WTI OIL PRICE PREDICTION MODELING AND FORECASTING

A Project

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

In

Economics

By

Halleh Bostanchi

2017

SIGNATURE PAGE

PROJECT: WTI OIL PRICE PREDICTION MODELING AND FORECASTING

AUTHOR: Halleh Bostanchi

DATE SUBMITTED: Fall 2017

Department of Economics

Dr. Carsten Lange _____
Project Committee Chair
Economics

Dr. Bruce C. Brown _____
Economics

Dr. Kellie Forrester _____
Economics

ACKNOWLEDGMENTS

I would like to thank everyone who helped me throughout many years, and made my education and eventually this thesis possible.

First of all I would like to thank my parents for their unconditional love and support. I am forever indebted to them. I also would like to thank all my teachers during my many years I have been attending school, particularly my professors at California Polytechnic University at Pomona. Next, I would like to thank the love of my life and my husband, Saeed, who supported me on every single step of my post-secondary education. Furthermore, I would like to thank my brothers and extended family for their continuous support and inspiration.

I would like to dedicate this work to all my loved ones in particular to my daughter, Sophia, whom I was pregnant with when working on and writing this thesis.

ABSTRACT

This work examines two different Baysian approaches to model short term oil price return for past decades and forecast it. We first built the multivariable linear regression model based on relevant explanatory variables. Then we build the univariate time series model using ARIMA models, followed by ARCH and GARCH models. Both methods are followed by required procedures and econometrics tests. The forecasting powers of time series approach perform better than linear regression and even structural models, yet linear approach is very relevant for knowing incapability of each variable to oil price.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Oil is one of the most important commodities that its price and volatility has huge impact on everyoneś life on the planet. Impact of oil price and its applications on daily life is undeniable, from all transportations, including flights, cars, and trains to daily consumer products such as tires, shampoo, paint and many more products. Oil is the major source of heating and energy in the world, which makes it challenge to replace it with other resources. Oil heavily effect the economic growth, from common consumer products to military and energy sectors. Unexpected movements on oil price has effects on economic stability for both supplier and producer countries, although more crucial for oil importing countries. Several studies concluded that it is harmful when the oil price has more volatility and is less predictable, which can have negative effect on many economic indicators (Bosler, 2010; LAM, 2013; Moshiri and Foroutan, 2006) .

It is beneficial to have accurate oil market forecasts in many sectors of economy. For instance both central banks and private sectors are using these forecasts in many cases to generate macroeconomics plans and also measuring risks. Some sectors are directly depend on these forecasts such as most of transportation manufacturer, utility companies

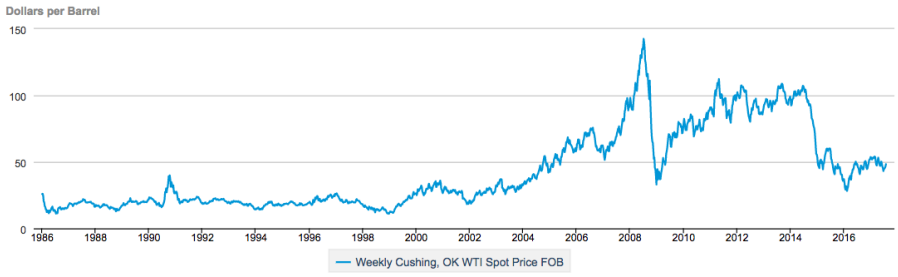

Figure 1.1: Historical price of crude oil, WTI futures (USDOE, 2017).

and even homeowners are all relay many of their decisions based on the oil prices (Ron, Lutz and Robert, 2011).

Modeling the price of oil is difficult, because of many fluctuations over time. As you can see in this historical graph shown in Figure 1.1, the oil price can dramatically change over short time, which makes it very difficult to predict. in past decade oil prices has fluctuated a lot,from $40 in 2003 to $140 in 2008. Oil demand and supply are quite inelastic in short term, that makes the price skyrocketed when the demand for oil exceeds supply. The unique feature of this market is that its supply is limited as cannot be renewed.

Furthermore, oil price is also heavily affected by political turbulences, which makes it even more difficult to predict oil price. For instance, in 1999 the Asian Financial Crisis (plummeting demand) and Iraq's decision to increase oil production (increased supply) caused oil prices to reach below $20 levels. The Dotom bubble in 2001 caused another round of economic panic, which caused oil price to stay low until early 2002. On the other hand and more recently due to financial crisis market volatility, oil price skyrocketed to $147 that shows how unpredictable price of oil is. Another challenge is organization of the petroleum exporting countries (OPEC) that is acting as a swing producer, covering the remaining demand after non-OPEC supply is used up, and hence takes the market price. However when OPEC produces at full capacity, it will become price maker as it will have competitive role in global market (Bosler, 2010).

Chapter 2

Literature Review

Modeling the price of oil is difficult, because of many fluctuating variables over time. As can be seen in Figure 1.1, the oil price can dramatically change over a short time, which makes it very difficult to predict. Oil demand and supply are quite inelastic in short time, meaning when the demand for oil exceeds supply, price will rise extremely high. On the other hand oil price is affected heavily by political turbulences. Such as 1999 the Asian Financial Crisis and Iraq deciding to increase oil production, which caused oil prices to reach a bottom (LAM, 2013). More recently and during 2008 financial crisis, market volatility skyrocketed to $147 and dropped to below $40 levels in less than a year, showing how difficult predicting oil price can be (USDOE, 2017).

In general there are three different approaches in forecasting oil prices: long term, medium term and short term (Alquist, Kilian and Vigfusson, 2011).

The long term forecasting looks into market for decades to come and is most used by central banks and governments to make macroeconomic policies. However, it requires so many variables and models that make it difficult to be implemented by individuals and more is being done by government organizations. The Energy Information Admin-

istration (EIA) of department of energy (DOE) have developed national energy model system (NEMS) tool for long-term forecasting of various energy related factors related to U.S. energy, including production, consumption, pricing, etc. The goal of this tool is to project national-level energy market in decades to come. Originally developed in 1993, this model has been used in recent annual energy outlook to predict a comprehensive list of energy related factors up to 2050 (DOE, 2017*a*). This computer-based model is based on macroeconomic and financial models and also includes many inputs and assumptions. It consists of many integrated modules that interact with each other as part of an equilibrium calculation (DOE, 2017*a*).

Medium term oil forecasting models focus on few years window. Central banks also use medium term models of oil price forecasting for Macroeconomic decision making. There most popular models to predict medium term oil price is Vector Autoregressive (VAR) models. The VAR model generally have high accuracy and the lower mean-square predication error with random walk for forecast horizon up to two years. The International Monterey Fund (IMF) working paper used VAR models to forecast the nominal price of oil benchmarks such as Brent and WTI instead of real oil prices. Recently there has been researches that used real time VAR models to forecast real price of oil for one year horizon. Baumeister and Kilian research has shown that the real price forecasts are more accurate than the forecasts based on future prices. Quarterly vector autoregressive models forecast estimate on monthly data. There are different approaches for the quarterly forecast. One way is to forecast the monthly real price of oil for each month and then convert them to quarterly average based on (Baumeister and Kilian, 2014).

This project focuses on the short term forecasting of future oil prices, specifically WTI crude oil. Forecasting such unpredictable economic series is stayed one of the main challenges for econometricians. In the literatures, there are several different models used

to predict and forecast short term oil prices. Historically linear structural models have not performed well for oil price forecasting and nonlinear time series models have performed much better in forecasting oil prices (LAM, 2013),(Bosler, 2010) and (Moshiri and Foroutan, 2006). F. Bosler examined a time series approach, which includes linear and nonlinear time series analysis and also structural models (LAM, 2013). He compared linear ARIMA model and neural network autoregressive model for nonlinear time series analysis and confirmed that thenonlinear models forecasts perform the better and follow the volatility of the oil price.

