CPI and Social Media forecasting

January 2020

By Simon Thiesen

Business Forecasting exam

Business Forecasting

Case 1

1. Plot figures and interpretation

Health Consumer Price Index

Deterministic components

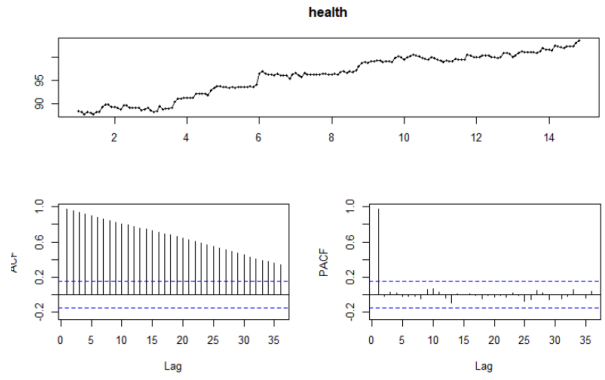
If we look at the health Consumer price index, it might seem like there is a long-term growth in the timeseries which will indicate a trend throughout the entire period. It's very difficult to state whether any seasonal behavior (repeating changes) is present in the data by first glance at the observed values. Cyclical behavior is determined by a longer wavelike fluctuation around the trend which is hard to determine by the visualization.

Stochastic component

Throughout the plot, it seems like there is some irregular or stochastic patterns in the data. To check for irregular stochastic patterns, an Augmented Dickey Fuller test has been conducted. As elaborated later in this section we cannot reject H_0 meaning that we have a unit root in the industrial production data.

Correlogram

The spike on the first lag in the PACF (partial correlation, one correlation at a time) diagram indicates a unit root in the time series. But it will be examined further by conducting an Augmented Dickey Fuller Test. Furthermore, the timeseries in general seems non-stationary and it seems like the ACF (autocorrelation, joint correlation) diagram are slowly decaying with signs of trending behavior in the patterns. The PACF have 1 significant spike. This would indicate an AR(1) model would be a good starting point for further analysis, but after the unit roots test an ARIMA (1,1,0) would be preferred.



H_0 : Unit root
H_1 : No Unit root
With a p-value on 0.5829 we cannot reject H_0 meaning we have a unit root.

```
        Augmented Dickey-Fuller Test

data:  health
Dickey-Fuller = -1.9842, Lag order = 5, p-value = 0.5829
alternative hypothesis: stationary
```

Food Consumer Price Index

Deterministic components

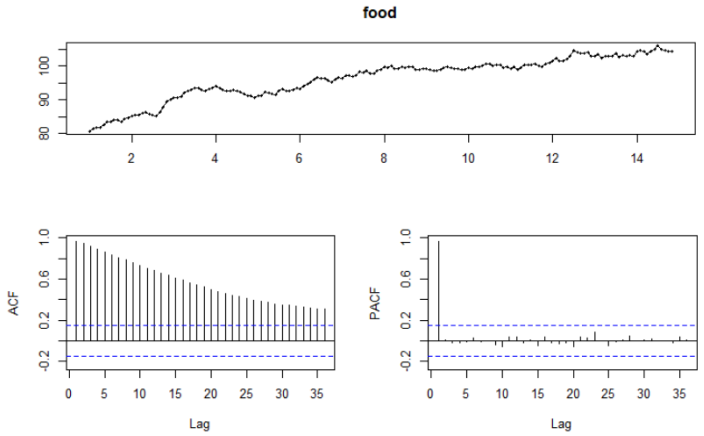
For the food consumer price index, it has some of the same identification points. As well as before It's very difficult to state whether any seasonal behavior is present in the data and the same with the cyclical component.

Stochastic component

There is a unit root in the timeseries meaning that we do have some stochastic trend components in the plot.

