Week 4 discussion

Victor Sim

10/2/2020

Context

Imagine that you are an investment actuary for a pension fund and you are asked to select a model to forecast future values of those economic and investment variables for future asset and liability modeling purposes. You have been told that the pension fund only invests in equities, bonds, and cash. Price inflation rates and wage inflation rates should also be incorporated in the model for liability modeling and real rates calculations. You are given the dataset to exploratory analysis and modeling.

Data Glossary

Variables: % effective rates per annum

P: Price Inflation Rate W: Wage Inflation Rate L: Long-term Interest Rate S: Short-term Interest Rate E: Equity index return B: Bond total return P_D4: Quarterly differenced price inflation rate (P(t+4)-P(t)) W_D4: Quarterly differenced wage inflation rate (W(t+4)-W(t)) L_D1: First differenced long-term interest rate (L(t+1)-L(t)) S_D1 First differenced short-term interest rate (S(t+1)-S(t))

Question 1: Explore the data and think about the question.

To model a forecast for asset and liability management (ALM) for a pension fund, we need to consider a time series analysis and its different components consisting of trend, seasonality and noise. Thus my question would be *"Does the model explain the variables well?"*

```
#Packages
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
suppressMessages(library(tidyverse))
suppressMessages(library(FitAR))
suppressMessages(library(reshape2))
suppressMessages(library(forecast))
```

```
#Import dataset

ts_data<- read.delim("C:/Users/user/Downloads/Model CC Quarterly 02Q2 12Q4.txt")

#Column names given
colnames(ts_data)<-c("Price","Wage","LT_Int","ST_Int",
                     "Equity","Bond","PQuarDiff",
                     "WQuarDiff","LTDiff1","STDiff1")

str(ts_data)
```

```
## 'data.frame':    43 obs. of  10 variables:
##  $ Price    : num  -1.505 0.592 0.399 1.388 -2.11 ...
##  $ Wage     : num  -1.5 4.07 35.74 -25.56 -1.83 ...
##  $ LT_Int   : num  0.637 0.759 0.786 0.717 0.752 ...
##  $ ST_Int   : num  0.483 0.566 0.579 0.489 0.557 ...
##  $ Equity   : num  8.47 -8.99 -16.48 9.69 -2.43 ...
##  $ Bond     : num  1.908 -0.902 0.289 0.766 0.476 ...
##  $ PQuarDiff: num  0.00374 0.09745 0.2 1.30314 -0.60494 ...
##  $ WQuarDiff: num  -1.836 -2.197 -1.625 1.929 -0.338 ...
##  $ LTDiff1  : num  -0.0673 0.1221 0.0267 -0.0688 0.0346 ...
##  $ STDiff1  : num  -0.000736 0.082745 0.013439 -0.090054 0.067082 ...
```

```
summary(ts_data)
```

```
##      Price              Wage             LT_Int           ST_Int      
##  Min.   :-2.1099   Min.   :-26.127   Min.   :0.6372   Min.   :0.2344  
##  1st Qu.:-0.1003   1st Qu.: -3.062   1st Qu.:0.7924   1st Qu.:0.4313  
##  Median : 0.8863   Median :  2.353   Median :0.8562   Median :0.5565  
##  Mean   : 0.6840   Mean   :  3.857   Mean   :0.8844   Mean   :0.6430  
##  3rd Qu.: 1.4351   3rd Qu.: 16.998   3rd Qu.:0.9537   3rd Qu.:0.7308  
##  Max.   : 3.0572   Max.   : 38.857   Max.   :1.2499   Max.   :1.6002  
##      Equity              Bond           PQuarDiff           WQuarDiff      
##  Min.   :-39.3656   Min.   :-3.1695   Min.   :-3.29601   Min.   :-3.7373  
##  1st Qu.: -8.9405   1st Qu.:-0.2292   1st Qu.:-0.54634   1st Qu.:-1.6258  
##  Median :  0.7554   Median : 0.6081   Median : 0.20235   Median :-0.2728  
##  Mean   :  1.3209   Mean   : 0.7787   Mean   : 0.07234   Mean   :-0.1259  
##  3rd Qu.: 11.5013   3rd Qu.: 1.4269   3rd Qu.: 0.90073   3rd Qu.: 1.5402  
##  Max.   : 37.7360   Max.   : 7.9029   Max.   : 2.49840   Max.   : 2.9593  
##     LTDiff1             STDiff1        
##  Min.   :-0.226982   Min.   :-0.83425  
##  1st Qu.:-0.050900   1st Qu.:-0.08021  
##  Median : 0.007844   Median : 0.02818  
##  Mean   : 0.004047   Mean   : 0.01484  
##  3rd Qu.: 0.059874   3rd Qu.: 0.12708  
##  Max.   : 0.251278   Max.   : 0.89764  
```