In another work, D. Lam modeled oil prices based on univariate time series using the Box-Jenkins methodology. Based on the ACF and PACF techniques ARIMA model was chosen, and followed by GARCH and APARCH as model residuals (Bosler, 2010). He also built a regression model to compare with his nonlinear model. The regression model was based on eight explanatory variables, including production, consumption, net import, ending stock, refinery utilization rate, U.S. interest rate, NYMEX oil futures contract 4 and S&P 500 index. But at the end the conclusion was that GARCH and APARCH perform the best.

S. Moshiri et. al. furthermore proposed another nonlinear model to forecast daily crude oil futures prices from listed in NYMEX. They used a nonlinear and flexible artificial neural network (ANN) model to forecast the series and concludedit will improve forecasting accuracy. This work claims that "If the data generating process is nonlinear, applying linear models could result in large forecast errors. Model specification in nonlinear modeling, however, can be very case dependent and time-consuming" (Moshiri and Foroutan, 2006).

A. Shabri et. al. have proposed an even more complicated forecasting model based on integrating wavelet transform and artificial neural network (WANN) (Shabri and Sam-

sudin, 2014). They decompose the price structure to various wavelet components and perform ANN model separately on each element. The conclusion was that WANN model provide better prediction for crude oil spot prices at lead times of 1 day for West Texas Intermediate (WTI) and Brent crude oil.

Chapter 3

Oil Price Prediction

This Project focuses on following models and compares their accuracy. It also implements some of these models to predict price of W&T Offshore index (WTI)

1. Structural models
    - Depends on fundamental data such as demand and supply
    - Implemented through the use of a linear regression
2. Time series-based models
    - Linear time series analysis, such as ARIMA
    - Nonlinear time series analysis, such as ARCH, GARCH
    - Autoregressive neural network (ANN) model

3.1 Data

The data for this project is taken from US Department of Enerdy, energy information administration (EIA) independent statistics and analysis (DOE, 2017*b*; USDOE, 2017),

where it has variety of information available. This includes prices, production levels, supplies, imports and many other relevant statistics to Cushing, OK west Texas intermediate (WTI) spot price freight on board (FOB).

For the use in structural modeling, this project will focus on the monthly oil related indicators such as US percent utilization of operable capacity, field production of crude oil, oil and petroleum imports, supplied products, crude oil future pricing, as well as US economy indicators such as SP500 stock price and ten year treasury constant maturity rate (DGS10) (figure 3.1 and figure 3.2). The data is available from EIA web site in Excel format, which is compatible with "read.csv" command in R for analysis (DOE, 2017*b*).

For the use in the time series model, we will only use the historical WTI oil price that was shown in figure 1.1. This data was also taken from EIA web services (DOE, 2017*b*).

3.2 Modeling

Oil price is predicted by two main approaches: Structural model and time series model. In structural model, the price is forecasted using many explanatory variables, such as demand and supply, consumption, production, available US stock of oil, net US oil imports, etc, as discussed by (LAM, 2013). The price is predicted using linear regression models and will predict with mean square error or mean absolute error at the end . On the other hand, the time series model can predict the price of oil, solely using the past price trend. Time series models predict using either linear (ARMA) or non-linear trends (GARCH or ANN) using the past time series rates. This project shows how to use first three models, meaning linear regression, ARIMA and GARCH models, to forecast WTI index price since 2006. Because of the complexity of ANN models, it will not be investigated in this

project.

3.2.1 Structural Models

Structural modeling approach uses explanatory variable and determines the statistical relation to response variable and make predictions. So let‘s start with the structural model. As mentioned this model uses co-integration and correlation between different time-series explanatory variables. It uses multivariable linear regression model, to find the best fit. It assumes that spot price of oil is linearly correlated with each of historical value of oil price and the explanatory variables. For this linear regression, different time

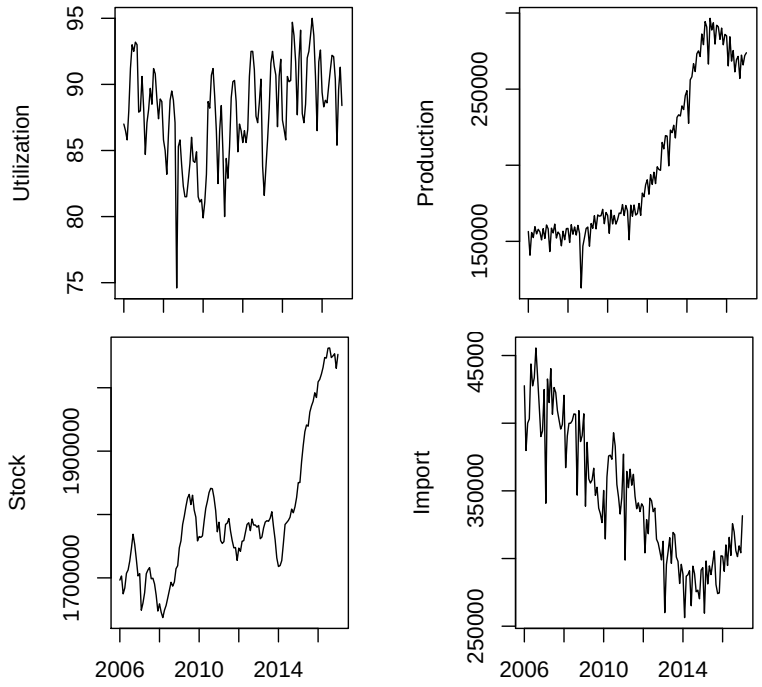

Figure 3.1: Historical U.S. oil related variables used in structural modeling: top left: utilization (%), top right: production (thousand Barrels), bottom left: stock price per thousands Barrels, and bottom right: imports (thousand Barrels) (USDOE, 2017).

spans can be regressed. Hence depending on available data, different time spans for linear regression is assumed. This section provides 12 month oil price prediction by considering 13 to 120 months of data.

In order to write regression formalism, we assume the price of oil at each given time period n is given by y_n and presented by $n \times 1$ matrix Y. At each given time x_{pn} refers to value of explanatory variable p at period n and presented by $p \times n$ matrix X. Now at a given time n the price of oil can be linearly regressed using the the explanatory variables at time $n\prime$ as (LAM, 2013),

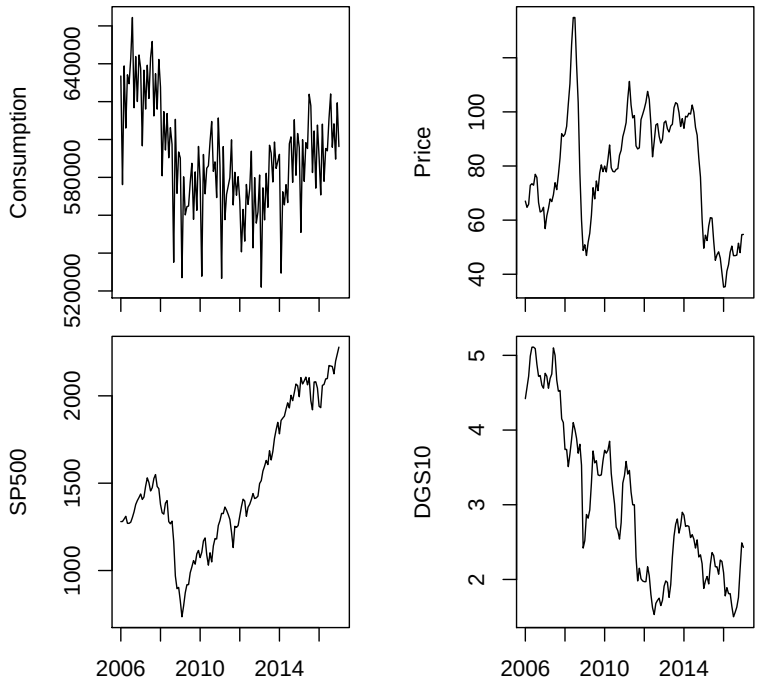

Figure 3.2: Historical variables used in structural modeling: top left: consumption (thousand Barrels), top right: crude future price (per Barrel), bottom left: S&P500 closing price, and bottom right: ten year U.S. treasury constant maturity rate (%).

$$y_n = \beta_0 + \sum_{i=1}^{n\prime} \beta_i x_{in\prime} + \varepsilon_n. \tag{3.1}$$

The set of regression equations in equation 3.1 for span of n events, can be summarized in matrix form as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{p,1} \\ 1 & x_{1,2} & \dots & x_{p,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n-1} & \dots & x_{p,n-1} \\ 1 & x_{1,n} & \dots & x_{p,n} \end{pmatrix} . \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{pmatrix}, \tag{3.2}$$

or

$$Y = X\beta + \varepsilon. \tag{3.3}$$

In equation 3.2, n refers to number of spot prices considered, and p refers to number of explanatory variables. First column of matrix X has unity values to denote the intercept of each line at given time. Matrix β will is solution matrix, where β_0 refers to expected value of fitted intercept and β_p refers to expected value of defined slope for explanatory variable p. Matrix ε contains error values ε_n, which is the difference between real and fitted prices at time n (LAM, 2013).

