Forecasting Crude Oil Price in United States

Zoe Zhu

11/20/2020

Summary

This report proposes a forecast method to predict crude oil price based on crude oil production, controlled for global demand for oil. MA (2) model is fitted to the residuals of the linear regression of difference in oil price on difference in oil production, time, and indicators for global demand. Estimted by this model, the 95% prediction intervals of forecasted values successfully cover the true observations in the testing set.

Introduction

Crude oil is a type of fossil fuel consists of a mixture of hydrocarbon deposits and other organic materials. They vary in color, composition and consistency. Different chemical compounds have different boiling temperatures, which permits crude oil to be refined and converted to more useful products such as gasoline and diesel. So far, crude oil remains the primary source of energy production worldwide.

Due to its non-renewable nature and vital significance in energy production, crude oil has become an important global commodity. Just like most commodities, oil price is driven by the supply and demand in the market. Therefore, crude oil production would directly influence the crude oil price. For the demand side, global economy plays a major factor on fluctuations of oil prices. Currently under the influence of COVID-19 pandemic, the demand of oil for commercial shipping decreases and oil price drops dramatically as a result.

This report aims to forecast the crude oil price in US with a time series ARIMA model. Specifically, this report focuses on predicting oil price using oil production, controlled for global demand, including the Great Recession (Dec 2007 to Jun 2009) and 2010s Oil Glut (Jun 2014 to 2016).

Data Preprocessing

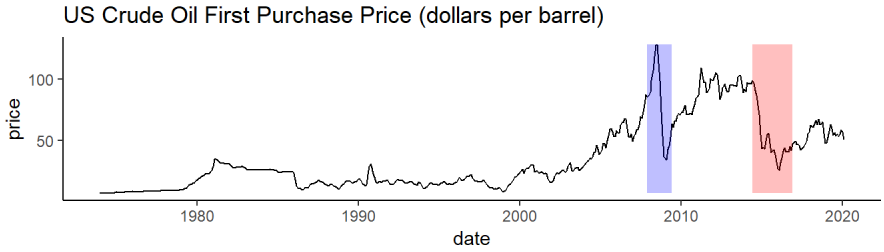
Crude oil production and price data were obtained from US Energy Information Administration. Crude oil production dataset has 1208 non-empty observations, ranging from January 1920 to August 2020. Crude oil price dataset has 560 non-empty observations, ranging from January 1974 to August 2020. Since there are no data for oil price before 1974, two datasets are matched on time and merged while rows with no observations for price are discarded.

The COVID-19 pandemic has a tremendous impact on various aspects, including crude oil price. However, COVID-19 is not a common factor that disrupts the supply and demand of oil regularly. Furthermore, COVID-19 pandemic is still ongoing and it is difficult to refer to a time period parallels the current situation. Taking this problem into consideration, crude oil price and production after February 2020 are excluded from the final dataset. To sum up, the final dataset contains 554 non-empty observations for crude oil price (in dollars per barrel) and production (in thousand barrels) per month per year, ranging from January 1974 to February 2020. The last 10 observations will be used as testing set while the remaining observations are training set.

Exploratory Data Analysis

Response Variable `price`

Generally, crude oil price has an upward trend with no obvious presence of seasonality and cyclic behavior. ACF plot provides evidence of serial correlation. There is a long-term fluctuation in the level of series which stays the same before 2000, sharp increase till early 2008, large decrease through 2008 and 2009, increase till early 2011, a period of fluctuations and then a large drop again from mid-2014 to late-2016, and increase slightly again. The final model would address the two large decreases, shaded in blue and red in the graph respectively, by adding indicators for these time periods.



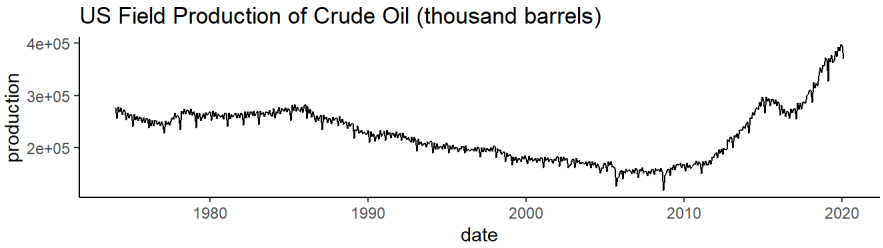
Indicator Variable `recession`

From the time series plot of crude oil price, the area shaded in blue indicates the decrease influenced by the Great Recession. The majority report provided by US Financial Crisis Inquiry Commission concluded that the causes of this recession were widespread failures in financial regulation, dramatic breakdowns in corporate governance, explosive bank runs and ill preparation for the crisis, resulting the most severe economic and financial meltdown since the Great Depression. According to the US National Bureau of Economic Research, the recession began in December 2007 and ended in June 2009. In this report, the indicator variable `recession` has value 1 for this time period and 0 otherwise.

