Test Exercise 6

Coursera/Erasmus U., Econometric Methods and Applications

Anthony Nguyen

Notes:

- See website for how to submit your answers and how feedback is organized.
- This exercise uses the datafile `TestExer6` and requires a computer.
- The dataset `TestExer6` is available on the website.

Goals and skills being used:

- Experience the process of practical application of time series analysis.
- Get hands-on experience with the analysis of time series.
- Give correct interpretation of outcomes of the analysis.

Questions

This test exercise uses data that are available in the data file `TextExer6`. The question of interest is to model monthly inflation in the Euro area and to investigate whether inflation in the United States of America has predictive power for inflation in the Euro area. Monthly data on the consumer price index (CPI) for the Euro area and the USA are available from January 2000 until December 2011. The data for January 2000 until December 2010 are used for specification and estimation of models, and the data for 2011 are left out for forecast evaluation purposes.

(a) Make time series plots of the CPI of the Euro area and the USA, and also of their logarithm, $log(CPI)$, and of the two monthly inflation series, $DP = \Delta log(CPI)$. What conclusions do you draw from these plots?

(b) Perform the Augmented Dickey-Fuller (ADF) test for the two $log(CPI)$ series. In the ADF test equation, include a constant (α), a deterministic trend term (β_t), three lags of $DP = \Delta log(CPI)$ and, of course, the variable of interest, $log(CPI_{t-1}$. Report the coefficient of $log(CPI_{t-1}$ and its standard error and t-value, and draw your conclusion.

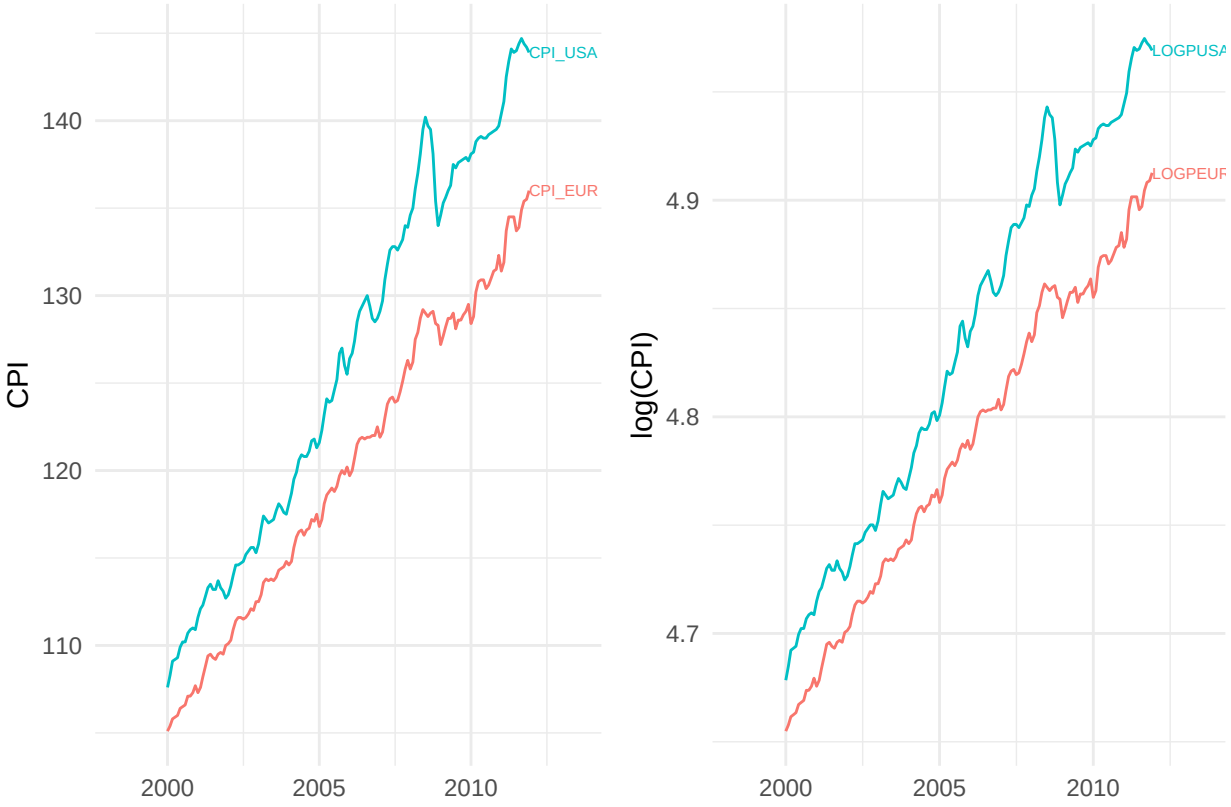
(c) As the two series of $log(CPI)$ are not cointegrated (you need not check this), we continue by modeling the monthly inflation series $DPEUR = \Delta log(CPIEUR)$ for the Euro area. Determine the sample autocorrelations and the sample partial autocorrelations of the this series to motivate the use of the following AR model: $DPEUR_t = \alpha + \beta_1 DPEUR_{t-6} + \beta_2 DPEUR_{t-12} + \varepsilon_t$. Estimate the parameters of this model (sample Jan 2000 - Dec 2010).

