Project II – Drug Sales Data Analysis

Li Hantao, G2101725H, MSAI, hli038@e.ntu.edu.sg

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. Build a good model to predict the drug sales.

I. Original Data Analysis

This project asks us to fit an appropriate model for the drug sales dataset. First of all, we will look in the original dataset, trying to find some attributes about the time series data. The original plot of data is shown in Fig. 1. We can effortlessly know that the maximum and minimum of data are 29.7 and 2.8, with a mean value of 10.7.

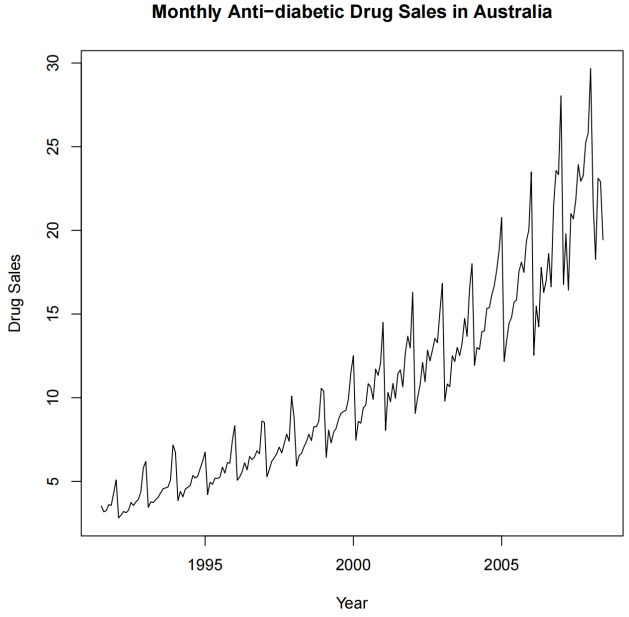

Fig. 1. Original Data Plot.

Observing the original data plot, we can find an apparent increasing trending component with a seasonal pattern. For the trend component, it may be non-linear. For the seasonality pattern, we can easily know that the period of it is 12 months by counting the peak in ten years. Firstly, we implement the seasonal decomposition on the original data to get the knowledge for more detailed information with the *slt()* function. The figure of them is shown in Fig. 2. Secondly, we can calculate the original data's ACF and PACF plots, shown in Fig. 3, which indicate that the data is non-stationary.

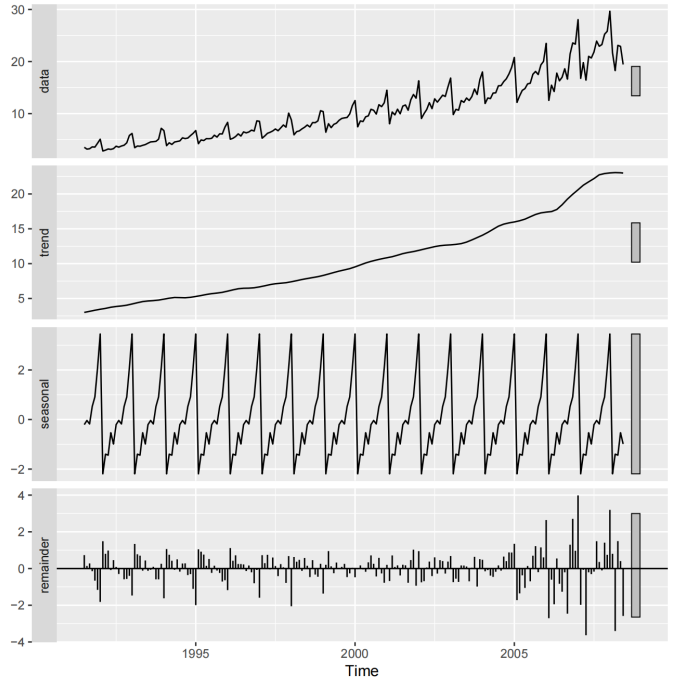

Fig. 2. Seasonal Decomposition on Original Data.

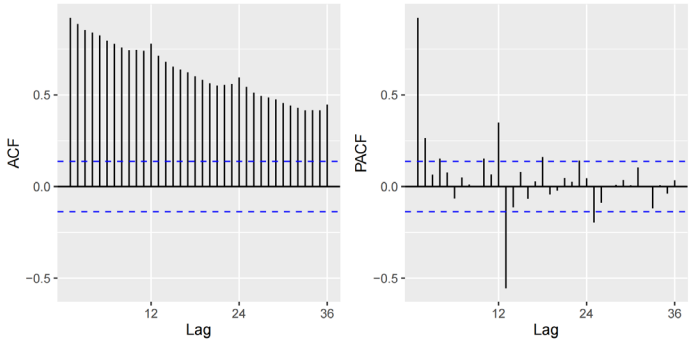

Fig. 3. ACF and PACF of Original Data.

