Spencer Yee

Final Project

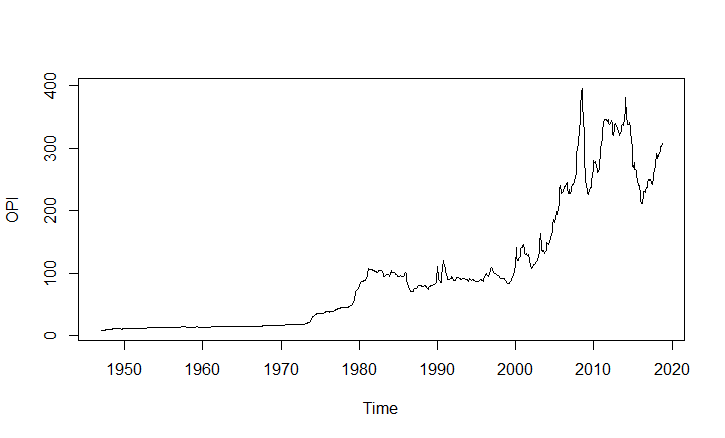
> library(readxl)

> OPI <- read\_excel("C:/Users/Spencer Yee/Desktop/EC245/OilPriceIndexUrbanConsumers.xlsx")

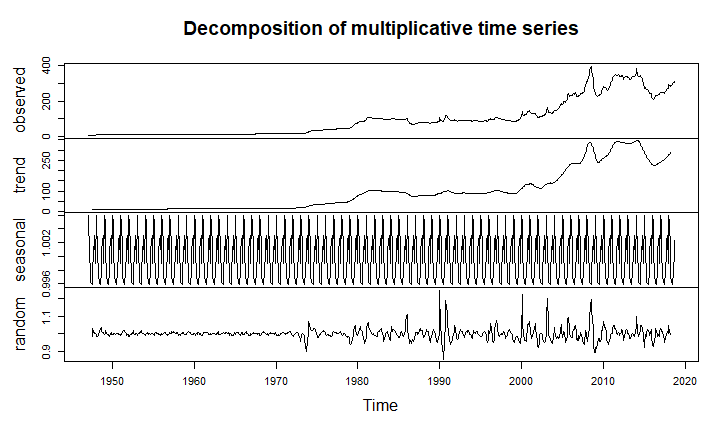
> View(OPI)

Data is already seasonally adjusted



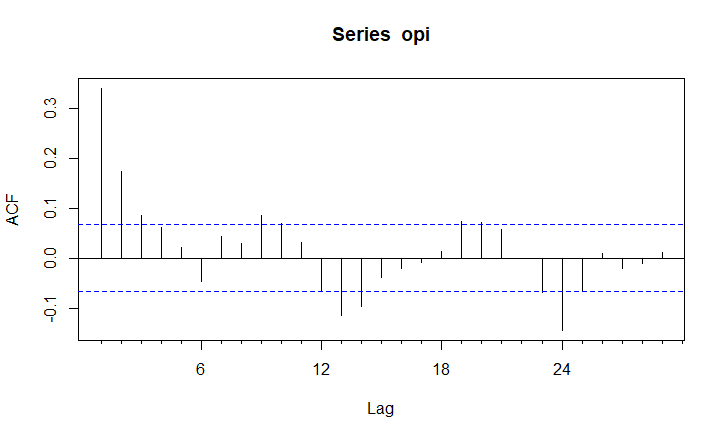
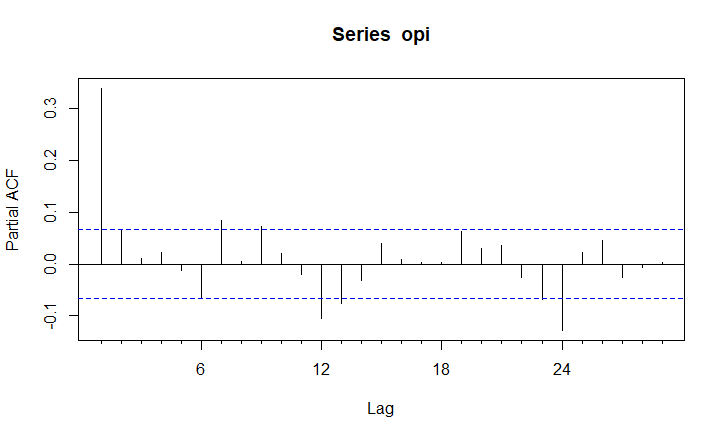


This is a time series variable for the oil/fuel price index for urban consumers with respect to time. The data is recorded monthly and dates all the way back to 1947. Based on the graph above, there appears to be a positive trend to the data. Decomposing the variable and viewing the charts tells us that there is indeed a positive trend and that the variable is seasonally stationary.