(d) Extend the AR model of part (c) by adding lagged values of monthly inflation in the USA at lags1, 6, and 12. Check that the coefficient at lag 6 is not significant, and estimate the ADL model $DPEUR_t = \alpha + \beta_1 DPEUR_{t-6} + \beta_2 DPEUR_{t-12} + \gamma_+ DPUSA_{t-1} + \gamma_2 DPUSA_{t-12} + \varepsilon_t$ (sample Jan 2000 - Dec 2010).

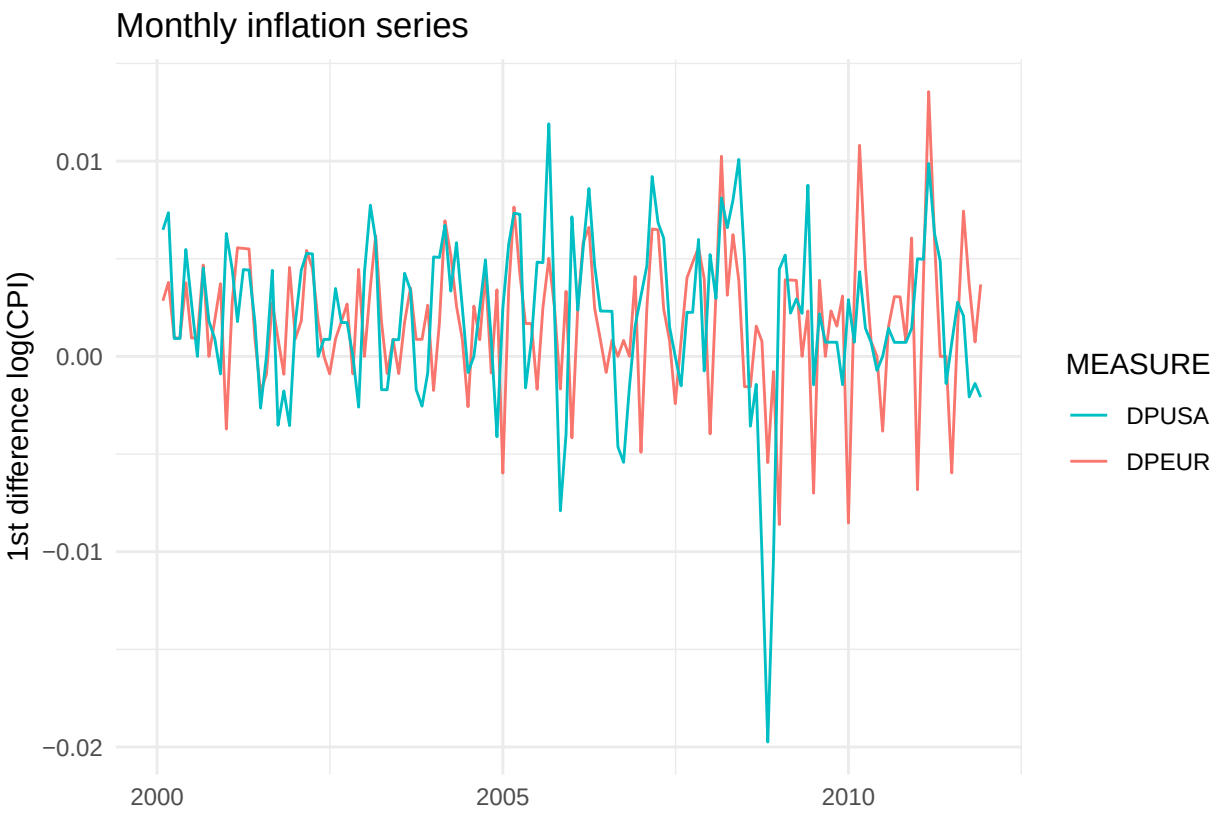
(e) Use the models of parts (c) and (d) to make two series of 12 monthly inflation forecasts for 2011. At each month, you should use the data that are then available, for example, to forecast inflation for September 2011 you can use the data up to and including August 2011. However, do not re-estimate the model and use the coefficients as obtained in parts (c) and (d). For each of the two forecast series, compute the values of the room mean squared error (RMSE), mean absolute error (MAE), and the sum of the forecast errors (SUM). Finally, give your interpretation of the outcomes.