The formalism listed above in 3.3 regresses the price of oil to the existing known events if n and $n\prime$ refer to the same time period. However in this work, in order to obtain a prediction to regress the price for $\kappa = 12$ steps in future, we shift the time span of n and $n\prime$ for this amount. Once we find the solution matrix β regressing price of today based on the data from κ steps in past, we apply it to current explanatory variable matrix to predict the price for κ steps in future.

The fitting can be optimized either using mean square error or mean absolute error (ε) methods. The expected value of solution of equation 3.2 is extracted as (LAM, 2013)

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \tag{3.4}$$

To model the WTI price, this project uses seven different variables. The data for these variables are extracted from US DOE EIA web site (USDOE, 2017).

- First is the utilization of the refineries that represents the utilization of all crude oil distillation units. It is calculated by dividing gross inputs to these units by the operable refining capacity of the unit, which is in turn defined as the amount of capacity that, at the beginning of the period, is in operation, or those which are not in operation and not under active repair, but capable of being placed in operation within 30 days. Utilization shows seasonality.

- Second explanatory variable is oil production, which ramps from 2012 to 2015, with a small drop after late 2015.

- Third variable is US oil stocks, which shows a ramp after 2014.

- The fourth variable is the oil import, which shows some small seasonality and has been decreasing since 2006 till 2014 and has remained low since then.

- The fifth parameters to predict price of WTI index using regression model is oil consumption, which has not changed much over time, due to oil‘s inelastic nature.

- Next variable is SP500 index prices as a measure of the market strength.

- The last variable is US treasury 10-year yield bonds as a measure of the economic strength.

The last two parameters are more economic variables and not oil related. They are intended to correlate the oil price to the economy. The data for each of these variable is depicted in figure 3.1 and figure 3.2. All oil variables are per thousand barrels per time period.

There are few important conditions for this model that has to be verified (LAM, 2013). First, absence of multicollinearity has to be met among explanatory variables. This means that parameters used for prediction should have small correlation of < 0.75. Table 3.1 summarizes the cross correlation between all 7 variables plus oil price itself (8 variables in total). As you can see all the selected variables satisfy the low cross correlation condition of being < 0.75 and hence the multicollinearity condition has met.

Table 3.1: Cross correlation between different variables affecting oil prices

	Utilization	**Production**	**Stock**	**Import**	**Consumption**	**SP500**	**Price**	**DGS10**
Utilization	1							
Production	0.28	1						
Stock	0.182	0.508	1					
Import	-0.096	-0.636	-0.380	1				
Consumption	0.253	-0.065	-0.093	0.328	1			
SP500	0.375	0.529	0.274	-0.396	0.128	1		
Price	0.102	-0.056	-0.324	-0.003	-0.201	-0.098	1	
DGS10	0.104	-0.562	-0.437	0.513	0.282	-0.283	0.059	1

The next condition is that the variables should also have constant variance in residual, which is also known as homoscedasticity. Uncorrelated errors assumption is the errors of the response variables are independent of each other. The errors must have normal distributinos, which can be checked by QQ plots test for normality of error. Lam (LAM, 2013) used this model to predict three points ahead, but here up to 12 point ahead pre-

diction is used. Later, we will see that time series methods can predict way more points than this and the result could be different with using different variables for regression.

If we load the matrix for the time series data in the matrix X in equation 3.2, the solution matrix β only provide the best linear regression results for the present data. In order to use this model to predict the price in future, the oil price matrix (Y) is shifted for amount of time in future that we would like to predict and the matrix X gets lagged from matrix Y for the number of points we would like to predict. Then solving equation 3.2 for present values of X in equation 3.2 results in Y matrix that predict the oil price in future.

To do this, there are two critical numbers that can change the prediction results. first one is span of time we choose for fitting, and the second is number of time spans we would like to predict in future. Figure 3.3 shows four different fitting conditions for 12 month oil price prediction, fitting data for consecutive 13, 24, 48 and 120 months. This graph shows the market WTI oil price as thin solid lines, estimated oil price as thick solid lines and estimated price as dashed lines.

As seen in Figure 3.3, the price prediction can vary significantly, changing from \$40 to \$80. In order to find the best fit, this project minimizes minimum least square error (MLSE), which is defined as

$$MLSE = \sqrt{\sum_{i=1}^{n} (Y - \hat{Y})^2}, \tag{3.5}$$

where n is number of time spans and $\hat{Y}$ is fitted price numbers. The results of calculated $MLSE$ value for fitting with various fitted time spans is shown in figure 3.4. As can be seen for the set of data we described above, the best fit is achieved by fitting 8 spans. Any fitting considering less spans results in very large error. Above this limit the error increases with square root of number of samples. Hence for the fitting of price of oil,

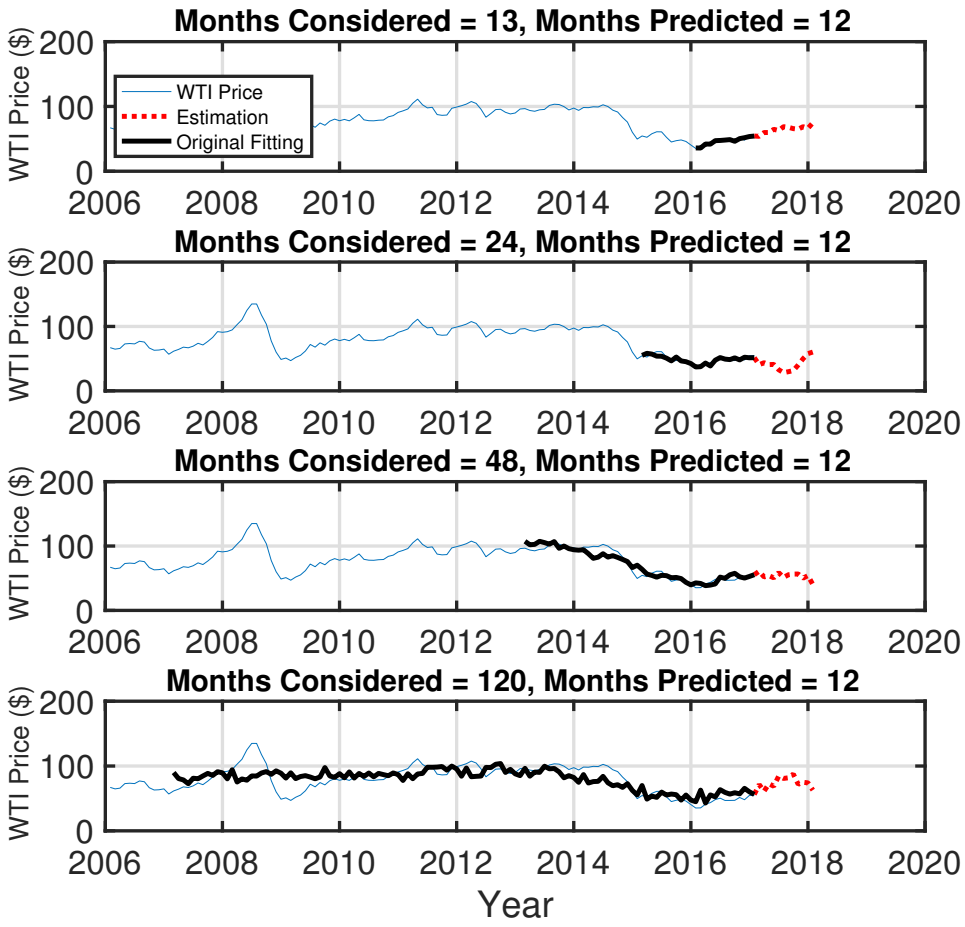

Figure 3.3: WTI monthly crude oil price prediction for 12 months shown in dotted lines and market data shown since 2006 in thin solid lines on each panel. Thick solid lines shows the fitted model that is based on considering various months for each panel, including 13, 24, 48 and 120 months.