From the data plot and remainder in Fig. 2, we can observe a noticeable increase in the seasonality variance. Thus, we should try to implement the Box-Cox transformation.

Before we begin the model selection, we first split the data into training set and validation set with the ratio of 4:1. After the splitting, there are 163 data points in the training set, while 41 data points in the validation set. We will not touch the validation data except the prediction results' comparison.

II. Transformation

Box-Cox transformation can stabilize the data with the fluctuation of variance. There is a hyper-parameter in Box-Cox, the lambda. This project will do the contrast experiment with a zero lambda and a non-zero lambda, 0.1313326, obtained by the function *BoxCox.lambda(data, 'guerrero')*. We try the zero lambda first, then re-implement the same model to the non-zero lambda situation. The data after the Box-Cox transformation is shown in Fig. 4. We can see the trend and remainder are more constant after logarithmic transformation.

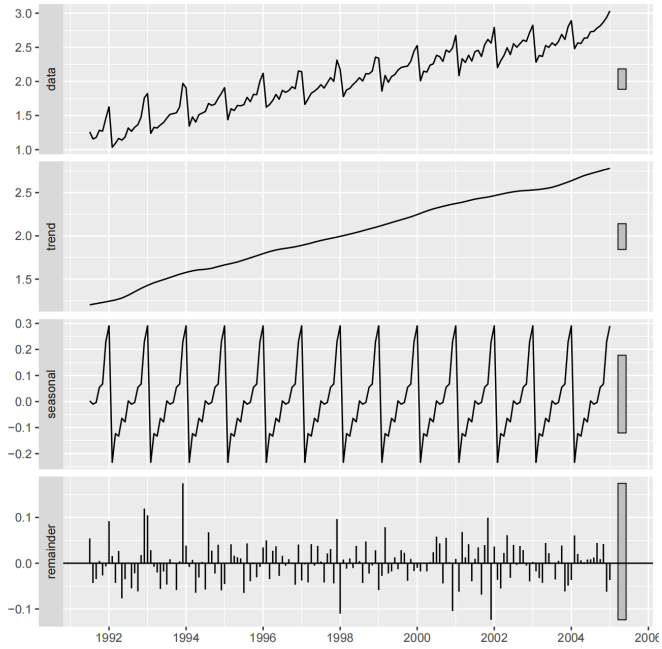

Fig. 4. Seasonal Decomposition on Data after Box-Cox.

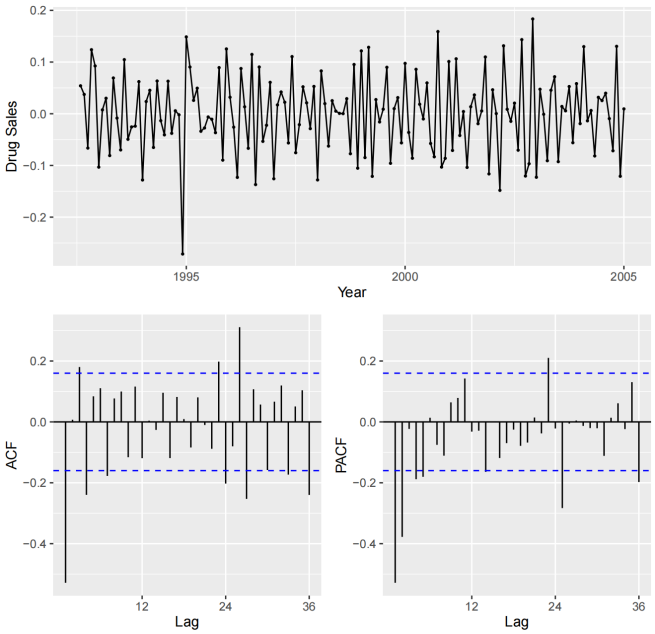

Fig. 5. Data after differencing.

After that, we utilize the differencing to make the data stable. In the observing before, we have known there are both trending components and seasonal differencing in the data.

Firstly, we try the first-order differencing on the data to remove the trending. Then, we try the lag-12 differencing to remove the seasonal component. Results are shown in Fig. 5.

III. SARIMA Model Fitting

From the analysis above and the information in the ACF plot and PACF plot in Fig. 5, we can determine appropriate parameters of the SARIMA model, which are p=5, d=1, q=4, P=0, D=1, Q=2. We will also try the AR and MA models in this project instead of only the ARMA model. Three models' parameters are:

1. SARIMA(5, 1, 4)(0, 1, 2)[12] (ARIMA model)
2. SARIMA(5, 1, 0)(0, 1, 0)[12] (AR model)
3. SARIMA(0, 1, 4)(0, 1, 2)[12] (MA model)

We can use the *tsdiag()* function and *checkresiduals()* function to implement the diagnostic check smoothly. Furthermore, we can plot the Normal Q-Q Plot of residuals. Fig. 6, 8-9 shows the results. (Due to the same sub-plot in the two different test functions, we rearrange the plot to delete the overlapping part and add the Q-Q plot in it.)

