ARIMA

> ARIMA models provide another approach to time series forecasting. Exponential smoothing and ARIMA models are the two most widely used approaches to time series forecasting, and provide complementary approaches to the problem. While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data.

a study notes of ARIMA

Stationary and Differencing

A stationary time serie data is time independent, any trend or seasonal are not stationary, noise is stationary, same as cyclic, because cycle has not fixed length.

> In general, a stationary time series will have no predictable patterns in the long-term

Differencing: differences between consecutive observations.

> ACF plot is also useful for identifying non-stationary time series. For a stationary time series, the ACF will drop to zero relatively quickly, while the ACF of non-stationary data decreases slowly.

one can do log transform before differencing(multiplicative?)

When differenced data are not stationary, you can try second-order differencing, which is differencing the differces.

random walk

> Random walk models are widely used for non-stationary data, particularly financial and economic data. Random walks typically have:

- long periods of apparent trends up or down
- sudden and unpredictable changes in direction.
- forecast from a random walk model is equal to last observation, naive it is then.

seasonal differencing

> A seasonal difference is the difference between an observation and the previous observation from the same season

if desire, one can combine seasonal differencing with second-order differencing and/or log transformation.

KPSS test, or unit root test

(Kwiatkowski, Phillips, Schmidt, & Shin, 1992)

Null hypothesis is data are stationary, samll p-value (< 0.05) suggest false, then differencing is required. `ruca` package `ur.kpss()`

```
goog %>% ur.kpss() %>% summary
```

```
##
## #######################
## # KPSS Unit Root Test #
## #######################
##
## Test is of type: mu with 7 lags.
##
## Value of test-statistic is: 10.7223
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

```
goog %>% diff() %>% ur.kpss() %>% summary()
```

```
##
## #######################
## # KPSS Unit Root Test #
## #######################
##
## Test is of type: mu with 7 lags.
##
## Value of test-statistic is: 0.0324
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

note the different of test statstics after diff

`ndiffs` function suggest number of difference in order to archive stationary.

Autogressive models

Why it called autogression? in traditional regression model, outcome predicted against predictor(s)

> we forecast the variable of interest using a linear combination of past values of the variable, The term autoregression indicates that it is a regression of the variable against itself.

`p` order autogressive model, or AR(p) defination: $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ...\phi_p y_{t-p} + e_t$, it is very much like a multi-variant regression.

> We normally restrict autoregressive models to stationary data. When `p is >= 3` the restrictions are much more complicated. R takes care of these restrictions when estimating a model.

moving average models

> Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model. # Appendix

`q` order of moving average, or MA(q) equation: $y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q}$

moving average models should not be confused with the moving average smoothing.

> A moving average model is used for forecasting future values, while moving average smoothing is used for estimating the trend-cycle of past values.

The fact that AR using observed value and MA using error, makes AR(p) model change scale of series not patterns, MA(q) change patterns not scale. AR and MA models are invertible.

non-seasonal ARIMA

Combine AR and MA, 'I' stands for integration, which is reverse of differencing, if $y^{'}$ is differenced series(either once or more). then `ARIMA(p,d,q)` is

$y^{'} = c + \phi_1 y^{'}_{t-1} + ...\phi_p y^{'}_{t-p} + \theta_1 e_{t-1} + ... + \theta_q e_{t-q} + e_t$

p is order of autogression, q is order or moving average, d is degree of first differencing involved.

```r
fit <- auto.arima(uschange[,"Consumption"], seasonal=FALSE)
summary(fit)
```

```
## Series: uschange[, "Consumption"]
## ARIMA(1,0,3) with non-zero mean
##
## Coefficients:
##          ar1      ma1     ma2     ma3    mean
##       0.5885  -0.3528  0.0846  0.1739  0.7454
## s.e.  0.1541   0.1658  0.0818  0.0843  0.0930
##
## sigma^2 estimated as 0.3499:  log likelihood=-164.81
## AIC=341.61   AICc=342.08   BIC=361
##
## Training set error measures:
##                       ME    RMSE       MAE      MPE    MAPE      MASE
## Training set 0.001051018 0.58356 0.4306226 49.39042 171.384 0.6746977
##                      ACF1
## Training set -0.001972116
```