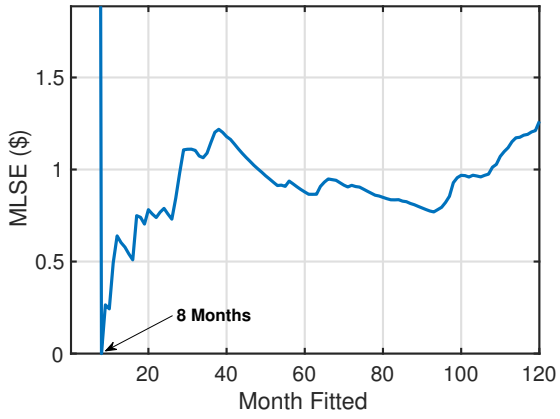

Figure 3.4: Minimum least square error for fitting WTI crude oil data for various spans of months considered. Fitting data for less than 8 month results in very large error.

fitting using 9 spans is used to predict price of oil for the next 8 months. Figure 3.5 shows the result of such fit.

Figure 3.5: WTI oil price prediction for next 8 months (thin lines), assuming fitting based on 9 consecutive months. Thick line shows historical WTI prices.

3.2.2 Box Jenkins Approach (ARIMA)

The next two methods for forecasting are using the time series approach. The structural model that was discussed before can only work, when there is less volatility in the data trend. But this can not be the case all the time. For example WTI oil price went through big volatility n 2009. Time series models are considered as they can deal with higher levels of volatility, particularly for financial time series with inherent time-varying volatility clustering. These models use auto-regression and moving averages of past events and past returns, which are processed to predict the future events (Enders, 2015; Box, Jenkins and Reinsel, 1994; Brockwell and Davis, 2006).

As mentioned, in the time series approach, we will use the similarity in trends of the past prices and returns and replicate it in future using auto-regressive approaches. The simplest and most common time series fitting approach is Box.Jenkins also known as ARIMA. ARIMA stands for autoregressive integrated moving average model and is a linear model. ARIMA can predict non-stationary time series, by multiple integration and converting them to stationary. In another word ARIMA model is integrated ARMA model, which is combination of moving average process with linear difference to get autoregressive moving average (Enders, 2015).

The ARIMA(p, d, q) parameter model for variable Y_t is shown in equation 3.6 and it is consist of p: autoregressive parameter, d: number of differencing, q: moving average parameters (Enders, 2015; Box, Jenkins and Reinsel, 1994; Brockwell and Davis, 2006).

$$Y_t = \phi_0 + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{3.6}$$

Here ϕ_is are auto regressive "AR" terms and represent lags of stationarized series and θ_js are moving average "MA" terms and represent lags of the forecast errors. Here εs represent a white noise process, with zero mean, zero correlation across time and inde-

pendent variables. ARMA model requires stationary series, which means constant mean and variance. Non-stationary series can become stationary and use ARIMA by taking d^{th} order difference. Box-Cox transformation is often used for this purpose. Usually Log or log-difference are used, which is nothing but law of return. This work uses R-package functions to calculate fitting ARIMA parameters (Enders, 2015; Box, Jenkins and Reinsel, 1994; Brockwell and Davis, 2006).

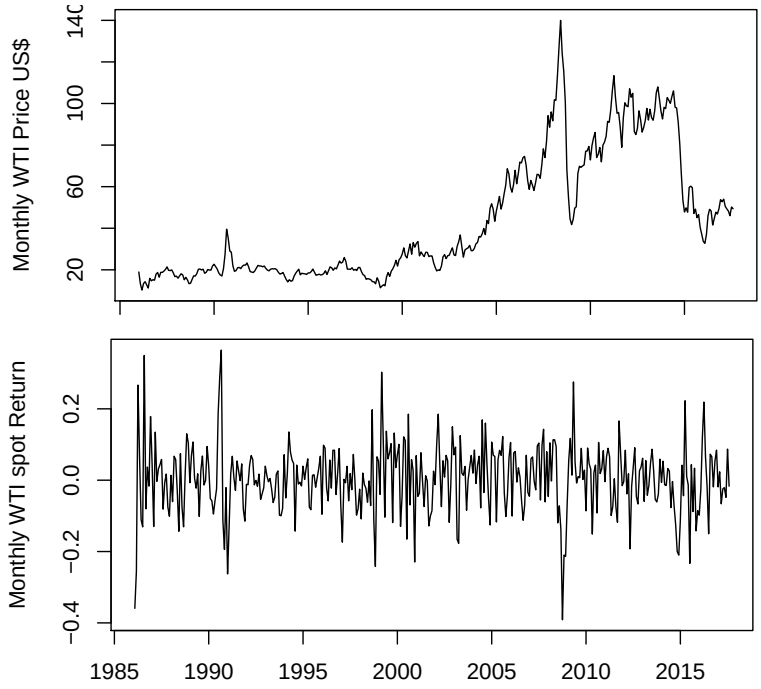

Figure 3.6: Top: Historical U.S. WTI stock price per thousands Barrels since 1986. Bottom: Log-return of oil price.

WTI monthly data since 1986 are considered in this section. The top panel of figure 3.6 shows linear WTI price (USDOE, 2017). The bottom panel of figure 3.6 shows logarithm of spot returns that is calculated by $Log(p_i/p_{i-1})$. In order to verify and test stationary nature of each of the data shown in figure 3.6, ACF and PACF tests was performed for both linear and log-return time series (Enders, 2015). First the results of this