Correlogram

The spike on the first lag in the PACF diagram indicates a unit root in the time series. Moreover, the timeseries in general seems non-stationary and it seems like the ACF diagram are slowly decaying with signs of trending behavior in the patterns. The PACF have 1 significant spike. This would indicate an ARIMA(1,1,0) model.




```
> adf.test(food)

        Augmented Dickey-Fuller Test

data:  food
Dickey-Fuller = -2.505, Lag order = 5, p-value = 0.3655
alternative hypothesis: stationary
```

Communication Consumer Price Index

Deterministic components

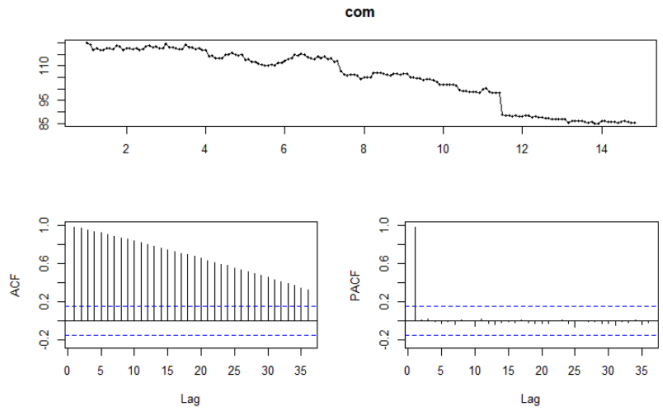
For the communication price index, it has some of the same identification points. However, there is a negative trend combined with the former plots. It's very difficult to state whether any seasonal behavior is present in the data and the same with the cyclical component.

Stochastic component

There is a unit root in the timeseries meaning that we do have some stochastic trend components in the plot.

Correlogram

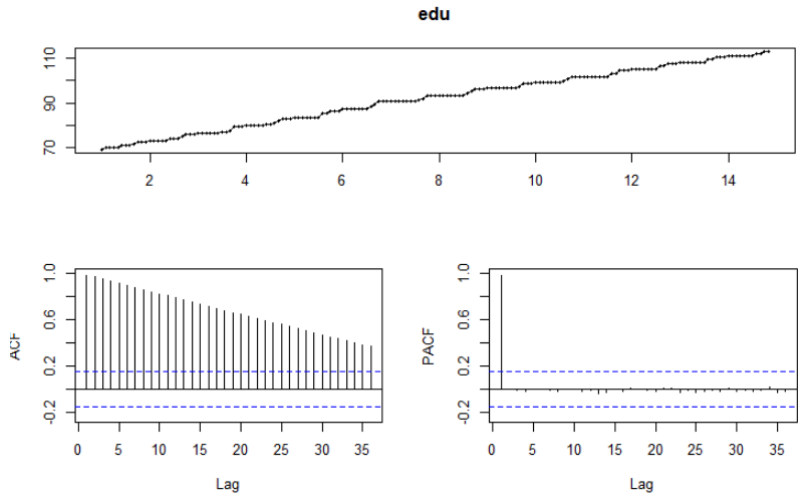
The spike on the first lag in the PACF diagram indicates a unit root in the time series. Moreover, the timeseries in general seems non-stationary and it seems like the ACF diagram are slowly decaying with signs of trending behavior in the patterns. The PACF have 1 significant spike. This would indicate an ARIMA(1,1,0) model.




```
        Augmented Dickey-Fuller Test

data:  com
Dickey-Fuller = -2.6152, Lag order = 5, p-value = 0.3195
alternative hypothesis: stationary
```

Education

Education have almost the same diagnostics as the HCPI plot. An ARIMA (1,1,0) is recommended due to a unit root and a spike in PACF and ACF slowly decay. There is a trending behavior in the observed value.




```
        Augmented Dickey-Fuller Test

data:  edu
Dickey-Fuller = -3.2209, Lag order = 5, p-value = 0.08672
alternative hypothesis: stationary
```

2. Dynamic model specification and forecasting

As mentioned in the description of the correlograms in assignment 1, an appropriate starting point would be an ARIMA(1,1,0) with Box-Jenkins methodology for HCPI. This model specification has been justified due to the significant lags in the PACF diagram and the decay in the ACF diagram and the unit root. Both of the forecast are plottet below.