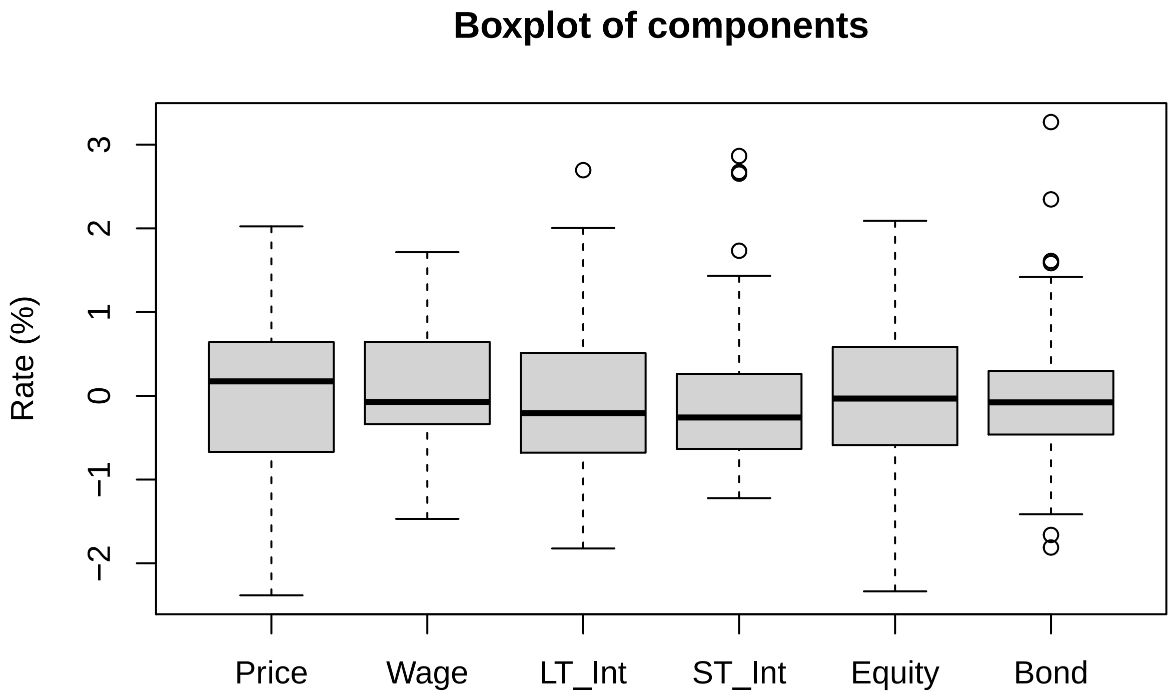
Question 2: What are your recommendations in terms of data manipulation and transformation before modelling?

Since all rates are effectively annualized, we may need to use log differences to capture the changes in the variables. We will also need to standardize the rates to compare among them. But first lets add a time column.

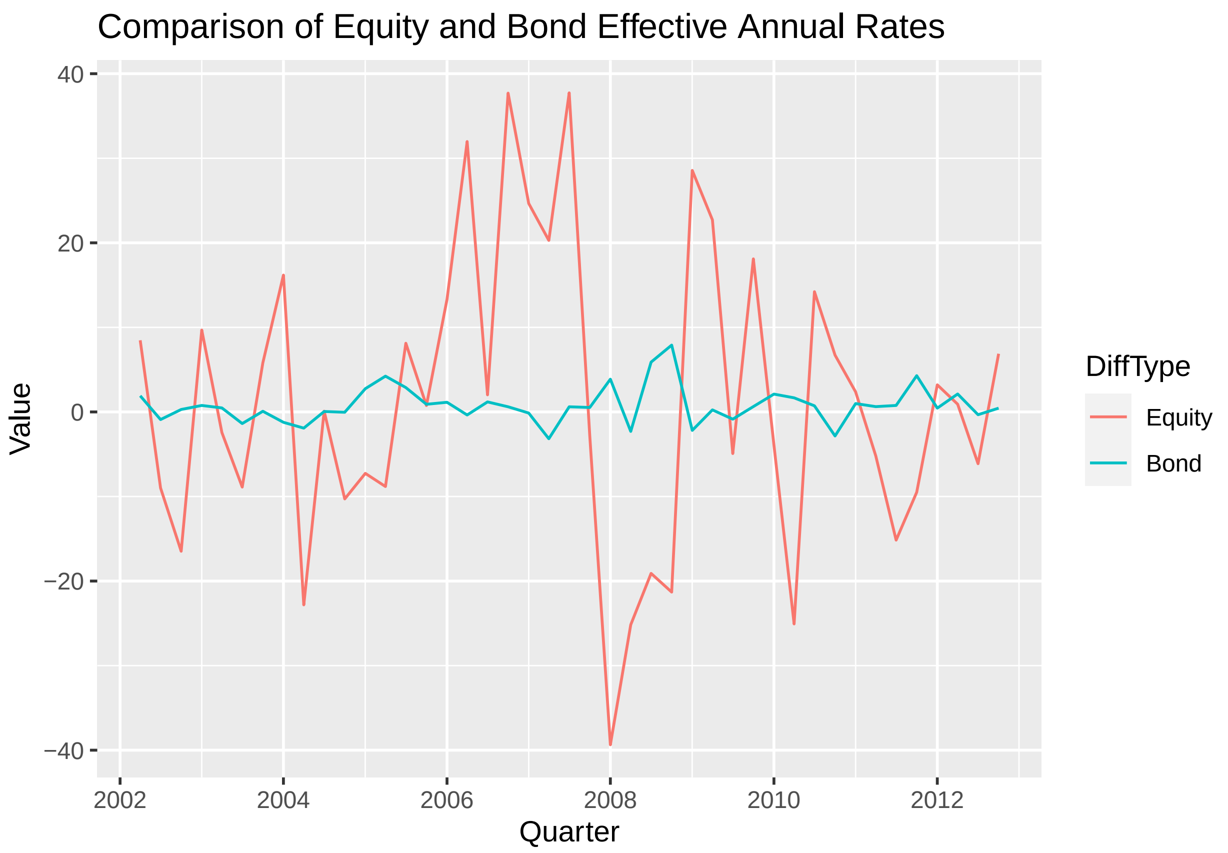
```
#Add quarter column for time
ts_data$Quarter<-seq(as.Date(’2002-04-01’),as.Date(’2012-12-31’),by=’quarter’)

#Add rearrange for quarter to come first
ts_data<-select(ts_data, Quarter, everything())

##Brief exploration
boxplot(scale(ts_data[,c(2:7)]), main="Boxplot of components", ylab="Rate (%)")
```