Indicator Variable `glut`

From the time series plot of crude oil price, the area shaded in red illustrates the decrease due to 2010s Oil Glut. Began in 2014, there was a serious surplus of crude oil caused by general increase in oil production in North America, slowed global growth, geopolitical rivalries and growing environmental concern. To control for 2010s Oil Glut, an indicator variable `glut` was created with value 1 starting June 2014 (estimated by US Bureau of Labor Statistics) till December 2016 (oil supply cut for OPEC and US to boost oil price), 0 otherwise.

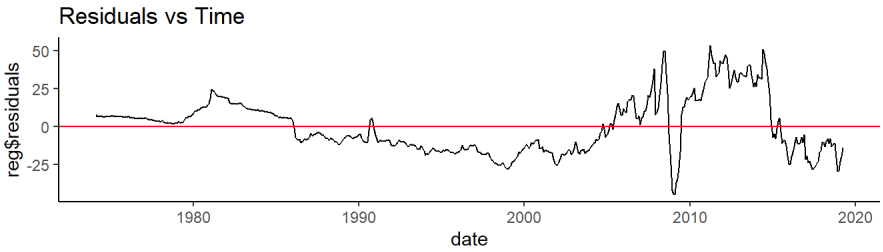
Predictor `production`

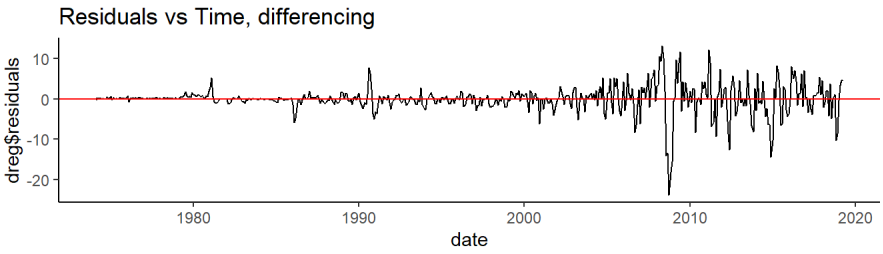


From the time series plot of crude oil production, there is a long-term fluctuation in the level of series which stays almost the same from 1974 to 1985, decreases till 2010, increases till 2015, decreases due to the supply cut as a response to 2010s Oil Glut and then increases again till now. In addition, there is presence of seasonality but no evidence of cyclic behavior. ACF plot provides evidence of serial correlation.

Model Selection

Differencing





Before introducing ARIMA model, the stationarity requirement must be satisfied. Based on previous discussion, crude oil price and production are not stationary. Fitting a linear regression of oil price on oil production, time, recession and glut, residuals are still not stationary. To meet the requirement for stationarity, differences between consecutive observations are computed and new variables `dprice` and `dprod` are used for model fitting. Therefore, fitting a linear regression of difference in oil price on difference of oil production, time, recession and glut, residuals seem to be stationary. Augmented Dickey-Fuller Test and KPSS Test confirm that residuals for this model are indeed stationary. Ljung–Box Test provides evidence of serial correlation. Detailed test results are in appendix.

Model selection is based on AIC with forward selection. ACF and PACF plots suggest that residuals might follow MA (2) process. Using `auto.arima()` with external regressors difference in oil production, time and indicators for the Great Recession and 2010s Oil Supply Glut, the function returns the best ARIMA model according to AIC value.