For these plots, we can implement a series of analyses similar to what we have done in the project I:

1. The residuals (standardized residuals) look random enough.
2. The histogram of residuals and Q-Q plot of residuals show that residuals are enough to be regarded as white noise.
3. The ARMA and MA model's ACF of the residuals cut off after lag 0. However, the AR model's ACF does not cut off after lag 0. Combined with the residuals checking, the AR model may not be a satisfactory model to fit the data.
4. The p-value for the Ljung-Box statistic is all high enough, which is high beyond 0.05, through the Box.test() function. The p-value of the MA model is apparently lower than the ARIMA model, even though still higher than 0.05. This may indicate that the MA model's performance will be lower than the ARIMA model.

We also calculate the AIC and BIC of these models, shown in Table I. Having concluded that those models may provide an adequate fit, we implement the forecasting to the following 20 data points and validation with the model. We can easily use the forecast() function to create the forecasting data, as shown in Fig. 10. The black curve is the forecasting result of the corresponding model, while the blue curve is the validation data (Ground truth). The validation length is 41, and the forecasting length is 20.

With the predicting result, we can calculate the RMSE and MAE of both the training set and validation set as two criteria to evaluate model fitting performance. The results are also shown in Table I. From the result, we can conclude that the ARMA model has the best performance in forecasting, while the AIC and BIC are only a little bit higher than the MA model. The forecasting results of the AR model are super bad, which coincides with our analysis above.

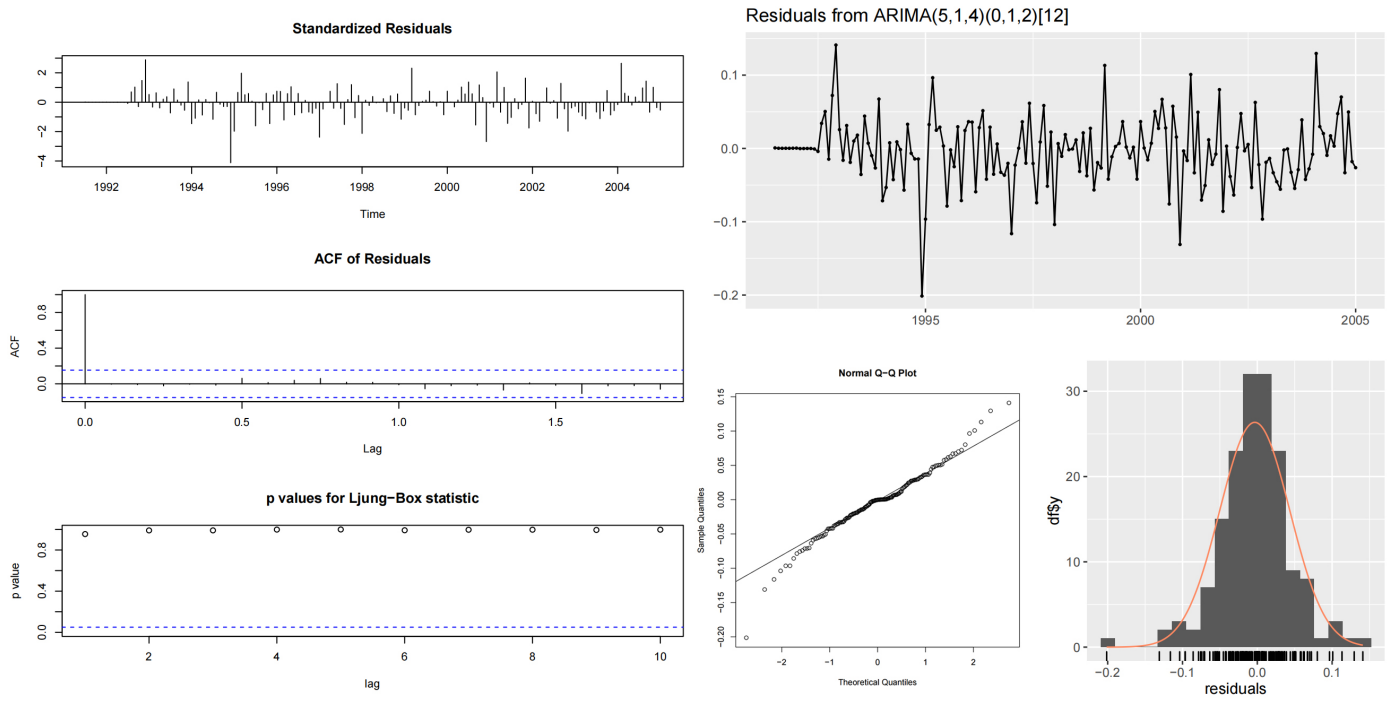

Fig .6. Diagnostic Check of SARIMA(5, 1, 4)(0, 1, 2)[12] (ARMA model)

