034

The additive Holt-Winters' model performs better with a lower RMSE, MAE, MPE and MAPE than the multiplicative model. The linear regression model outperforms both the Holt-Winters' models. However, the seasonal ARIMA ARMA and MA models still performs better than all of the alternative models in terms of forecasting.

6. Conclusion

The time series data was first plotted in a time plot to have a overview of the data. It was noticed that the data has a trending component and a seasonal pattern. The data was also deemed as non-stationary. In order to calculate the accuracy of the forecast models, the dataset was spilt into 80% training and 20% test set. To reduce the seasonal variance, Box-Cox transformation was applied on the training set. Trend and seasonal differencing were both performed on the training set. The output's ACF and PACF indicates that the suitable values of p, q, P, Q are 5, 4, 0, 2 respectively. It was found out that a seasonal ARMA(p,q) model has the lowest RSME value as its forecast accuracy. The diagnostic check performed on the model also deemed the model as an adequate fit.

Two alternative models were also experimented. They are the Holt-Winters' additive and multiplicative model as well as a linear regression model with trend and seasonality. The linear regression model fared better than the Holt-Winters' models, but still does not perform as well as the seasonal ARMA(p,q) model.

After experimenting with two lambdas (zero and non-zero), it is worth noting that that a zero lambda (log) transformation generally leads to models that performs better than the lambda suggested by the R function *BoxCox.lambda()* (0.13).