Final Model

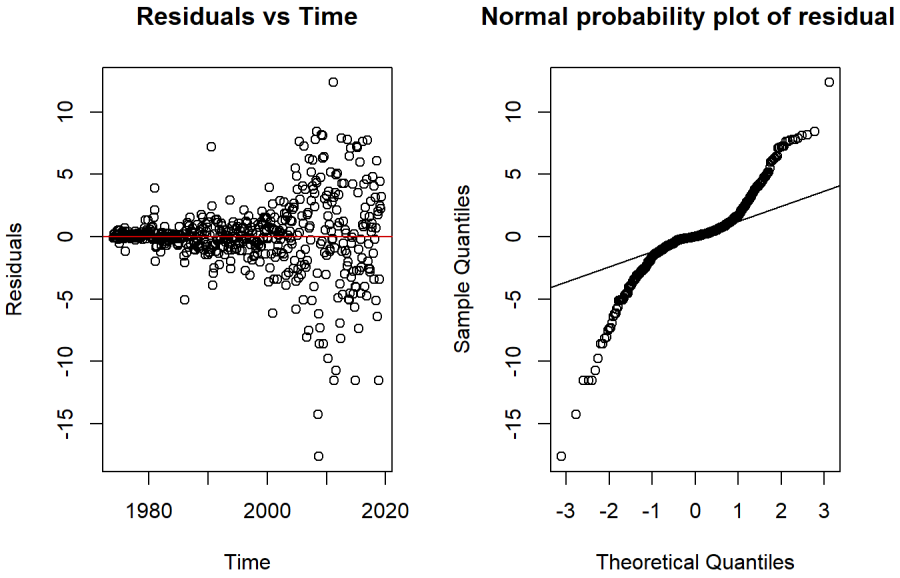
Fitting a linear regression of difference in price on difference in production and time, controlled for the Great Recession and 2010s Oil Supply Glut, residuals follow MA(2) process, which means the difference in price at time t depends on the value of the deviation from the error term at time t-1 and t-2:

$$y_t = \epsilon_t + 0.4837\epsilon_{t-1} + 0.2632\epsilon_{t-2}; \epsilon_t \sim N(0, \sigma^2)$$

	ma1	ma2	intercept	dprod	date	recession	glut
	0.4837	0.2632	2.6888	0	0	-1.1691	-1.6743
s.e.	0.0417	0.0406	1.5811	0	1e-04	1.1597	0.9880

sigma^2 estimated as 8.701: log likelihood = -1358, aic = 2731.99

Model Diagnostics

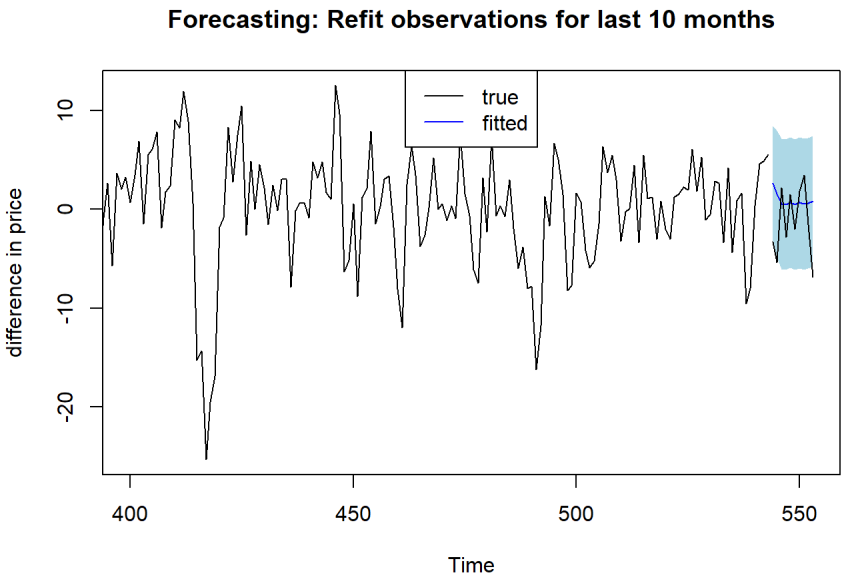


According to the Residual vs Time plot, residuals are uncorrelated and randomly scattered along 0. Before 2000, residuals are spread closely to 0 while after 2000 residuals are more spread out, implying the assumption of constant variance is violated. From the Normal QQ plot, less than half of the points are on the line, upper and lower tails are long and deviated away from the 45-degree line, indicating the assumption of normality is violated.

Diagnostic plots demonstrate that differencing successfully improves the forecasting method by satisfying the two important properties: 1. Residuals are uncorrelated. 2. Residuals have zero mean. The other two properties, constant variance and normality, are not satisfied. These two properties make the calculation of prediction intervals easier. However, these two properties are not necessary and there is usually little improvement can be done to do to ensure the residuals have constant variance and a normally distribution. We will use this model to forecast the oil price for last 10 months (May 2019 to February 2020).

Forecasting

Forecasted oil price for last 10 months are displayed below, bounded by 95% prediction intervals shaded in light blue.



Most of the observations are within the 95% prediction intervals, while the observation for February 2020 is out of bound, likely resulted from the influence of COVID-19 pandemic. RMSE for forecasted values is 3.7071. Compare to other ARIMA models, final model gives the smallest RMSE and performs well in forecasting the future difference in crude oil price.