Answers

(a) Make time series plots of the CPI of the Euro area and the USA, and also of their logarithm, $log(CPI)$, and of the two monthly inflation series, $DP = \Delta log(CPI)$. What conclusions do you draw from these plots?



Looking at the plots for the time series for CPI and log(CPI), we can see that both plots are very similar. For both the CPI and log(CPI) plots, we can see that prices seem to steadily increase over time and that there appears to be a correlation between USA and EURO prices.



For the monthly inflation series plot, again we can see there appears to be a correlation between US and EU inflation, but different regions have greater swings up and down at different times–i.e. both regions move in fairly similar ways early on, but the U.S. experiences greater fluctuation in the middle period, and the E.U. experiencing greater fluctuation towards the end of this series. All in all, the series looks rather stationary.

(b) Perform the Augmented Dickey-Fuller (ADF) test for the two $log(CPI)$ series. In the ADF test equation, include a constant (α), a deterministic trend term (β_t), three lags of $DP = \Delta log(CPI)$ and, of course, the variable of interest, $log(CPI_{t-1}$. Report the coefficient of $log(CPI_{t-1}$ and its standard error and t-value, and draw your conclusion.

As indicated in the lectures, the Augmented Dickey-Fuller (ADF) test for data with a clear trend direction takes the form:

$$\Delta y_t = \alpha + \beta t + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + ... + \gamma_L \Delta y_{t-L} + \varepsilon_t$$

Where we reject the hull hypothesis (H_o) of non-stationarity if the test statistic < -3.5.

When we perform the ADF test for $log(CPI)$ for the US series, we get:

```
# manually, we can check for the coefficient of interest
mod1a <- lm(DPUSA~TREND+lag(LOGPUSA,1)+
              lag(DPUSA,1)+lag(DPUSA,2)+lag(DPUSA,3),
            data = dat)
summary(mod1a)$coefficients
```

```
##                    Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)      0.349440600 1.272072e-01  2.74701838 6.841675e-03
## TREND            0.000151383 5.723046e-05  2.64514756 9.141541e-03
```

```
## lag(LOGPUSA, 1) -0.074337308 2.718524e-02 -2.73447271 7.093288e-03
## lag(DPUSA, 1)    0.609115127 8.404262e-02  7.24769353 3.026606e-11
## lag(DPUSA, 2)   -0.151263557 9.649875e-02 -1.56751824 1.193521e-01
## lag(DPUSA, 3)   -0.006444181 8.622768e-02 -0.07473448 9.405374e-01
```

```
# here we can double check with the adf.test() function
adf.test(dat$LOGPUSA, alternative = "stationary", k = 3)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  dat$LOGPUSA
## Dickey-Fuller = -2.7345, Lag order = 3, p-value = 0.2706
## alternative hypothesis: stationary
```