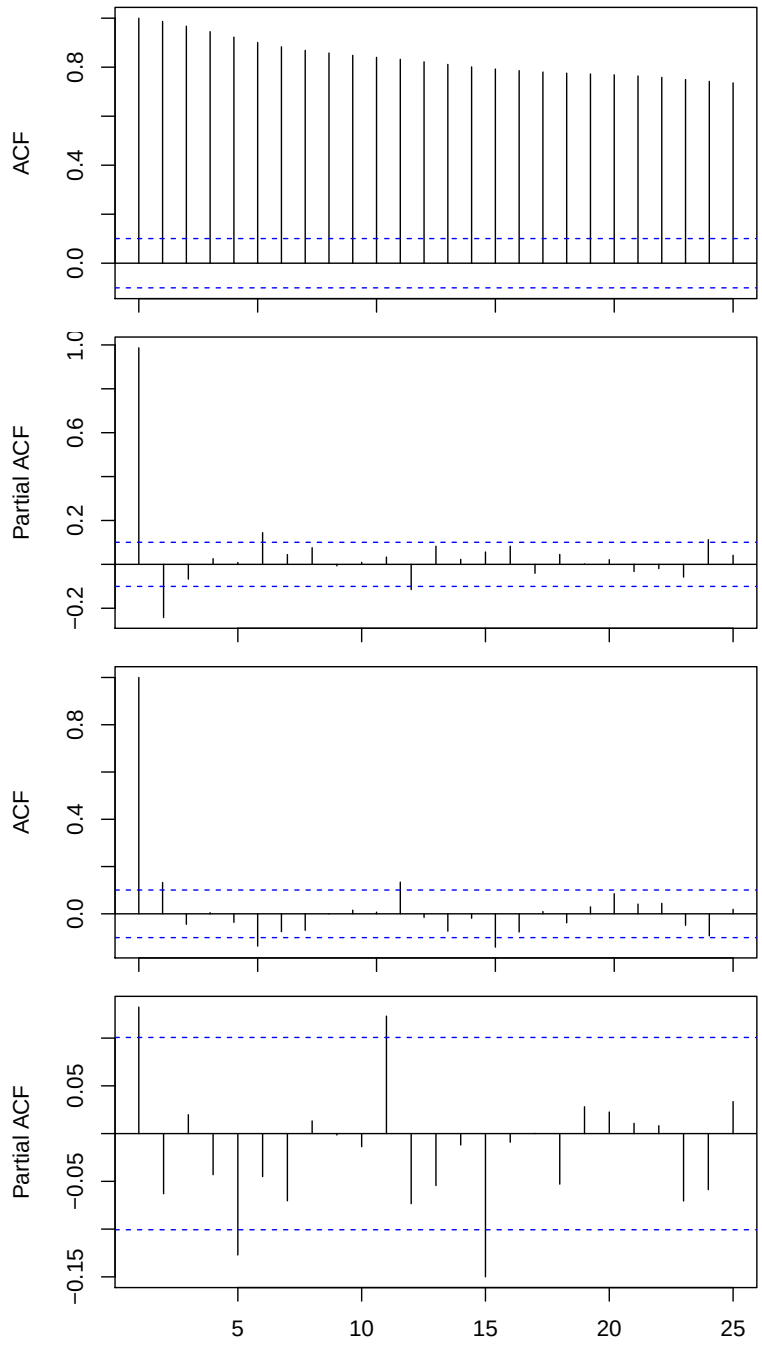

Figure 3.7: Top: ACF and second form top: PACF of WTI oil price; third from top: ACF and bottom: PACF for log return of WTI oil price

test was performed on WTI price and the results are shown in two top panels of figure 3.7. As it is visible the tail of ACF is not diminishing and PACF value exceeds the 0.1 limits. This means that the price of WTI monthly oil data is not stationary and can not be used in ARIMA model with $d = 0$. Next the log return time series data are considered. The bottom two panels in figure 3.7 shows ACF and PACF calculation for WTI log returns. Based on the data shown in bottom two panels, the ACF seems to be diminishing and PACF is also well below 0.1 limits. Hence it is concluded that log difference is indeed stationary. So for forecasting these data either ARIMA(p,0,q) on the log return data will be used in this section and next section.

Another method to validate the data is to use "Augmented Dickey-Fuller" test (Enders, 2015; Box, Jenkins and Reinsel, 1994; Brockwell and Davis, 2006). This test has been run using R. The results for WTI price is listed below. It clearly shows that the stationary null hypothesis is rejected ($P > 0.04$), proving that WTI oil price is not stationary.

```
Augmented Dickey-Fuller Test
data:  log(oilspot.ts1)
Dickey-Fuller = -1.9076, Lag order = 7, p-value = 0.6161
alternative hypothesis: stationary
```

Additionally the result for WTI price log return is also listed below. It is obvious that the null hypothesis is passed ($P < 0.04$), proofing that the log return of the data is stationary.

```
Augmented Dickey-Fuller Test
data:  diff(as.vector(log(oilspot.ts1)))
Dickey-Fuller = -7.8864, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Next this section uses program R's "autoArima" function to find the best order of p, d and q parameters. Based on stationary test, we expect the best fit having $d = 0$. One

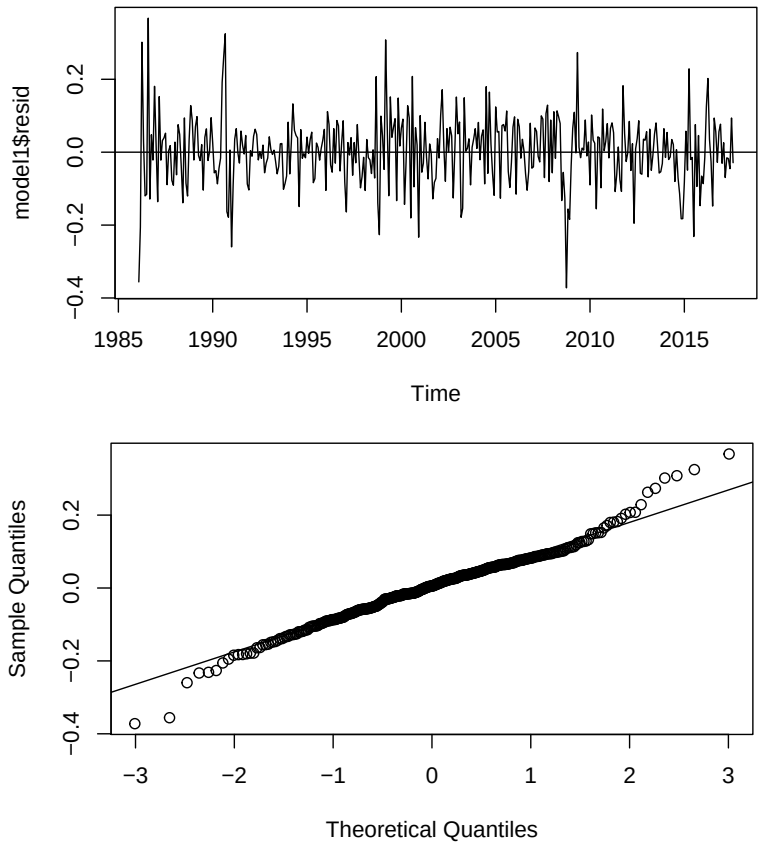

Figure 3.8: Top: ARIMA 100 residual for WTI price log return; bottom: QQ plot of ARIMA100 residuals.

of the conditions of ARIMA function is for input data to have zero mean. In order to also satisfy this conditions for ARIMA, the mean of log of return is subtracted from the log-return variables to get zero mean series. The result of "autoArima" function is listed below. It shows that ARIMA(1,0,0) for the log rerun (equivalent of ARIMA(1,1,0) for the oil price) return the best fit, which is first-order autoregressive model. This means that log-return data can be predicted as a multiple of its own previous value, plus a constant. The forecasting equation in this case shown in equation 3.7, meaning that log-return is

regressed on itself lagged by one period.

$$\hat{Y}_t = \phi_0 + \phi_1 Y_{t-1} \tag{3.7}$$

```
ARIMA(2,0,2)(1,0,1)[12] with non-zero mean : -703.929
ARIMA(0,0,0) with non-zero mean : -684.0895
ARIMA(1,0,0)(1,0,0)[12] with non-zero mean : -725.2799
ARIMA(0,0,1)(0,0,1)[12] with non-zero mean : -680.0588
ARIMA(0,0,0) with zero mean     : -689.7679
ARIMA(1,0,0) with non-zero mean : -698.4504
ARIMA(1,0,0)(2,0,0)[12] with non-zero mean : -718.4278
ARIMA(1,0,0)(1,0,1)[12] with non-zero mean : -720.9864
ARIMA(1,0,0)(2,0,1)[12] with non-zero mean : -714.0095
ARIMA(0,0,0)(1,0,0)[12] with non-zero mean : -719.259
I(1,0,0)[12] with non-zero mean : -721.2429
ARIMA(1,0,1)(1,0,0)[12] with non-zero mean : -720.997
ARIMA(2,0,1)(1,0,0)[12] with non-zero mean : -715.5432
ARIMA(1,0,0)(1,0,0)[12] with zero mean     : -730.882
ARIMA(1,0,0) with zero mean     : -703.9654
ARIMA(1,0,0)(2,0,0)[12] with zero mean     : -723.9279
ARIMA(1,0,0)(1,0,1)[12] with zero mean     : -726.5976
ARIMA(1,0,0)(2,0,1)[12] with zero mean     : -719.5286
ARIMA(0,0,0)(1,0,0)[12] with zero mean     : -724.8728
ARIMA(2,0,0)(1,0,0)[12] with zero mean     : -726.9369
ARIMA(1,0,1)(1,0,0)[12] with zero mean     : -726.573
ARIMA(2,0,1)(1,0,0)[12] with zero mean     : -721.2148

Best model: ARIMA(1,0,0)(1,0,0)[12] with zero mean
```

The rest of this section focuses on ARIMA(1,0,0). Referring to equation 3.7 , the desired fit has no θ_j and only has ϕ_1. Again ARIMA function in R is used to investigate the detail of this fit. The result of ARIMA (1,0,0) is listed below. The "ar1" variable is in fact ϕ_1 used in equation 3.6. The *s.e.* the results is also very small, which is due to the fact that we deducted mean of the data before the fit.