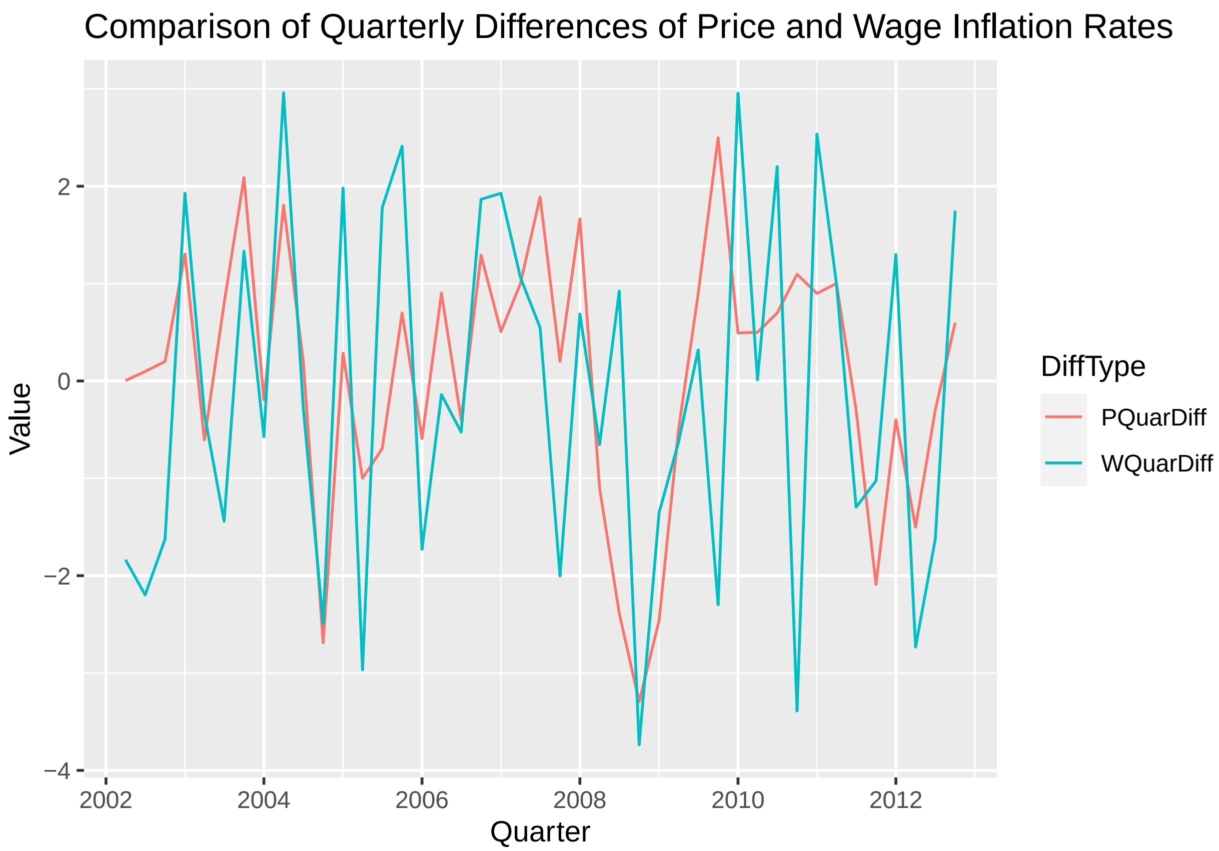


```
#Equity vs Bond
EvsB_compare<-melt(select(ts_data,Quarter,Equity,Bond),
                   id.vars = "Quarter",value.name = "Value",variable.name = "DiffType") %>%
  ggplot(aes(x=Quarter,y=Value,colour=DiffType))+
  geom_line()+ggtitle("Comparison of Equity and Bond Effective Annual Rates")
EvsB_compare
```

