2016

FORECAST ELECTRICITY RETAIL SALES IN THE US BY END-USE SECTORS

Jing Han
Michigan Technological University, jinghan@mtu.edu

FORECAST ELECTRICITY RETAIL SALES IN THE US BY END-USE SECTORS

By

Jing Han

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Applied Natural Resource Economics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2016

This thesis has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Applied Natural Resource Economics.

School of Business and Economics

Thesis Advisor: *Dr. Latika G. Lagalo*

Committee Member: *Dr. Gary A. Campbell*

Committee Member: *Dr. Yeonwoo Rho*

School Dean: *Dr. Gene Klippel*

Contents

List of Figures

List of Tables

Abstract

This paper forecasts electricity retail sales using monthly data by sectors from January 2001 through December 2014 and compares the results to the actual data from January 2015 to April 2015. This forecasting shows electricity sales have a significant seasonal pattern. Three models are developed to capture this pattern and all of them are proved to be appropriate for cyclical data. These three models are the model of regression with dummy variables and ARMA disturbances, the autoregressive distributed lags model, and seasonal difference model. AutoRegressive Distributed Lag model helps us know how current and lagged values of average retail sales price, population, and the Industrial Production Index affect the current retail sales data.

1. Introduction

Electricity is a secondary form of energy, not like other primary energies such as coal, natural gas which are directly extracted from nature, it requires complex processes to convert other energy resources into the kind of energy we need. Unlike other energy resources, electricity is not able to be storage under technological concern though a tiny part can be saved by batteries with high cost. Demand and supply unbalance will cause damage voltage surges or drops that lead to system failure in seconds, thus supply and demand must match all the time. This requires operating plants must be able to modify their output by bring on-line or take off-lines to meet the demand needs. Consequently, it becomes necessary for us to forecast the future electricity demand that people in electricity industry are able to make marketing decisions in advance.

This study uses the monthly electricity sales data. Because it is technically required that the demand and supply must match all the time, we are able to forecast the electricity demand from using retail sales data from supply side so. The electricity retail sales data behaves the strong seasonal pattern so that we aim to capture this seasonality and other trends. In order to forecast the retail sales of electricity, we develops three time series models incorporate dummy variables, current and lagged value of average retail price of electricity, population, and Industrial Production index. The study also works on different end-use sectors and the result proves that this approach is necessary. The forecast results are also compared to actual values for each sector.

1.1. Brief history of electrical utility industry

It wasn't until Benjamin Franklin's famous kite experiment in 1752, people were aware of that lightning was electricity. His discover then encouraged those science talents to start their own research on electricity. The invention of electric motor by Michael Farady in 1821 as long as the mathematical analysis of electrical circuit by Georg Ohm in 1827 are regarded as the beginning of modern age of electricity. Farady's discover established the principle of electricity generation which made large scale electricity generation possible and his principle was also used in commercial area in the later 19th century (Davidson, 2003). Demand of electricity increased rapidly after Thomas Edison developed the incandescent light bulb in 1879. Three years later Edison started the first private generation facility using his direct current (DC) system in New York City which provided service to the Wall Street area. However, the DC system has its own technical limitation, the generation had to be very close by the end users. The General Electric switched their system to Nikola Tesla's alternating current (AC) system several years later and it dominated the market rapidly as this new system made transform more efficient and made long distance distribution possible with very low losses. It dominant the market rapidly. Tesla later sold his patent to Westinghouse Electric Company which became a main competitor of General Electric. The competition between DC and AC system ended up with AC's overwhelming. By 1900's, electricity had been widely used in both the urban and industrialized area of the US. The Rural Electrification Administration (REA) created by President Franklin Roosevelt in 1935 provided loan funds for farmers who live in rural areas to install expensive line lines.

1.2. Three basic segments of electricity industrial

The electricity industry is a complex system which can be divided into three basic segments: power generation, transmission, and distribution. Because of the high fix cost of transmission line construction and maintenance, it is impossible for those competing companies to build different systems. Consequently, the electricity industry tends to be natural monopolies. There are several regulations were established by the government in order to prevent electric companies from taking advantage of their marketing powers. Power plants are mostly fueled not only by coal, oil, natural gas, hydro, nuclear, but also can by wind, solar or geothermal. They are operated by electricity generators. Transmission lines carry relatively high voltages electricity energy over long distance, from one region to another region. A distribution system is sometimes called retailers deliver electricity power directly to customer at the require voltage.

1.3. Four end-use sectors and cyclical demand of electricity

Figure 1-1 shows retail sales of electricity in United States by sector from 2000 to 2014. It also shows that relatively similar amounts of electricity are used by residential, commercial, and industrial end use sector, but a very small amount is consumed in transportation sector. There was a large drop in consumption in industrial sector between 2008 and 2010 and it was due to the economic downturn from late 2007 through 2009. The electricity consumption in the US is expected to grow slowly because of complex reasons such as slower population growth, efficiency improvements. It is projected that total U.S. electricity use grows by an average of 0.8% per year from 2013 to 2040.

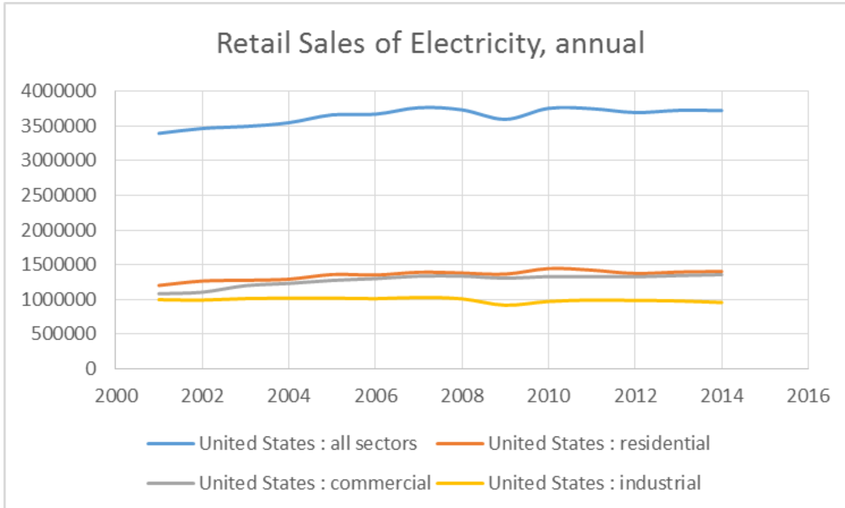

Figure 1-1Retail sales of electricity by sectors (million kilowatthours), annual 2000-2014

It is easy to conclude the seasonal cyclical demand of electricity through Figure 1-2. The retail sales shows several peaks and valleys. This periodical pattern can exhibit daily, weekly, quarterly and seasonally. For residential sectors, demand starts grow around 6 a.m. as people get up and peaks around noon then declines until a secondary peak in the late afternoon. The demand drops off in the late evening as people start to go to bed. This pattern exists everyday but is somewhat different on weekends. The consumption on weekdays is higher than weekend consumption. This is because most the business buildings have limited activity on weekend especially on Sunday. In addition, lighting is the largest use of electricity in the commercial sector. Machine drives are the largest use of electricity in industrial sector.

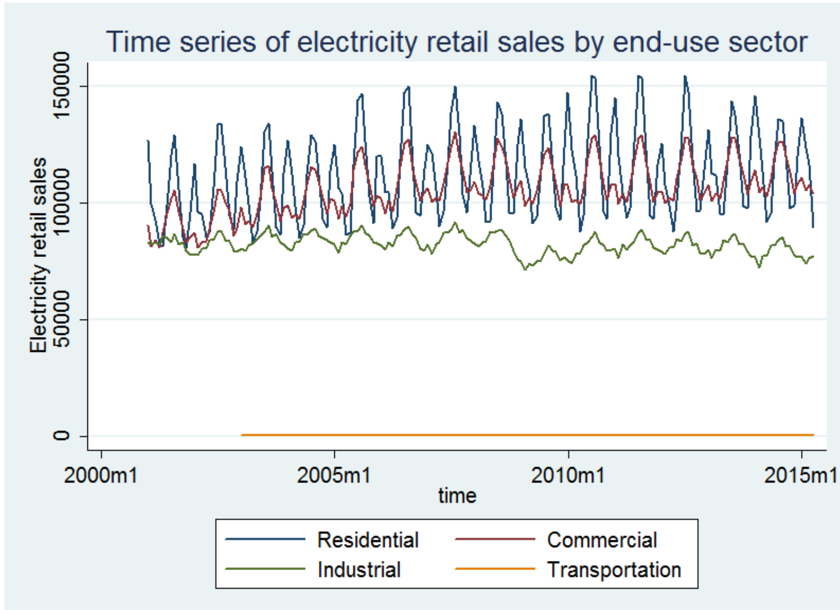

Figure 1-2 Retail sales of electricity by sectors, monthly

All three main end use sectors, residential, commercial and industrial response to differently to seasonal changes in weather. Figure 1-2 also shows that the residential sector has the largest seasonal variance, with two significant peaks in every summer and winter. This is due to households in the US use electricity for air conditioning and lighting. The commercial sector has less variance with two spikes in summer and winter. Except for lighting, the major uses of electricity in this sector are for cooling, powering office equipment, ventilation and space heating. The industrial sector is relatively flat because the majority of electricity is used to power machines and only a small portion of the energy use for heating and cooling in this sector.

2. Literature Review

Many models have been developed for forecasting electricity demand. Beenstock, Goldin and Nabot (1999) used quarterly data for Israel to develop three econometrics models by households and industrial companies. Three methods are Dynamic Regression

Model, OLS, and Maximum likelihood. They find the price elasticities in each model are similar. Hondroyiannis (2004) examines the monthly demand of electricity by residential sector in Greek. This study finds out that in the long-run, the residential demand is related to residential income, price level and wether condition. Bentzen and Engsted (1993) also estimated demand elasticity using Danish annual data and found that in long-run, the price elasticity is typically small. In 2001 they developed a standard autoregressive distributed lag model and error-connection models to estimate energy demand relationships and they get similar results by fitting these two models using Danish residential consumption data. This research also concluded that the model is still appropriate if explain variables are not stationary series. A number of models studied various economic variables that affect energy demand such as population (Egelioglu, 2001), price (Harris, 1993), temperature (Yan, 1998, Mirasgedis et al., 2007), GDP (Fung, 1993). Fung (1993) found that electricity price, gross domestic product, deflated export exports, and population are correlated with electricity consumption in Hong Kong. Holt provides a model for forecasting seasonal and trends which is named exponential weighted moving averages model. Erdogdu (2007) reports that in Turkey, both income elasticity and price elasticity are quite limited by using cointegration and autoregressive integrated moving average (ARIMA) model.

Mohamed and Bodger (2005) forecast electricity consumption in New Zealand using GDP, price and population as independent variables. However, they used annual data as much as 40 observations so they failed to take seasonality of electricity consumption into consideration. This paper will forecast the electricity consumption in the United States by four end-use sectors using monthly data. Dummy variables will be included as well. Yang

(2004) investigated the econometric regression model using various economic indicators and forecasts the demand of electricity in rural China. Lakhani and Bumb (1978) developed two multiple linear regression models for residential and nonresidential demand for electricity in Maryland respectively. The function of residential sector incorporated price, price of substitute, capital income and a lagged demand variable. The function for nonresidential sector incorporated employment, price and lagged demand. Weron (2001) modeling electricity loads in California using the combination of ordinary linear regression and Vasicek model. Vasicek model was introduced when the author failed to remove annual cycles from the system load with generalized Ornstin-Uhlenbeck type process. Contreras (2003) estimated the next day electricity price in Spain and California by using ARIMA model. Besides, the AutoRegressive Integrated Moving Average (ARIMA) is also appropriate in the study of forecasting electricity consumption and price. (Pappas, 2008). However, they used annual data including 40 observations which failed to take seasonality of electricity consumption into consideration.

Researchers from other areas also try to forecast the electricity demand in short-run and long-run. These techniques includes artificial neural networks (ANNs) (Ringwood 2001, Abraham 2001), logic approach (Kucucali, 2010), Grey prediction (Akay, 2007) and so on. In conclusion, these papers have already done lots of work but most of them developed models using annual data with small data sample set and they didn’t capture the seasonal cyclical character, one of the most important character of electricity sales. Also, most of previous research fail to talk about demand electricity by sectors or simply divide the sectors as residential and non-residential sector. Therefore, three different time series

models will be developed for each end-user sector and compare forecast data and actual data of Jan 2015 to April 2015. These three models are, linear regression model with ARMA errors, ADL model, and Seasonal-difference model.

3. Methodology

Time series analysis generally means to understand mechanism. It require us to find patterns that lead us to find appropriate models for observed series and then to forecast the future values of the observed series.

3.1. Fundamental concepts

3.1.1. Stochastic process

For a stochastic series, a series of random variables, $\{Y_t: t = 0, \pm 1, \pm 2, \ldots\}$, the mean function is defined by

$$\mu_t = E(Y_t) \quad \text{for } t = 0, \pm 1, \pm 2, \ldots \tag{3-1}$$

and the autocorrelation function is defined as

$$\rho_{t,s} = Corr(Y_t, Y_s) = \frac{Cov(Y_t, Y_s)}{\sqrt{Var(Y_t)Var(Y_s)}} = \frac{\gamma_{t,s}}{\sqrt{\gamma_{t,t}\gamma_{s,s}}} \tag{3-2}$$

Where $\gamma_{t,s} = Cov(Y_t, Y_s)$ for $t, s = 0, \pm 1, \pm 2, \ldots$ is called autocovariance function that

$$cov(Y_t, Y_s) = E[(Y_t - \mu_t)(Y_s - \mu_s)] = E(Y_t Y_s) - E(Y_t)E(Y_s) = E(Y_t Y_s) - \mu_t \mu_s \tag{3-3}$$

Both autocovariance and autocorrelation measure the linear dependence between two random variables. If value of $\rho_{t,s}$ =0, Y_t and Y_s are uncorrelated or independent with each other. If the value of $\rho_{t,s} = \pm 1$, Y_t and Y_s are positively/negatively linear dependence.

Therefore, if value of $\rho_{t,s}$ closer to ± 1, the stronger dependence. If value of $\rho_{t,s}$ closer to 0, the weaker the dependence is.

3.1.2. Stationarity

Suppose $e_1, e_2, \ldots$ is a white noise process of an independent, identically distributed random variables with zero mean and variance σ_e^2.

Stationarity is one of the most important assumptions in time series analysis. A stationary process means the behavior of the process does not change over time. Strictly stationary process $\{Y_t\}$ is defined as the joint distribution of $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ is the same as the joint distribution of $Y_{t_1-k}, Y_{t_2-k}, \ldots, Y_{t_n-k}$ for all choices of time point $t_1, t_2, \ldots, t_n$ and all choices of time lag k. (Cryer & Chen 2008). For the stationary process, both the mean function and the variance hold constant over time.

If the mean function of series is constant and the variance doesn't change over time, the process is regarded as a weakly stationary process.

3.1.3. Cyclical trend

As we introduced above, in time series analysis the mean function must be constant over time, however, in fact not all series have a constant mean, for example, it could have a general upward trend. In our case, as we have introduced in the introduction chapter, the electricity retail sales data in the United State shows a cyclical trend. A possible model might be

$$Y_t = \mu_t + X_t \tag{3-3}$$

Where X_t is the unobserved variation and $E(X_t) = 0$. μ_t is the mean function which represents the deterministic trend function as $\mu_t = \mu_{t-12}$, as the data behaves a periodic with period 12.

Another important assumption for μ_t with period 12 is that there are 12 parameters giving the expected average monthly retail sales for each 12 months. It generally defined as:

$$\mu_t = \begin{cases} \beta_1 \;\; for\; t = 1, 13, 25, \ldots \\ \beta_2 \;\; for\; t = 2, 14, 26, \ldots \\ \quad\quad\quad . \\ \quad\quad\quad . \\ \quad\quad\quad . \\ \beta_{12} \;\; for\; t = 12, 24, 36, \ldots \end{cases} \tag{3-4}$$

The software we use will automatically generate intercept term which is not concluded in the above function, so it simply leave out one of the parameters (December coefficient in our case) and use an intercept term instead. Therefore, the January coefficient represents the difference between January and December average retail sales, the February coefficient represents the difference between February and December average retail sales, and so on. This model is called the seasonal means model.

3.1.4. Residual analysis

We expect our least square estimates are BLUE, best linear unbiased estimates. For seasonal means model, we predict $\{X_t\}$ by the residual

$$\widehat{X_t} = Y_t - \widehat{\mu_t} \qquad (3\text{-}5)$$

And $\widehat{X_t}$ is the residual of the tth observation. If the standard residual is generated from the regression, we need to examine residual plots. If the data is well fitted or modeled, we expect the residual is normally distributed and independent from each other. The most commonly used test for normality is called Shapiro-Wilk test. The null hypothesis of this test is: "the series is normally distributed" while the alternative hypothesis is: "the series is not normally distributed". This test will give a test statistic of W with p-value. Compared to the p-value to 0.05, if p-value is greater than 0.05, we fail to reject the null hypothesis that the residual of this model is normally distributed.

Similar like normality test, the runs test is used to examine if the independence. It counts how many runs are above or below their median and neither too many nor too few runs may lead us to reject the independence. The counted runs will lead to a p-value and if the p-value is greater than 0.05, we cannot reject the null hypothesis that the independence of the residuals.

3.1.5. The Sample autocorrelation (ACF) and the sample partial autocorrelation (PACF)

Besides plots and tests we introduced in section 1.4., another important technique to examine dependence of residuals is sample autocorrelation function. We define sample autocorrelation function r_k, at lag k as

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \quad for\ k = 1, 2, \ldots \qquad (3\text{-}6)$$

We are interested in discovering the sample autocorrelation function of the standard residuals. The r_k is estimates of ρ_k and if all values of r_k are within the significance level, $\pm 2\sqrt{n}$, it indicates that the standard residual series is a white noise series. Otherwise, if there is one or there are several spikes that exceed the significance level, the series is not a white noise process and it is not the result what we expect.

Levinson (1947) and Durbin (1960) introduced the sample partial autocorrelation function. The estimate of $\widehat{\phi_{kk}}$ is solved as:

$$\widehat{\phi_{kk}} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j}\, \rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j}\, \rho_{k-j}} \tag{3-7}$$

where $\phi_{k,j} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j} \quad for\ j = 1, 2, \dots, k-1$

Followed the estimate of ϕ_{kk}, we may calculate values of ϕ_{kk}. Quenoulle (1949) argues that if the null hypothesis AR(P) is appropriate if the most recent pth lags are greater than the critical limits $\pm 2\sqrt{n}$. Therefore, we can know the right order of p by counting how many lags exceed the boundaries after we draw sample partial autocorrelation using software STATA.

3.2. Model Identification

3.2.1. Moving Average (MA) Process

First introduced by Slutsky (1927), the moving average of q is defined as

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{3-8}$$

Where $\gamma_0 = \left(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2\right)\sigma_e^2$

And

$$\rho_k = \{ \begin{array}{ll} \dfrac{-\sigma_k + \sigma_1\sigma_{k+1} + \sigma_2\sigma_{k+2} + \cdots + \sigma_{q-k}\sigma_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2} & for\ k = 1, 2, \dots, q \\ 0 \quad for\ k > q & \end{array} \quad (3\text{-}9)$$

Therefore, the autocorrelation function equals 0 after q th lags. In sample autocorrelation plot, all values should be within the significance boundary after the lag q. The MA (1) model is: $Y_t = e_t - \sigma_1 e_{t-1}$. It has no autocorrelation after lag 1 but the higher-order correlation may exisit.

The MA (2) model can be expressed as : $Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$. The first lag shows fairly strong autocorrelation and the second lag shows fairly weak autocorrelation and lags beyond three lack of autocorrelation.

3.2.2. Autoregressive (AR) Process

Yule (1926) first introduced autoregressive process $\{Y_t\}$ of pth order which is defined as

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \quad (3\text{-}10)$$

Y_t is a process that combined of its p most recent past values and an term e_t that is independent of $Y_t, Y_{t-1}, Y_{t-2}, \dots$

In general, the autocovariance $\gamma_k = \phi^k \frac{\theta_e^2}{1-\phi^2}$ and the autocorrelation is $\rho_k = \frac{\gamma_k}{\gamma_0} = \phi^k\ for\ k = 1, 2, 3, \dots$ The denominator of autocovariance γ_k indicates that $|\phi| < 1$. Therefore AR (1) model satisfies $Y_t = \phi_1 Y_{t-1} + \varepsilon_t$. The AR (1) is stationary if and only if $|\phi| < 1$. AR(2) model satisfies $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$

Following formula (3-11) The AR (2) characteristic equation is defined as $1 - \phi_1 x - \phi_2 x^2 = 0$

The AR (2) process is a stationary process if and only if the absolute value of roots of characteristic equation are greater than 1.