The first difference of the log of OPI was calculated to transform the variable to ensure it is stationary. Viewing the summary, the mean is 0.004168. This is the average percentage rate of change for the variable per time period.

1. From the ACF and PACF graphs I conclude that 4 potential models are MA3, MA10, AR2, and AR7 (Lag of order 3 and 10 for moving average and 2 and 7 for autoregressive). I am not ruling out a mixed model.

ACF PACF

All of these potential models must be tested for significant values (coefficient different from zero), undergo a box test for residuals, tested for inverted roots within the unit circle, and determine the AIC and BIC. The results of these tests are displayed in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Significant Values** | **AIC | BIC** | **p-value** | **Stationary? (y/n)** |
| MA3 | 2 of 3 | -3831.7 | -3807.91 | 0.0002 | Y |
| MA10 | 5 of 10 | -3837.17 | -3780.07 | 0.1966 | Y |
| AR2 | 1 of 2 | -3836.14 | -38.17.11 | 0.0005 | Y |
| AR7 | 2 of 7 | -3836.46 | -3793.66 | 0.0167 | Y |
| AR3 | MA3 | 3 of 6 | -3836.89 | -3798.83 | 0.0009 | N |
| AR10 | MA10 | 4 of 20 | -3849.62 | -3744.94 | 0.9615 | Y |

Based on these results, it is clear that the two best models are MA10 and AR10 | MA10. They pass the most tests out of all of the models. The AIC and BIC of both models are the lowest, the p-value’s are the highest, and both are stationary. The only concern is the significant values for the AR10 | MA10 model where only 4 of 20 were significant. The MA10 model is in better stance with 5 of 10 being significant. These results were determined using the ARIMA, LJ box test, and plot functions.



MA10: I will begin by splitting the first of the models (MA10) into a prediction and estimation sample. The prediction sample (I called m1f) contains 86 observations (approx. 10%) and the estimation sample (I called m1) contains the rest.

* 1. t.test was conducted with the m1f$residuals which showed that the average forecast error was not equal to zero. P-value = 0.988
  2. Box.test with df = 8 was conducted with the estimation sample (m1) to show that the p -value was very large (> 0.05) suggesting that the forecast errors are white noise.
  3. A summary of the prediction sample shows that the RMSE = 0.0236 and the MAE = 0.0169.
  4. After creating a regression equation with these variables (called eff), the very low p-value 4.795e-06 indicates the null hypothesis should be rejected and that the coefficient is not statistically different from zero. The values 0.946 and 4.8e-06 suggest that α0 and α1 are not zero.
  5. Running a linear hypothesis test with eff and m1f$fitted indicates that we should reject the null hypothesis and the intercept and slope coefficient do not appear to be jointly equal to 0 from this test.

AR10 | MA10: This model will also be split into a prediction and estimation sample called m2f and m2 respectively.

1. t.test was conducted with the m2f$residuals which showed that the average forecast error was not equal to zero. P-value = 0.93
2. Box.test with df = 8 was conducted with the estimation sample (m2) to show that the p -value was very large (> 0.05) suggesting that the forecast errors are white noise.
3. A summary of the prediction sample shows that the RMSE = 0.0205 and the MAE = 0.0142.
4. After creating a regression equation with these variables (called eff1), the very low p-value 9.219e-12 indicates the null hypothesis should be rejected and that the coefficient is not statistically different from zero. The values 0.987 and 9.22e-12 suggest that α0 and α1 are not zero.
5. Running a linear hypothesis test with eff1 and m2f$fitted indicates that we should reject the null hypothesis and the intercept and slope coefficient do not appear to be jointly equal to 0 from this test.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Error = 0** | **White Noise** | **RMSE & MAE** | **p-value** | **Jointly = 0** |
| MA10 | No | Yes | RMSE = 0.0236 MAE = 0.0169 | 4.795e-06 | No |
| AR10 | MA10 | No | Yes | RMSE = 0.0205 MAE = 0.0142 | 9.219e-12 | No |