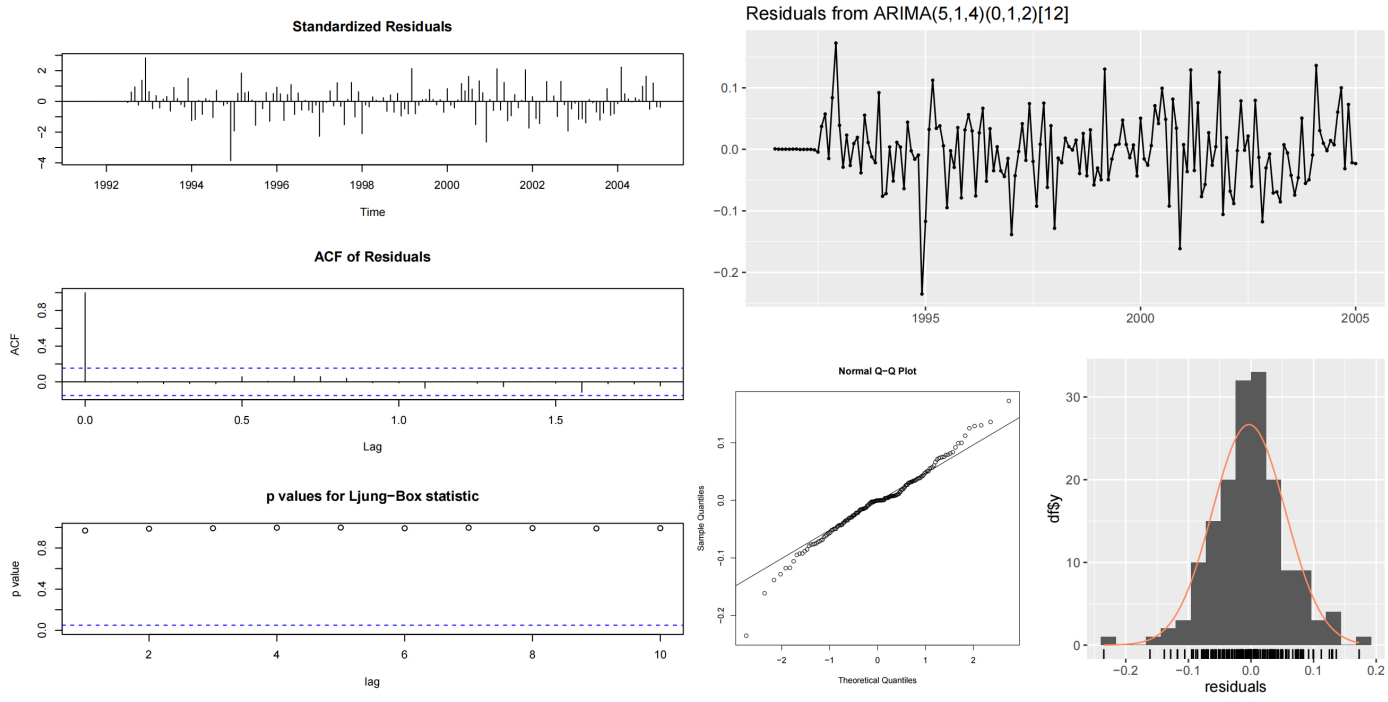

Fig .7. Diagnostic Check of (5, 1, 4)(0, 1, 2)[12] (ARMA model) with auto lambda.