3.2.3. Percentage changes and logarithms

Economic time series are often analyzed after transforming the original series into logarithms forms. (Stock and Watson, 2003). One reason is that many economic series have an approximate constant percentage change over time, taking the logarithm of the series helps transform the exponential growth process into a linear series. Logarithm form also gives the elasticity of the observed series. Another reason is after the logarithms transformation, the standard deviation is approximately constant. It also can give us information about elasticity of variables.

Mathematically, suppose a process $\{\mathrm{Y_t}\}$ tends to have a constant percent changes over time and can be written as $\mathrm{Y_t} = (1 + X_t)Y_{t-1}$ where $\mathrm{X_t}$ is the percentage change from $\mathrm{Y_{t-1}}$ to $\mathrm{Y_t}$. We take logarithms form of both sides of the equation and we get

$$\log(\mathrm{Y_t}) = \log(1 + X_t) + \log(Y_{t-1})$$

$$\log(\mathrm{Y_t}) - \log(Y_{t-1}) = \log(1 + X_t) = \log(\frac{\mathrm{Y_t}}{Y_{t-1}}) \tag{3-11}$$

If the percentage changes are less than $\pm$ 20%, approximately $\log(1 + X_t) \approx X_t$

Thus the new process

$$\nabla \log(Y_t) \approx X_t \tag{3-12}$$

is a relatively stationary process for model and future forecast.

Figures in section 4.2.2 supports that after taking the first difference of logarithms series, the new series is a stationary process.

3.2.4. The Dickey- Fuller Unit-Root Test

The null hypothesis of Dickey-Fuller Unit-Root Test is that the AR characteristic polynomial has a unit root and the alternative hypothesis is that it has no unit roots. In other words, the null hypothesis is that the series is non-stationary and the alternative hypothesis is that the series is a stationary process. The AR order can be first estimated based on some information criteria before applying ADF test as the AR order increases with the sample size. (Said & Dickey, 1984). Therefore, we must determine k before ADF test by using AIC criteria. (AIC will be introduced in Section 3.4.1).

Table *3-1* indicates that before taking difference of logarithm, except for the retail sales data of industrial sector, all other processes are not stationary because p-values are greater than 0.05, thus fail to reject the null hypothesis that the process is not stationary. After taking difference of logarithm, all p values are less than 0.05 and this means we reject the null hypothesis, in other words, we have evidence to show that processes are stationary. The results of ADF test future support us to use the log-difference form of monthly electricity retails sales for model and forecast.

Table 3-1 the augmented Dickey-Fuller (ADF) test Result

Variable	Suggest lag k	ADF Statistic	P-value	Comment
The United States	14	-2.591	0.329	Fail to Reject H_0
Residential	14	-2.594	0.328	Fail to Reject H_0
Commercial	13	-2.441	0.392	Fail to Reject H_0
Industrial	14	-4.031	0.010	Reject H_0
Transportation	14	-2.638	0.311	Fail to Reject H_0
Differencing of logarithm form				
The United States	16	-4.007	0.010	Reject H_0
Residential	16	-4.948	0.010	Reject H_0
Commercial	12	-5.261	0.010	Reject H_0
Industrial	19	-4.488	0.010	Reject H_0
Transportation	12	-4.359	0.010	Reject H_0

3.3. Model Specification.

3.3.1. Regression with dummy variables and ARIMA errors (Model 1)

Recall the seasonal means model equation (3-4 and displayed as following

$$\mu_t = \begin{cases} \beta_1 \ \ for\ t = 1, 13, 25, \dots \\ \beta_2 \ \ for\ t = 2, 14, 26, \dots \\ \quad\quad\quad\quad . \\ \quad\quad\quad\quad . \\ \quad\quad\quad\quad . \\ \beta_{12} \ \ for\ t = 12, 24, 36, \dots \end{cases}$$

After we fit linear regression with dummy variables, the residual always have time series structures. It is possible to adjust the regression by incorporating standard errors. For example, a simple linear regression model with autoregressive errors can be written as

$$y_t = \beta_0 + \beta_1 x_1 + \epsilon_t \tag{3-13}$$

Where y, t, and X_1 are variables and $\epsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \omega_t$and ω_t is a white noise series.

This model allows us to use relevant variables to produce accurate result, however, it fails to provide time series dynamic information that can be addressed with ARIMA models. To include extend ARIMA models, we can simply combine regression model and ARIMA models to give regression with ARIMA errors, thus we will allow the error term to contain autocorrelations. We'll replace the error term u by $u_t$in the previous equation. The error series e_t is assumed to follow an ARIMA (p, d, q) model. The new model is given:

$$Y_t = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + \cdots b_{11} X_{11} + u_t \tag{3-14}$$

$$\emptyset B(1-B)^d u_t = \theta B e_t$$

Where $\{e_t\}$ is zero mean white noise series with variance var (e_t)=σ_e^2. In this notation, $\{u_t\}$ follows an autoregressive integrated moving average model with d^{th} differences. The letter B refers to the backshift operator. The autoregressive (AR) and moving average (MA) characteristic operators are:

$\emptyset(B) = (1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p)$, Autoregressive characteristic operator

$\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$, Moving average characteristic operator.

The model introduced above produces two error terms – the error from the regression donated by u_t and the error term from ARIMA model which we donated by e_t

Therefore, we can first develop an ordinary regression and store the residuals. Then we examine if the residual has time series structures. If it appears to have time series structures, we can figure out what type of the time series error it is by checking ACF and PACF plot and then redevelop the model incorporate these time series disturbances. Repeat this procedure until the residual is a white noise series.

3.3.2. Autoregressive Distributed Lag (ADL) Model (Model 2)

The finite distributed lag (FDL) model could allow us to incorporate with one or more variables to affect the dependent variables with lags. For example, if it only contains one independent variable, it could be written as:

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_q X_{t-q} + \mu_t \qquad (3\text{-}15)$$

The equation *(3-15)* recognize that, how dependent variable y is resulted from X variable immediately and behaviors of X in the past q periods. As β_0 is the immediate change in y due to one-unit change in x at time t, β_0 is named impact multiplier.

Also, the dependent y may be influenced from its own past performance, in other words, the dependent variable can be not only react to changes in independent variable with lags, but also react to its own changes in the past. We can rewrite the equation *(3-15)* as:

$$Y_t + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_q y_{t-q} \\ = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_q X_{t-q} + \mu_t \tag{3-16}$$

In our case, recall the work done by Mohamed and Bodger we reviewed in Chapter 2, GDP, price and population should be used as independent variables. However, GDP is a quarterly released data, therefore we use a similar monthly released economic indicator called Industrial Production (IP) Index instead of using GDP. As we mentioned in the section 3.3.1, dummy variables are also widely used in the time series analysis. We create dummy variable $d1$ which indicates January thus $d1$ equals 1 when it is January, and zero otherwise. Similarly, we generate another 10 dummy variables $d2, d3, \dots, d11$ to represent individual month. Our model will be:

$$Y_t = \sum_{i=1}^{12} \alpha_i Y_{t-i} + \sum_{i=1}^{11} \zeta_i d_i + \sum_{i=0}^{12} \beta_i Price_{t-i} + \sum_{i=0}^{12} \gamma_i Pop_{t-i} + \\ \sum_{i=0}^{12} \delta_i IP_{t-i} + \varepsilon_t \tag{3-17}$$

Where Y is the dependent variable, electricity retail sales. $d_1, d_2, \dots, d_{11}$ are dummy variables that represent month. $price_0$ is the average sale price of electricity of current month and $price_{t-i}$ is the electricity retail price in i month ago. Pop_0 is the current population data in the United States and Pop_{t-i} is the population in ith month ago. IP_0 is the IP index of current month and IP_{t-i} is the IP index of the ith previous month.

The model is displayed as equation *(3-17)* allows us to know how electricity sales is influenced by past electricity sales data, current and past value of average retail price of the electricity, current and past population in the United States, current and past industry growth (IP index), and seasonal patterns. Considering previous research results, we expect that in the United States the retail sales of electricity is inelastic of its price but is effected positively by population growth. Compare to other sectors, the sales in residential sector would be easier to be affected by price change in the long run since customer may adjust their electricity consumption habit under financial concern. Increasing population could also bring electricity sales growth in residential sector as new consumer growth but little influence on commercial and industrial sectors because business and industry may need much longer time to react to the population growth. IP index is expected to influence industrial sector and commercial sector the most and have relatively little effect to residential sector. The higher IP index indicates more business enter the industry which will lead to increase consumption of electricity.

3.3.3. Seasonal-difference approach (Model 3)

If the series is nonstationary, the seasonal difference for the series $\{Y_t\}$ could be defined as

$$\nabla_s Y_t = Y_t - Y_{t-s} \tag{3-18}$$

In our case, we consider the changes from January to January, February to February and so forth for successive years. And the data length will changed from n to $\mathrm{n} - \mathrm{s}$. We can remove the trend by taking seasonal difference which is donated by

$$\nabla \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-12}) \tag{3-19}$$

Besides seasonal pattern, there is another substantial correlation needed to be removed as we discussed in 3.2.3. We need to take a first difference of the seasonal difference of logarithm series. The new series $\{W_t\}$ is denoted by

$$W_t = [\log(Y_t) - \log(Y_{t-12})] - [\log(Y_{t-1}) - \log(Y_{t-13})] \tag{3-20}$$

After taking first and seasonal difference of electricity retail sales, the seasonality trend has been removed and the new series looks somewhat stationary. Figure 3-1, gives an example of electricity sales in United States. It appears that most of the seasonality is removed and new series looks stationary. The Dickey- Fuller Unit-Root Test displayed in *Table 3-2* also supports that the new series is stationary. Figure 3-2, Figure 3-3, Figure 3-4, and Figure 3-5 displays the time series plot of first and seasonal difference of electricity sales in residential sector, commercial sector, industrial sector, and transportation sector respectively.

Table 3-2 The Dickey- Fuller Unit-Root Test of first and seasonal difference series.

Variable	**Suggest lag k**	**ADF Statistic**	**P-value**	**Comment**
The United States	12	-5.009	0.010	Reject H_0
Residential	12	-5.667	0.010	Reject H_0
Commercial	12	-5.077	0.010	Reject H_0
Industrial	15	-3.835	0.020	Reject H_0
Transportation	0	-13.042	0.010	Reject H_0

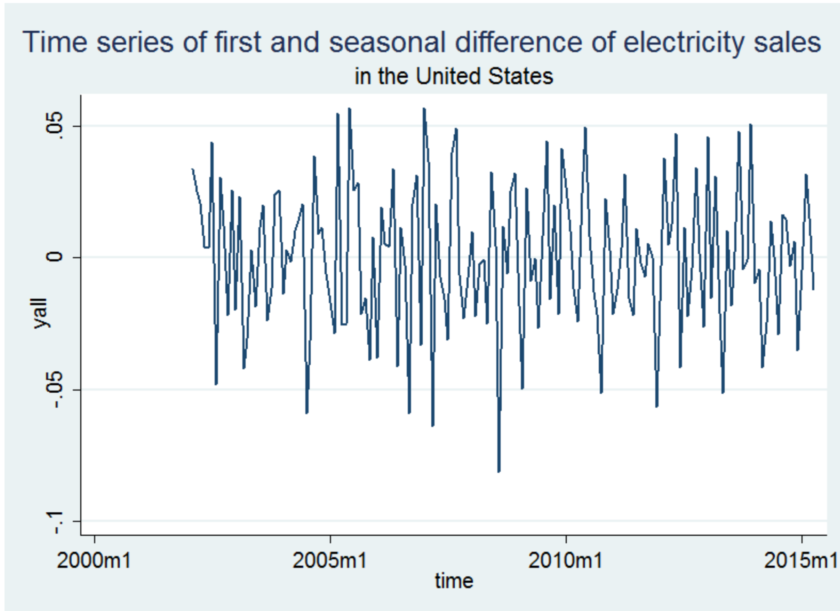

Figure 3-1 Time series plot of first and seasonal difference of electricity retail sales, the United States

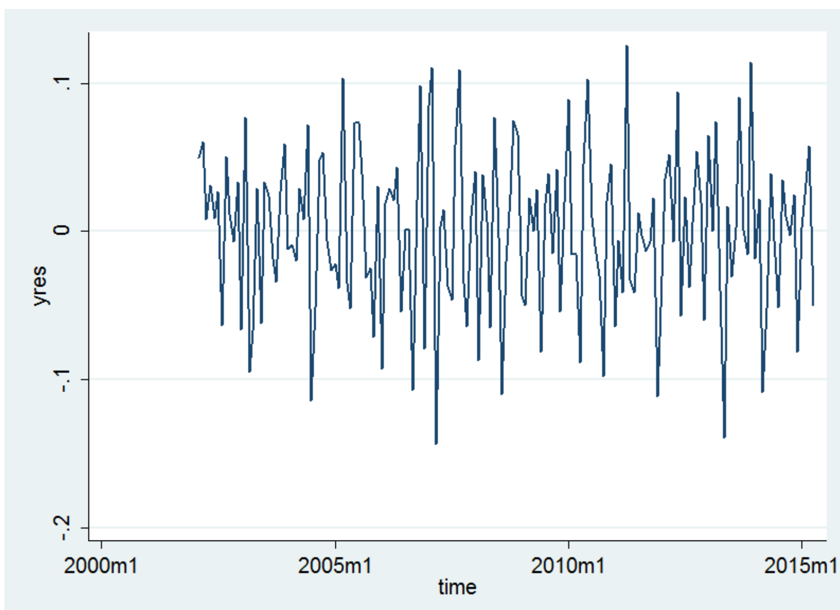

Figure 3-2 Time series plot of first and seasonal difference of electricity retail sales, Residential Sector

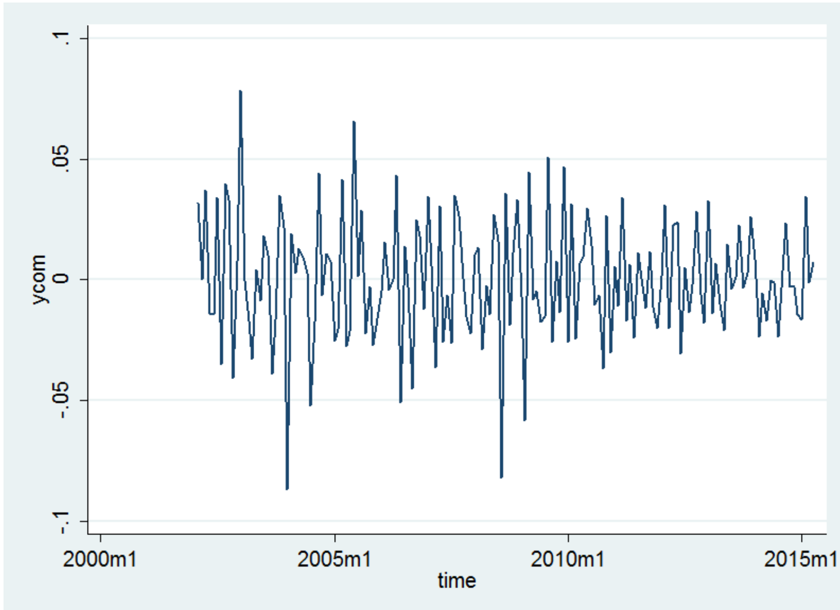

Figure 3-3 Time series plot of first and seasonal difference of electricity retail sales, Commercial Sector

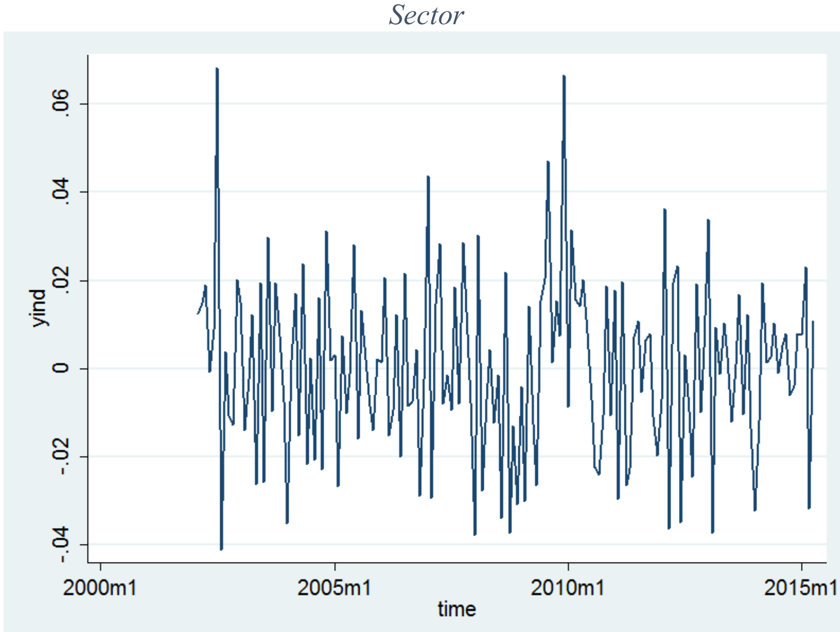

Figure 3-4 Time series plot of first and seasonal difference of electricity retail sales, Industrial Sector

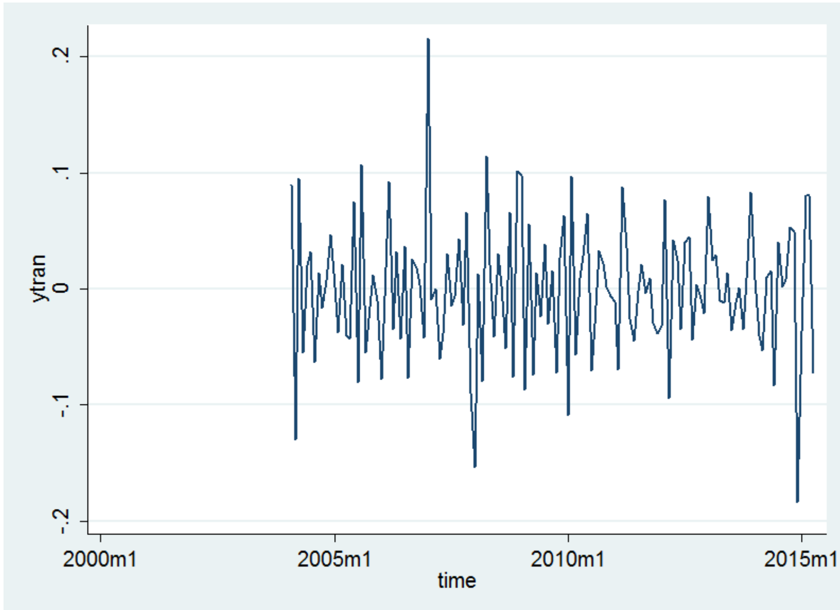

Figure 3-5 Time series plot of first and seasonal difference of electricity retail sales, Transportation Sector

3.3.4. Conclusion

Model 1 is from 3.3.1, the equation (1) (Model 1-1) helps us identify dummy variables. The equation (2) (Model 1- 2) is the time series structure of residual generated by the first equation. Combined two equations, we can get a model incorporate dummy variables and time series errors.

(Model 1)

$$Y_t = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + \cdots b_{11}X_{11} + u_t \qquad (1)$$

$$\emptyset \mathrm{B}(1 - \mathrm{B})^d u_t = \theta B e_t \qquad (2)$$

Model 2 is from section 3.3.2, it uses electricity retail sales data as dependent variable, average retail price of electricity, population, IP index, and dummy variables as independent variable.

(Model 2) $$Y_t = \sum_{i=1}^{12} \alpha_i Y_{t-i} + \sum_{i=1}^{11} \zeta_i d_i + \sum_{i=0}^{12} \beta_i Price_{t-i} + \sum_{i=0}^{12} \gamma_i Pop_{t-i} + \sum_{i=0}^{12} \delta_i IP_{t-i} + \varepsilon_t$$

Model 3 is from section 3.3.3, the equation (1) in model 3 (Model 3-1) identifies data preparation for this model that we take first and seasonal difference of the electricity retail sales. The equation (2) (Model 3-2) indicates that we identify the time series model for the new series.

(Model 2)

$$\mathrm{W_t} = [\log(\mathrm{Y_t}) - \log(Y_{t-12})] - [\log(\mathrm{Y_{t-1}}) - \log(Y_{t-13})] \qquad (1)$$

$$\emptyset \mathrm{B}(1 - \mathrm{B})^d W_t = \theta B e_t \qquad (2)$$

In Model Fitting and Forecast Result part, the electricity retail sales data of each sector is used to fit Model 1, Model2, and Model 3.

3.4. Model Selection and Diagnose

3.4.1. Akaike's Information Criterion and Bayesian Information Criterion

Akaike's (1973) Information Criterion (AIC) is used to select the model that minimize

$$AIC = -2\log(maximum\ likehood) + 2k \qquad (3\text{-}21)$$

Where $= p + q + 1$. If the model includes an intercept term and $k = p + q$

Schwarz Bayesian Information Criterion (BIC) is used to select a model that minimize

$$BIC = -2\log(maxmimum\ likelihood) + klog(n) \qquad (3\text{-}23)$$

These two criterions are used to select a model that minimize AIC and BIC value. Thus when we get similar models that fit the same series, we can select the best model by comparing the AIC and BIC value.