Based on these tests it can be concluded that the forecast performance of these models is pretty good. The average forecast error is not zero and the actual forecast errors are white noise. The RMSE and MAE values for both models are very low which is desired. Though, the AR10 | MA10 model has slightly lower values indicating better performance. However the p-values of the tests suggest that α0 and α1 for both models are not jointly equal to zero. Despite this, the efficiency of these forecasts is respectable. As both of these models are considered to be good, the MSE and MAE must be tested.

In order to determine that the MSE of both models are different, I ran a t.test to check if the mean of the residuals of both models were different. The test calculated the mean of x to be 0.00055 and the mean of y to be 0.00041 thus indicating that the MSE of the models are indeed different.

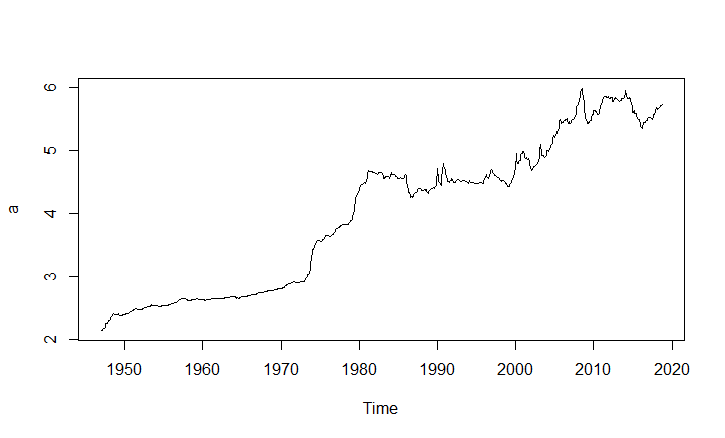
For the MAE, I similarly took the absolute value of both model’s residuals and ran the same t.test. It concluded that these values were also different with the mean of x being 0.017 and the mean of y being 0.014.

The default null hypothesis that both were equal was rejected. From this I conclude that the best model to carry on with is AR10|MA10.

1. After choosing the best model, I re-estimated the model using the Arima function to capture the entire time period. I then forecasted 1 period (1 month) into the future. The forecast value for November 2018 is 0.0068. Running a 95% confidence interval for this forecast, the value is expected to vary by -0.043 and +0.057 for the month of November. I then ran a density forecast for the requested quantiles using the R function qnorm and determined the corresponding z-values to be:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0.10 | 0.25 | 0.40 | 0.60 | 0.75 | 0.90 |
| -1.282 | -0.674 | -0.253 | 0.253 | 0.674 | 1.282 |

1. Taking the log of OPI, I proceeded to plot a time series graph of the data:

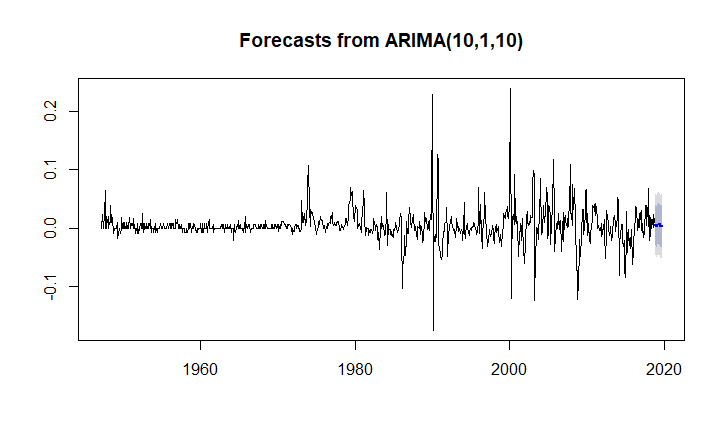


From this graph we can see that it is definitely possible that the time series has a unit root. The data is trending upwards up until about 1980 when it experiences a slight downward trend. Therefore the alternative hypothesis would be that this series has a unit root.