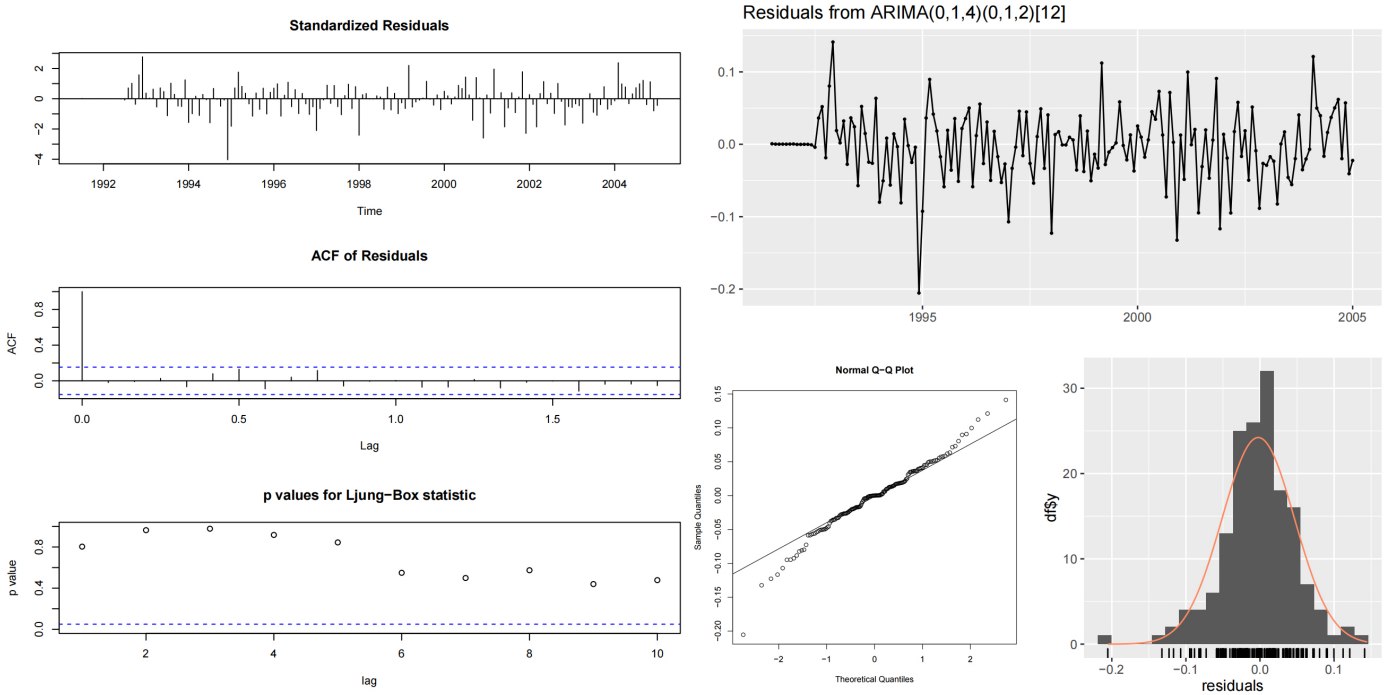

Fig .8. Diagnostic Check of SARIMA(0, 1, 4)(0, 1, 2)[12] (MA model)

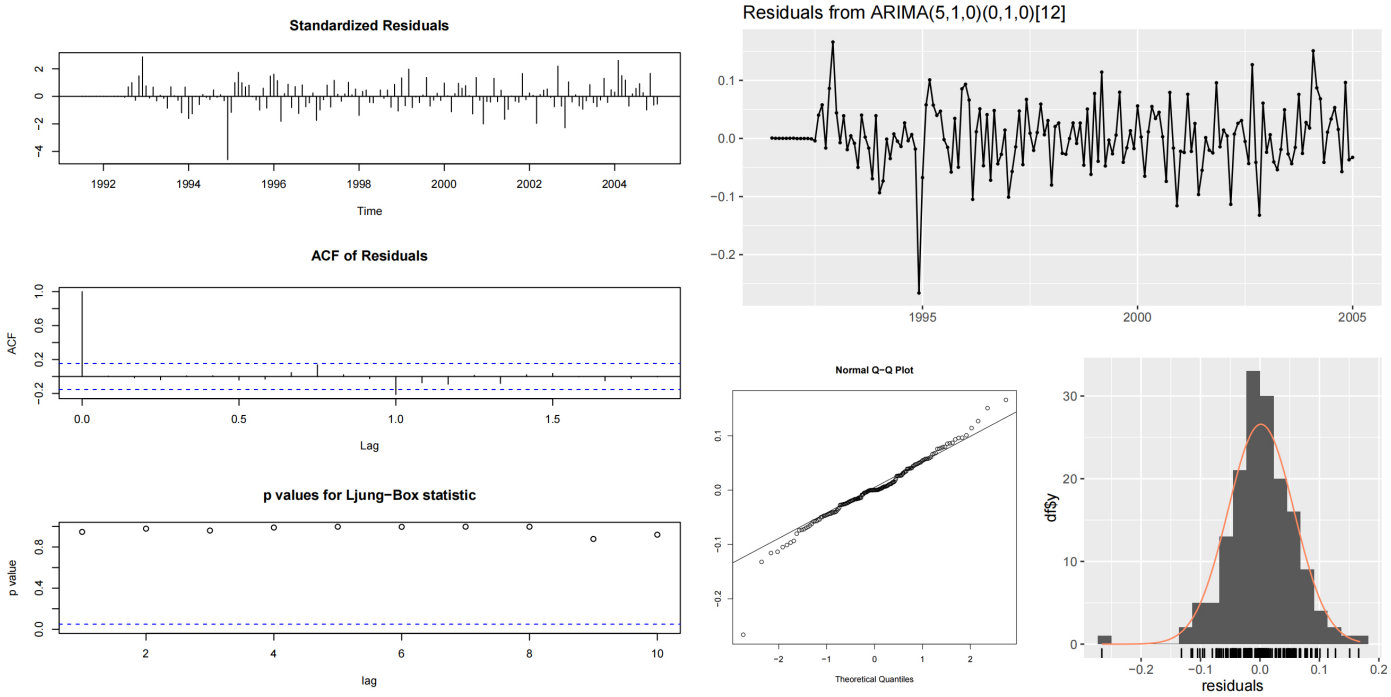

Fig .9. Diagnostic Check of SARIMA(5, 1, 0)(0, 1, 0)[12] (AR model)