```
arima(x = log.return1, order = c(1, 0, 0), include.mean = FALSE)
Coefficients:
         ar1 0.1380
```

```
         s.e.  0.0518
sigma^2 estimated as 0.009167:  log likelihood = 351.37,
aic = -700.74
```

Figure 3.8 shows the residual values for the fit above as well we the QQ plot. The QQ plot displays a comparison of the sample quantiles to the corresponding theoretical quantiles (Enders, 2015; Box, Jenkins and Reinsel, 1994; Brockwell and Davis, 2006). The rule of thumb is if the points in a this plot depart from a straight line, then the assumed distribution is questionable. Based on the bottom panel of figure 3.8, we can see that the QQ plot is fairly linear, particularly for theoretical quantiles between -2 and 2. This manifests good quality of the fit we found. Figure 3.9 shows the ACF and PACF stationary test on the residual of above ARIMA(1,0,0) fit. It clearly shows that the ACF data on residuals vanished and PACF is mostly within 0.1 band. Hence the residual of this fit is stationary.

3.2.3 Non-linear Time Series Models (GARCH)

As discussed before, time series models provide better fit for data with volatility clustering. GARCH model stands for Generalized Autoregressive Conditional Heteroskedasticity, and is calculated by fitting a nonlinear function (second degree polynomial) on past events. The condition is that there are some data points in a series for which the variance of the current error term is a square of the actual sizes of the previous time periods' error terms. The GARCH(p,q) model is given by (Enders, 2015; Engle, 1982)

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2, \tag{3.8}$$

where following assumptions are made:

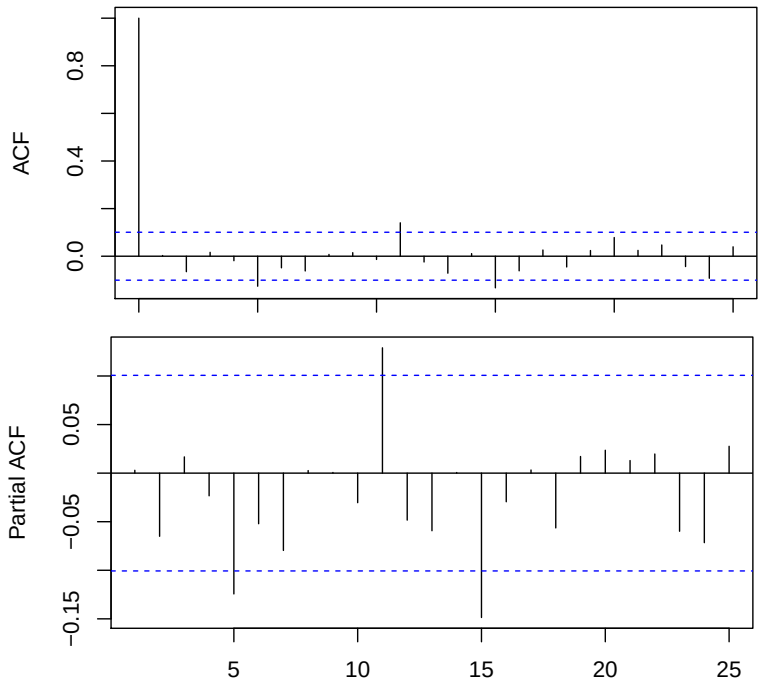

Figure 3.9: Top: ACF and bottom: PACF of ARIMA100 residuals.

1. ε_t denotes the return residuals or error terms. It can be also described as $\varepsilon_t = \sigma_t \, z_t$, assuming z_t is a white noise and σ_t is time dependent standard deviation.

2. $\omega > 0$, $\alpha_i > 0$, $\beta_i > 0$, and $i > 0$

3. $\sum_{i=1}^{q} \alpha_i + \sum_{i=1}^{p} \beta_i < 1$

This last condition is to satisfy the stationary nature of the time series. One step ahead can be forecasted using the coefficient in the equation above and following term. More steps ahead can be calculated recursively from here. Hence it is expected as we predict more points ahead of use, the errors accumulate and the accuracy of the forecast decreases (Enders, 2015; Box, Jenkins and Reinsel, 1994; Brockwell and Davis, 2006). This work uses R-package functions to calculate fitting GARCH parameters.

For this “uGARCH” function from “fGARCH” package from R is used to forecast log-return of monthly WTI data. When we used in the quarterly data, it was too coarse to return reasonable fit as GARCH fitting model requires to have at least 100 data points in time series. A for-loop was used to check AIC of various fits and find best GARCH(x,y). It was found that GARCH(1,1) returns the lowest AIC indicating this one returns the best fit. Here we evaluate quality of forecast for GARCH(1,1) model on ARMA(0,0) to predict, which is combination of first order GARCH and first order ARCH models. Below is the summary of GARCH(1,1) forecast.

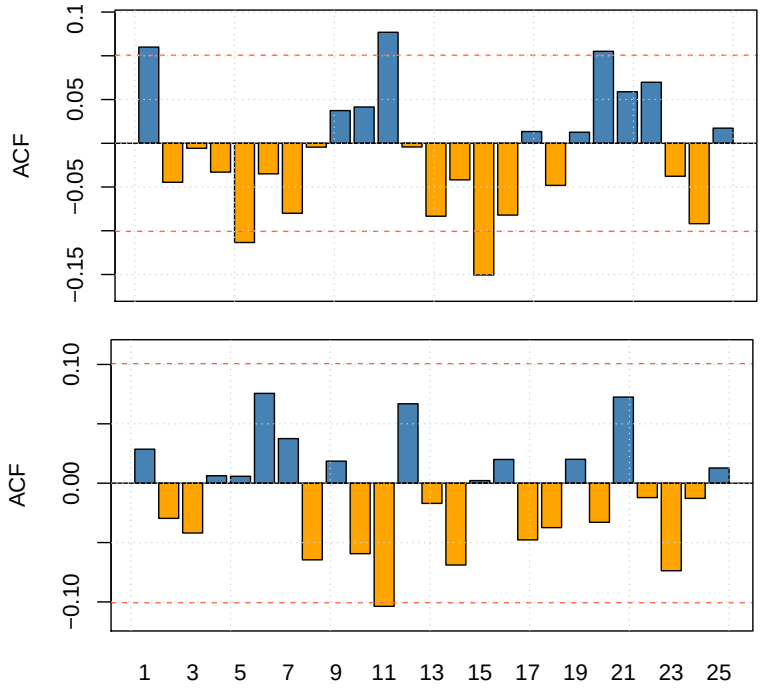

Figure 3.10: ACF of uGarch(1,1)’s Top: residual and bottom: residual squared.