Running an Augmented Dickey-Fuller test, the test statistic is determined to be -2.201. The critical values were calculated to be -3.96, -3.41, and -3.12 for 1%, 5%, and 10% significance levels respectively. Since -2.201 > -3.12, we cannot reject the null hypothesis at the 10% significance level. The log of seasonally adjusted OPI has a unit root. We must now test to see if the first difference of log(OPI) has a unit root and so on.

After taking the first difference of the of log(OPI), the Dicker-Fuller test calculated the test statistic to be -8.3657. The critical values were determined to be -3.43, -2.86, and -2.57 for 1%, 5%, and 10% significance levels respectively. Since -8.3657 < -2.57 we can reject the null hypothesis at even the 10% significance level and conclude that the first difference of the log of OPI is stationary. The series is I(1) as the first difference does not have a unit root.

1. As I concluded that OPI does have a unit root of order 1, I will call the Arima function to generate a new model that incorporates this differencing which looks like (10,1,10). After forecasting this model 12 period ahead, the value for October 2019 is 0.0033. I have also provided a graph of the forecast below.



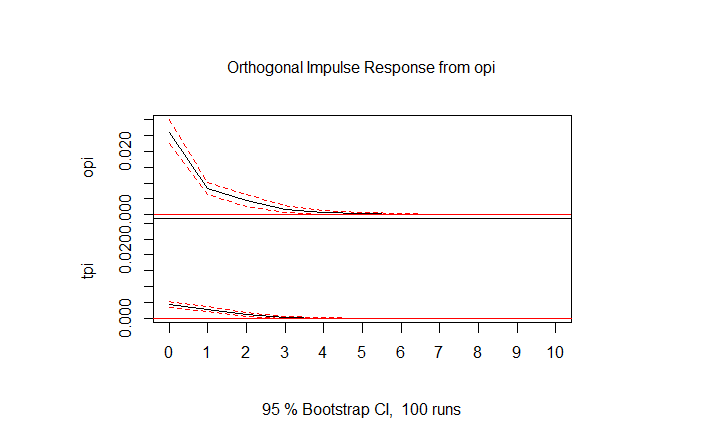
The lighter shade of blue on the graph represents the 80% confidence interval while the darker blue is 95%.

1. The variable/data I have chosen is the transportation price index for urban consumers. I have very good reason to believe that it is related to the urban consumer price index for oil/fuel. This is because most transportation requires oil/fuel. The frequency of the data is also monthly and is of the exact same time period. After loading the data and creating a time series variable, I transformed the data by taking the first difference of the log of the variable (TPI). I then called the data.frame and VARselect function to determine that the order of the VAR model is 2, which minimizes AIC.

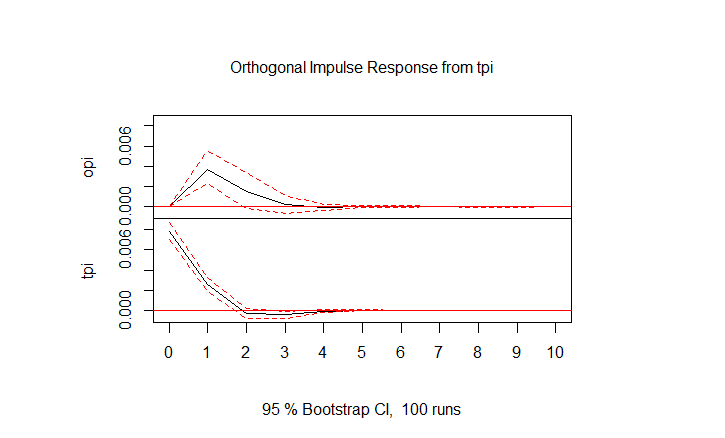
Continuing with the Granger Causality tests, the first one testing OPI’s effect on TPI, the p-value is incredibly low at 5.717e-06. This suggests we can reject the null hypothesis that OPI does not Granger-cause TPI. OPI is useful in explaining TPI.