TABLE I
ACCURACY

Set	Model	AIC	BIC	RMSE	MAE
Training	SARIMA(5, 1, 4)(0, 1, 2)[12]	-441.2474	-405.1198	**0.04691566**	**0.03377543**
	SARIMA(5, 1, 4)(0, 1, 2)[12] – with auto lambda	*-366.7601*	*-330.6324*	*0.05841829*	*0.04262249*
	SARIMA(5, 1, 0)(0, 1, 0)[12]	-416.2912	-398.2274	0.05547478	0.04048461
	SARIMA(0, 1, 4)(0, 1, 2)[12]	**-441.8861**	**-420.8116**	0.04889734	0.03598504
Test	SARIMA(5, 1, 4)(0, 1, 2)[12]	-	-	**1.500039**	**1.217112**
	SARIMA(5, 1, 4)(0, 1, 2)[12] – with auto lambda	-	-	1.664577	1.37995
	SARIMA(5, 1, 0)(0, 1, 0)[12]	-	-	*3.400101*	*2.956103*
	SARIMA(0, 1, 4)(0, 1, 2)[12]	-	-	1.545632	1.257072

The best result in each criterion of data is marked in bold and the worst result is marked in italics.

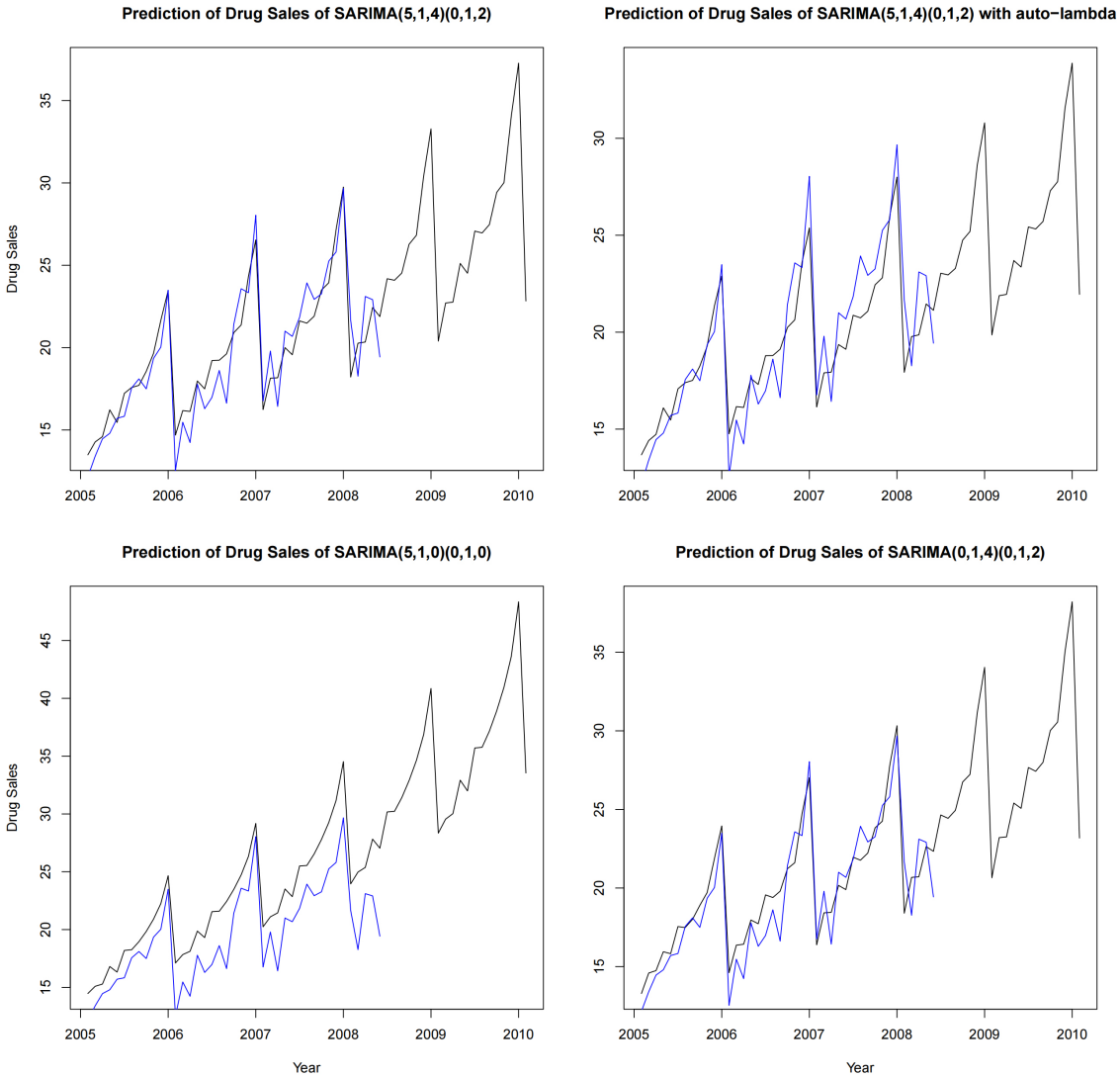

Fig .10. Predictions of different models.

Thus, in the three models with zero lambda, the ARMA model SARIMA(5, 1, 4)(0, 1, 2)[12] has the best forecasting performance.

After obtaining the best model with zero lambda Box-Cox transformation, we re-implement the same experiment on the data gained from non-zero lambda, 0.1313326, Box-Cox transformation. The results are shown in Fig. 7 and Table I. We can see that the performance of the non-zero lambda model is worse than the zero lambda model. Thus, in this project, we will use zero-lambda Box-Cox transformation.

IV. Brute-Force to Finding Parameters

In the previous discussion, we discussed three possible models. The parameters of these models are from the analysis of data characteristics. After analyzing these three models, we found that they effectively fit the existing data. However, we did not find a "best" model parameter. Therefore, this section will use the brute force method to traverse and search the model parameters and record their AIC, BIC, RSME, and MAE of the validation set.

Because we want to avoid the high complexity of the model, we limit the model parameters to less than six while only searching for the first-order differencing. The search results are listed in Table II and Table III.