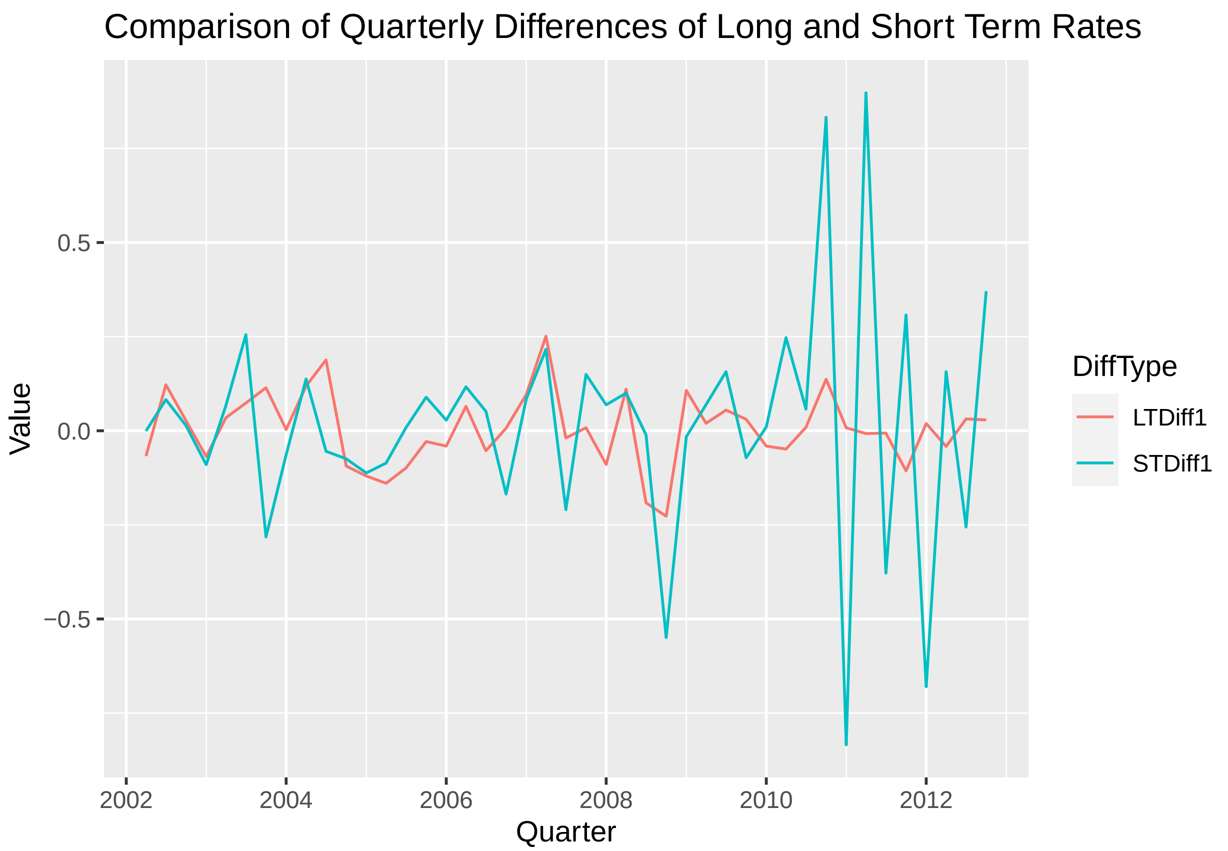


```
## Equity is more volatile with a significant drop due to the impact of the GFC while Bond remains stab


#Quarterly Differences
PvsWInf_compare<-melt(select(ts_data,Quarter,PQuarDiff,WQuarDiff),
                      id.vars = "Quarter",value.name = "Value",variable.name = "DiffType") %>%
  ggplot(aes(x=Quarter,y=Value,colour=DiffType))+geom_line()+ggtitle("Comparison of Quarterly Differenc
PvsWInf_compare
```



```
## The inflation rates are cyclical but price inflation appears more volatile
LvsS_compare<-melt(select(ts_data,Quarter,LTDiff1,STDiff1),
                   id.vars = "Quarter",value.name = "Value",variable.name = "DiffType") %>%
  ggplot(aes(x=Quarter,y=Value,colour=DiffType))+geom_line()+ggtitle("Comparison of Quarterly Differenc
LvsS_compare
```