3.4.2. The Ljung-Box Test

Besides the residual analysis we have discussed in 3.1.4 that if the model is adequate then its residual series should be a white noise series. In other word, if the forecast model is appropriate, the residuals are expected to be an identified series that is normally and independently distributed. The sample ACF and sample PACF are expected to show no significant autocorrelation in the residuals.

It is necessary to have a test that check residual correlation. Box and Pierce (1970) proposed the statistic

$$Q = n(\hat{r}_1^2 + \hat{r}_2^2 + \cdots + \hat{r}_K^2) \tag{3-24}$$

to address this possibility. The modified Box- Pierce is given by Ljung-Box

$$Q_* = n(n+2)\left(\frac{\hat{r}_1^2}{n-1} + \frac{\hat{r}_2^2}{n-2} + \cdots + \frac{\hat{r}_K^2}{n-K}\right) \tag{3-25}$$

The null hypothesis of Ljung-Box test is H_0 : The model is appropriate. If the p-value is above 0.05, we have no evidence to reject the null hypothesis, thus the model is appropriate.

4. Data Preparation

4.1. Data Sources

The present study used electricity retail sales and average retail price data are released by U.S. Energy Information Administration (EIA) and downloaded from EIA official website. Mid-month population data is from economic research provided by Federal Reserve Bank of St. Louis. IP index is obtained from Federal Reserve Bank data download program. All these data are monthly data ranged from January 2001 to April

2015. Except transportation sector that contains 148 observations, other variables consists of 172 observations. The last four observations, that is to say observation from January 2015 to April 2015 are left for future diagnosis of the accuracy of a model which will be specified in the following study. All statistics analyses have been performed in Stata.

4.2. Data statistics and time series plots

4.2.1. Summary statistics

Retail Sales of electricity is measured in million kilowatt-hours.

Table *4-1* shows that the most of electricity in the U.S. were sold to residential, commercial and industrial customers. There was a sector called others category which rolled into a new sector named transportation in 2002, that's why there are only 148 observations in the transportation sector. Average retail price of electricity is measured in cents per kilowatt-hours. Table 4- 1 also shows that the variance of electricity retails sales is different among four end-use sectors. Residence, transportation users, and commercial users enjoy the highest average electricity price and users in industrial sector enjoy the lowest price among all users. The average price in industrial sector is 42% lower than average price users pay in residential sector. This also supports us to study electricity sales by sectors. The variance of IP index is relatively large and this is because of the economic recession period from 2007 to 2009 and the economy starts to recovery in recent years. The population also has very large standard deviation because the population in the U.S. has been growing throughout years.

Table 4-1 Summary Statistics of Electricity Sales by sector

Dependent Variables	**Sample size**	**Mean**	**Std. Dev**	**Min**	**Max**
Sales: the United States	172	303369.8	29236.4	253033.7	373364.8
Sales: Residential	172	112738.9	19053.5	80806.7	154728.9
Sales: Commercial	172	106320.2	11382.8	80951.2	130474.7
Sales: Industrial	172	82497.1	4111.0	71358.0	92115.2
Sales: Transportation	148	629.9	42.1	519.4	769.0
Independent Variables	**Sample size**	**Mean**	**Std. Dev**	**Min**	**Max**
Price: the United States	172	8.96	1.20	6.75	11.02
Price: Residential	172	10.59	1.49	7.73	13.05
Price: Commercial	172	9.41	1.07	7.25	11.17
Price: Industrial	172	6.18	0.82	4.71	7.72
Price: Transportation	148	9.69	1.26	6.62	11.84
IP Index	172	98.18	5.12	86.46	108.26
Population	172	303122.3	10946.4	283960	320975
Notes: Retail sales: million kilowatt-hours, Average retail price: cents per kilowatt-hours, Population: thousands					

4.2.2. Time series plots

Figure 4-1 shows that electricity retail sales by all end-use sectors from January 2001 to April 2015. The series shows strong seasonality which concludes dual peaks and valleys. From January 2001 to April, 2015, it peaks in July or August due to high load from air-conditioning equipment and sub-peaks in December or January because of shorter periods of daylight and electric heating. Two valleys mainly happen in April and November respectively. Peak sales rises from 2001 to 2007 and begins to go up again after a drop in 2008 and 2009 but it declines again steadily since 2011 till now. It is not a stationary process as it is required that the mean and variance are constant throughout the time. The difference of the logarithms of the electricity retail sales values are displayed in Figure 4-2. Compared to Figure 4-1, both mean and variance are more stable and we may consider the difference of the logarithms series is a stationary series and appropriate to model. A

quantitative test is introduced in the next session which also supports the stationarity of the difference of logarithm form series.

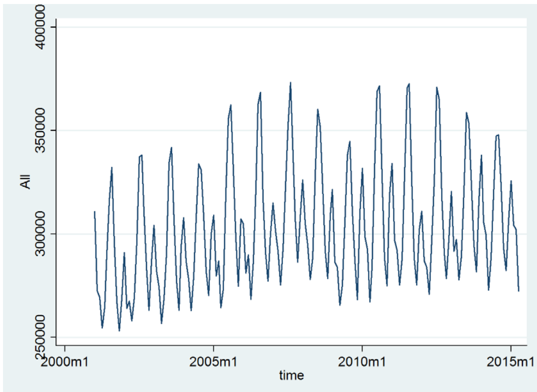

Figure 4-1 Retail sales of electricity in the United States, monthly

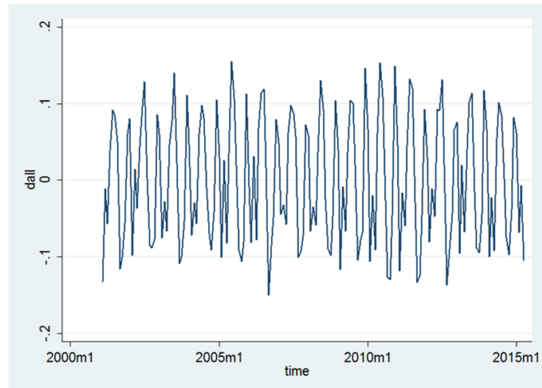

Figure 4-2 Difference of Logarithm for Electricity Sales in the United States, Monthly

Figure 4-3 shows the electricity monthly retail sales in residential sector. Because the dominant end use of electricity in this sector is for lightning and space heating, it displays strong seasonality. The peak consumption moves towards the same trend with all sectors, mainly because it contributes the largest portion in electricity consumptions. The variance in Figure 4-4 is relatively constant compared to Figure 4-3.

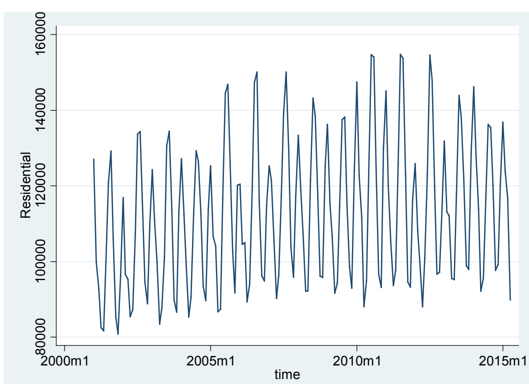

Figure 4-3 Retail sales of electricity in Residential Sector, monthly

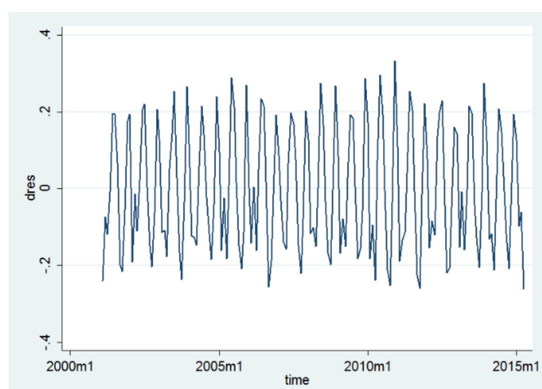

Figure 4-4 Difference of Logarithm for Electricity Sales in Residential Sector, monthly

Figure 4-5 verifies what we conclude in the previous graphs. Similar with residential sector, the electricity retail sales in commercial sector includes not only large seasonal

variance but also an increasing trend. Compare to residential sector, the difference between peak load and off-peak load in this sector is relatively larger. This is because commercial users are unlikely to turn their devices off during peak hours but few home owner may shut down their air conditioners at hottest times. Figure 4-6 indicates after taking difference of logarithm series, the seasonal pattern and increasing trend are removed.

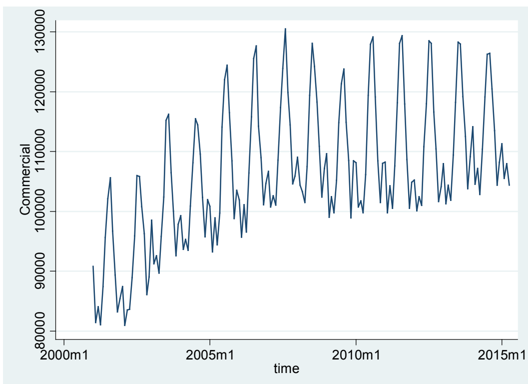

Figure 4-5 Retail sales of electricity in Commercial Sector, monthly

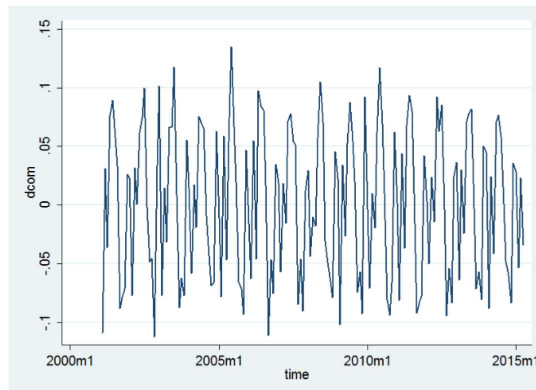

Figure 4-6 Difference of Logarithm for Electricity Sales in Commercial Sector, monthly

The majority amount of electricity in industrial sector is for operate machinery which lasts throughout the year. That's one of the reason why retail sales industrial sector tends to have smaller seasonality than other sectors. In addition, lower energy rate in off - peak hours encourages industrial users to shift their operating times to off-peak times, for example, during the night. Among all sectors, industrial users are more easily to be affected by economic conditions. It is very easy to tell the drop happened from 2007 to 2009 from Figure 4-7, it was due to the recession which is considered to be the worst recession since World War II. Though some points still hang in together after taking the first difference of logarithm series, the ADF test we have applied in section 3.2.4. proves that the new series doesn't have unit root and it is a stationary process and it is adequate for modelling. It

needs to mention that the original series is also a stationary series, in order to standard the study, we use the first difference of logarithm series.

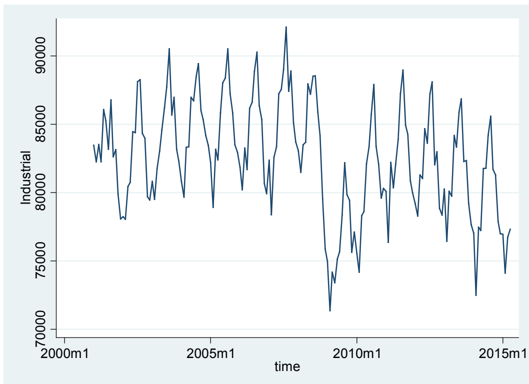

Figure 4-7 Retail sales of electricity in Industrial Sector, monthly

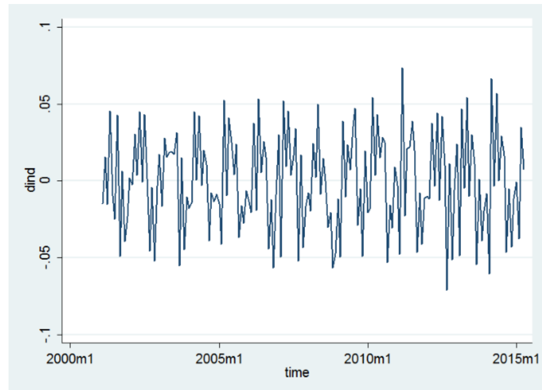

Figure 4-8 Difference of Logarithm for Electricity Sales in Industrial Sector, monthly

A tiny amount of electricity is consumed in transportation sectors, its seasonality is weaker than seasonality in other sectors, showed in Figure 4-9. Though it also peaks in winter and summer, the winter consumption is higher than consumption in summer at most of the time. The Figure 4-10 shows a stationary series that has a relatively constant mean and variance.

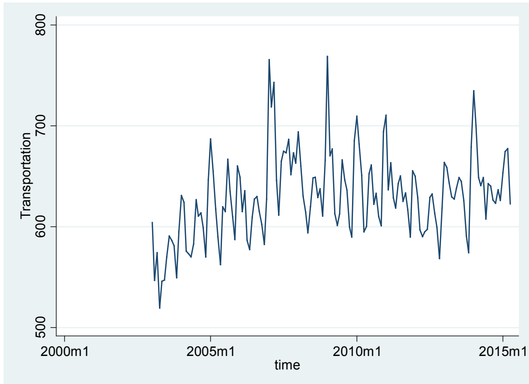

Figure 4-9 Retail sales of electricity in Transportation Sector, monthly

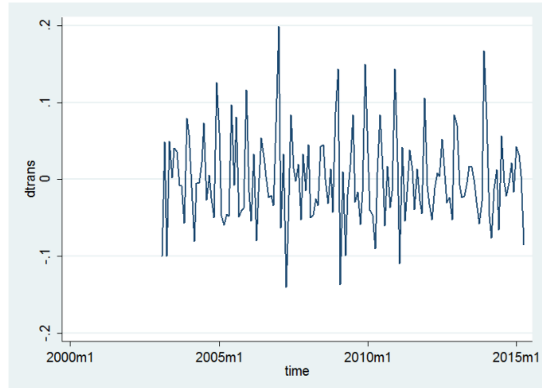

Figure 4-10 Difference of Logarithm for Electricity Sales in Transportation Sector, monthly

Customers in the United States experienced large increase in retail electricity during past years. The price increased quickly from 2001 to 2009 but at a slower pace after 2010. The price peaks in summer and valleys in the winter. Electricity price keeps rising for various reasons such as increasing cost of plant generation, utility investment in transmission and distribution, economic conditions, efficiency of energy use, increasing operating costs. Figure 4-11 indicates that average retail prices in three different sectors have very similar characters. Price peaks in summer but sometimes in September. The peak price keeps increasing since 2002 and it continue rising but in slower pace after a slight decline in 2009. Residential customers account the most expensive electricity price among all users while industrial customers enjoy the lowest rate. Residential peak price has risen 44% during past 14 years from 9.07 cents per kilowatthours in Jun, 2010 to 13.05 cents per kilowatthours in July, 2014. For commercial users, they experienced the highest peak price 11.17 cents per kilowatthours in July 2014 during the period, 3.61% higher than price in July 2013 and 32.8% higher than peak price in 2001. Compared to other sectors, industrial peak price drops as much as 8.03% in July 2009 because of the economic crisis.

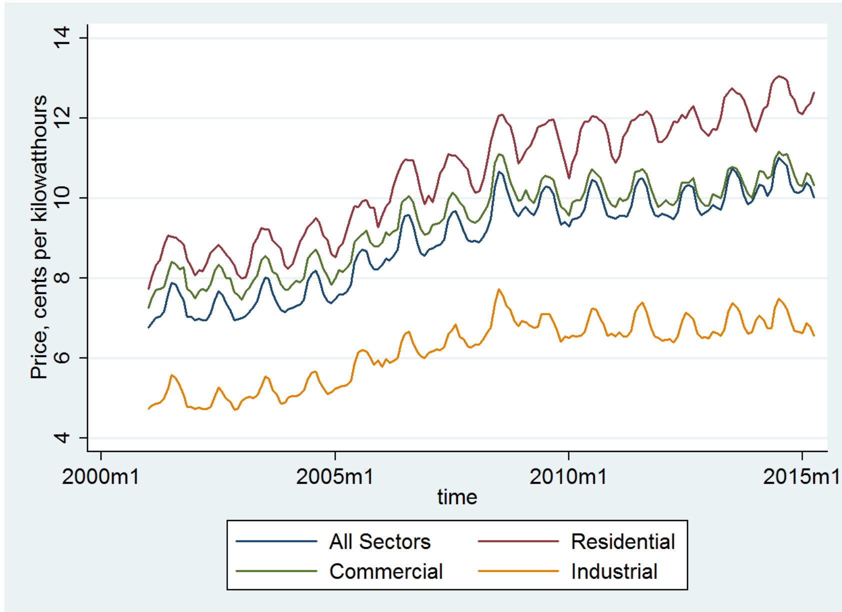

Figure 4-11 U.S. Average retail electricity price by sectors, monthly (2000m1 – 2015m4)

It's difficult to find strong seasonality in transportation price. It increases not only in summer but also sometimes rises largely in winter. In addition, the transportation price boost significantly before 2009 but unlike prices go up in other sectors, it keeps decreasing very slowly since 2010.

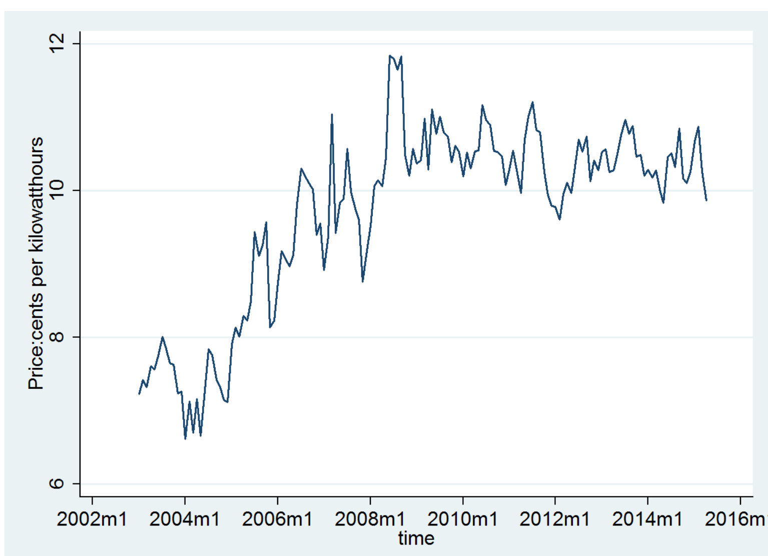

Figure 4-12 Electricity average sale price of transportation, monthly (2002m1 – 2015m4)

Industrial Production (IP) index is a widely used economic indicator which measures the monthly output from several industrial branches including the manufacturing, mining, and utilities. It is released monthly by the Federal Reserve Board. The indicator It attempts to measure the increases and decrease in production output. Figure 4-13 illustrates industrial production index from January 2001 to April 2015. In this graph, though some fluctuations exist, IP index reached a peak in late 2007. Since that time it continued decreasing until late 2009 from 106.85 to 86.46 (23.58%), it started to recover and to increase steadily after 2010. The graph also shows the IP index has already exceed the previous peak in 2007. This is a positive sign that suggests the start of recovery from the previous economic recession, however, it is reported by Annual Revision of IP index in July 2015 that the increase rate of IP index has been slower than reported earlier and it is estimated to fall back to its pre-recession peak in May, 2014.

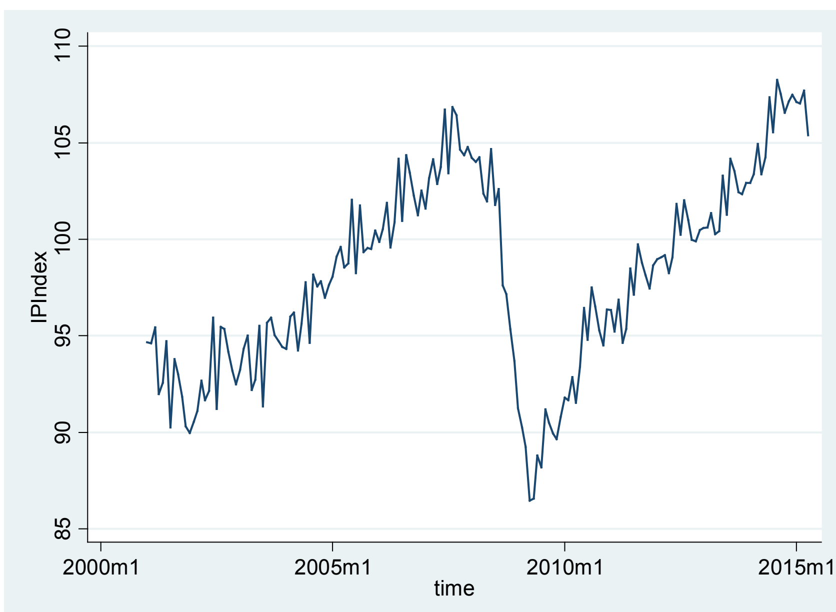

Figure 4-13 Historical records of Industrial Production Index

The Figure 4-14 shows monthly population in the United States from 2001 to April 2015. Population have been rising in average 0.07% per month since January 2001, driven by a high level of immigration. The population in April 2015 reached 321.2 million, which makes the United States of America rank the third most populous country following China (1.4 billion) and India (1.27 billion).