From Table II, we can get a new pair of parameters (2, 1, 3) with the lowest AIC and BIC value. Even if them does not seems better in the RSME and MAE, we still try to test them and get the forecasting results, shown in Fig. 11.

TABLE II
AIC and BIC Value of Different Parameters' Models

AIC*		q						
		0	1	2	3	4	5	6
p	0	N.A.	443.47	442.03	440.94	441.88	441.60	440.22
	1	417.65	442.10	440.12	440.57	441.02	439.79	440.32
	2	437.09	440.27	439.11	**451.45**	445.54	443.56	442.69
	3	435.21	443.02	442.00	445.41	443.15	442.82	**446.33**
	4	437.27	442.56	440.69	443.48	441.58	440.97	439.09
	5	442.51	440.99	439.86	443.24	441.24	439.15	436.10
	6	440.52	442.16	440.41	**446.37**	439.89	445.46	438.96

BIC		q						
		0	1	2	3	4	5	6
p	0	N.A.	**431.43**	426.97	422.87	420.81	417.51	413.13
	1	405.61	**427.05**	422.06	419.50	416.93	412.70	410.22
	2	422.04	422.20	418.04	**427.37**	418.44	413.46	409.58
	3	417.15	421.95	417.92	418.31	413.04	409.70	410.20
	4	416.19	418.48	413.59	413.38	408.46	404.85	399.96
	5	418.43	413.90	409.75	410.12	405.11	400.01	393.95
	6	413.42	412.05	407.29	410.25	400.75	403.31	393.80

The three most minor (best) values in each table are marked in bold.
*AIC and BIC are all negtive, we omit the '-' in the table.

TABLE III
RMSE and MAE of Different Parameters' Models

RMSE		q						
		0	1	2	3	4	5	6
p	0	N.A.	1.6427	1.6493	1.6203	1.5456	1.6584	1.5741
	1	1.5736	1.6482	1.6475	1.5644	1.5659	1.6439	1.5956
	2	1.5337	1.6368	1.6523	1.6820	1.5651	1.5700	1.5371
	3	1.5373	1.5186	1.5153	1.5540	**1.5034**	1.5540	1.6401
	4	1.5983	1.6367	1.6677	1.5697	1.5700	1.5518	1.5599
	5	1.6560	1.6796	1.6706	**1.5002**	**1.5000**	1.5411	1.5510
	6	1.6583	1.5830	1.5626	1.7069	1.5378	1.5677	1.5822

MAE		q						
		0	1	2	3	4	5	6
p	0	N.A.	1.3326	1.3397	1.3143	1.2570	1.3480	1.2744
	1	1.2843	1.3389	1.3383	1.2744	1.2701	1.3362	1.3001
	2	1.2447	1.3286	1.3419	1.3934	1.2740	1.2777	1.2491
	3	1.2478	1.2328	1.2292	1.2649	**1.2263**	1.2663	1.3604
	4	1.2934	1.3287	1.3554	1.2783	1.2774	1.2632	1.2689
	5	1.3511	1.3672	1.3596	**1.2173**	**1.2171**	1.2570	1.2617
	6	1.3530	1.2891	1.2716	1.4047	1.2517	1.2841	1.2961

The three most minor (best) values in each table are marked in bold.