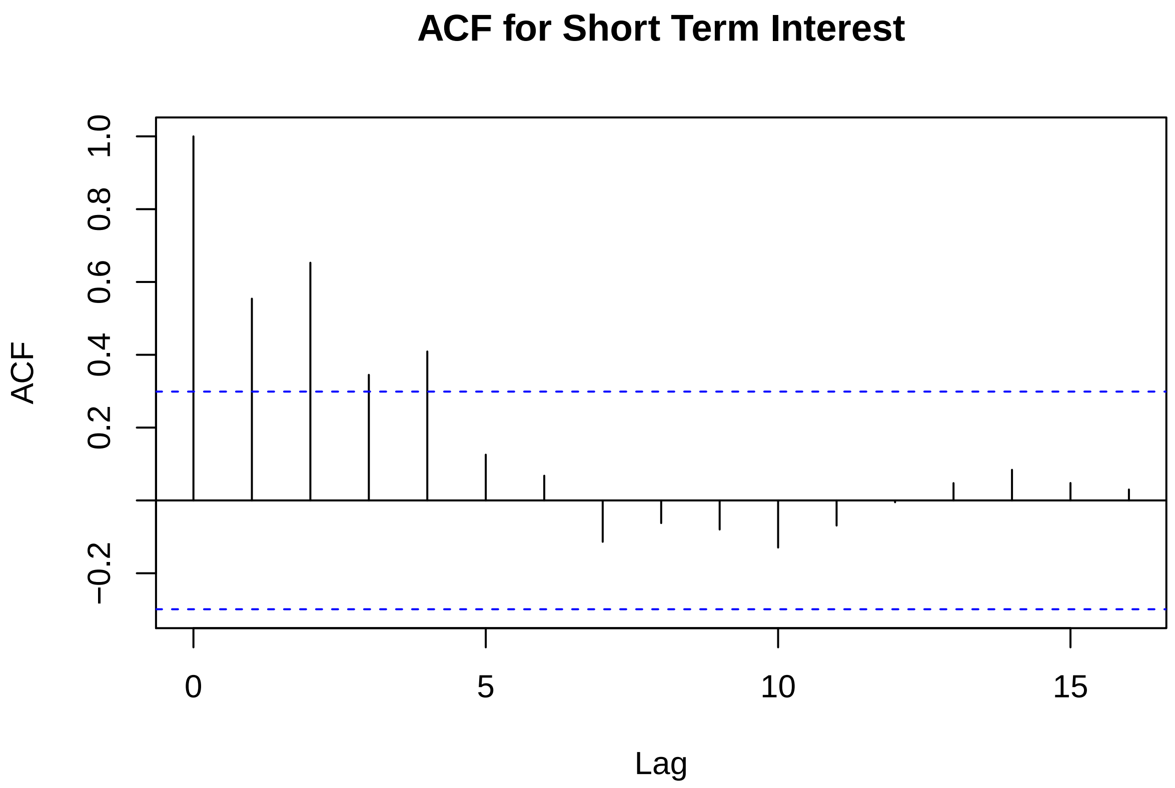
```
## Short term and long term rates are equally until the GFC,
## when short term rates start to fluctuate since then.
```

Question 3:

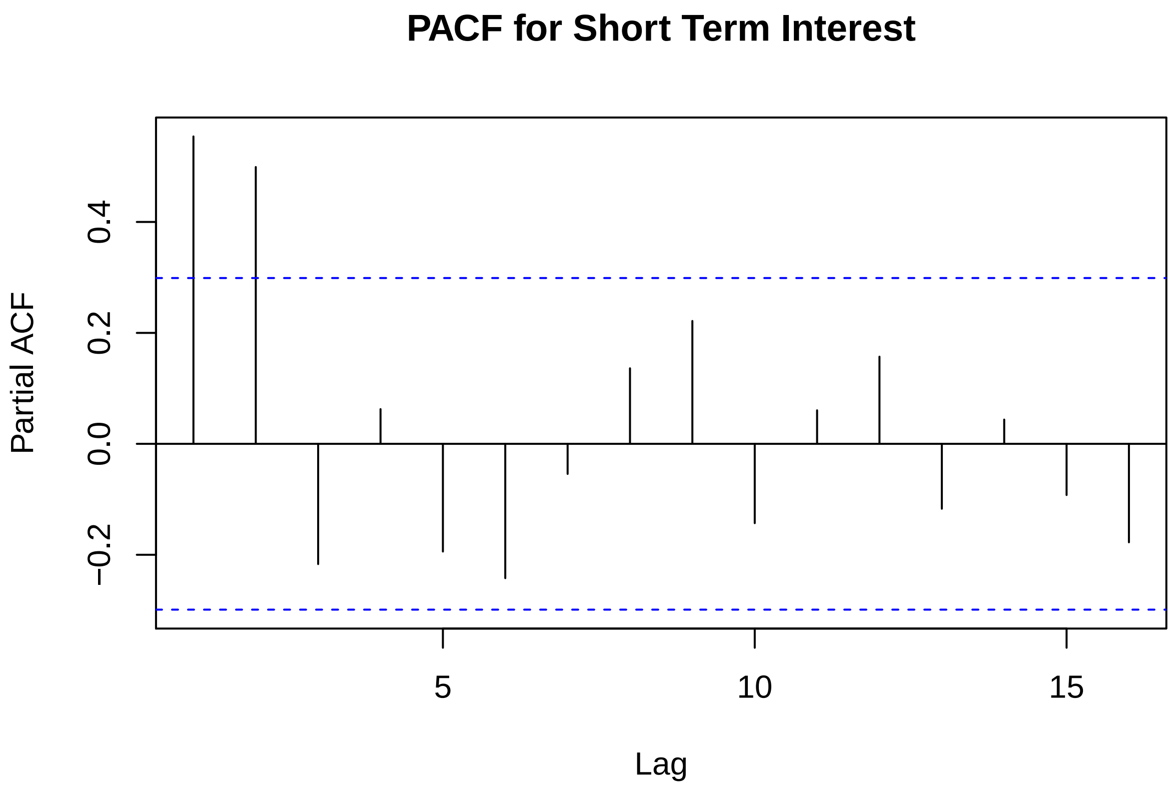
1.Select one variable from the data. Which model would you consider for this variable?

```
#Time Series Analysis

#I choose price inflation
attach(ts_data)
acf(ST_Int, main="ACF for Short Term Interest")
```



```r
pacf(ST_Int, main="PACF for Short Term Interest")
```



```
## Both tail off thus an ARMA is required.

#Model Selection
tsmodel<-auto.arima(ST_Int)
summary(tsmodel)
```

```
## Series: ST_Int
## ARIMA(1,1,0)
##
## Coefficients:
##           ar1
##       -0.6098
## s.e.   0.1211
##
## sigma^2 estimated as 0.05981:  log likelihood=-0.17
## AIC=4.35   AICc=4.65   BIC=7.82
##
## Training set error measures:
##                      ME      RMSE       MAE       MPE     MAPE      MASE
## Training set 0.01826198 0.2388052 0.1725775 -6.827117 31.03855 0.8515079
##                    ACF1
## Training set 0.07460092
```

```
## Resulting model selected is ARIMA (1,1,0)
```

The model selected is differenced first-order autoregressive model. It helps reduce the correlation of the errors of a random walk model by adding one lag of the dependent variable.

$$Y_t = \mu + Y_{t-1} + \phi_1(Y_{t-1} - Y_{t-2})$$

2. What are the features and how to do feature selection for time-series models?

The features for time series model is the lag selection for the model. The easiest way to perform it is to try and fit the lag terms sequentially using stepwise selection and choose the model with the lowest AIC. Although computational expensive, with the size of the dataset we are provided (sample size = 43), it should be sufficient for further analysis.

3. What are the measures you will consider for model selection and assessment?

For model selection, an important criteria will be AIC and BIC to assess the model fit by penalizing parameter with low parsimony. The p-values of the coefficients fitted for the model will be observed. If less than 5%, then they are significant and included for the final model selection. With the residuals, I will use ACF and QQ plot to support the choice of model selection in the lag significance. On top of that, I will use the Ljung Box Test and this will determine whether the autocorrelation for residual errors are non-zero i.e. lack of model fit.

Statistic for Ljung Box Test:

$$Q(m) = n(n+2)\sum_{j=1}^{m}\frac{r_j^2}{n-j}$$

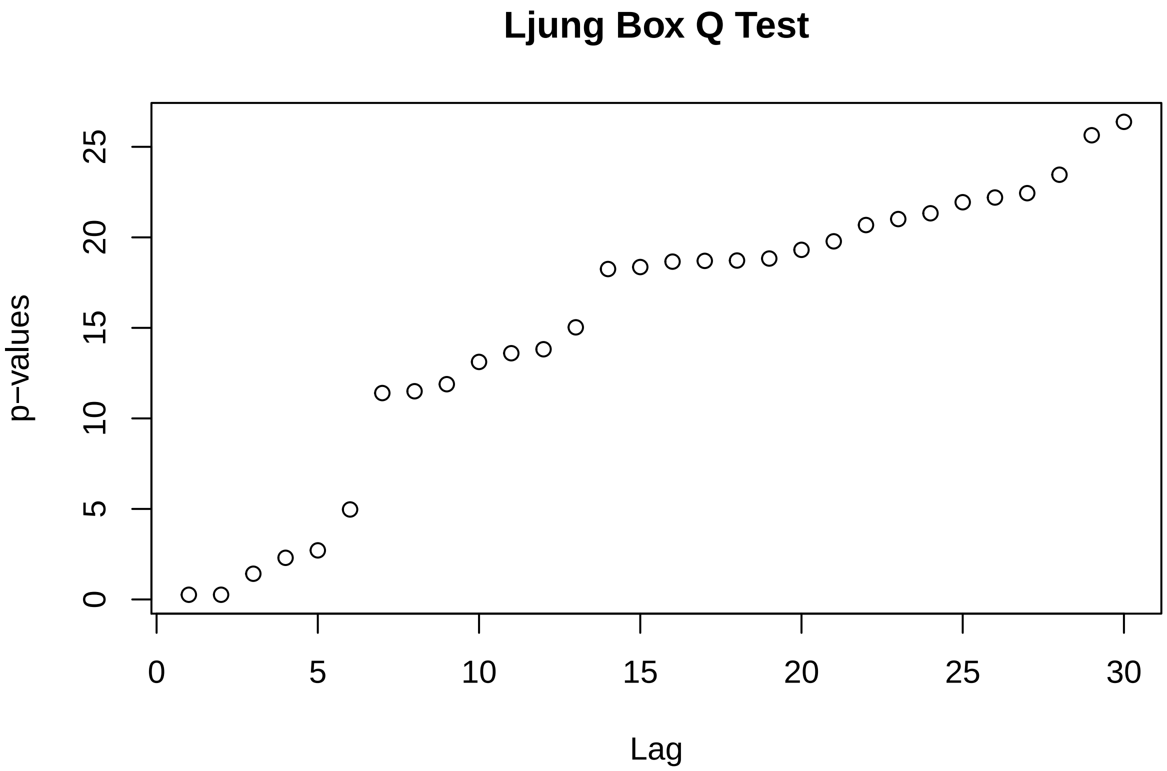
```
#AIC and BIC
tsmodel$aic
```

```
## [1] 4.347023
```

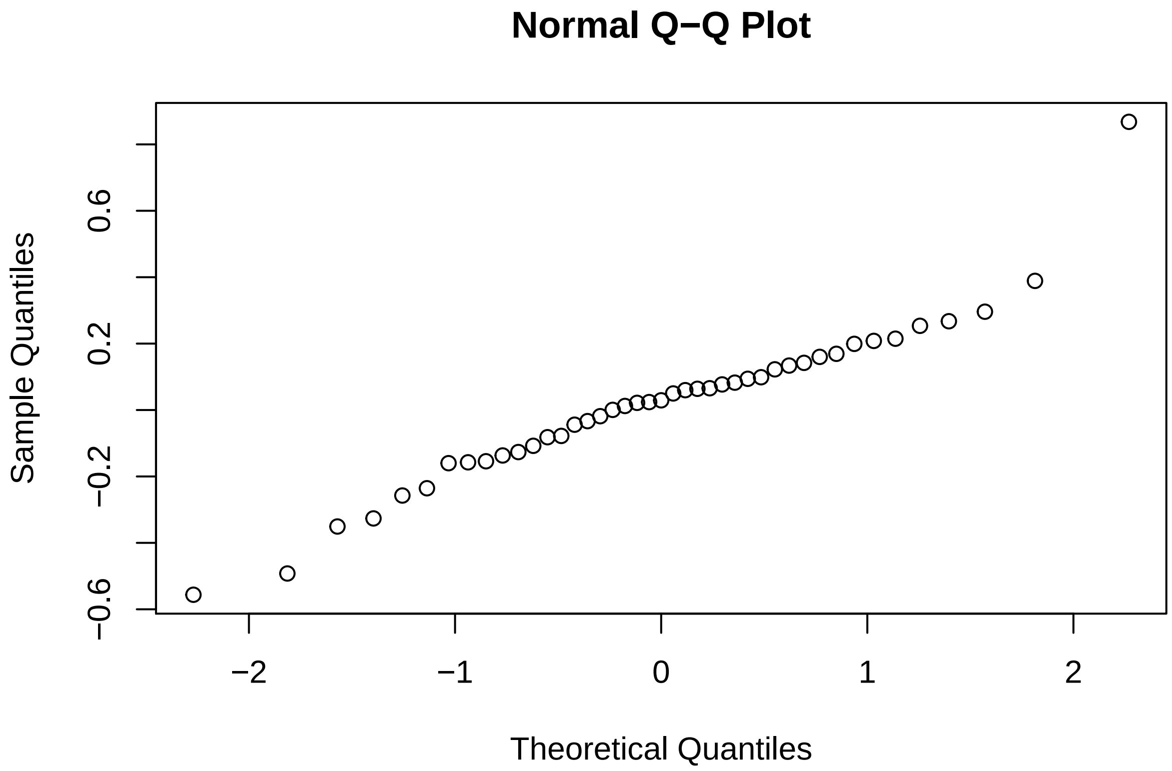
```
tsmodel$bic
```

```
## [1] 7.822362
```