Conclusion

Fitting a linear regression of difference in price on difference in production and time, controlled for the Great Recession and 2010s Oil Supply Glut, residuals follow MA (2) process. 95% prediction intervals proposed by the forecasting MA (2) model successfully cover the true observations of difference in oil price from May 2019 to February 2020.

Limitation

In model diagnostics, properties of equal variance and normality are not satisfied. Since these two properties are not necessarily required and Box-Cox transformation did not improve significantly, the final model just takes difference in price and production with no data transformation. In the future, we could examine more methods for transformation and try alternative approachs to obtain prediction intervals. In this case, the assumption of normality for forecast errors is violated, one alternative could be using bootstrapped residuals to obtained bootstrapped prediction interval.

Crude oil price is notoriously difficult to predict and influenced by various factors. This report narrows the scope and only focuses on crude oil production, time and indicators for the Great Recession and 2010s Oil Supply Glut with 554 monthly observations. With more background research, we could expand and include more relevant variables and hopefully address the influence of COVID-19 pandemic with advanced approaches.

Appendix

Test for stationarity and serial correlation

Augmented Dickey-Fuller Test: `tsresid`

Test statistic	Lag order	P value	Alternative hypothesis
-9.75	8	0.01 * *	stationary

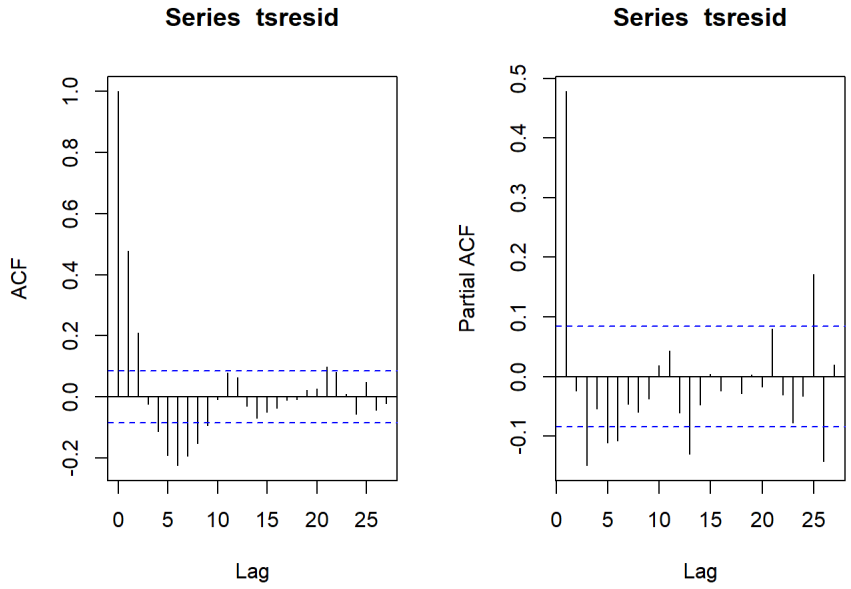
KPSS Test for Level Stationarity: `tsresid`