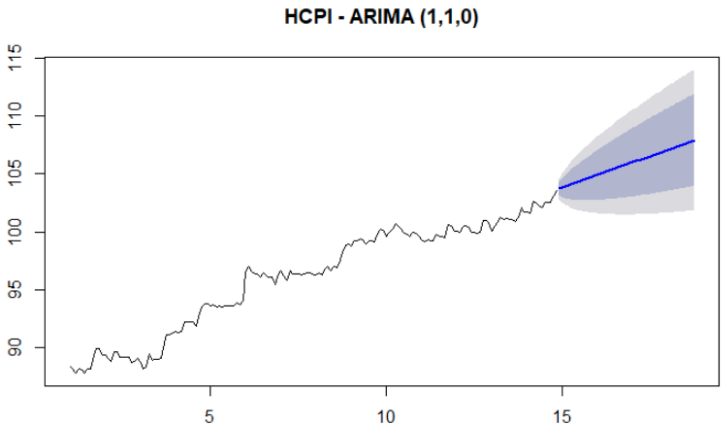
AR definition: ACF should be exponential decaying towards zero exponentially fast

Typical PACF will spike in **P** lags (**P** significant spikes)

MA definition: Typical ACF will spike in **q** lags (**q** significant spikes - above the significant level)
PACF = exponential decaying towards zero exponentially fast

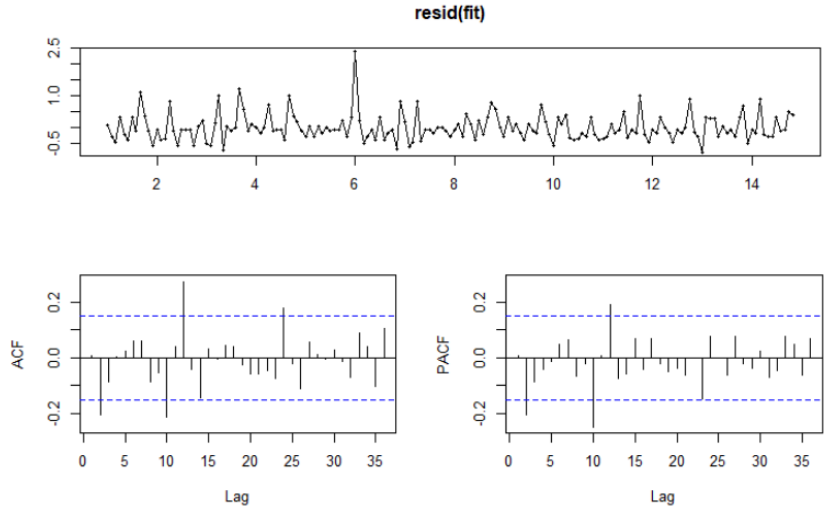
And the **d** in the ARIMA function are the regular differences.

ARIMA(p,d,q)



The following RMSE for the forecast compared to the out sample.

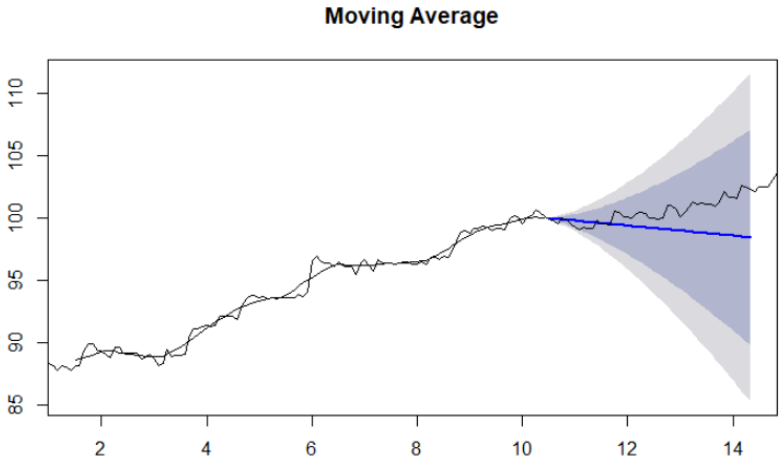
	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	0.000599412	0.4306173	0.3128905	-0.0002126925	0.3265438	1.055687	0.007543535	NA
Test set	-3.520681099	3.5316285	3.5206811	-3.5098042273	3.5098042	11.878721	0.129367284	14.70604



The post regression residuals are somewhat close to white noise, but not entirely. There are still some significant spikes in the ACF and PACF curve. This would not satisfy the OLS assumptions.