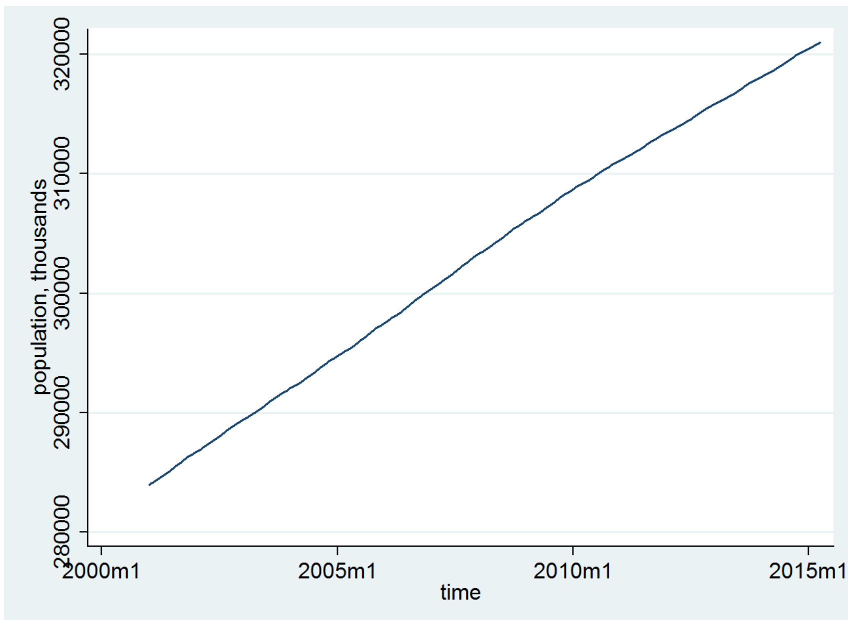

Figure 4-14 Population in the US, monthly (2002m1 – 2015m4)

5. Model Fitting and Forecast Result

5.1. The United States

5.1.1. Regression with dummy variables and ARIMA errors

We first apply regression with dummy variables followed by Model 1-1, the result is displayed in Table 5-1. Before incorporating ARMA process, the standard residual should be a stationary process. Otherwise, it needs to be transformed into a stationary

series. The ADF test of standard residual series suggests that this series is stationary and doesn't have unit roots. The $p-$value of the ADF test on residual is 0.000 means the residual is a stationary process.

Table 5-1 Result of Model 1-1

	Coefficient	Standard Error	P>t
d1	-0.056	0.007	0.000
d2	-0.197	0.007	0.000
d3	-0.114	0.007	0.000
d4	-0.171	0.007	0.000
d5	-0.054	0.007	0.000
d6			
d7			
d8	-0.095	0.007	0.000
d9	-0.211	0.007	0.000
d10	-0.204	0.007	0.000
d11	-0.160	0.007	0.000
_cons	0.105	0.003	0.000
AIC: -679.795 ; BIC: -650.166; Adjusted R^2= 0.931			

The sample ACF and PACF of residual predicted by model 1-1 are plotted in Figure 5-1and Figure 5-2 respectively. In Figure 5-1 ACF plot shows that lag (2) and lag (7) are out of the confidence intervals. It inform us that we should bring first 7 lag of MA error terms back into our model. Similarly, the PACF displayed in Figure 5-2 suggests us to bring first 7 lag of AR error terms to the linear model.

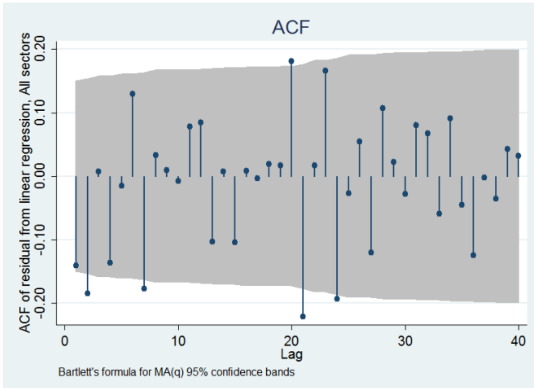

Figure 5-1 ACF plot of residuals from model 1-1

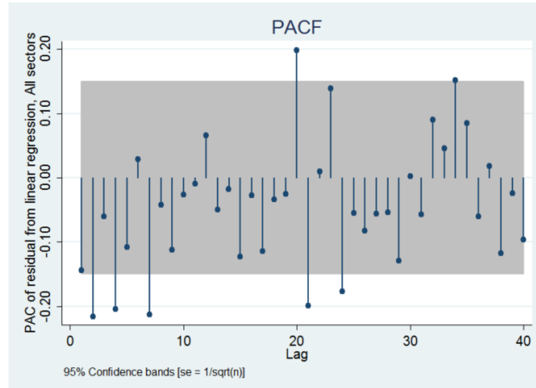

Figure 5-2 PACF plot of residuals from model 1-1

Therefore, we first bring pure 7 lags of AR term back to the model. After removing lags that the coefficient is statistically insignificant, we have two candidate models displayed in Table 5-2. The p-value of coefficient of AR lag 7 in candidate model 1(b) is 0.062, above 0.05. We still keep this term because if we remove AR lag 7, there is a spike out of the confidence interval in PACF. The AIC and BIC value of two models are very close and the candidate model 1 (b) is better than the other one which indicates that the candidate model 1(b) is a better fit.

Table 5-2 Two candidate models fit model 1

Candidate Model 1 (a)				Candidate Model 1(b)			
	Coefficient	Standard Error	P>t		Coefficient	Standard Error	P>t
d1	-0.058	0.006	0.000	d1	-0.056	0.007	0.000
d2	-0.198	0.006	0.000	d2	-0.198	0.006	0.000
d3	-0.115	0.006	0.000	d3	-0.115	0.006	0.000
d4	-0.172	0.008	0.000	d4	-0.172	0.009	0.000
d5	-0.054	0.009	0.000	d5	-0.055	0.009	0.000
d6				d6			
d7				d7			
d8	-0.096	0.007	0.000	d8	-0.096	0.006	0.000
d9	-0.211	0.006	0.000	d9	-0.211	0.006	0.000
d10	-0.205	0.008	0.000	d10	-0.204	0.008	0.000
d11	-0.161	0.010	0.000	d11	-0.160	0.010	0.000
_cons	0.106	0.003	0.000	_cons	0.106	0.003	0.000
AR.				AR.			
L2.	-0.242	0.084	0.004	L7.	-1.547	.0828	0.062
L4.	-0.255	0.085	0.003				
L7.	-0.159	0.079	0.045				
MA.				MA.			
L1.	-0.225	0.096	0.019	L1.	-0.263	0.084	0.000
				L2.	-0.267	0.091	0.007
				L4.	-0.181	0.083	0.0190
AIC: -821.041 BIC: -774.271				AIC: -823.088 BIC: -776.318			

In order to diagnose the model, besides tests displayed in Table 5-5, the ACF and PACF in Figure 5-3 Figure 5-4 Figure 5-5 Figure 5-6. Figure 5-3 and Figure 5-4 suggest that model 1(a) is appropriate as there is no correlation exist in residual series. Though there are one spike in lag 21 in ACF and two lags in PACF, most of the lags are within confidence interval so that we can regard there is no significant correlation in residual. Figure 5-5 and Figure 5-6 suggest model 1 (b) is also appropriate.

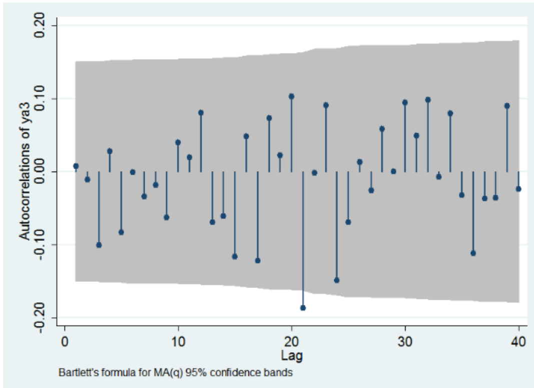

Figure 5-3 ACF of residuals from model 1 (a)

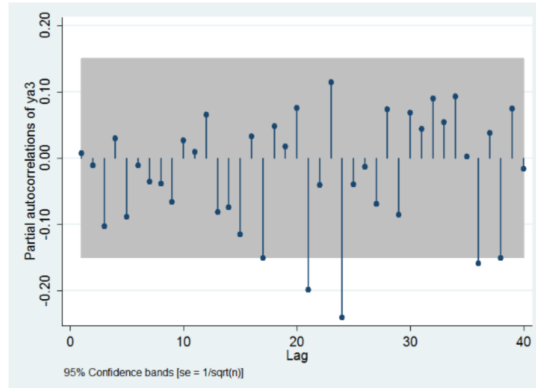

Figure 5-4 PACF of residuals from model 1(a)

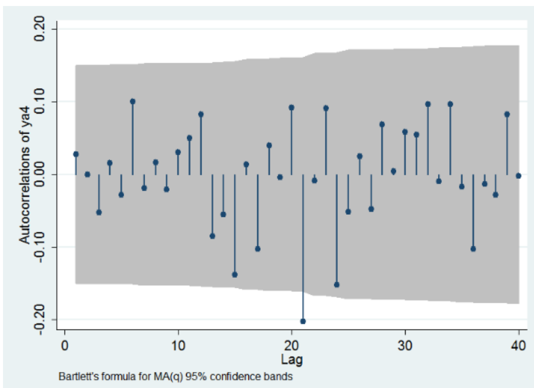

Figure 5-5 ACF of residuals from model 1 (b)

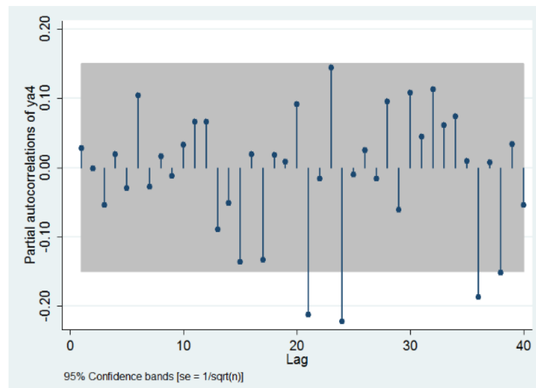

Figure 5-6 PACF of residuals from model 1(b)

5.1.2. ADL model

The ADL model result is displayed in Table 5-3. We take current and 12 most recent value of dependent variable and all independent variables into consideration. The result suggests that electricity consumers in the United States are price insensitive in both short-run and long-run. Population is correlated to the electricity sales in short run as the coefficients of current and the most recent three months are significantly different from zero. IP index is also correlated to electricity sales in short run and long run. In this study, the first three months are regarded as the short run and the period from three months to twelve months is regarded as the long run.

Table 5-3 ADL model (model 2) fitting.

	Coefficient	Standard Error	P>t
Retail Sales			
L1.	-0.212	0.070	0.003
L2.	-0.151	0.075	0.045
L3.	-0.203	0.086	0.019
L4.	-0.183	0.081	0.026
L5.	-0.378	0.084	0.000
L7.	-0.222	0.082	0.008
L8.	-0.177	0.071	0.013
Dummy Variables			
d1			
d2	-0.187	0.020	0.000
d3	-0.125	0.025	0.000
d4	-0.187	0.024	0.000
d5	-0.097	0.020	0.000
d6	-0.048	0.020	0.020
d7			
d8	-0.062	0.022	0.006
d9	-0.184	0.025	0.000
d10	-0.158	0.023	0.000
d11	-0.126	0.018	0.000
Average Price			
Population			
Current	313.691	101.909	0.003
L1.	-593.984	179.049	0.001
L2.	748.149	195.896	0.000
L3.	-687.966	196.051	0.001
L4.	414.995	183.151	0.025
L5.	-303.774	117.551	0.011
L7.	134.686	65.989	0.043
IP index			
Current	0.493	0.183	0.008
L6.	0.410	0.159	0.011
_cons	0.080	0.018	0.000
AIC: -785.405 BIC: -705.613			

Figure 5-7 is the sample ACF of residual from ADL model (Model 2). It suggest there is no moving average structures in the residual series. Figure 5-8 shows though the second lag is out of the boundaries, the PACF still suggest there is no correlation exist in

residual series. Thus, the model 2 we fit is appropriate for future forecast. The test results in Table also support it.

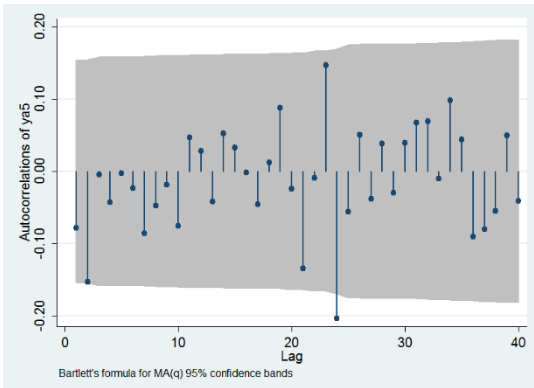

Figure 5-7 ACF plot of residuals from model 2

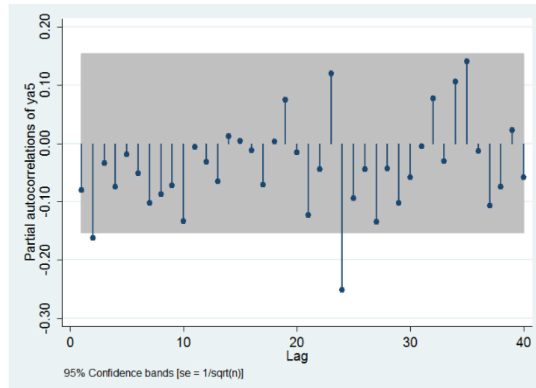

Figure 5-8 PACF plot of residuals from model 2

5.1.3. Seasonal Difference

Similar with what we have done in section 5.1.1., we check ACF (Figure 5-9) and PACF (Figure 5-10) to select the right order for MA model and AR model. The ACF suggests to choose first four MA lags and the PACF also suggests to choose first AR lags. However, calculated by STATA, all coefficients from the MA (4) model are insignificant, neither MA (3). Finally we select MA (2) instead as both coefficient of lag 1 and lag2 are significantly different from zero. Similarly AR (2) should be selected but the PACF of residual still recommends to include the AR lag 4. Therefore AR (4) is selected instead. Figure 5-13 and Figure 5-14 also supports that AR (4) model is an appropriate model. Figure 5-11 and Figure 5-12 provide evidence that MA (2) can be used to forecast as the residual series is independent. Two candidate models are displayed in the Table 5-4.

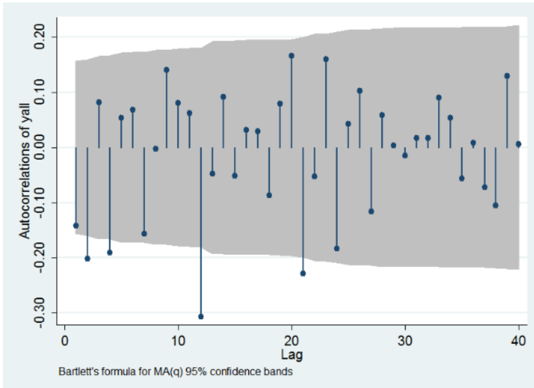

Figure 5-9 ACF of first and seasonal difference series

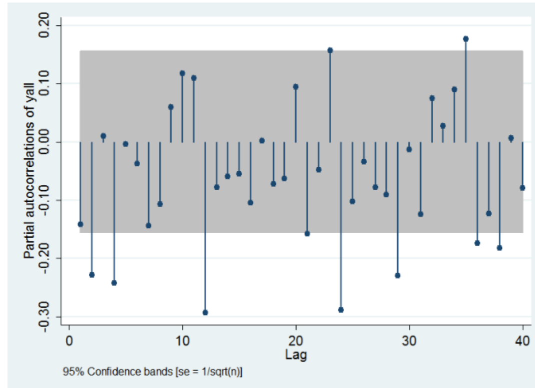

Figure 5-10 PACF of first and seasonal difference series

Table 5-4 Model 3: Seasonal Difference Fitting

Model 3 (a): MA (2) model				Model 3 (b): AR (4) model			
	Coefficient	Standard Error	P>t		Coefficient	Standard Error	P>t
Cons.	0.000	0.001	0.951	Cons	0.000	0.001	0.925
MA.				AR.			
L1.	-0.198	0.085	0.019	L1.	-0.175	0.086	0.043
L2.	-0.339	0.077	0.000	L2.	-0.276	0.087	0.002
				L3.	-0.032	0.091	0.720
				L4.	-0.243	0.081	0.003
AIC: - 679.045 BIC: -666.872				AIC: -679.433 BIC: -661.173			

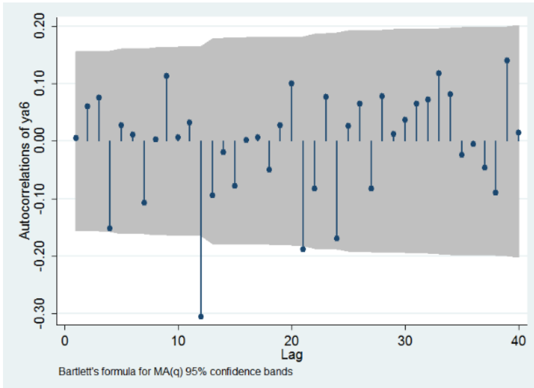

Figure 5-11ACF of MA(2) model residual

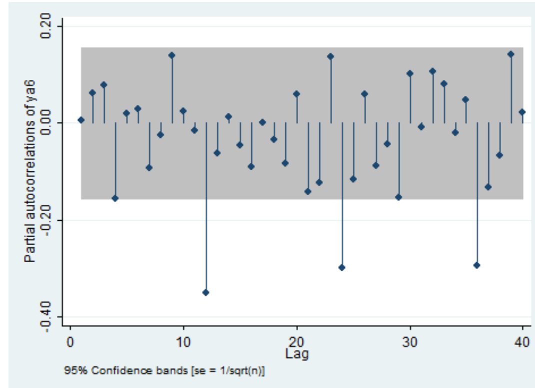

Figure 5-12 PACF of MA(2) model residual

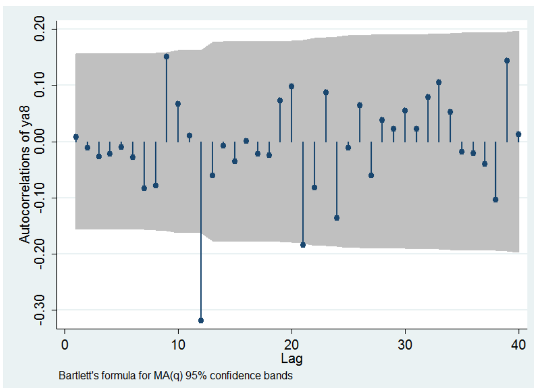

Figure 5-13 ACF of AR(4) model residual

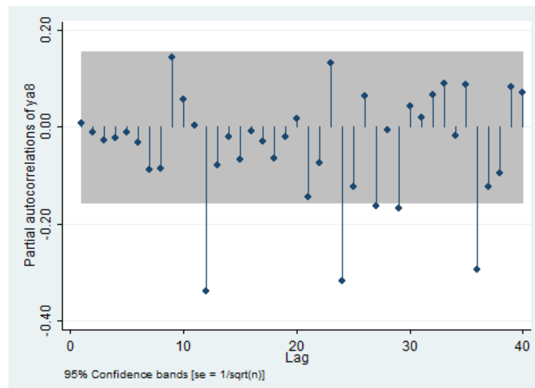

Figure 5-14 PACF of AR(4) model residual

Table 5-5 concludes three model diagnose tests result. All candidate models fail to reject the assumption that residuals are normally distributed. Model 2 ADL model reject the assumption that residuals are independent, the Ljung-Box test however, recommends that this model is an appropriate fitting. Considering the residual ACF and PACF in Figure 5-7 and Figure 5-8 respectively, we still select ADL model as a forecasting model. The forecast results also provides evidence that this model is adequate. Therefore, for first and seasonal difference data, we select model 3 (b) AR (4) model. The forecast plots are displayed in Figure 5-15 and Figure 5-16. In Figure 5-15, it suggests that three candidate models all tend to move in the same direction and are all fit well. Especially the forecast of model 2 ADL model of January 2015 is very close to actual value. Figure 5-16 indicates

that from the March to April, the actual value falls but the forecast slightly increase while other forecast fit the actual value well.