When we perform the ADF test for $log(CPI)$ for the EURO series, we get:

```
# manually, we can check for the coefficient of interest
mod1b <- lm(DPEUR~TREND+lag(LOGPEUR,1)+
              lag(DPEUR,1)+lag(DPEUR,2)+lag(DPEUR,3),
            data = dat)
summary(mod1b)$coefficients
```

```
##                      Estimate   Std. Error   t value    Pr(>|t|)
## (Intercept)      0.6420383970 2.263206e-01  2.836853 0.005264196
## TREND            0.0002374236 8.495751e-05  2.794616 0.005959120
## lag(LOGPEUR, 1) -0.1373730025 4.860536e-02 -2.826294 0.005430624
## lag(DPEUR, 1)    0.1442453918 8.665410e-02  1.664611 0.098326947
## lag(DPEUR, 2)   -0.0902171792 8.520917e-02 -1.058773 0.291607965
## lag(DPEUR, 3)   -0.1128027357 8.565361e-02 -1.316964 0.190098379
```

```
# here we can double check with the adf.test() function
adf.test(dat$LOGPEUR, alternative = "stationary", k = 3)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  dat$LOGPEUR
## Dickey-Fuller = -2.8263, Lag order = 3, p-value = 0.2324
## alternative hypothesis: stationary
```

In both cases the test-statistic is greater that -3.5, so we fail to reject the null hypothesis of non-stationarity.

(c) As the two series of $log(CPI)$ are not cointegrated (you need not check this), we continue by modeling the monthly inflation series $DPEUR = \Delta log(CPIEUR)$ for the Euro area. Determine the sample autocorrelations and the sample partial autocorrelations of the this series to motivate the use of the following AR model: $DPEUR_t = \alpha + \beta_1 DPEUR_{t-6} + \beta_2 DPEUR_{t-12} + \varepsilon_t$. Estimate the parameters of this model (sample Jan 2000 - Dec 2010).

First, we can determine the sample autocorrelations and the sample partial autocorrelations for the entire $DPEUR$ column vector by using the `acf()` and `pacf()` functions in the `tseries` package:

```
# subset `dat` to keep observations from Jan 2000 - Dec 2010
dat_sample <- slice(dat, 1:132)

# convert relevant column to numeric
dat_sample$DPEUR <- as.numeric(dat_sample$DPEUR)

# create table to store autocorrelation (ac) results
n <- nrow(dat_sample) - 1
ac <- tibble(lag = 1:n)
ac$AC <- NA
ac$PAC <- NA

# loop through $DPEUR with acf() and pacf() and store results in table (ac)
for (i in 1:n) {
    acf <- acf(dat_sample$DPEUR, lag.max = i, na.action = na.pass,
               plot = FALSE, demean = TRUE)

    pcf <- pacf(dat_sample$DPEUR, lag.max = i, na.action = na.pass,
                plot = FALSE, demean = TRUE)

    ac[i, 1] <- i
    ac[i, 2] <- acf$acf[i + 1, 1, 1]
    ac[i, 3] <- pcf$acf[i + 0, 1, 1]
}
```

After determining the sample autocorrelations (AC) and partial autocorrelations (PAC), we can check to see what are the largest values:

```
ac %>% group_by(PAC) %>% arrange(desc(PAC)) %>% head(., 10)
```

```
## # A tibble: 10 x 3
## # Groups:   PAC [10]
##      lag       AC    PAC
##    <int>    <dbl>  <dbl>
##  1    12  0.554   0.398
##  2     6  0.403   0.374
##  3    24  0.516   0.187
##  4    36  0.487   0.114
##  5    81 -0.0483  0.103
##  6    85  0.0310  0.0906
##  7    67  0.00806 0.0860
##  8    94 -0.0278  0.0837
##  9     1  0.0833  0.0833
## 10    50 -0.0125  0.0701
```

Here we can see that lag 12 and lag 6 have the highest PAC values.