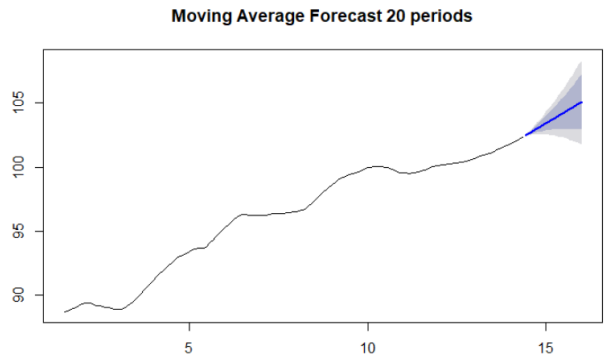
The simple moving average

If we compare the 2 models on the RMSE we would have a more precise forecast result using the moving average method. Therefor the Moving Average will be the one executing the projection.




```
> accuracy(fcast1, outsamp)
                      ME       RMSE        MAE           MPE       MAPE      MASE       ACF1 Theil's U
Training set -0.0009798621 0.03484671 0.02368682 -0.0008349437 0.02525856 0.1988485  0.5170152        NA
Test set      0.2374995822 0.28665257 0.23749958  0.2367809779 0.23678098 1.9937853 -0.1219702 0.6247419
> 
```

Forecasting 20 months ahead



3. Multiple regression and OLS assumptions

In a regression setting you can see that the relationship between HCPI, CCPI, EDCPI and FCPI are all significant meaning that the they are related. If CCPI increases by 1 unit the IP has a 0.19277-unit increase and vice-versa. You could in this case run a multicollinearity test to see if the multiple regression setting is better off without one of the variables. Adjusted R-square are adjusting for the number of independent variables in the model and shows how well the regression model predicts responses for new observations

and penalizing overfitting. In this case 97.3% of the variability in the dependent variable (HCPI) can be explained by the other variables. The regression model for Industrial Production would look like this:

Multiple regression
$ß_0$: Intercept (When X_1 , X_2 and X_3 equals 0)
$ß_1$: Partial regression coefficient (average change in Y per unit change in the specific independent variable holding the other variables constant)

Statistical evaluation - Jointly significance
F-testing = joint hypothesis testing (summary of the fits - are testing entire significant of regression) and in this case the entire regression is significant.

H_0: $b_1 = 0 \ldots = b_k = 0$
H_1: at least one $b_i \neq 0$

The regression equatation:

$$y(HCPI) = 37.49 - 0.18FCPI + \ 0.19277\text{CCPI} + 0.61055\text{ECPI}$$

```
Call:
lm(formula = health ~ food + com + edu)

Residuals:
     Min       1Q   Median       3Q      Max
-2.20846 -0.44287  0.02639  0.49777  1.60610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.48901    2.86314  13.094  < 2e-16 ***
food        -0.18365    0.03695  -4.971 1.68e-06 ***
com          0.19277    0.01753  11.000  < 2e-16 ***
edu          0.61055    0.02875  21.238  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7533 on 163 degrees of freedom
Multiple R-squared:  0.9735,    Adjusted R-squared:  0.973
F-statistic:  1997 on 3 and 163 DF,  p-value: < 2.2e-16
```

In general, you will need OLS assumptions to be satisfied to have a valid regression. Otherwise the outcome of the regression would not be reliable.

OLS assumptions:
1. The residuals are normally distributed
2. The residuals have constant variance (homoscedastic)
3. The residuals are uncorrelated (no autocorrelation)
4. The residuals have a linear relationship