Table 5-5 Model diagnose test results

Test	Model	P-value	Comment
The Shapiro-Wilk test	Model 1 (a)	0.519	Residuals are normally distributed
	Model 1 (b)	0.503	Residuals are normally distributed
	Model 2	0.340	Residuals are normally distributed
	Model 3 (a)	0.518	Residuals are normally distributed
	Model 3 (b)	0.544	Residuals are normally distributed
The Runs Test	Model 1 (a)	0.400	Residuals are independent
	Model 1 (b)	0.590	Residuals are independent
	Model 2	0.020	Reject independent assumption
	Model 3 (a)	0.580	Residuals are independent
	Model 3 (b)	0.690	Residuals are independent
Ljung-Box Test	Model 1 (a)	0.931	The model is appropriate
	Model 1 (b)	0.862	The model is appropriate
	Model 2	0.623	The model is appropriate
	Model 3 (a)	0.728	The model is appropriate
	Model 3 (b)	0.985	The model is appropriate

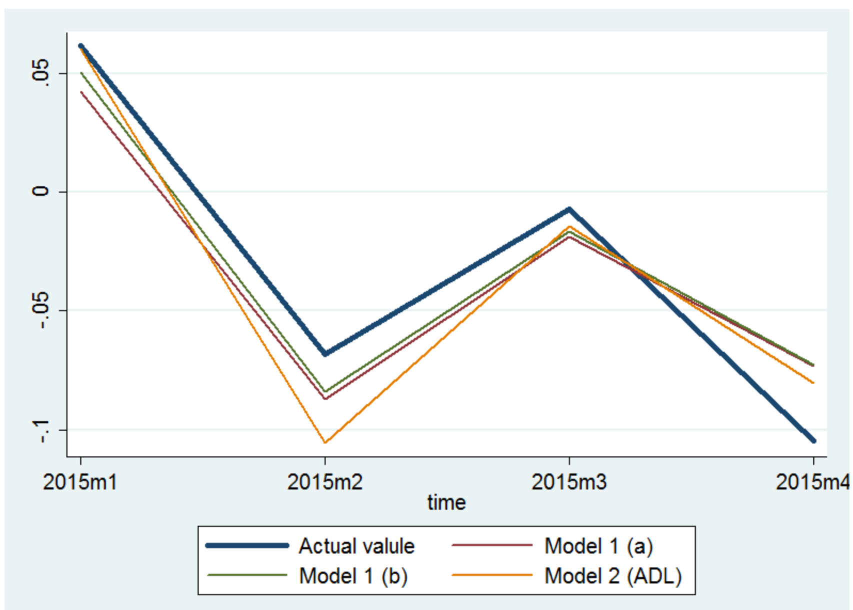

Figure 5-15 Forecast of first different of logarithm series

Figure 5-16 Forecast of first and seasonal different

5.2. Residential Sector

5.2.1. Regression with dummy variables and ARIMA errors

Follow the steps we developed in section 3.3.1, the linear regression with dummy variables for first difference of logarithm of residential retail sales data is displayed in table 5-6. Figure 5-17 and Figure 5-18 shows ACF and PACF of residuals from regression model with dummy variables Model 1-(1) and it shows that the residual have time series structure. It also indicates that we need to take MA (2), AR (2,4,7,8,9) into consideration.

Table 5-6 Result of Model 1-1, Residential Sector

	Coefficient	Standard Error	P>t
d1	-0.116	0.014	0.000
d2	-0.379	0.013	0.000
d3	-0.308	0.013	0.000
d4	-0.386	0.013	0.000
d5	-0.189	0.013	0.000
d6			
d7	-0.034	0.013	0.013
d8	-0.225	0.013	0.000
d9	-0.403	0.013	0.000
d10	-0.439	0.013	0.000
d11	-0.265	0.013	0.000
_cons	0.229	0.008	0.000
AIC: -581.794 ; BIC: -574.451; Adjusted R^2= 0.934			

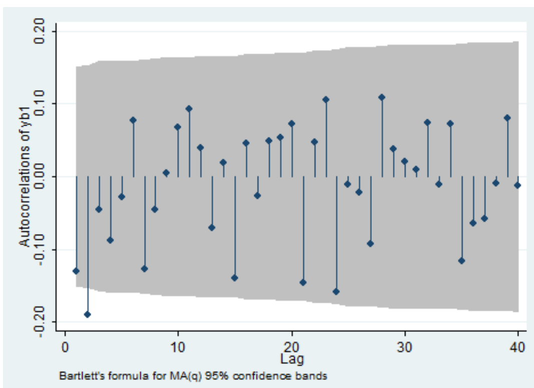

Figure 5-17ACF of residuals from model 1-1

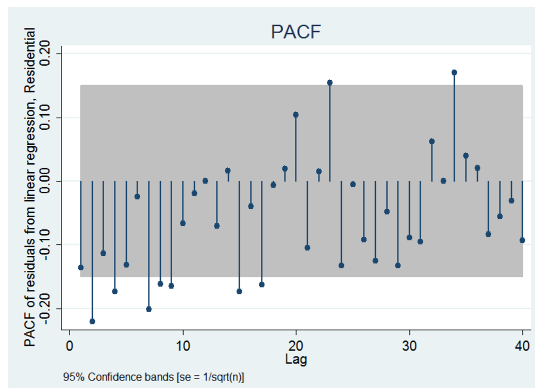

Figure 5-18 PACF of residuals from model 1-1

Deleting lags that have insignificant coefficients and repeat the procedures, we obtain a model with AR (1) and MA (1) errors, showed in table 4-4. Compared to simple linear regression, both absolute value of AIC and BIC increase and this is the evidence that the model with ARMA disturbance is improved. Residual ACF and PACF provide evidence that this model is adequate as there is no lags are out of the confidence interval.

Table 5-7 Model 1, Residential Sector

	Coefficient	Standard Error	P>t
d1	-0.113	0.012	0.000
d2	-0.378	0.011	0.000
d3	-0.308	0.013	0.000
d4	-0.387	0.016	0.000
d5	-0.191	0.015	0.000
d6	-0.036	0.014	0.010
d7	-0.226	0.013	0.000
d8	-0.404	0.014	0.000
d9	-0.439	0.016	0.000
d10	-0.264	0.019	0.000
d11	0.230	0.007	0.000
_cons	-0.113	0.012	0.000
AR.			
L1	0.563	0.090	0.000
MA.			
L1	-0.951	0.035	0.000
AIC: -607.463; BIC: -563.811;			

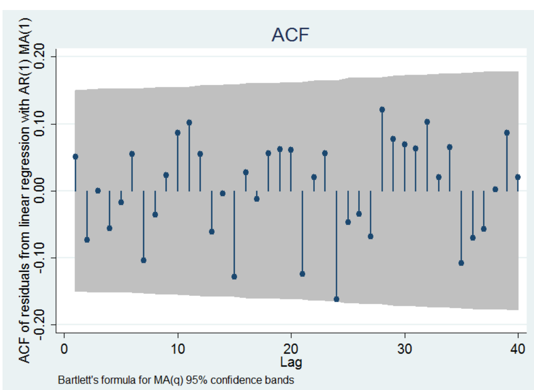

Figure 5-19 ACF of Model 1 residuals

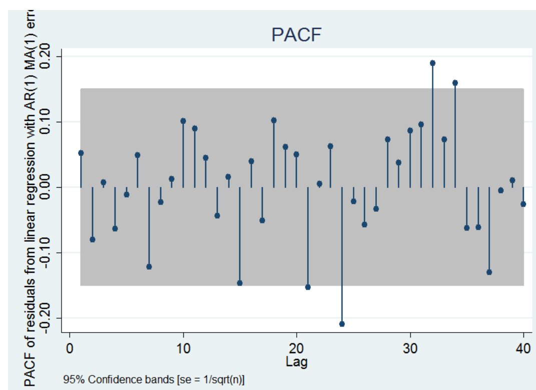

Figure 5-20 PACF of Model 1 residuals

5.2.2. ADL Model

The ADL model result is displayed in Table 5-8. The result suggests that consumers in residential sectors are more sensitive to electricity price in short run and long run. This is because compare to consumers in other sectors, people in residential sector are more easily to adjust their consumption habit when the electricity price changes. The electricity retail sales in this sector is correlated with the population growth in the long run. Current

IP index is significant but his doesn't make much sense in economics. ACF in Figure 5-21 and PACF in Figure 5-22 provide evidence that the model is adequate as there is no significant correlation exist in residual series.

Table 5-8 ADL model (model 2) fitting, Residential Sector

	Coefficient	Standard Error	P>t
Retail Sales			
L1.	-0.136	0.056	0.017
L2.	-0.139	0.057	0.017
L4.	-0.123	0.058	0.036
L5.	-0.284	0.064	0.000
L7.	-0.131	0.052	0.014
Dummy Variables			
d1	-0.049	0.018	0.006
d2	-0.315	0.023	0.000
d3	-0.303	0.030	0.000
d4	-0.344	0.031	0.000
d5	-0.132	0.024	0.000
d6			
d7			
d8	-0.205	0.022	0.000
d9	-0.416	0.029	0.000
d10	-0.430	0.026	0.000
d11	-0.264	0.022	0.000
Average Price			
Current	-1.815	0.174	0.000
L3.	0.470	0.185	0.012
L5.	-0.447	0.222	0.046
L11.	-0.567	0.191	0.003
Population			
L11.	359.575	132.137	0.007
L12.	-308.745	135.727	0.025
IP index			
Current	0.462	0.230	0.047
_cons	0.175	0.026	0.000
AIC: -643.376 BIC: -576.420			

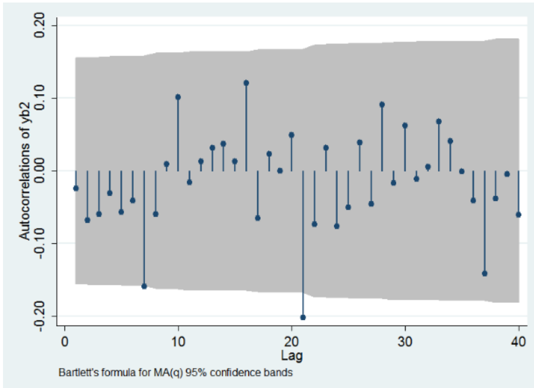

Figure 5-21 ACF of residuals from model 2

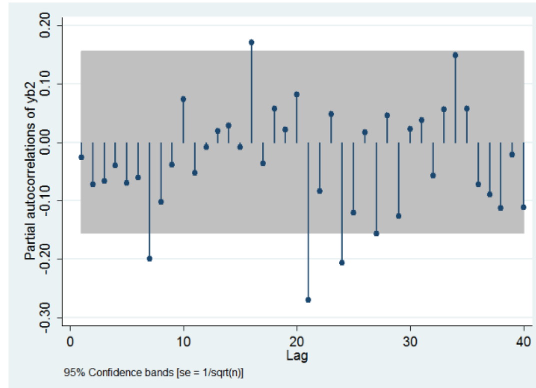

Figure 5-22 PACF of residuals from model 2

5.2.3. Seasonal Difference

The ACF and PACF of first and seasonal difference series displayed in Figure 5-23 and Figure 5-24. They are used to select the right order for MA model and AR model. The ACF suggests to choose first two MA lags that the model might be MA (2) and the PACF suggests to choose first AR lag. However, calculated by STATA, the first lag of MA (2) model is omitted and the MA (1) residual ACF and PACF show correlation. There is no pure MA model. AR (4) is adequate thought the third lag is insignificant. The residual ACF and PACF supports AR (4) model and the Model result is displayed in Table 5-9 and ACF and PACF are displayed in Figure 5-25 and Figure 5-26.

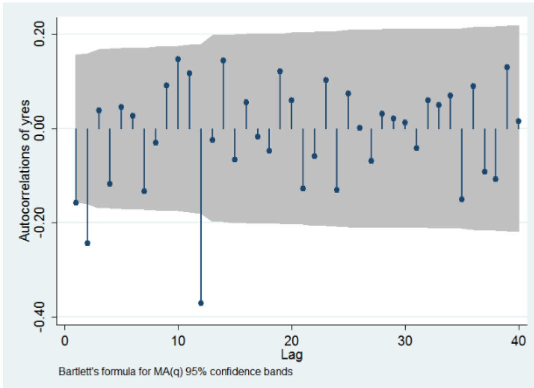

Figure 5-23 ACF of first and seasonal difference series

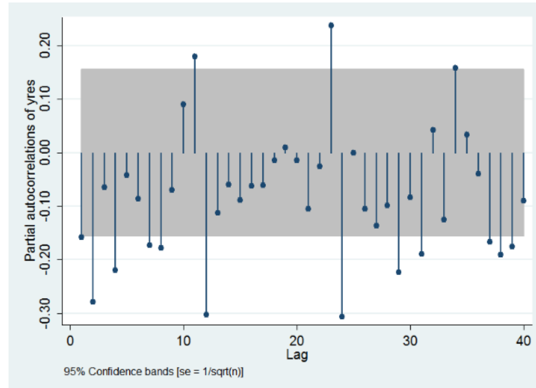

Figure 5-24 PACF of first and seasonal difference series

Table 5-9 Model 3: Seasonal Difference Fitting

	Coefficient	Standard Error	P>t
Cons.	0.000	0.022	0.985
AR.			
L1.	-0.233	0.091	0.010
L2.	-0.356	0.900	0.000
L3.	-0.106	0.783	0.177
L4.	-0.227	0.081	0.005
AIC: -470.063 BIC: -451.803			

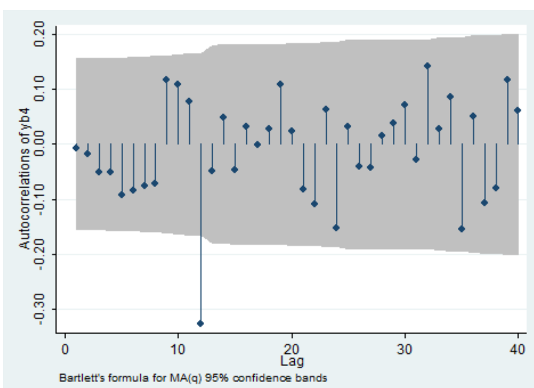

Figure 5-25 ACF of AR(4) model residual

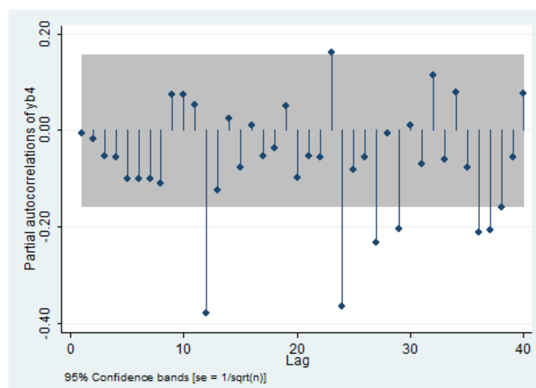

Figure 5-26 PACF of AR(4) model residual

Table 5-10 displays model diagnose tests. All residual series are normally and independently distributed supported by Shapiro-Wilk test and the Runs test. Ljung-box test fail to reject hypothesis when all p-value are above 0.05 wchih means model 1, model 2 and model 3 are adequate model fitting. Figure 5-28 displays the forecast result and only

the forecast trend of January to February moves the same direction with actual value. The forecast value falls from February to March while the actual value increases. The forecast is relatively flat while the actual value falls. Figure 5-27 is the comparison of the actual value and forecast from model 1 and model 2. It indicates that Model 1 and Model 2 provides reliable forecast result especially the forecast of model 1 is very close to actual value in January and the forecast of model 2 is very close to actual value in March,

Table 5-10 Model diagnose test results

Test	Model	P-value	Comment
The Shapiro-Wilk test	Model 1	0.792	Residuals are normally distributed
	Model 2	0.303	Residuals are normally distributed
	Model 3	0.513	Residuals are normally distributed
The Runs Test	Model 1	0.250	Residuals are independent
	Model 2	0.940	Residuals are independent
	Model 3	0.470	Residuals are independent
Ljung-Box Test	Model 1	0.424	The model is appropriate
	Model 2	0.780	The model is appropriate
	Model 3	0.126	The model is appropriate

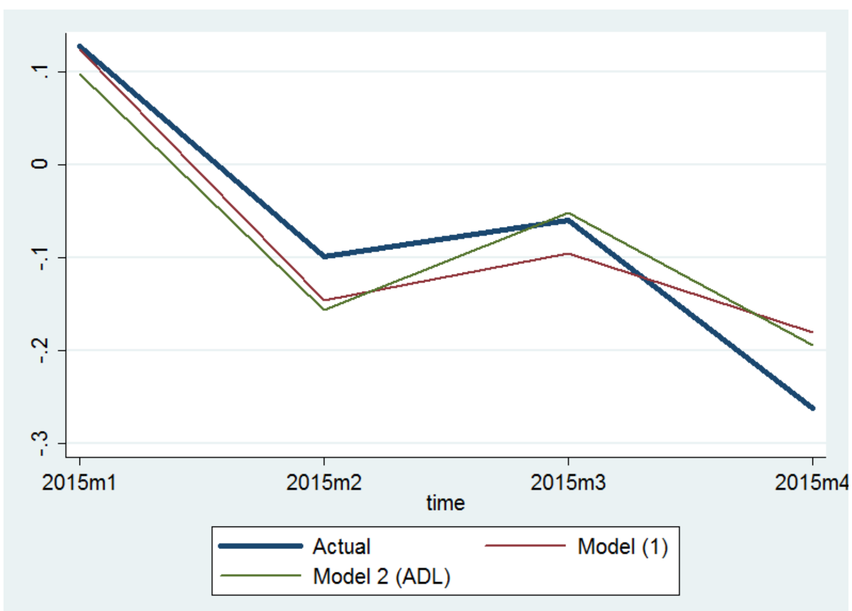

Figure 5-27 Forecast of first difference of logarithm series, residential sector

Figure 5-28 Forecast of first and seasonal different, residential sector

5.3. Commercial Sector

5.3.1. Regression with dummy variables and ARIMA errors

Table 5-11 displays the regression result of Model 1-1. The value of adjusted R^2 is 0.913 indicates that the dummy variables displayed in table 5-11 explains 91.3% the variability of the response data around its mean. However, both the residual ACF and PACF suggest that residual is not an independent series and it has time series structures. In order to get the forecast model, we need to identify this time series errors.

Table 5-11 Result of Model 1-1, Commercial Sector

	Coefficient	Standard Error	P>t
d1	-0.024	0.007	0.001
d2	-0.117	0.007	0.000
d3	-0.017	0.007	0.016
d4	-0.071	0.007	0.000
d5	0.025	0.007	0.000
d6	0.042	0.007	0.000
d7	0.031	0.007	0.000
d8	-0.036	0.007	0.000
d9	-0.119	0.007	0.000
d10	-0.108	0.007	0.000
d11	-0.127	0.007	0.000
_cons	0.045	0.005	0.000
AIC: -843.510 BIC: -806.094; Adjusted R^2= 0.913			

The ACF (Figure 5-29) shows the MA lag (1), MA lag (6), and MA lag (7) are out of the confidential interval. We need to take first seven MA lags into consideration in order to identify the right order of MA model. The PACF (Figure 5-30) indicates that residual is significantly correlated in the first two lags and lag (7). We need to check the right AR order from lag 7. We first consider the pure MA process. After removing all insignificant lags, the MA (1, 2, 6) is selected however the residual is not a white noise process. If we start with a pure AR process, we get model 1 which is displayed in Table 5-12. The AIC is improved from -843.510 to -871.690 and BIC is improved from -806.094 to -821.802. This means the model is improved by incorporating time series errors. Residual ACF (Figure 5-31) and PACF (Figure 5-32) suggest that there is no correlation in residual process thereby the Model 1 is adequate.