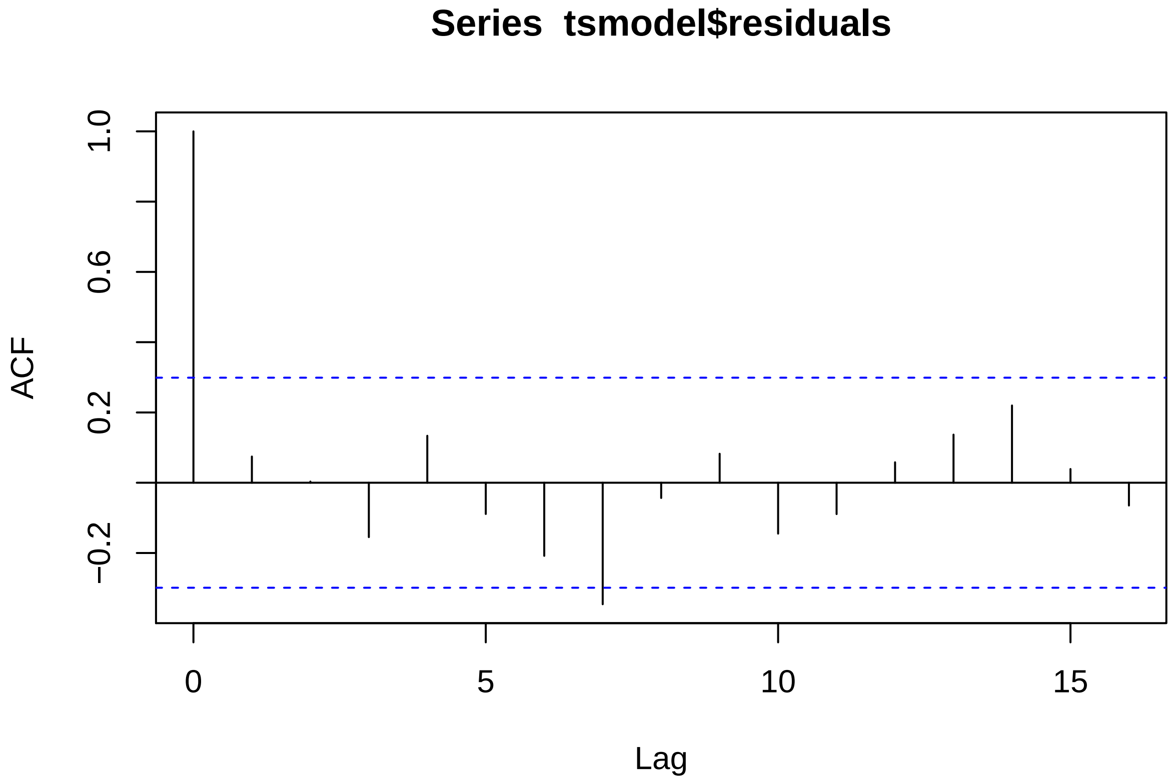
```
#residual analysis
res_plot<-LjungBoxTest(tsmodel$residuals) %>%
  plot(main="Ljung Box Q Test", xlab="Lag",ylab="p-values")
```



```
##QQplot
qqnorm(tsmodel$residuals)
```



```
#ACF of residual
acf(tsmodel$residuals)
```



AIC and BIC are low with respective scores of 4.35 and 7.82. Other than the first two lags, the rest has produced high p-values. The Q-Q plot is almost fitted linearly and ACF shows near-zero auto-correlation for residual. These support that model selection is appropriate for forecasting.

4. Perform model fitting and model selection for this variable.

As performed above.

5. (optional)You could also try other variables and think about how to do multivariate time series modelling considering all variables interested together.

Applying the same procedure with Price and Wage (hidden to save space and results not material for this discussion), the logical way to proceed is to construct a matrix with vector of univariate time series models for each variable and perform model selection and fitting for the corresponding parameters in vectors.

Question 4: If you are now given a much bigger dataset including 100 more macroeconomic related variables, similar to the Australian database. What is your suggestion now for the potential models to use?

From the previous question, a multivariate time series (MTS) will be implemented instead and one such method will be using vector auto regression (VAR). This essentially incorporates both time series properties i.e. the effect of previous measures and the relationships among variables.