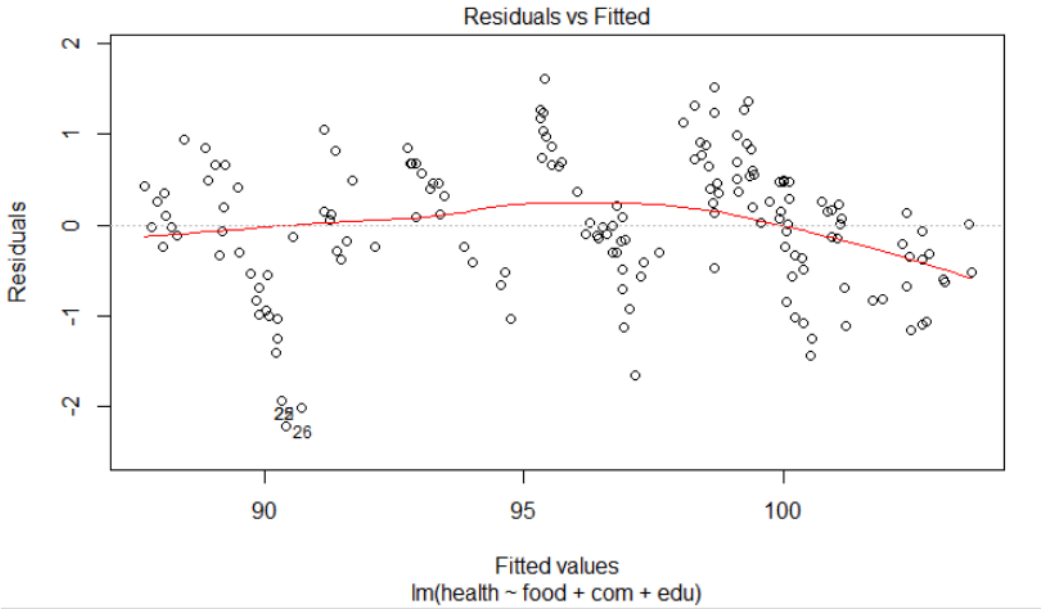
If all assumptions are satisfied it means that Epsilon (error term) is **white noise.**

To check if you have satisfied the OLS assumptions you will conduct a goodness of fit test which is to check if the error term (epsilon) are normal distributed, you will need a constant variance (Homoskedasticity) and epsilon cannot be autocorrelated. The intuition about OLS assumptions on the regression are stated below.

Residuals vs. Fitted

Residuals vs. fitted will plot the residuals against the explanatory variables. This means that the magnitude of the datapoint should be around 0. A spread around 0 will indicate a that the variances of the error terms are equal (homoscedastic). Furthermore, if you can see that the residuals are "randomly" plotted around the 0 line, this will further indicate that the assumptions about the relationship being linear is reasonable.

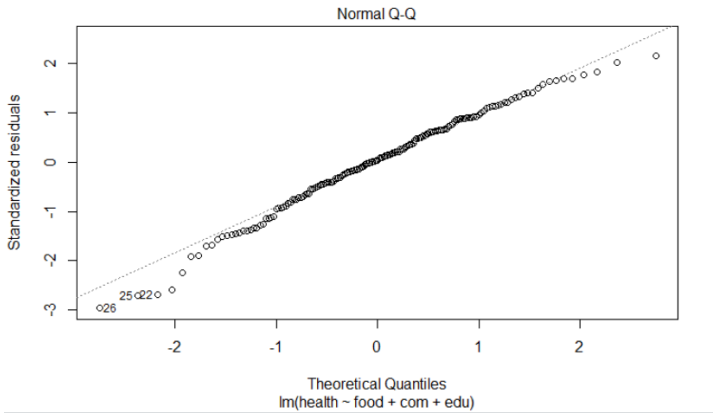
In this case it looks like the variance is changing over time, and the spread does not seem to be equal (to the red line) across the plot. There are also some residuals who "stands out" (in the bottom left corner) which suggest some outliers in the data. The curved red line indicates a non-linear relationship.



Normal Q-Q

In the normal Q-Q plot you can check whether or not epsilon is normal distributed, if the datapoints follows the line, it indicates a normal distribution.