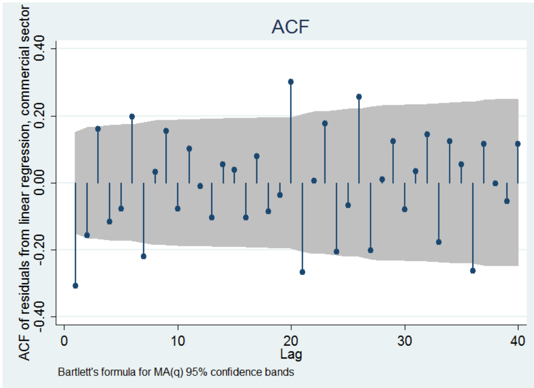

Figure 5-29 ACF of residuals from model 1-1

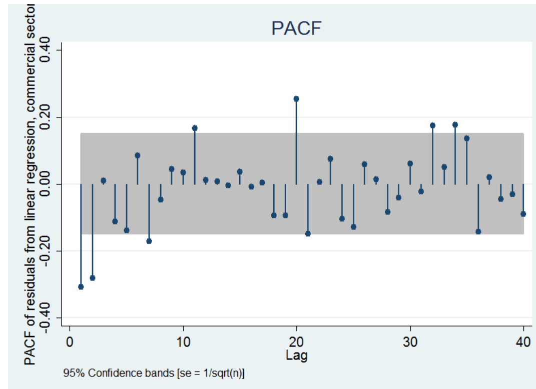

Figure 5-30 PACF of residuals from model 1-1

Table 5-12 Model 1, Commercial Sector

	Coefficient	Standard Error	P>t
d1	-0.023	0.007	0.001
d2	-0.118	0.008	0.000
d3	-0.018	0.008	0.018
d4	-0.072	0.008	0.000
d5	0.025	0.009	0.005
d6	0.042	0.006	0.000
d7	0.031	0.008	0.000
d8	-0.037	0.007	0.000
d9	-0.119	0.006	0.000
d10	-0.109	0.008	0.000
d11	-0.128	0.011	0.000
_cons	0.045	0.005	0.000
AR.			
L1	-0.374	0.084	0.000
L2	-0.290	0.084	0.001
L7	-0.161	0.074	0.030
AIC: -871.690 ; BIC: -821.802			

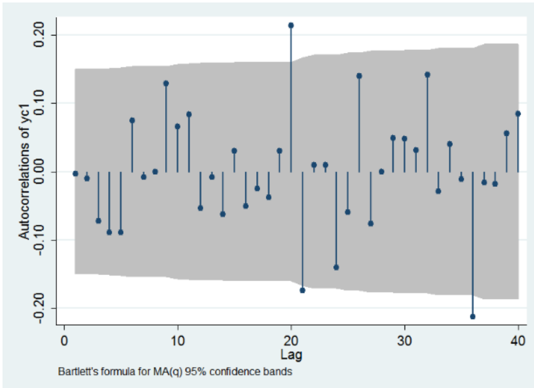

Figure 5-31 ACF of Model 1 residuals

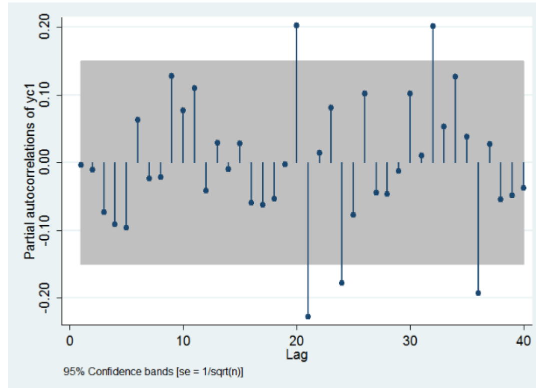

Figure 5-32 PACF of Model 1 residuals

5.3.2. ADL model

The Autoregressive distributed lag (ADL) model result is displayed in Table 5-13. The value of adjusted R^2 is 0.940 which means that independent variables can explain 94% of retail sales of electricity in commercial sector. From Table 5-12, we can know that the electricity sales in this sector is correlated with its past value in both short-run and long-run. Consumers in this sector is very insensitive to price change and this makes sense since the majority electricity consumption in this sector is for heating spaces and lighting in business buildings. It is almost impossible that people shut down the business only because the electricity price increases. Population growth is correlated to electricity sales in short-run. IP index is correlated to electricity sales in commercial sector in the long run though the lag 5 is significant. Figure 5-33 and Figure 5-34 indicate Model 2 is adequate as the residual series has no significant correlation.

Table 5-13 ADL model (model 2) fitting, Commercial Sector

	Coefficient	Standard Error	P>t
Retail Sales			
L1.	-0.429	0.064	0.000
L2.	-0.285	0.058	0.000
L3.	-0.221	0.068	0.001
L5.	-0.181	0.073	0.015
L9.	0.181	0.072	0.013
L10.	0.185	0.058	0.002
L11.	0.256	0.053	0.000
Dummy Variables	-0.315	0.023	0.000
d1			
d2	-0.104	0.014	0.000
d3	-0.040	0.014	0.005
d4	-0.099	0.014	0.000
d5			
d6	0.098	0.012	0.000
d7	0.130	0.014	0.000
d8	0.115	0.012	0.000
d9			
d10	-0.053	0.013	0.000
d11	-0.126	0.013	0.000
Average Price			
Population			
Current	348.024	77.456	0.000
L1.	-704.517	131.952	0.000
L2.	626.225	148.362	0.000
L3.	-468.869	134.443	0.001
L4.	224.922	75.096	0.003
IP Index			
L5.	0.447	0.125	0.000
_cons	-0.010	0.012	0.418
AIC: -840.102 BIC: -773.005 Adjusted $R^2 = 0.940$			

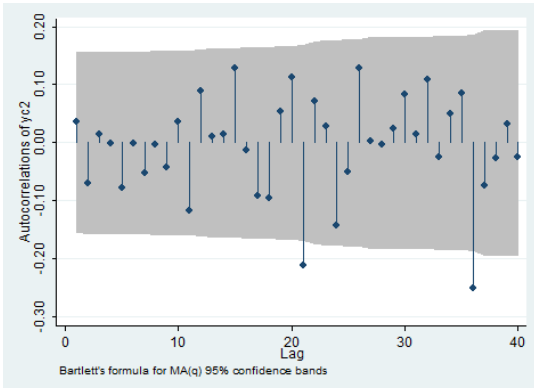

Figure 5-33 ACF of residuals from model 2

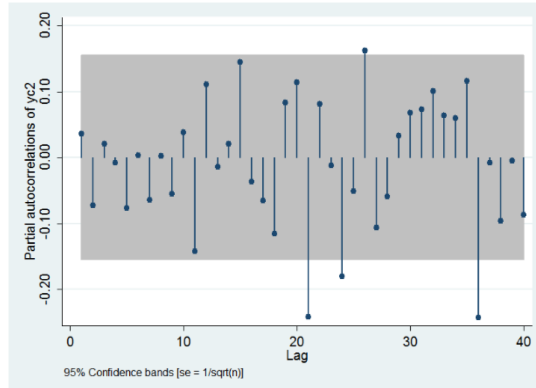

Figure 5-34 PACF of residuals from model 2

5.3.3. Seasonal Difference

In order to specify the right order of MA and AR model, we need to check ACF and PACF first. Figure 5-35 ACF supports a MA (1) model because the first lag is out of the confidence bands. Figure 5-36 PACF indicates that AR (2) model may be appropriate. Therefore, we have two candidate models that are displayed in Table 5-13. AIC and BIC value of two models are very close and MA (1) is slightly better as it has one less lag than AR (2) model. Figure 5-37 and Figure 5-38 supports that MA (1) model is adequate while Figure 5-39 and Figure 5-40 supports that AR (2) model is also a good model fit.

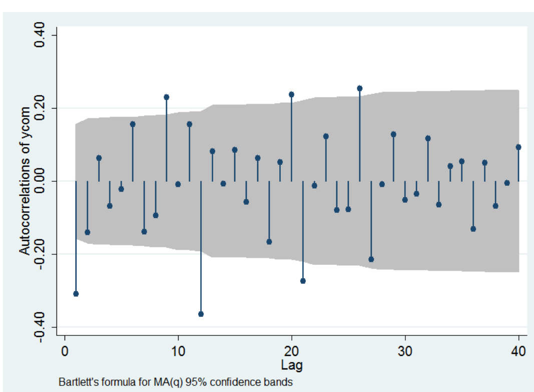

Figure 5-35 ACF of first and seasonal difference series

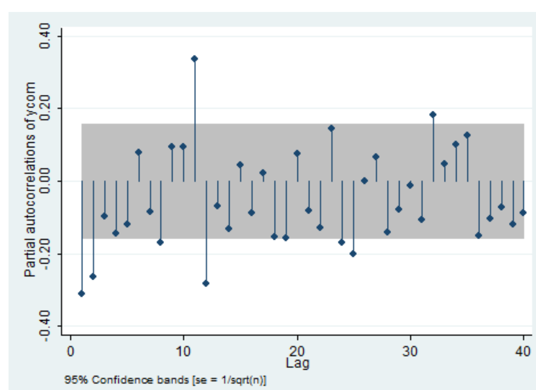

Figure 5-36 PACF of first and seasonal difference series

Table 5-14 Model 3: Seasonal Difference Fitting, Commercial Sector

Model 3 (a)				Model 3 (b)			
	Coefficient	Standard Error	P>t		Coefficient	Standard Error	P>t
Cons.	0.000	0.001	0.914	Cons	0.000	0.001	0.899
MA.				AR.			
L1.	-0.539	0.072	0.000	L1.	-0.396	0.086	0.000
				L2.	-0.265	0.076	0.001
AIC: -714.576 BIC: -705.446				AIC: -711.137 BIC: -698.963			

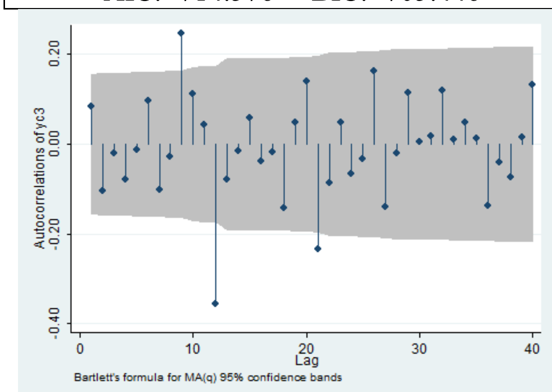

Figure 5-37ACF of MA(1) model residual

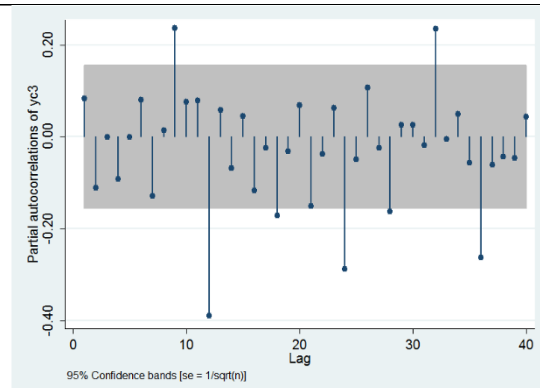

Figure 5-38 PACF of MA(1) model residual

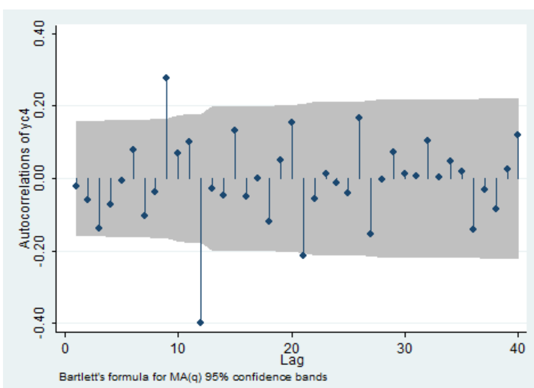

Figure 5-39 ACF of AR(1) model residual

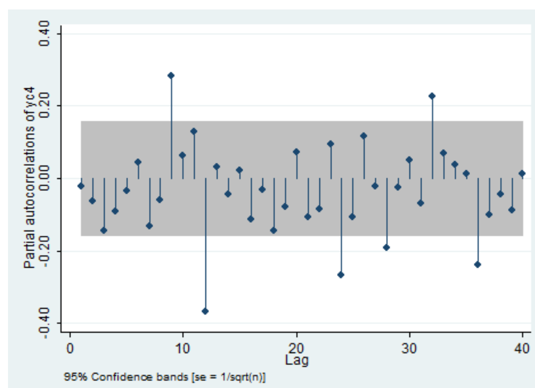

Figure 5-40 PACF of AR(1) model residual

Table 5-14 concludes all model diagnose test results. Model 1 reject the null hypothesis that the residual is normally distributed, however, the Ljung-Box test recommends the residual series is a white noise series. The Shapiro-Wilk test, the runs test, and Ljung-Box test all suggest that model 2, model 3 (a), and model (b) are appropriate models. The forecast result for difference of logarithm series is displayed in Figure 5-41, both model (1) and model (2) forecast the correct moving direction with actual value. The

forecast of model (1) in January, March, and April are very close to the actual value. Thus the model (1) is recommended as a forecasting model for electricity retail sales in commercial sector. Figure 5-42 displays the forecast results of first and seasonal difference series. Forecast result of two models derivate from model 3 are very similar and MA (1) is slightly better, this is in accordance with the model select criterion AIC and BIC we've applied. Though the slope of forecast from January to February is flatter than actual value, the rest forecasts move parallel with actual values.

Table 5-15 Model diagnose test results, Commercial Sector

Test	Model	P-value	Comment
The Shapiro-Wilk test	Model 1	0.032	Reject normality assumption
	Model 2	0.512	Residuals are normally distributed
	Model 3 (a)	0.612	Residuals are normally distributed
	Model 3 (b)	0.801	Residuals are normally distributed
The Runs Test	Model 1	0.940	Residuals are independent
	Model 2	0.630	Residuals are independent
	Model 3 (a)	0.390	Residuals are independent
	Model 3 (b)	0.940	Residuals are independent
Ljung-Box Test	Model 1	0.880	The model is appropriate
	Model 2	0.726	The model is appropriate
	Model 3 (a)	0.238	The model is appropriate
	Model 3 (b)	0.480	The model is appropriate

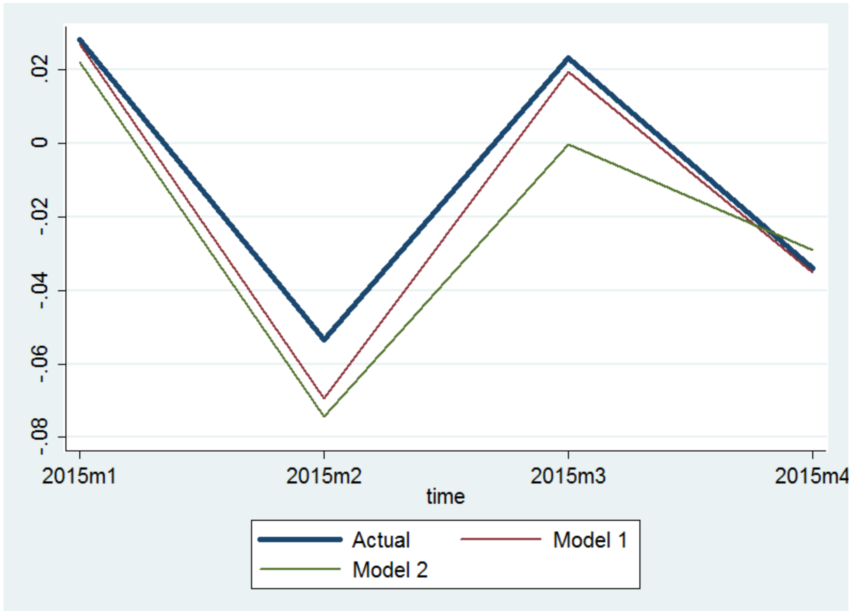

Figure 5-41 Forecast of first difference of logarithm series, commercial sector

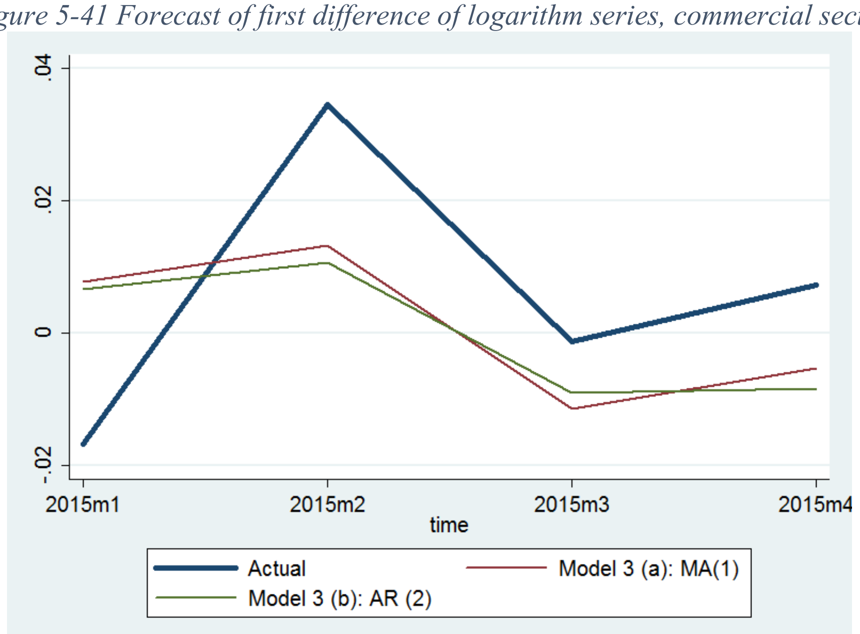

Figure 5-42 Forecast of first and seasonal different, commercial sector

5.4. Industrial Sector

5.4.1. Regression with dummy variables and ARIMA errors

As we analyzed in the introduction part, the seasonality in industrial sector and transportation sector are not as strong as the other two sectors, thus the adjusted R^2 falls in this sector. Table 4-7 displays the simple linear regression model fitting electricity retail sale of industrial sector. The adjusted R^2 is 0.794 which indicates dummy variables can explain 79.5% of the retail sales series in industrial sector. The ACF and PACF in Figure 5-43 and Figure 5-44 suggest us to incorporate AR (1) and MA (1) lags, thus MA (1) and AR (1) can be selected.

Table 5-16 Result of Model 1-1, Industrial Sector

	Coefficient	Standard Error	P>t
d1			
d2	-0.024	0.004	0.000
d3	0.049	0.004	0.000
d4			
d5	0.047	0.004	0.000
d6	0.009	0.004	0.035
d7	0.029	0.004	0.000
d8	0.026	0.004	0.000
d9	-0.041	0.004	0.000
d10	0.002	0.004	0.579
d11	-0.036	0.004	0.000
_cons	-0.006	0.002	0.013
AIC: -935.034 BIC: -903.854 Adjusted R^2= 0.794			

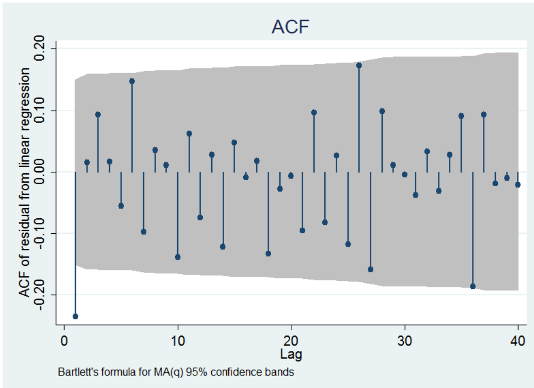

Figure 5-43ACF of residuals from model 1-1

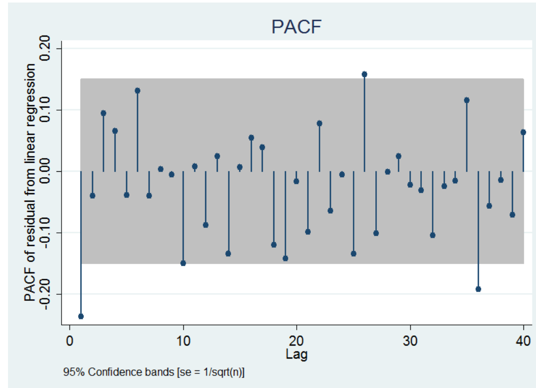

Figure 5-44 PACF of residuals from model 1-1

The selected model is displayed in Table 5-16. The AIC is improved from -935.034 to -942.288 (AR (1)) and -942.001 (MA (1)) and the BIC is improved from -903.854 to -907.990 (AR (1)) and -907.703 (MA (1)). Only from AIC and BIC value, it is difficult to tell which model is better. ACF and PACF of residual predicted by Model 1 (a) are plotted in Figure 5-45 and Figure 5-46, ACF and PACF of Model 1 (b) are plotted in Figure 5-45 and Figure 5-46. There is no evidence to show the two residual series have correlations. Thus, two models are selected.