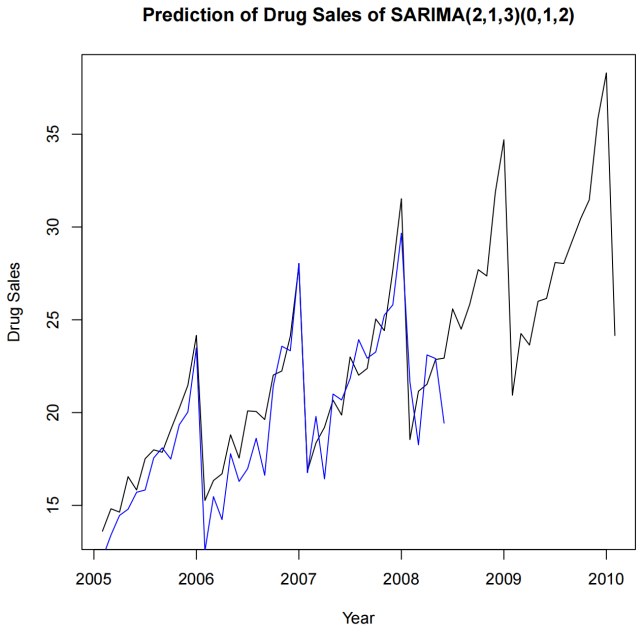

Fig. 11. Predictions of SARIMA(2,1,3)(0,1,2)[12].

In the above sections, we select two models. On the one hand, SARIMA(5,1,4)(0,1,2)[12] has the best performance with the lowest RSME and MAE when testing the forecasting result on the validation set; on the other hand, SARIMA(5,1,4)(0,1,2)[12] has the best AIC and BIC values which also shows it may be a good choice to fit the data.

V. Holt-Winter Model

When the data have both trending and seasonal components, the Holt-Winter model may have adequate performance in fitting and forecasting it. This section tries to fit the data with the Holt-Winter model with additive seasonal components and multiplicative seasonal components after implementing Box-Cox transforming with zero-lambda.

We can easily utilize the *hw()* function to fitting the data. The prediction results are shown in Fig. 12.

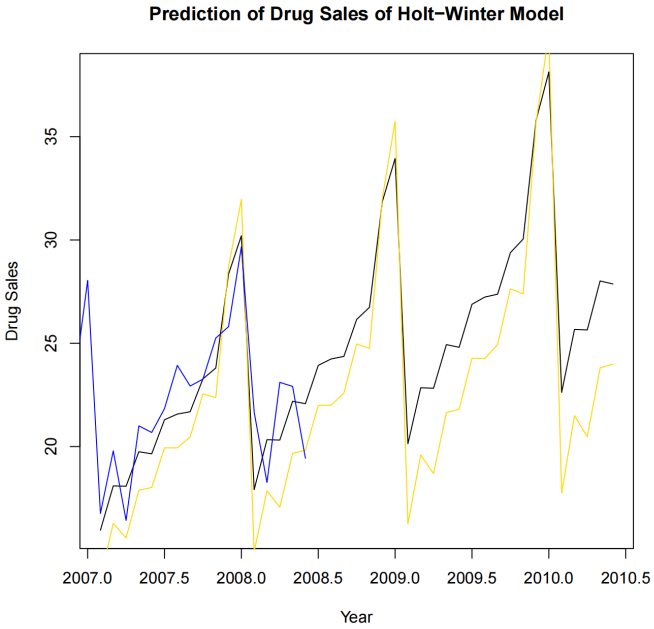

Fig. 12. Predictions of Holt-Winter Models.

In Fig. 12, the black curve is the prediction of additive seasonal components Holt-Winter model, the gold curve is the prediction of multiplicative seasonal components Holt-Winter model, and the blue curve is the validation data (GT). We can notice that the additive model fits better than the multiplicative model.

We also calculate the RMSE and MAE of these two models, shown in Table IV.

TABLE IV
Accuracy

Set	Model	RMSE	MAE
Test	Additive Holt-Winter	1.857903	1.593702
	Multiplicative Holt-Winter	*3.258563*	*2.777955*
	SARIMA(5, 1, 4)(0, 1, 2)[12]	**1.500039**	**1.217112**

The best result in each criterion of data is marked in bold and the worst result is marked in italics.

It can be known in Table IV that the additive model fits better than the multiplicative model of Holt-Winter, which coincides with our analysis in Fig. 12. Moreover, both of the Holt-Winter perform worse than the SARIMA model we selected in the above sections, which means the best model is still the SARIMA model in this project report.