so in this ARIMA(1,0,3), the formular is below

$y_t = c + 0.589 y_{t-1} - 0.353 e_{t-1} + 0.0846 e_{t-2} + 0.174 e_{t-3} + e_t$

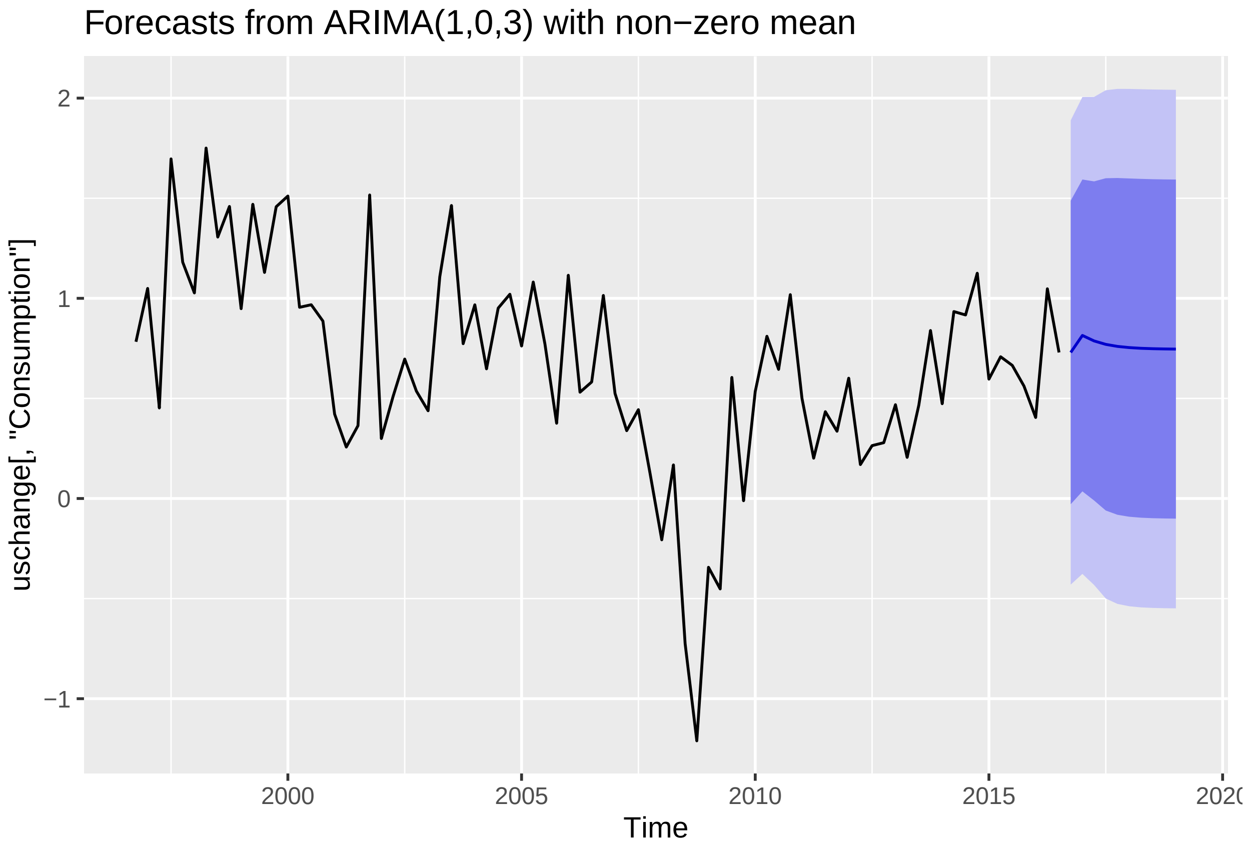
where mean times (1-ar1) $c = 0.745(1 - 0.589)$, e^2 is 0.35

constant `c` effect on long-term foreast:

- If c = 0 and d = 0 , the long-term forecasts will go to zero.

- If c = 0 and d = 1 , the long-term forecasts will go to a non-zero constant.
- If c = 0 and d = 2 , the long-term forecasts will follow a straight line.
- If c 0 and d = 0 , the long-term forecasts will go to the mean of the data.
- If c 0 and d = 1 , the long-term forecasts will follow a straight line.
- If c 0 and d = 2 , the long-term forecasts will follow a quadratic trend.

```
fit %>% forecast(h=10) %>% autoplot(include=80)
```



ACF

Did you forget what is test statistic old man?

significance level (alpha) is chance of H0 get wrongfully rejected, 0.05 is most common. in a bell curve plot(assuming two sided hypothesis test), alpha is cut off area of both tail. the critial calue are cut-off value of tail region, so the test statistic is within the rejection region.

p-value is area of test statistic area on both tail (2 sided), so if p-value > alpha, H0 faled to reject.

a good illustration here