Finally, we can estimate the parameters of our model using the `ar()` function:

```
ar(dat_sample$DPEUR[2:131], order.max = 12, method = "ols")
```

```
##
```

```
## Call:
## ar(x = dat_sample$DPEUR[2:131], order.max = 12, method = "ols")
##
## Coefficients:
##       1        2        3        4        5        6        7        8
##  0.0590   0.0014  -0.0972   0.0082  -0.1393   0.1943  -0.0567  -0.1271
##       9       10       11       12
## -0.0431  -0.1074   0.0603   0.5168
##
## Intercept: -2.317e-05 (0.000223)
##
## Order selected 12  sigma^2 estimated as  5.858e-06
```

From these results, are parameter estimates for this model would be:

$$DPEUR_t = \alpha + 0.194 DPEUR_{t-6} + 0.517 DPEUR_{t-12} + \varepsilon_t$$

(d) Extend the AR model of part (c) by adding lagged values of monthly inflation in the USA at lags 1, 6, and 12. Check that the coefficient at lag 6 is not significant, and estimate the ADL model $DPEUR_t = \alpha + \beta_1 DPEUR_{t-6} + \beta_2 DPEUR_{t-12} + \gamma_+ DPUSA_{t-1} + \gamma_2 DPUSA_{t-12} + \varepsilon_t$ (sample Jan 2000 - Dec 2010).

```
mod2 <- lm(DPEUR~lag(DPEUR, 6) + lag(DPEUR, 12) +
            lag(DPUSA, 1) + lag(DPUSA, 6) + lag(DPUSA, 12),
           data = dat_sample)

summary(mod2)$coefficients
```

```
##                     Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)     0.0004407009 0.0002853312  1.544524 1.252574e-01
## lag(DPEUR, 6)   0.2029830907 0.0785534689  2.584012 1.104053e-02
## lag(DPEUR, 12)  0.6367562986 0.0874784299  7.279009 4.784952e-11
## lag(DPUSA, 1)   0.2264302703 0.0511299287  4.428527 2.202571e-05
## lag(DPUSA, 6)  -0.0560494826 0.0547668405 -1.023420 3.082952e-01
## lag(DPUSA, 12) -0.2300590312 0.0541713545 -4.246876 4.470166e-05
```

According the regression results, we can see that the coefficient for $DPUSA_{t-6}$ has a p-value of 0.308 and is not significant, so we can drop this term from our model. When we re-run the regression without this term, we have:

```
# here we drop the insignificant term from the previous model and re-run
mod3 <- lm(DPEUR~lag(DPEUR, 6) + lag(DPEUR, 12) +
             lag(DPUSA, 1) + lag(DPUSA, 12),
           data = dat_sample)

summary(mod3)$coefficients
```

```
##                    Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)     0.000339085 0.0002675538  1.267353 2.076120e-01
## lag(DPEUR, 6)   0.168727693 0.0710802780  2.373762 1.927994e-02
## lag(DPEUR, 12)  0.655163572 0.0856272424  7.651345 6.928678e-12
## lag(DPUSA, 1)   0.232646311 0.0507784408  4.581596 1.187047e-05
## lag(DPUSA, 12) -0.226506091 0.0540712331 -4.189031 5.547523e-05
```

(e) Use the models of parts (c) and (d) to make two series of 12 monthly inflation forecasts for 2011. At each month, you should use the data that are then available, for example, to forecast inflation for September 2011 you can use the data up to and including August 2011. However, do not re-estimate the model and use the coefficients as obtained in parts (c) and (d). For each of the two forecast series, compute the values of the room mean squared error (RMSE), mean absolute error (MAE), and the sum of the forecast errors (SUM). Finally, give your interpretation of the outcomes.