Table 5-17 Two candidate models fit model 1

Candidate Model 1 (a)				Candidate Model 1(b)			
	Coefficient	Standard Error	P>t		Coefficient	Standard Error	P>t
d1				d1			
d2	-0.024	0.004	0.000	d2	-0.024	0.004	0.000
d3	0.049	0.004	0.000	d3	0.048	0.004	0.000
d4				d4			
d5	0.047	0.005	0.000	d5	0.047	0.005	0.000
d6	0.009	0.005	0.071	d6	0.009	0.005	0.074
d7	0.028	0.004	0.000	d7	0.028	0.004	0.000
d8	0.026	0.004	0.000	d8	0.026	0.004	0.000
d9	-0.041	0.005	0.000	d9	-0.041	0.005	0.000
d10				d10			
d11	-0.037	0.005	0.000	d11	-0.038	0.005	0.000
_cons	-0.005	0.002	0.004	_cons	-0.005	0.002	0.004
MA.				AR.			
L1.	-0.225	0.083	0.006	L1.	-0.237	0.086	0.006
AIC: -942.001 BIC: -907.703				AIC: -942.288 BIC: -907.990			

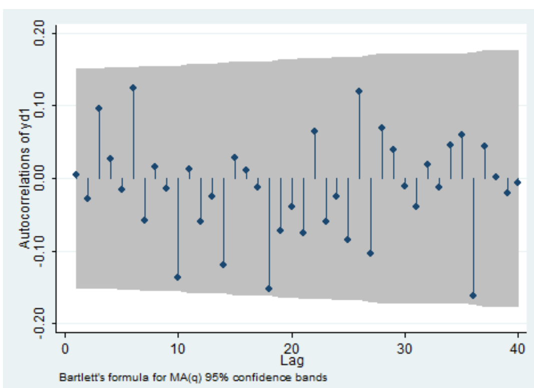

Figure 5-45ACF of residuals from model 1 (a)

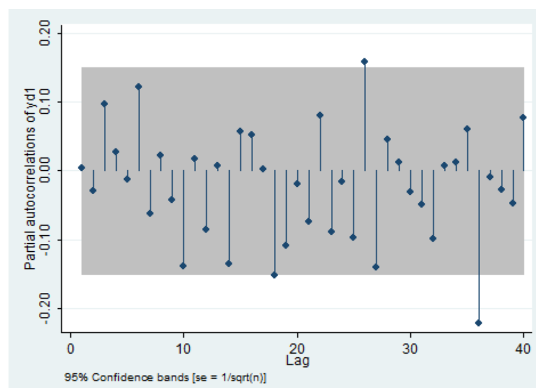

Figure 5-46 PACF of residuals from model 1(a)

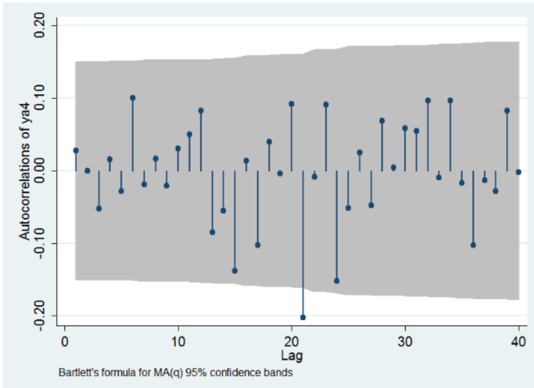

Figure 5-47 ACF of residuals from model 1 (b)

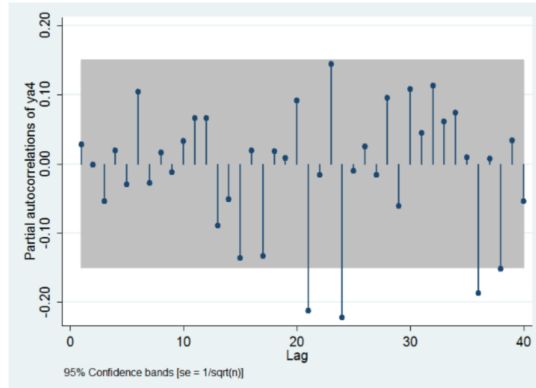

Figure 5-48 PACF of residuals from model 1(b)

5.4.2. ADL Model

Table 5-17 displays the Model 2 fit. The value adjusted R^2 is 0.867 which means independent variables explain 86.7% of electricity sales in industrial sector. Correlated to its historic value in short run, the electricity sales in industrial sector is also correlated to the average price in short-run and long-run. The consumers in this sector are highly motivated in saving operating cost as electricity bill is one of operating cost item. The factory can adjust their work hours in order to save cost. Statistically, the electricity sales in industrial sector has little correlation with population. It is expected that IP index is highly correlated to electricity consumption in industrial sector, however, the result indicates that we lack of evidence to support this expectation as they are only correlated in the short run. ACF and PACF in Figure 5-49 and Figure 5-50 also support model (2).

Table 5-18 ADL model (model 2) fitting, Industrial Sector

	Coefficient	Standard Error	P>t
Retail Sales			
L1.	-0.316	0.064	0.000
L2.	-0.176	0.068	0.010
L4.	-0.151	0.060	0.013
L5.	-0.246	0.060	0.000
Dummy Variables			
d1			
d2	-0.026	0.005	0.000
d3	0.030	0.006	0.000
d4			
d5	0.059	0.007	0.000
d6	0.017	0.008	0.027
d7	0.060	0.011	0.000
d8	0.050	0.008	0.000
d9	-0.065	0.008	0.000
d10			
d11	-0.054	0.007	0.000
Average Price			
L1.	-0.030	0.011	0.008
L3.	0.034	0.009	0.000
L12.	-0.033	0.011	0.004
Population			
IP Index			
Current	0.376	0.111	0.001
L1.	0.559	0.116	0.000
L3.	0.296	0.108	0.007
_cons	-0.007	0.003	0.007
AIC: -924.411 BIC:-866.586 Adjusted $R^2 = 0.867$			

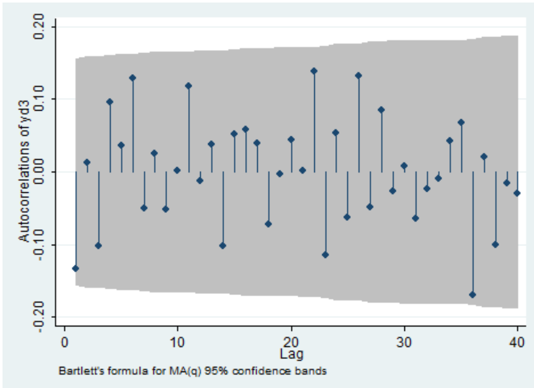

Figure 5-49 ACF of residuals from model 2

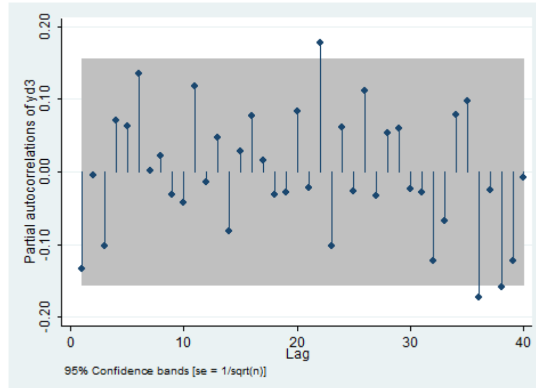

Figure 5-50 PACF of residuals from model 2

5.4.3. Seasonal Difference

ACF and PACF are plotted in Figure 5-51 and Figure 5-52. Figure 5-35 ACF supports a MA (1) model because the first lag is out of the confidence bands. Figure 5-36 PACF indicates that AR (1) model may be appropriate. When we approach the MA (1) model, the residual ACF and PACF suggests that both MA lag2 and AR lag2 should be selected. However, the coefficient of MA (2) and AR (2) are insignificant. Therefore, we have one candidate model and it is displayed in Table 5-18. Figure 5-53 and Figure 5-54 supports that AR (1) model is adequate

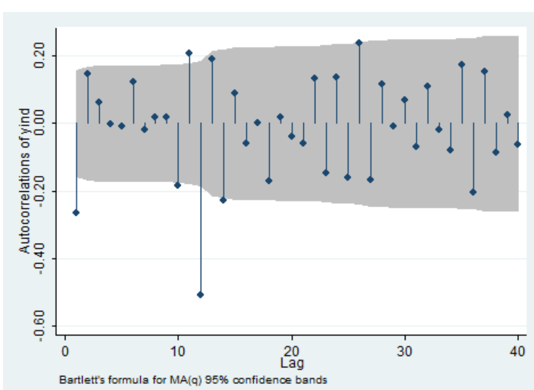

Figure 5-51 ACF of first and seasonal difference series

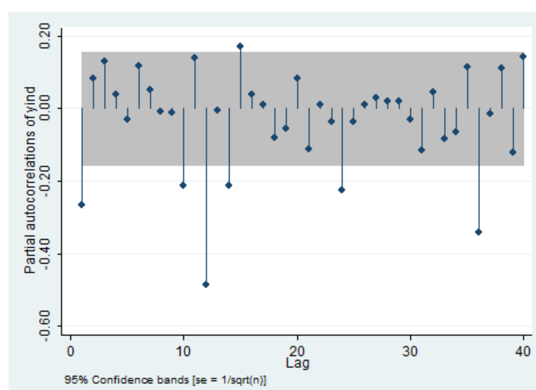

Figure 5-52 PACF of first and seasonal difference series

Table 5-19 Model 3: Seasonal Difference Fitting, Industrial Sector

Model 3			
	Coefficient	Standard Error	P>t
Cons.	0.000	0.001	0.795
AR.			
L1.	-0.258	0.089	0.004
AIC: - 775.458 BIC: -766.328			

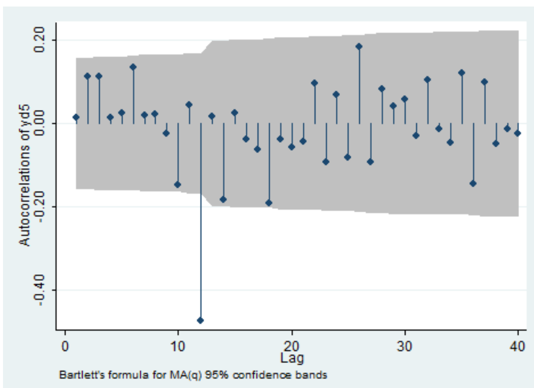

Figure 5-53 ACF of AR (1) model residual

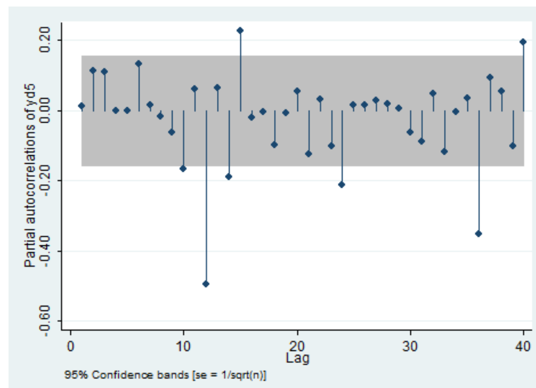

Figure 5-54 PACF of AR (1) model residual

Table 5-20 concludes all model diagnose test results. Shapiro-Wilk test result of Model 1 and model 3 reject the null hypothesis that the residual is normally distributed, however, the Ljung-Box test recommends that both the residual series is a white noise series. The Shapiro-Wilk test, the runs test, and Ljung-Box test all suggest that model 2 is appropriate model. The forecast result for difference of logarithm series is displayed in Figure 5-415, forecast of model 1 (a) and model 1 (b) are too close that the two forecast lines overlap. The forecast of model 1 (a) and model 1(b) in January are very close to the actual value while the forecast of model 2 are closer to actual value in February, March, and April. All model 1 (a), model 1 (b), and model 2 are recommended. Figure 5-56 displays the forecast results of first and seasonal difference series. Forecast has smaller variance than the actual value which makes the forecast looks flatter. We lack of evidence

to say that model 3 is a good forecast model for electricity sales in industrial sector and this is probably because that the seasonality in this sector is not as strong as other sectors we've studied.

Table 5-20 Model diagnose test results, Industrial Sector

Test	Model	P-value	Comment
The Shapiro-Wilk test	Model 1 (a)	0.048	Reject normality assumption
	Model 1 (b)	0.063	Residuals are normally distributed
	Model 2	0.407	Residuals are normally distributed
	Model 3	0.007	Reject normality assumption
The Runs Test	Model 1 (a)	0.190	Residuals are independent
	Model 1 (b)	0.490	Residuals are independent
	Model 2	0.470	Residuals are independent
	Model 3	0.940	Residuals are independent
Ljung-Box Test	Model 1 (a)	0.777	The model is appropriate
	Model 1 (b)	0.880	The model is appropriate
	Model 2	0.291	The model is appropriate
	Model 3	0.384	The model is appropriate

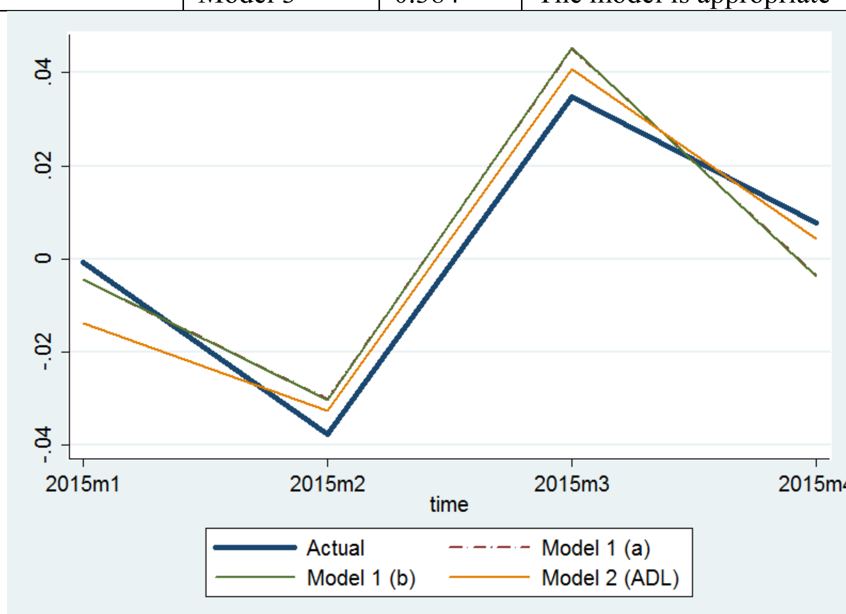

Figure 5-55Forecast of first difference of logarithm series, industrial sector

Figure 5-56 Forecast of first and seasonal different, industrial sector

5.5. Transportation Sector

5.5.1. Regression with dummy variables and ARIMA errors

The amount of electricity sales from transportation end-user sector is smallest among all sectors and retail sales in this sector has the smallest seasonal variance. The adjusted R^2 is 0.522, much less than adjusted R^2 of residential and commercial sectors. The result of model 1-1 is displayed in Table 5-20. The first spike in ACF (Figure 5-57) indicates that there may conclude a MA(1) error and there are 3 spikes in PACF (Figure 5-58) which suggest us to take AR lags errors into consideration.

Table 5-21 Result of Model 1-1, Residential Sector

	Coefficient	Standard Error	P>t
d1			
d2	-0.136	0.015	0.000
d3	-0.096	0.015	0.000
d4	-0.135	0.015	0.000
d5	-0.088	0.015	0.000
d6	-0.045	0.015	0.003
d7	-0.042	0.015	0.004
d8	-0.081	0.015	0.000
d9	-0.091	0.015	0.000
d10	-0.106	0.015	0.579
d11	-0.105	0.015	0.000
_cons	0.078	0.009	0.013
AIC: -498.786 BIC: -466.195 Adjusted R^2= 0.522			

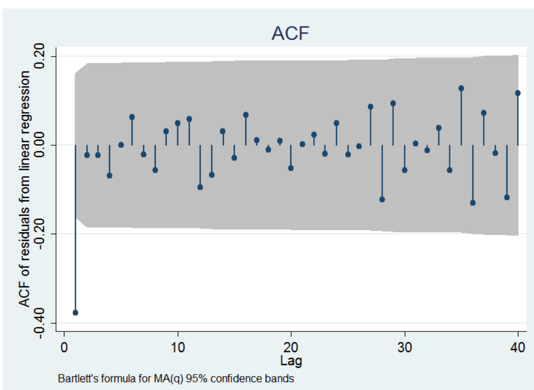

Figure 5-57 ACF of residuals from model 1-1

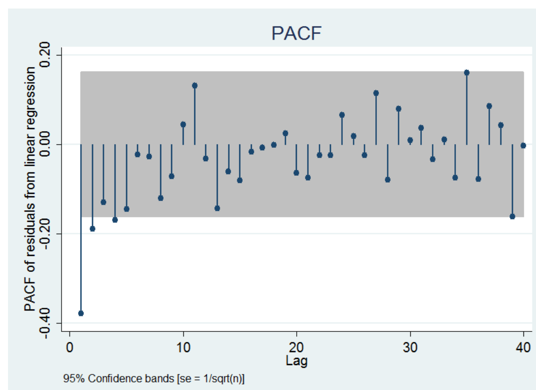

Figure 5-58 PACF of residuals from model 1-1

The ACF (Figure 5-59) and PACF (Figure 5-60) indicate that there is no correlation exist in model 1 with MA (1) error. However, when the AR (4), AR (3), and AR (2) all concludes insignificant lags, the residual ACF of AR(1) error still has a spike in order 2. Thus, the first model is a regression model with dummy variables and AR (1) and MA (2) errors. Along with MA (1) model, two models are displayed in Table 5-21. Coefficients of these two models are close and AIC and BIC value are also very similar. ACF and PACF are displayed from Figure 5-59 to Figure 5-62.

Table 5-22 Two candidate models fit model 1

Candidate Model 1 (a)				Candidate Model 1(b)			
	Coefficient	Standard Error	P>t		Coefficient	Standard Error	P>t
d1				d1			
d2	-0.142	0.012	0.000	d2	-0.141	0.012	0.000
d3	-0.098	0.011	0.000	d3	-0.098	0.011	0.000
d4	-0.137	0.014	0.000	d4	-0.137	0.014	0.000
d5	-0.090	0.013	0.000	d5	-0.090	0.014	0.000
d6	-0.046	0.014	0.001	d6	-0.046	0.014	0.001
d7	-0.044	0.020	0.027	d7	-0.044	0.020	0.027
d8	-0.083	0.016	0.000	d8	-0.083	0.016	0.000
d9	-0.093	0.022	0.000	d9	-0.093	0.022	0.000
d10	-0.108	0.020	0.000	d10	-0.108	0.020	0.000
d11	-0.103	0.017	0.000	d11	-0.103	0.018	0.000
_cons	0.080	0.005	0.000	_cons	0.080	0.005	0.000
				AR.			
				L1.	-0.573	0.076	0.000
MA.	-0.579	0.062	0.000	MA.			
L1.	-0.142	0.012	0.000	L2.	-0.348	0.080	0.000
AIC: -529.740 BIC: -491.223				AIC: -527.778 BIC: -486.298			

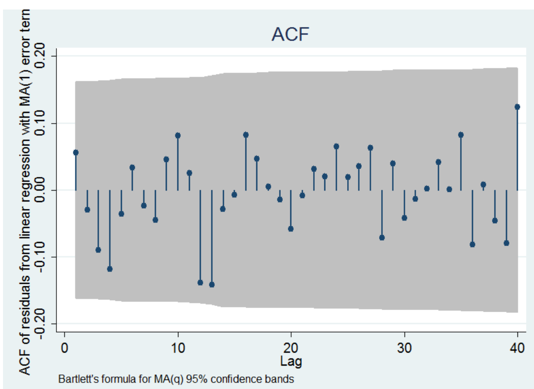

Figure 5-59ACF of residuals from model 1 (a)

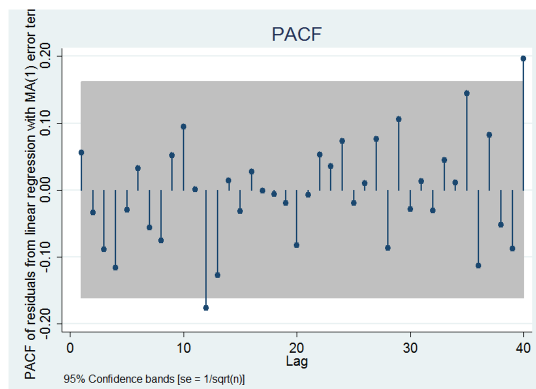

Figure 5-60 PACF of residuals from model 1(a)

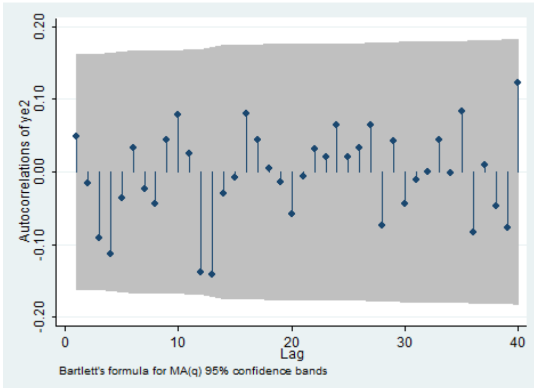

Figure 5-61 ACF of residuals from model 1 (b)

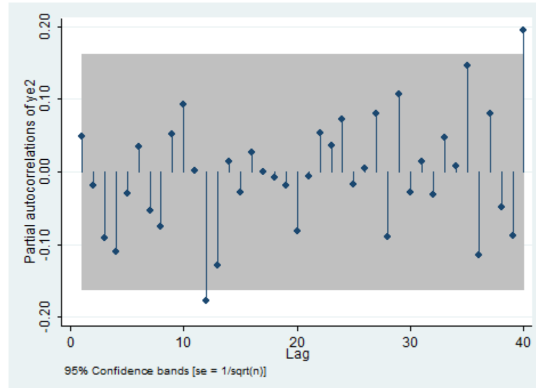

Figure 5-62 PACF of residuals from model 1(b)

5.5.2. ADL Model

The electricity sales in transportation sector has relatively weaker seasonality pattern. The adjusted R^2 falls dramatically to 0.637 which means dummy variables, price, population, IP index can only explain 63.7% of electricity sales in transportation sector. From Table 5-22, we know that statistically the sales is correlated to historical sales data in the short run, it is influenced by the electricity average price in six month and seven month ago. It is not correlated to population and there are no enough evidence to conclude the correlation between electricity sales to consumers in transportation sector and IP index. Figure 5-63 and Figure 5-64 shows that most of lags in ACF and PACF fall within the confidence boundaries.