The second one testing TPI’s effect on OPI, the p-value is also very low at 0.0001056. We can also reject the null hypothesis and conclude TPI is useful in explaining OPI.

As it is evident that both variables Granger-cause the other, I will analyze the graphs of the Impulse Response Functions:



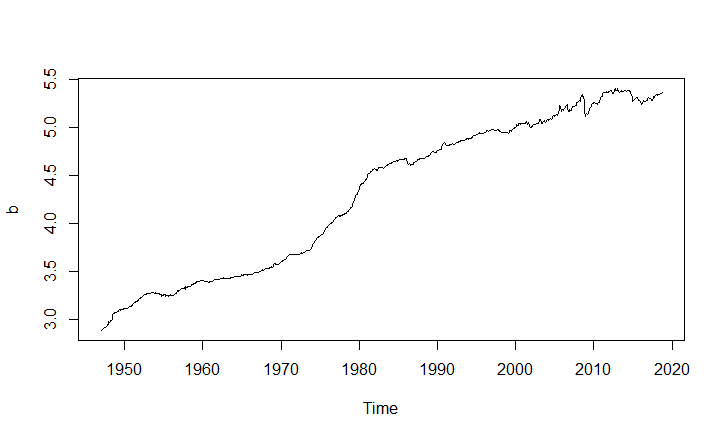
From this plot we can see that the changes in OPI affect the values of TPI. The variable TPI changes at the same instances as OPI.



From this plot we can see that the changes in TPI affect the values of OPI. The variable OPI changes at the same instances as TPI.

Given the results of this VAR model (Granger-Causality tests & Impulse Response Functions), it is safe to assume that a multivariate model would be better at forecasting than the original univariate model. There is reasonable association between these two variables.

1. I will now test to see if TPI also has a unit root of order I(1). Taking the log of TPI, I plotted the variable below:



There does indeed appear to be a unit root as the data, which trends upwards, experiences slight drops.

Running the Augmented Dickey-Fuller test with this variable, the test statistic is determined to be -0.7176. The critical values were calculated to be -3.96, -3.41, and -3.12 for 1%, 5%, and 10% significance levels respectively. Since -0.7176 > -3.12, we cannot reject the null hypothesis. The log of seasonally adjusted TPI has a unit root. We can now check to see if the first difference of log(TPI) has a unit root.

After taking the first difference of the of log(TPI), the Dicker-Fuller test calculated the test statistic to be -7.747. The critical values were determined to be -3.43, -2.86, and -2.57 for 1%, 5%, and 10% significance levels respectively. Since -7.747 < -2.57 we can reject the null hypothesis at even the 10% significance level and conclude that the first difference of the log of TPI is stationary. The series is I(1) as the first difference does not have a unit root and is thus of the same order as OPI.

As both OPI and TPI are the same order, I will run a Johanssen test to see if the two variables are co-integrated. The test statistic for this test is 599.59 and the critical values for 0.10, 0.05, and 0.01 are 17.85, 19.96, and 24.60 respectively. Since the test statistic of 599.59 > 17.85 we can reject the null hypothesis of r=0 at the 0.10 percent significance level. This indicates OPI and TPU are co-integrated.

I will now estimate a VEC model of the two variables. The error correction term for OPI is very small at 2.13e-05 indicating significance. This also tells us that changes in OPI affect changes in TPI. Likewise, the error correction term for TPI is also significant and even smaller at 2e-16. Changes in TPI do affect OPI. This supports the results of all of the tests above.

Based on the results from these tests, it can be concluded that a multivariate model may be better in the long-run in forecasting OPI. OPI and TPI are both I(1) as displayed by the Dickey-Fuller test, are both co-integrated as shown by the Johanssen test, and are directly related as determined by the Granger-Causality tests and VEC model.