```
GARCH ORDER 11 on arma00

*---------------------------------*
*          GARCH Model Fit        *
*---------------------------------*
```

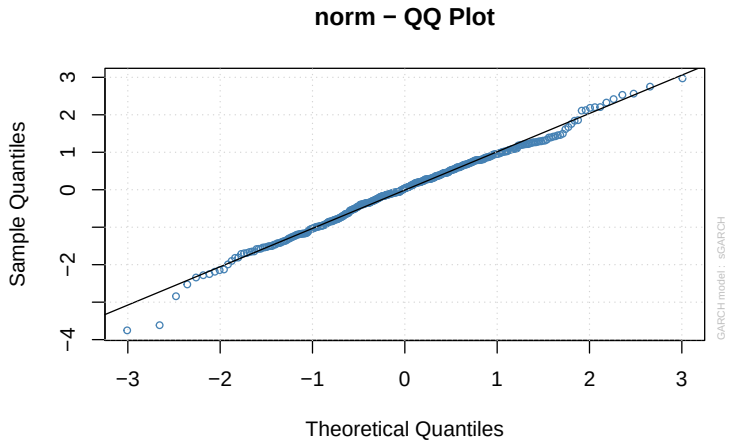

Figure 3.11: QQ plot for uGARCH(1,1).

```
Conditional Variance Dynamics
-----------------------------------
GARCH Model : sGARCH(1,1)
Mean Model : ARFIMA(0,0,0)
Distribution : norm

Optimal Parameters
------------------------------------
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.003259    0.003929  0.82954 0.406797
omega   0.000918    0.000354  2.59050 0.009584
alpha1  0.246215    0.060784  4.05069 0.000051
beta1   0.658708    0.060412 10.90362 0.000000

Robust Standard Errors:
        Estimate  Std. Error  t value Pr(>|t|)
mu      0.003259    0.003804  0.85674 0.391588
omega   0.000918    0.000341  2.68983 0.007149
alpha1  0.246215    0.066083  3.72585 0.000195
beta1   0.658708    0.056520 11.65448 0.000000

LogLikelihood : 376.2167

Information Criteria
------------------------------------
```

```
Akaike       -1.9642
Bayes        -1.9226
Shibata      -1.9644
Hannan-Quinn -1.9477

Weighted Ljung-Box Test on Standardized Residuals
------------------------------------
                        statistic p-value
Lag[1]                      4.615 0.03169
Lag[2*(p+q)+(p+q)-1][2]     4.996 0.04077
Lag[4*(p+q)+(p+q)-1][5]     6.393 0.07259
d.o.f=0
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals
------------------------------------
                        statistic p-value
Lag[1]                     0.3131  0.5758
Lag[2*(p+q)+(p+q)-1][5]    0.9971  0.8603
Lag[4*(p+q)+(p+q)-1][9]    2.6975  0.8077
d.o.f=2

Weighted ARCH LM Tests
------------------------------------
            Statistic Shape Scale P-Value
ARCH Lag[3]    0.6677 0.500 2.000  0.4138
ARCH Lag[5]    0.6872 1.440 1.667  0.8273
ARCH Lag[7]    2.1566 2.315 1.543  0.6842

Nyblom stability test
------------------------------------
Joint Statistic:  0.4727
Individual Statistics:
mu     0.10550
omega  0.13441
alpha1 0.04379
beta1  0.14143

Asymptotic Critical Values (10% 5% 1%)
Joint Statistic:       1.07 1.24 1.6
Individual Statistic:  0.35 0.47 0.75
```

```
Sign Bias Test
-----------------------------------
                   t-value   prob sig
Sign Bias           0.6859 0.4932
Negative Sign Bias  0.3493 0.7270
Positive Sign Bias  0.4590 0.6465
Joint Effect        3.210
```

Figure 3.10 shows the ACF of the residual and residual squared of above fit. It clearly shows that the residual squared data has more stationary behavior, meaning that GARCH model would return better fit compared to linear time series models, i.e. ARIMA. Also figure 3.11 shows the QQ analysis of the above GARCH(1,1) fit. It is much more linear compared to QQ analysis of best ARIMA model shown in figure 3.8, particularly on third quadrants. This means again indicates that GARCH(1,1) forecast has better performance compared to ARIMA(1,1) shown in previous section (Mikosch, 2011).

Chapter 4

Conclusions

This work studied and evaluated various forecasting methods for monthly and quarterly WTI oil index data. It evaluated and compared few structural and time-series methods, including multivariable linear regression model, linear time series model (ARIMA) and nonlinear time-series modes (GARCH).

Section 3.2.1 showed details of structural model and validated 8 different variables for using in this model. These variable were later used to regress various length of data to predict up to 12 month in advance. Based on MLSE analysis it was found that at least 8 months of data has to be fitted to return small MLSE. Based on this oil price forecast for 9 months were performed. It indeed predicted drop of the oil price in the first half of 2017 and increase of the price towards the end of the year. However, this model suffers, when the data has big chance of volatility clustering, like what happened in 2009.

To address data volatility clustering, section 3.2.2 and section 3.2.3 investigated linear and nonlinear time series based models, respectively. ARIMA model, which is a linear model, showed that applying ARIMA(1,0,0) on log-return data (equivalent of ARIMA (1,1,0) on original data) returns best performing fit. However, better fit found on residuals

squared compared to residuals suggesting advantage of the nonlinear models. Hence section 3.2.3 investigated nonlinear GARCH Model forecasting quality. It was found that GARCH(1,1) model provides best performing fit. The details of time-series models and forecasting also discussed in this report.

In conclusion, due to high volatility nature of oil price, it is found that non-linear time series based forecasting provide the best forecasting. Nevertheless and due to complicated cost dynamics of the oil, even the best models can have very large error in predicting real price of oil in the future, particularly when it is forecasting for longer stretches of time.

Bibliography

Alquist, Ron, Lutz Kilian, and Robert Vigfusson. 2011. "Forcasting of the Price of Oil." *International Finance Discussion Papers*, 1022.

Baumeister, Christiane, and Lutz Kilian. 2014. "Real-time analysis of oil price risks using forecast scenarios." *IMF Economic Review*, 62: 119–145.

Bosler, Fabian Torben. 2010. "Models for oil price prediction and forecasting." PhD diss. San Diego State University, Department of Mathematics.

Box, Geroge, Gwilym Jenkins, and Gregory Reinsel. 1994. *Time series analysis forecasting and control.* New Jersey:Prentice-Hall Internationall Inc.

Brockwell, Peter, and Richard Davis. 2006. *Time series: Theory and methods.* New York:Springer.

DOE, US. 2017*a*. "Annual Energy Outlook 2017."

DOE, US. 2017*b*. "US energy information administration: independent statistics and analysis."

Enders, Walter. 2015. *Applies Econometric Time Series.* New York:John Wiley.

Engle, Robert. 1982. "Autoregressive conditional heteroscedasticity with estimates of the variance of united Kingdom Inflation." *Econometrica*, 50: 987–1008.

LAM, Derek. 2013. "Time series modeling of monthly WTI crude oil returns." PhD diss. University of Oxford, Mathematical institute.

Mikosch, Thomas. 2011. "Is it really a long memory we see in financial returns?" *IDEAS Working Paper Series from RePEc*, 55(1): 869–889.

Moshiri, Saeed, and Faezeh Foroutan. 2006. "Forecasting nonlinear crude oil futures prices." *The Energy Journal*, 27: 81–96.

Ron, Alquist, Kilian Lutz, and Vigfusson Robert. 2011. "Forecasting the Prices of Oil." *IDEAS Working Paper Series from RePEc*, 55(1): 869–889.

Shabri, Ani, and Ruhaidah Samsudin. 2014. "Daily crude oil price forecasting using hybridizing wavelet and artificial neural network model." *Mathematical Problems in Engineering*, 2014.

USDOE, Web. 2017. "Cushing, OK WTI Spot Price FOB."

Appendices

A Matlab Code for Structural Modeling