Table 5-23 ADL model (model 2) fitting, Transportation Sector

	Coefficient	Standard Error	P>t
Retail Sales			
L1.	-0.430	0.075	0.000
L2.	-0.251	0.097	0.011
L3.	-0.197	0.098	0.046
L4.	-0.217	0.090	0.017
Dummy Variables			
d1			
d2	-0.103	0.018	0.000
d3	-0.097	0.021	0.000
d4	-0.088	0.021	0.000
d5	-0.109	0.016	0.000
d6	-0.081	0.018	0.000
d7			
d8	-0.092	0.018	0.000
d9	-0.066	0.016	0.000
d10	-0.089	0.016	0.000
d11	-0.098	0.014	0.000
Average Price			
L7.	0.217	0.080	0.008
L8.	0.226	0.081	0.006
Population			
IP Index			
Current	0.937	0.314	0.003
_cons	0.068	0.008	0.000
AIC: -502.418 BIC: -453.028 Adjusted $R^2 = 0.637$			

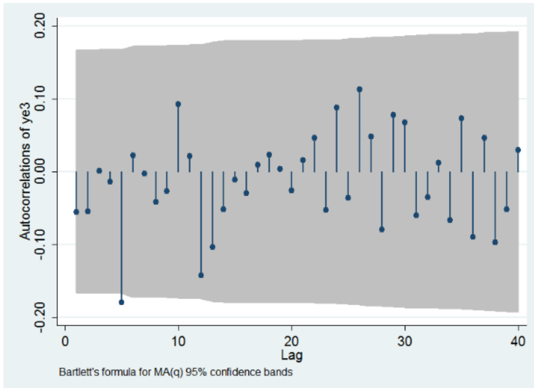

Figure 5-63 ACF of residuals from model 2

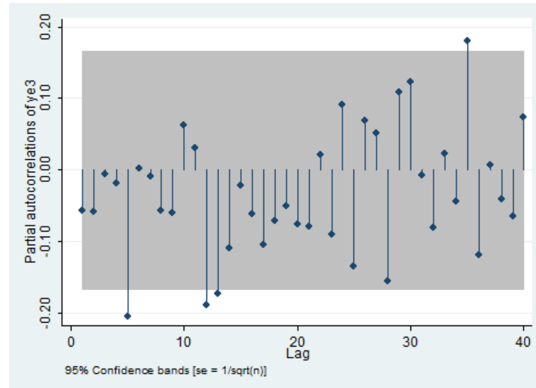

Figure 5-64 PACF of residuals from model 2

5.5.3. Seasonal Difference

After taking seasonal difference of log-transformed electricity sales data in transportation sector, the series is stationary. In sample ACF, the first three lags are out of the confidence bands and in sample PACF, the first lag is out of the confidence bands. The pure AR (1) model doesn't exist as the residual have several spikes in sample ACF and PACF. Except for lag 1 of AR (2) model, coefficient of lag 2 is insignificant. This happens to AR (3) and AR (4) models as well. The second lag is out of the confidence interval in both residual ACF and PACF, however, coefficients in ARMA (1, 1), ARMA (1, 2), ARMA (2, 1), ARMA (2, 2) are all insignificant. Therefore we only have MA (1) model in this sector for first and seasonal difference series. It is displayed in Table 5-23. Residual ACF and PACF displayed in Figure 5-67 and Figure 5-68 shows that most of the lags fall within confidence intervals.

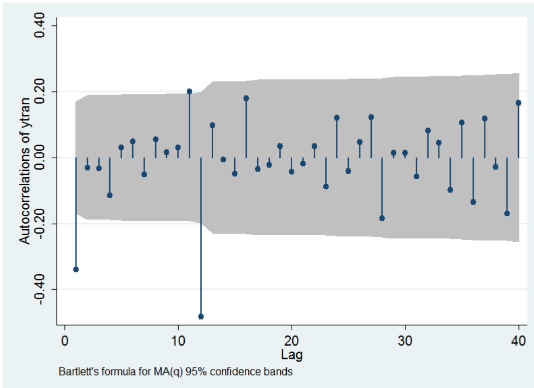

Figure 5-65 ACF of first and seasonal difference series

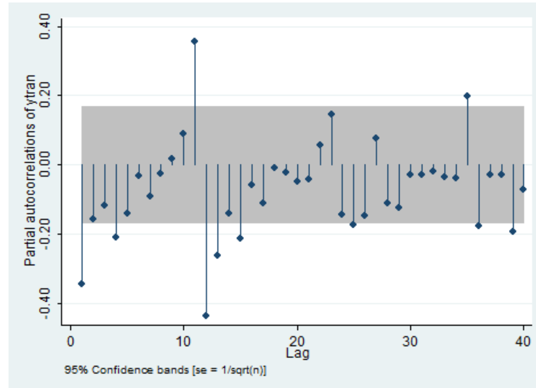

Figure 5-66 PACF of first and seasonal difference series

Table 5-24 Model 3: Seasonal Difference Fitting, Transportation Sector

Model 3			
	Coefficient	Standard Error	P>t
Cons.	0.000	0.0012	0.799
MA.			
L1.	-0.535	0.059	0.000
AIC: -396.001 BIC: -387.376			

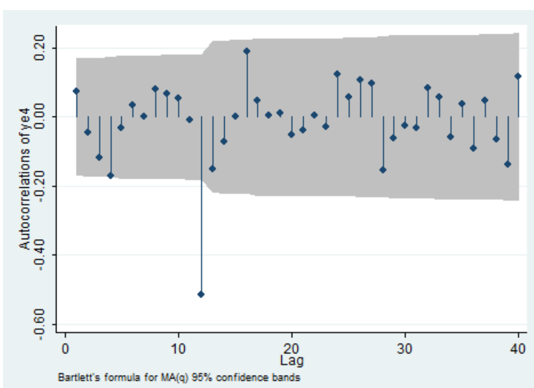

Figure 5-67 ACF of MA(1) model residual

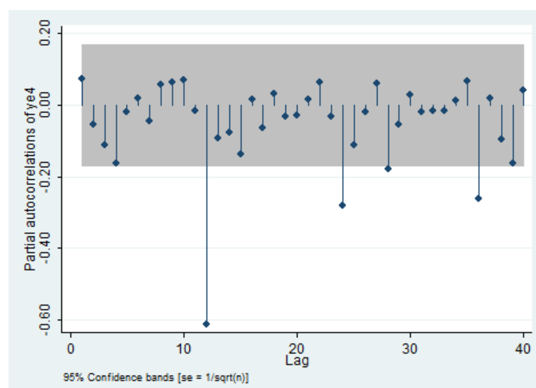

Figure 5-68 PACF of MA(1) model residual

Table 5-24 displays the tests results. All models past Shapiro-Wilk test which means each residual series is normally distributed. The runs test indicates each residual series is an independent series. The Ljung-Box test proves that each residual series is a white noise process. Therefore, all models are appropriate for modeling and forecasting electricity sales in transportation sector. The forecast displayed in Figure 5-69 shows that

the model 1 (a) and model 1 (b) get almost the same forecast. The forecast results of model 1 (1), model 1 (b), and model 2 are higher than actual value in January and March and forecast are smaller than actual value in February. In April, the model 1 (a) and model 1 (b) forecast is very close to the actual value while the model 2 forecast is larger than actual value. Forecast of first and seasonal difference is displayed in Figure 5-70. Model 3 forecast is higher than actual value in January and April while the forecast is smaller than actual value in the other two month.

Table 5-25 Model diagnose test results, Transportation Sector

Test	Model	P-value	Comment
The Shapiro-Wilk test	Model 1 (a)	0.502	Residuals are normally distributed
	Model 1 (b)	0.559	Residuals are normally distributed
	Model 2	0.112	Residuals are normally distributed
	Model 3	0.055	Residuals are normally distributed
The Runs Test	Model 1 (a)	0.680	Residuals are independent
	Model 1 (b)	0.800	Residuals are independent
	Model 2	0.430	Residuals are independent
	Model 3	0.540	Residuals are independent
Ljung-Box Test	Model 1 (a)	0.642	The model is appropriate
	Model 1 (b)	0.694	The model is appropriate
	Model 2	0.724	The model is appropriate
	Model 3	0.293	The model is appropriate

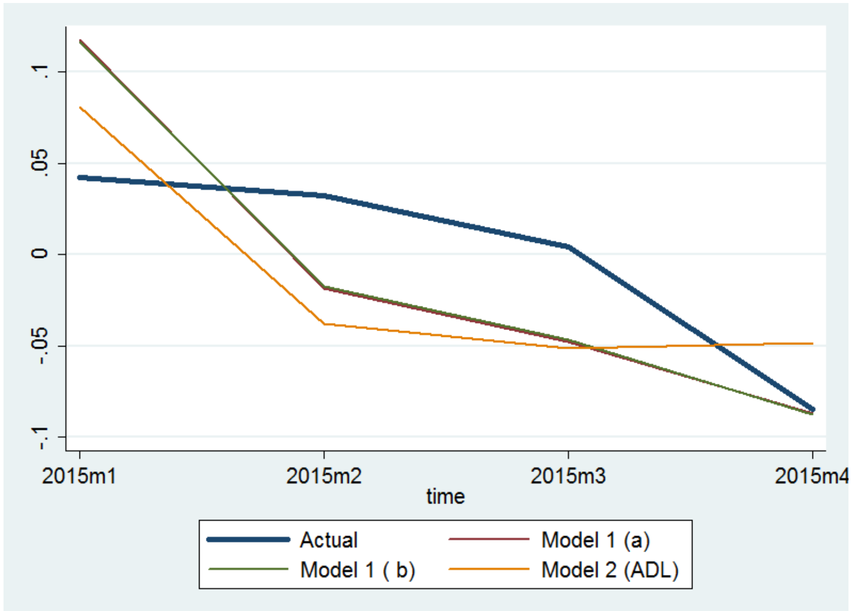

Figure 5-69 Forecast of first difference of logarithm series, transportation sector

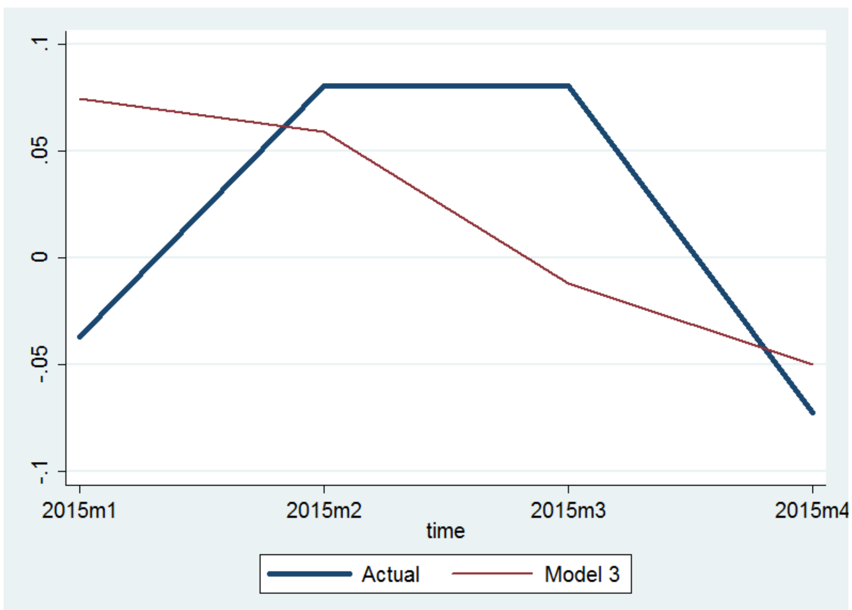

Figure 5-70 Forecast of first and seasonal different, transportation sector

6. Conclusion

To evaluate the forecast accuracy, one of the most widely used technique is mean absolute errors which is donated by $\text{MAE} = \text{mean}(|e_i|)$ where e_i is the error of actual value and forecast value. Table 6-1 displays the mean absolute errors results by each sector and it shows that in residential sector, the model 1(b) is the best. In residential sector, the ADL model (Model 2) is slightly better. In Commercial sector, the model with dummy variables and ARMA (1,1) disturbance is better. In industrial sector, the ADL model is the best. In transportation sector, the model 1(b) is the best. The result displayed in table 6-1 still suggests that forecast results of model1 and model2 are quite similar and comparable statistically, however, the autoregressive distributed lags model is more meaningful in economics field as it implies the how independent variables along with its past values are correlated with the dependent variable. In this scenario, the ADL model is recommended when two models have similar results.

Table 6-1 Mean absolute error by sectors

	Model	Mean Absolute Errors
The United States	Model 1 (a)	0.0204
	Model 1 (b)	0.0171
	Model 2 (ADL)	0.0208
Residential	Model 1	0.0419
	Model 2 (ADL)	0.0409
Commercial	Model 1	0.0056
	Model 2 (ADL)	0.0140
Industrial	Model 1 (a)	0.0079
	Model 1 (b)	0.0079
	Model 2 (ADL)	0.0069
Transportation	Model 1 (a)	0.0451
	Model 1 (b)	0.0443
	Model 2 (ADL)	0.0503

During past a few years, a large number of demand and demand elasticities of electricity has been studied based on time series approach. The major objective of this paper is to forecast the electricity sales using regression model with ARMA errors, autoregressive distributed lags model, and seasonal difference model. Then we compare these forecast results to the actual data. Another objective is to apply these models to different end-use sectors.

The results suggest that it may be wise to continue using the linear regression with ARMA errors approach in estimating electricity retail sales by sectors. In contrast to earlier studies, most of them only forecast the electricity demand of residential users, however, users in industrial sector and transportation sector have relatively less seasonality pattern.

The result also suggests us that it is meaningful and necessary to analyze and forecast electricity data by sectors. Customers in the United State as a whole lack of price elasticity and the population growth is significantly correlated to the electricity sales while the economy condition doesn't show significant influence on the electricity sales. This conclusion is similar with others' research we review in chapter 2 which suggests that the demand of electricity is very inelastic in the short run. In our case, the average price is omitted as none of coefficients of current and lagged variables are significant in 0.05 value. The population is correlated to electricity sales in the US in short run and long run. This is the similar result with Fung's (1993) work that it is reasonable to population as one of independent variables to forecast the electricity consumption. Similarly, IP index is correlated to electricity retail sales as we treat it as a replacement of GDP for monthly data.

Both Fung's work and Mohamed's work conclude that GDP is correlated to consumption of electricity.

This conclusion will be slightly different in different sectors. In residential sectors, consumers are relatively more sensitive to electricity price compared to the whole level. Population growth will affect residential electricity consumption in the long run and current IP index is correlated to the retail sales of electricity. In commercial sector, price variable is omitted as none of coefficients are significant. This suggests that consumers in this sector is very insensitive to price in short run and long run. This makes sense because electricity in this sector is majorly used for space heating and lightning. For example, business is unlikely to turn off lights or the air conditioners when the price of electricity goes up. Current population and its most recent four lagged value is correlated to the electricity sales Consumers in commercial sectors are influenced by earlier economy environment. In industrial end-use sector, population growth has little correlation with electricity sales. Consumers are price sensitive in both the short run and long run. The current and economy condition in short run are significantly related to retail sales of electricity. As the major use of electricity in transportation sector is for electrical trains, the sales in this sector is correlated to average price in long run and current IP Index.

All three models are very comparable. The forecast results of linear regression with ARMA model incorporating dummy variables and ADL model are very similar, the former models without economic explain variables in the United States, in the commercial sector, and in transportation sector have better results than the latter one. ADL model gives us idea about the relationship between electricity sales and other economic variables. This paper

mainly attempts to discuss appropriate forecast models that capture the seasonal cyclical of electricity sales. The seasonal difference model could also be used when the time series data has other cyclical period rather than 12. It is effectively performed and implies its significance in forecasting time series data with cycles such as electricity retail sales data. Comparisons have been made with the actual retail sales by each sector. Mean absolute errors is used to evaluate the forecast accuracy. Every comparison made by all these three models are vary comparable with actual values. Average price of electricity, population, Industrial Production Index are proved to be correlated to electricity sales in the United States.

7. Reference

Akay, D., & Atak, M. (2007). Grey prediction with rolling mechanism for electricity demand forecasting of Turkey. Energy, 32(9), 1670-1675.

Beenstock, M., Goldin, E., & Nabot, D. (1999). The demand for electricity in Israel. Energy Economics, 21(2), 168-183.

Bentzen, J., & Engsted, T. (1993). Short-and long-run elasticities in energy demand: a cointegration approach. Energy Economics, 15(1), 9-16.

Bentzen, J., & Engsted, T. (2001). A revival of the autoregressive distributed lag model in estimating energy demand relationships. Energy, 26(1), 45-55.

Contreras, J., Espinola, R., Nogales, F. J., & Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. Power Systems, IEEE Transactions on, 18(3), 1014-1020.

Cryer, J. D., & Kellet, N. (1986). Time series analysis (Vol. 101). Boston: Duxbury Press.

Egelioglu, F., Mohamad, A. A., & Guven, H. (2001). Economic variables and electricity consumption in Northern Cyprus. Energy, 26(4), 355-362.

Erdogdu, E. (2007). Electricity demand analysis using cointegration and ARIMA modelling: A case study of Turkey. Energy policy, 35(2), 1129-1146.

Fung, Y. H., & Tummala, V. R. (1993, December). Forecasting of electricity consumption: a comparative analysis of regression and artificial neural network models. In Advances in Power System Control, Operation and Management, 1993. APSCOM-93., 2nd International Conference on (pp. 782-787). IET.

Hamilton, J. D. (1994). Time series analysis (Vol. 2). Princeton: Princeton university press.

Harris, J. L., & Liu, L. M. (1993). Dynamic structural analysis and forecasting of residential electricity consumption. International Journal of Forecasting, 9(4), 437-455.

Harvey, A. C., & Harvey, A. C. (1993). Time series models (Vol. 2). New York: Harvester Wheatsheaf.

Silk, J. I., & Joutz, F. L. (1997). Short and long-run elasticities in US residential electricity demand: a co-integration approach. Energy Economics, 19(4), 493-513.

Lakhani, H. G., & Bumb, B. (1978). Forecasting demand for electricity in Maryland: an econometric approach. Technological Forecasting and Social Change, 11(3), 237-259.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. Biometrika, 65(2), 297-303.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. Journal of econometrics, 54(1), 159-178.

Kucukali, S., & Baris, K. (2010). Turkey's short-term gross annual electricity demand forecast by fuzzy logic approach. Energy Policy, 38(5), 2438-2445.

Mirasgedis, S., Sarafidis, Y., Georgopoulou, E., Lalas, D. P., Moschovits, M., Karagiannis, F., &

Mohamed, Z., & Bodger, P. (2005). Forecasting electricity consumption in New Zealand using economic and demographic variables. Energy, 30(10), 1833-1843.Papakonstantinou D. (2006). Models for mid-term electricity demand forecasting incorporating weather influences. Energy, 31(2), 208-227.

Pappas, S. S., Ekonomou, L., Karamousantas, D. C., Chatzarakis, G. E., Katsikas, S. K., & Liatsis, P. (2008). Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models. Energy, 33(9), 1353-1360.

Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. Biometrika, 75(2), 335-346.

Ranjan, M., & Jain, V. K. (1999). Modelling of electrical energy consumption in Delhi. Energy, 24(4), 351-361.

Ringwood, J. V., Bofelli, D., & Murray, F. T. (2001). Forecasting electricity demand on short, medium and long time scales using neural networks. Journal of Intelligent and Robotic Systems, 31(1-3), 129-147.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 591-611.

Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A comparative study of various tests for normality. Journal of the American Statistical Association,63(324), 1343-1372.

Stock, J. H., & Watson, M. W. (2003). Introduction to econometrics (Vol. 104). Boston: Addison Wesley.

Wiener, N. (1949). Extrapolation, interpolation, and smoothing of stationary time series (Vol. 2). Cambridge, MA: MIT press.

Yan, Y. Y. (1998). Climate and residential electricity consumption in Hong Kong. Energy, 23(1), 17-20.

Yang, M., & Yu, X. (2004). China's rural electricity market—a quantitative analysis. Energy, 29(7), 961-977.

Weron, R., Kozłowska, B., & Nowicka-Zagrajek, J. (2001). Modeling electricity loads in California: a continuous-time approach. Physica A: Statistical Mechanics and its Applications, 299(1), 344-350.