For 2011, the actual values of $DPEUR$ are already given in the original data for this exercise. We can add the forecasted values for 2011 for the AR and ADL model by using the formulas we derived in parts (c) and (d), where:

$$\widehat{DPEUR}_t = 0 + 0.194 \cdot DPEUR_{t-6} + 0.517 \cdot DPEUR_{t-12} + \varepsilon_t \qquad \text{AR model}$$

$$\widehat{DPEUR}_t = 0 + 0.169 \cdot DPEUR_{t-6} + 0.655 \cdot DPEUR_{t-12} + 0.233 \cdot DPUSA_{t-1} - 0.226 \cdot DPUSA_{t-12} + \varepsilon_t \qquad \text{ADL model}$$

```
# First we create a table to store the results of our respective models
nn <- 133:144

dat_2011 <- tibble(TREND = dat$TREND[nn], DPEUR_actual = dat$DPEUR[nn])
dat_2011$AR_predicted <- NA
dat_2011$ADL_predicted <- NA

# Next we can loop through the relevant `TREND` numbers for 2011 and save
# the predicted values for the AR and ADL models to our table
for (i in 1:length(nn)) {
  AR_temp <- (0 + (0.194*dat$DPEUR[nn[i]-6]) + (0.517*dat$DPEUR[nn[i]-12]))

  ADL_temp <- (0 + (0.169*dat$DPEUR[nn[i]-6]) + (0.655*dat$DPEUR[nn[i]-12]) +
                 (0.233*dat$DPUSA[nn[i]-1]) - (0.226*dat$DPUSA[nn[i]-12]))

  dat_2011[i, 3] <- AR_temp

  dat_2011[i, 4] <- ADL_temp
}
```

```
# display dat_20111  results
knitr::kable(dat_2011)
```

TREND	DPEUR_actual	AR_predicted	ADL_predicted
133	-0.0068260	-0.0051527	-0.0065560
134	0.0037980	0.0019054	0.0032973
135	0.0135544	0.0061825	0.0077777
136	0.0059657	0.0029685	0.0055018
137	0.0000000	0.0005427	0.0019336
138	0.0000000	0.0011767	0.0023222
139	-0.0059657	-0.0033028	-0.0039839
140	0.0014948	0.0015291	0.0014826
141	0.0074405	0.0042106	0.0047778
142	0.0036996	0.0027336	0.0033265
143	0.0007383	0.0003933	-0.0001473

TREND	DPEUR_actual	AR_predicted	ADL_predicted
144	0.0036832	0.0031357	0.0033260

From here, we can easily calculate the RMSE, MAE and SUM of our models:

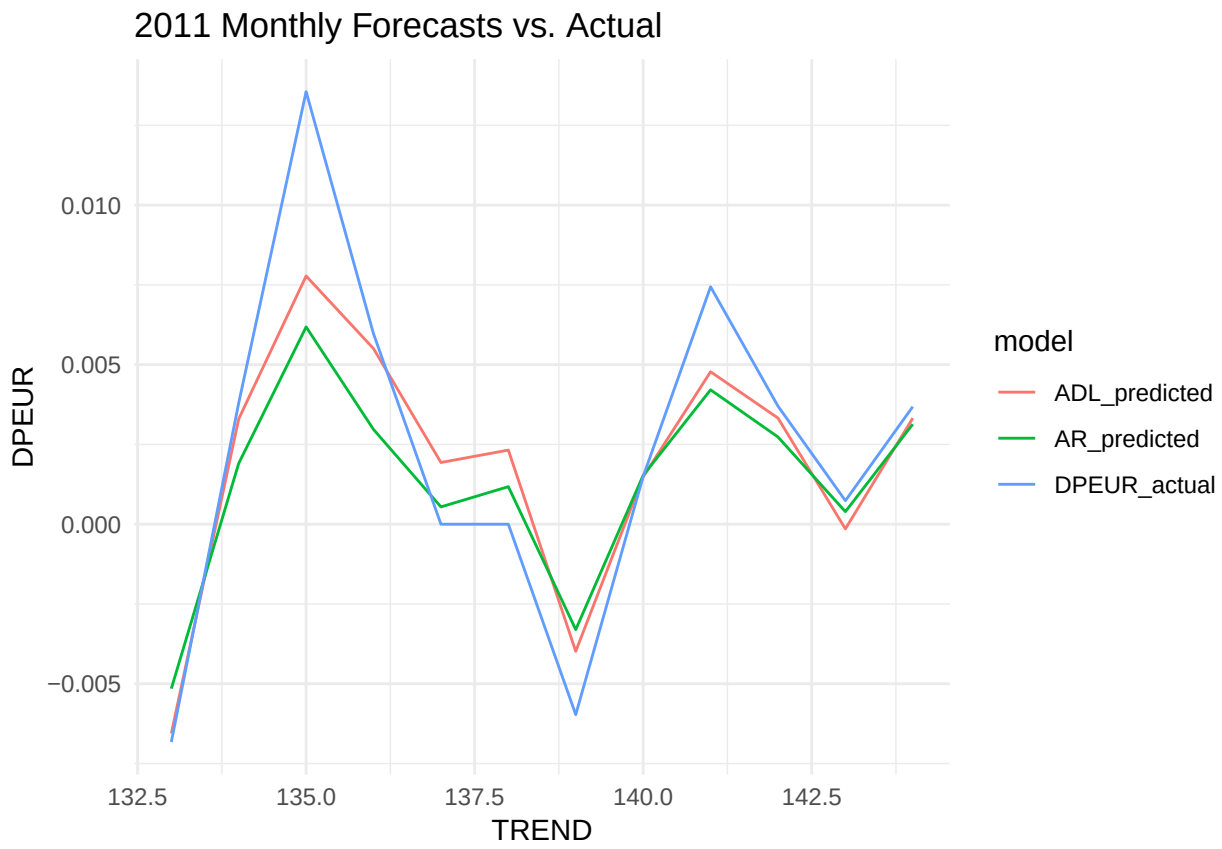
```
# calculate values for AR model
RMSE_AR = sqrt(mean((dat_2011$DPEUR_actual - dat_2011$AR_predicted)^2))
MAE_AR =  mean(abs(dat_2011$DPEUR_actual - dat_2011$AR_predicted))
SUM_AR =  sum(dat_2011$DPEUR_actual - dat_2011$AR_predicted)

# calculate values for ADL model
RMSE_ADL = sqrt(mean((dat_2011$DPEUR_actual - dat_2011$ADL_predicted)^2))
MAE_ADL =  mean(abs(dat_2011$DPEUR_actual - dat_2011$ADL_predicted))
SUM_ADL =  sum(dat_2011$DPEUR_actual - dat_2011$ADL_predicted)
```

Thus, for our AR model, we find that:
| RMSE = 0.0027
| MAE = 0.002
| SUM = 0.0113

For our ADL model, we find that:
| RMSE = 0.0021
| MAE = 0.0015
| SUM = 0.0045

Comparing the two models, it appears that the ADL model performs better as it has lower error values across all three measures.

```
plot_2011 %>% ggplot(aes(y=DPEUR, x=TREND, color = model)) +
  geom_line() +
  theme_minimal() +
  ggtitle("2011 Monthly Forecasts vs. Actual")
```


2011 Monthly Forecasts vs. Actual
DPEUR
0.010
0.005
0.000
−0.005
132.5
135.0
137.5
140.0
142.5
TREND
model
ADL_predicted
AR_predicted
DPEUR_actual