```
clear all; close all; clc;
set(0,'DefaultAxesLineWidth',2)
set(0,'DefaultAxesFontSize',20)
i=0;
mt2=13:48; % Total months fitted looping
for mt=mt2
    i=i+1;
    [MLSE(i),MLAE(i)]=Multivariable_regression2(mt,12);
end
figure(10); plot(mt2,MLSE); hold on;
xlabel('Month Fitted'); ylabel('MLSE ($)'); grid on;
% plot(mt2,MLAE,'r')

function [MLSE,MLAE]=Multivariable_regression2(mt,ma)
warning off;
% mt=120; % Total months fitted
% ma=6; % Total months predicted after last point
year=(2006+(1:133+ma)./12)';
dl=importdata('Oil_Data_csv.csv'); %read data
data=dl.data;
%Cushing OK Crude Oil Future Contract 4 (Dollars per Barrel)
Y=data(134-mt:end,7);
X=[data(134-mt-ma:end-ma,1) data(134-mt-ma:end-ma,2) ...
data(134-mt-ma:end-ma,3)  data(134-mt-ma:end-ma,4) ...
data(134-mt-ma:end-ma,5) data(134-mt-ma:end-ma,6)...
   data(134-mt-ma:end-ma,8) ]; % all other variables with a lag to Y
X=[ones(length(Y),1) X]; %Adding first coloumn of X matrix
beta=inv(X'*X)*X'*Y; %solving regression
```

```
Y_estimate=X*beta; %Calculating fitted Y
figure(1); %subplot(4,1,4);
plot(year(1:end-ma),data(:,7),'LineWidth',3); hold on;
xlabel('Year'); ylabel('WTI Price ($)'); grid on;
error=Y-Y_estimate; %calculating spot error
MLSE=sqrt(sum(error.^2))./length(error);
MLAE=sum(abs(error))./length(error);
% figure(2); plot(year(134-mt:end-ma),error)
% xlabel('Year'); ylabel('Estimation Error ($)');
X2=[data(134-mt:end,1) data(134-mt:end,2) data(134-mt:end,3) ...
    data(134-mt:end,4) data(134-mt:end,5) data(134-mt:end,6) ...
    data(134-mt:end,8) ]; %constructing X matrix without lag
X2=[ones(length(Y),1) X2]; %Adding first coloumn of X matrix
Y_estimate2=X2*beta; % Use the regression solution to predict
figure(1); %subplot(4,1,4)
plot(year(133:end),Y_estimate2(end-ma:end),'r');
plot(year(134-mt:end-ma),Y_estimate,'k','LineWidth',3)
legend('WTI Price','Estimation','Original Fitting')
title(['Months Considered = ' num2str(mt) ', ...
Months Predicted = ' num2str(ma)])
end
```

B R Code for Non-Linear Modeling

```
setwd("/Users/halleh/Desktop/masters/spring 2017/Thesis")
rm(list=ls())

library(quantmod); library(fGarch);
#library(rugarch); #library(sarima);
#library(FinTS);
library(tseries);
library(forecast); library(stats); library(fBasics);
library(car); library(PerformanceAnalytics); library(TSA);
library(astsa);

WTI_data=read.csv("Cushing_OK_WTI_Spot_Price_FOB.csv")
WTI = xts(WTI_data[,-1], order.by=as.Date(WTI_data[,1], "%m/%d/%Y"))
WTI=to.monthly(WTI)
colnames(WTI)
start(WTI)
end(WTI)

# extract adjusted closing prices
WTI = WTI[, "WTI.Close", drop=F]

oilspot.ts1<-ts(WTI,start=c(1986,1),frequency=12)
#Set up realised values for validation
#realised.ts1<-diff(log(ts(WTI,start=c(1986,1),frequency=12)))
summary(oilspot.ts1);
#BoxCox plots to show the need of log transformation
ts.plot(BoxCox(oilspot.ts1,lambda = seq(-2, 2, 1/10)));
plot(log(oilspot.ts1),
     ylab="Quarterly WTI spot price US$"); abline(v=2003)
#ADF test indicating non-stationarity
adf.test(log(oilspot.ts1),alternative=c("stationary"))
#ADF test indicating non-stationarity
adf.test(diff(as.vector(log(oilspot.ts1))),alternative=c("stationary"))

#log return plot
plot(diff(log(oilspot.ts1)),start=c(2005,1), frequency=4,
     ylab="Monthly WTI spot Return US$")
abline(v=2003,h=0)
log.return1<-diff(log(oilspot.ts1))
#ACF and PACF plots
```

```
par(mfrow=c(1,2));
acf(as.vector(log.return1),drop.lag.0=FALSE)
pacf(as.vector(log.return1))
summary(log.return1)
log.return1=log.return1-mean(log.return1)

#best arima fit
fit1<- auto.arima(log.return1, trace=TRUE, test="kpss", ic="bic")
Box.test(fit1$residuals^2,lag=10, type="Ljung-Box")

#Models determined
model1<-arima(log.return1,order=c(2,0,2),include.mean=FALSE)
#Residual analysis
plot(model1$resid);abline(h=0)
mean(model1$resid);
#Residual ACF and PACF plots
acf(as.vector(model1$resid),drop.lag.0=FALSE)
pacf(as.vector(model1$resid))
acf(as.vector(model1$resid^2),drop.lag.0=FALSE)

#Normality tests
qqnorm(residuals(model1)); qqline(residuals(model1))
jarque.bera.test(model1$resid);
#Independence tests
McLeod.Li.test(,model1$resid,gof.lag=20)

#GARCH model fitted
#model.garch1<-garch(model1$resid,order=c(1,1),trace=F)
#model.garch1.res<-resid(model.garch1)[-1]
#acf(model.garch1.res,drop.lag.0=FALSE)
#pacf(model.garch1.res)
#acf(model.garch1.res^2,drop.lag.0=FALSE)
#pacf(model.garch1.res^2)
#For comparing forecast accuracy, fit GARCH/APARCH models
#gfit1<-garchFit(formula=~arma(0,1)+garch(1,1),
#                data=log.return1,trace=FALSE,include.mean=FALSE)
#gfit11<-garchFit(formula=~arma(0,1)+aparch(1,1),
#                 data=log.return1,trace=FALSE,include.mean=FALSE)
#gfit2<-garchFit(formula=~arma(2,2)+garch(1,1),
#                data=log.return1,trace=FALSE,include.mean=FALSE)
#gfit22<-garchFit(formula=~arma(2,2)+aparch(1,1),
```

```
#                 data=log.return1,trace=FALSE,include.mean=FALSE)

#garch11.spec = ugarchspec(variance.model = list(garchOrder=c(1,1)),
mean.model = list(armaOrder=c(0,0)), fixed.pars=list(mu = 0, omega=0.1,
alpha1=0.1,beta1 = 0.7))

garch11.spec = ugarchspec(variance.model = list(garchOrder=c(1,1)),
mean.model = list(armaOrder=c(0,0)))

oil.garch11.fit = ugarchfit(spec=garch11.spec, data=log.return1,
solver.control=list(trace = 1))
class(oil.garch11.fit)
slotNames(oil.garch11.fit)
names(oil.garch11.fit@fit)
names(oil.garch11.fit@model)

# show garch fit
oil.garch11.fit

# use extractor functions

# estimated coefficients
coef(oil.garch11.fit)
# unconditional mean in mean equation
uncmean(oil.garch11.fit)
# unconditional variance: omega/(alpha1 + beta1)
uncvariance(oil.garch11.fit)
# persistence: alpha1 + beta1
persistence(oil.garch11.fit)
# half-life:
halflife(oil.garch11.fit)
# residuals: e(t)
plot.ts(residuals(oil.garch11.fit), ylab="e(t)", col="blue")
abline(h=0)
# sigma(t) = conditional volatility
plot.ts(sigma(oil.garch11.fit), ylab="sigma(t)", col="blue")
# illustrate plot method
par(mfrow=c(3,3))
#plot(oil.garch11.fit)
#plot(oil.garch11.fit, which=1)
plot(oil.garch11.fit, which="all")
```

```
#plot(oil.garch11.fit, which=9)
# simulate from fitted model
oil.garch11.sim = ugarchsim(oil.garch11.fit, n.sim=nrow(log.return1),
rseed=12, startMethod="unconditional")
class(oil.garch11.sim)
slotNames(oil.garch11.sim)
```