R Input Code:

> library(readxl)

> OilPriceIndexUrbanConsumers <- read\_excel("C:/Users/Spencer Yee/Desktop/EC245/OilPriceIndexUrbanConsumers.xlsx")

> View(OilPriceIndexUrbanConsumers)

> library(readxl)

> OPI <- read\_excel("C:/Users/Spencer Yee/Desktop/EC245/OilPriceIndexUrbanConsumers.xlsx")

> View(OPI)

> OPI = ts(OPI$Price\_Index, start = c(1947, 1), frequency = 12)

> ts.plot(OPI)

> deOPI = decompose(OPI, type = "multi")

> plot(deOPI)

> opi = diff(log(OPI))

> summary(opi)

> model1 = Arima(opi, c(0,0,3))

> summary(model1)

> Box.test(model1$residuals, 12, "Lj", 8)

> plot(model1)

> model2 = Arima(opi, c(0,0,10))

> summary(model2)

> Box.test(model2$residuals, 12, "Lj", 8)

> plot(model2)

> model3 = Arima(opi, c(2,0,0))

> summary(model3)

> Box.test(model3$residuals, 12, "Lj", 8)

> plot(model3)

> model4 = Arima(opi, c(7,0,0))

> summary(model4)

> Box.test(model4$residuals, 12, "Lj", 8)

> plot(model4)

> model5 = Arima(opi, c(3,0,3))

> summary(model5)

> Box.test(model5$residuals, 12, "Lj", 8)

> plot(model5)

> model6 = Arima(opi, c(10,0,10))

> summary(model6)

> Box.test(model6$residuals, 12, "Lj", 8)

> plot(model6)

> m1 = Arima(opi[1:775], c(0,0,10))

> m1f = Arima(opi[775:862], c(0,0,10))

> t.test(m1f$residuals)

> Box.test(m1$residuals, 8, "Lj")

> summary(m1f)

> eff = lm(opi[775:862]~m1f$fitted)

> summary(eff)

> library("car", lib.loc="~/R/win-library/3.5")

> lht(eff, c("(Intercept)","m1f$fitted"))

> m2 = Arima(opi[1:775], c(10,0,10))

> m2f = Arima(opi[775:862], c(10,0,10))

> Box.test(m2$residuals, 8, "Lj")

> eff1 = lm(opi[775:862]~m2f$fitted)

> summary(eff1)

> lht(eff1, c("(Intercept)","m2f$fitted"))

> e1 =m1f$residuals

> e1sq = e1^2

> abse1 = abs(e1)

> e2 = m2f$residuals

> e2sq = e2^2

> abse2 = abs(e2)

> t.test(e1sq, e2sq)

> t.test(abse1, abse2)

> m10 = Arima(opi, c(10,0,10))

> forecast(m10, 1)

> forecast(m10, 1, .95)

> qnorm(c(0.10, 0.25, 0.40, 0.60, 0.75, 0.90))

> a = log(OPI)

> ts.plot(a)

> library("urca", lib.loc="~/R/win-library/3.5")

> df1 = ur.df(a, "trend", 12, "AIC")

> summary(df1)

> a1 = diff(log(OPI))

> df2 = ur.df(a1, "drift", 12, "AIC")

> summary(df2)

> M = Arima(opi, c(10,1,10))

> forecast(M, 12)

> fcast = forecast(M, 12)

> plot(fcast)

> library(readxl)

> TPIUC <- read\_excel("C:/Users/Spencer Yee/Desktop/EC245/TransportationPriceIndexUrbanConsumers.xlsx")

> View(TPIUC)

> TPI = ts(TPIUC$Price\_Index, start = c(1947, 1), frequency = 12)

> tpi = diff(log(TPI))

> library("vars", lib.loc="~/R/win-library/3.5")

> opi\_tpi = data.frame(opi, tpi)

> VARselect(opi\_tpi)

> varM = VAR(opi\_tpi, 2)

> causality(varM, "opi")

> causality(varM, "tpi")

> imp = irf(varM)

> plot(imp)

Hit <Return> to see next plot:

> b = log(TPI)

> ts.plot(b)

> df3 = ur.df(b, "trend", 12, "AIC")

> summary(df3)

> b1 = diff(log(TPI))

> df4 = ur.df(b1, "drift", 12, "AIC")

> summary(df4)

> jt = ca.jo(opi\_tpi, type = "trace", ecdet = "const", K = 2)

> summary(jt)

> ve = cajorls(jt, 1)

> summary(ve$rlm)