Test statistic	Truncation lag parameter	P value
0.02805	6	0.1

Box-Ljung test: `diff(tsresid)`

Test statistic	df	P value
47.39	10	8.035e-07 * * *

ACF & PACF plot for residuals



```
knitr::opts_chunk$set(echo = F)
knitr::opts_chunk$set(fig.align="center")
knitr::opts_chunk$set(out.width = '65%')
library(tseries)
library(forecast)
library(astsa)
library(ggplot2)
library(gridExtra)
library(Metrics)
library(sjPlot)
# Import data
oil <- read.csv("D:/MIDS/fall 2020/702 modeling/dataset/eia crude oil.csv",header=TRUE, col.n
ames = c('date','production','price'),nrows = 1208)
oil$date <- as.Date(paste(oil$date,"-01",sep=""),format="%B-%Y-%d")
# oil price as response, oil production as predictor, subset for valid rows
# Remove time after feb 2020 - the influence of covid-19
oil_data = oil[649:1202,]
rownames(oil_data) <- NULL
ggplot(oil_data,aes(x=date,y=price))+geom_line()+theme_classic()+theme(aspect.ratio=1/5)+
  ggtitle('US Crude Oil First Purchase Price (dollars per barrel)')+
  geom_ribbon(aes(xmin=as.Date("2007-12-01"), xmax=as.Date("2009-06-01")), fill = 'blue',alph
a=0.25)+
  geom_ribbon(aes(xmin=as.Date("2014-06-01"), xmax=as.Date("2016-12-01")), fill = 'red',alpha
=0.25)
prevprod <- c(NA, oil_data$production[1:553])
prevprice <- c(NA, oil_data$price[1:553])

oil_data$prevprod <- prevprod
oil_data$prevprice <- prevprice

oil_data$dprod <- oil_data$production - prevprod
oil_data$dprice <- oil_data$price - prevprice

oil2 = oil_data[2:554,]
rownames(oil2) <- NULL
# Add indicator for great recession dec 2007-jun 2009
oil2$recession <- 0
oil2$recession[407:425] <- 1
oil2$recession = factor(oil2$recession)
# Add indicator for global supply glut jun 2014-dec 2016(OPEC agreement)
oil2$glut <- 0
oil2$glut[485:515] <- 1
oil2$glut = factor(oil2$glut)
ggplot(oil_data,aes(x=date,y=production))+geom_line()+theme_classic()+theme(aspect.ratio=1/5)
+
  ggtitle('US Field Production of Crude Oil (thousand barrels)')
n <- length(oil2$dprice)
y_train <- oil2[1:(n-10),]
y_test <- oil2[(n-9):n,]
reg <- lm(price ~ production + date + recession + glut, data = y_train)
dreg <- lm(dprice ~ dprod + date + recession + glut, data = y_train)

p1 <- ggplot(y_train,aes(x=date,y=reg$residuals))+geom_line()+theme_classic()+theme(aspect.ra
tio=1/5)+
  ggtitle('Residuals vs Time')+geom_hline(yintercept=0,color='red')
p2 <- ggplot(y_train,aes(x=date,y=dreg$residuals))+geom_line()+theme_classic()+theme(aspect.r
atio=1/5)+
```

```
  ggtitle('Residuals vs Time, differencing')+geom_hline(yintercept=0,color='red')

grid.arrange(p1,p2,ncol=1)
#auto.arima(y_train$dprice,xreg = cbind(y_train$dprod,y_train$date,y_train$recession,y_train
$glut),
#           stationary = TRUE,ic = c("aic"))
model <- arima(y_train$dprice, order = c(0,0,2),xreg = cbind(y_train$dprod,y_train$date,y_tra
in$recession,y_train$glut))
p1 <- ggplot(y_train, aes(x=dprod, y=dreg$residual)) + theme_classic()+
   geom_point(alpha = .5,colour="blue4") +
   geom_hline(yintercept=0,col="red3") + labs(title="Residuals vs Difference in Production")

p2 <- ggplot(y_train, aes(x=date, y=dreg$residual)) + theme_classic() +
  geom_point(alpha = .5,colour="blue4") +
  geom_hline(yintercept=0,col="red3") + labs(title="Residuals vs Time")

grid.arrange(p1,p2,ncol=1)
par(mfrow=c(1,2))
plot(y_train$date,model$residuals,main = "Residuals vs Time",xlab='Time',ylab='Residuals')
abline(h=0, col="red3")

qqnorm(model$residuals, main = "Normal probability plot of residuals")
qqline(model$residuals)
# refit the last 10 months
# split data
n <- length(oil2$dprice)
y_train <- oil2[1:(n-10),]
y_test <- oil2[(n-9):n,]

# prediction
h <- 10
m <- n-h
fcast <- predict(model, n.ahead=h,newxreg = cbind(y_test$dprod,y_test$date,y_test$recession,y
_test$glut))

upper <- fcast$pred+1.96*fcast$se
lower <- fcast$pred-1.96*fcast$se

fit <- fcast$pred

# plot
plot.ts(y_train$dprice, xlim =c(400,n),
        main = "Forecasting: Refit observations for last 10 months",ylab = 'difference in pri
ce')
polygon(x=c(m+1:h,m+h:1), y=c(upper,rev(lower)), col='lightblue', border=NA)
lines(x=m+(1:h), y=fit,col='blue')
lines(x=m+(1:h), y=y_test$dprice,col='black')
legend("top", legend =c("true","fitted"), lty=c(1, 1), col =c("black","blue"))
#rmse(y_test$dprice,fit)
tsresid <- ts(dreg$residual)
#ts.plot(tsresid,col="blue4")

adf_test<- adf.test(tsresid,alternative = 'stationary')
pander::pander(adf_test)

kpss_test <- kpss.test(tsresid)
pander::pander(kpss_test)
```

```
pander::pander(Box.test(diff(tsresid), lag=10, type="Ljung-Box"))
par(mfrow=c(1,2))
acf(tsresid);pacf(tsresid)
```