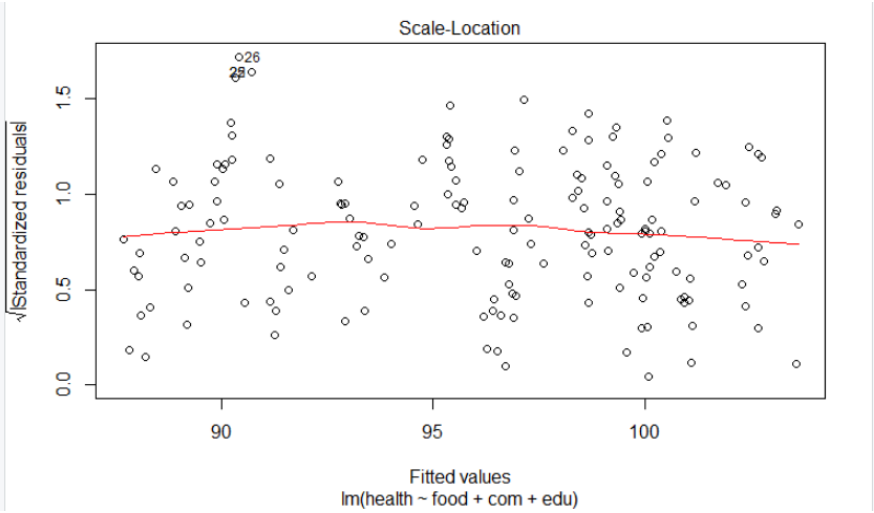
In this case it seems like the data is somewhat close to normal distribution even though the ends are a bit skewed away from the line but enough that the normal distribution assumptions do not seems satisfactory.



Scale-Location

Also called the spread location plot, this plot will show if the residuals are equally spread along the predicted value. This is how you get an intuition about equal variance and if homoskedasticity is present or not. A good fit would be a horizontal line with equally spread of points.

In this case it indicates very random movement away from the red line, it's difficult in this plot to state something very clearly but it seems that the variance is not constant, and the constant variance assumption is also violated.



Durbin-Watson test
To test for autocorrelation a DW test is conducted.

Hypothesis:
H_0: No autocorrelation
H_1: Autocorrelation

According to the P-value of the DW test we can reject H_0 even with a 1% confidence interval, meaning that autocorrelation is detected in the data. A DW score of 0.45109 is considered auto correlation, and it might indicate a positive auto correlation because the value is below 2, which is also confirmed on the coefficients.

```
        Durbin-Watson test

data:  linmod
DW = 0.45109, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Breusch-Pagan test
To check for homo-/heteroskedasticity a Breusch-Pagan test has been conducted with the following hypothesis:

H_0: Variance is constant (Homoskedastity)
H_1: Heteroskedasticity
According to the P-value of the bp test we can reject H_0, meaning that the variance is not constant (heteroskedastic).

```
        studentized Breusch-Pagan test

data:  linmod
BP = 8.6323, df = 3, p-value = 0.0346
```

To sum up, it seems like all the OLS assumptions has been violated and should therefore be fixed.

4. Transformation

If the data are trending (non-stationary), then some form of trend removal is required.
Two common trend removal or de-trending procedures are

- First differencing
- Time-trend regression.

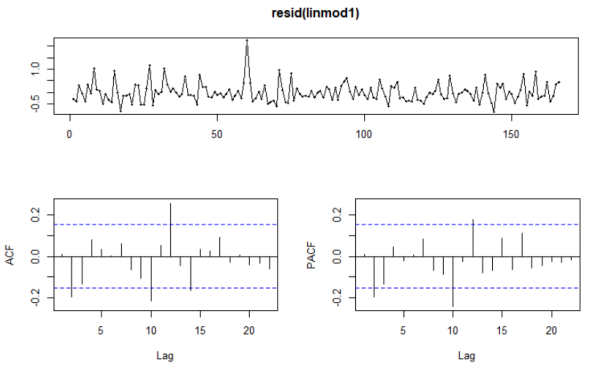
Unit root tests can be used to determine if trending data should be first differenced or regressed on deterministic functions of time to render the data stationary.

When a unit root is present in the residuals of the model a normal procedure would be to differentiate the timeseries, so you'll remove (or lower) the stochastic trend component. In this case this is the right way to do because we have a very low p-value on the ADF test meaning we can reject H_0 and therefor there is a unit root in the residuals.

```
        Augmented Dickey-Fuller Test

data:  resid(linmod)
Dickey-Fuller = -4.069, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

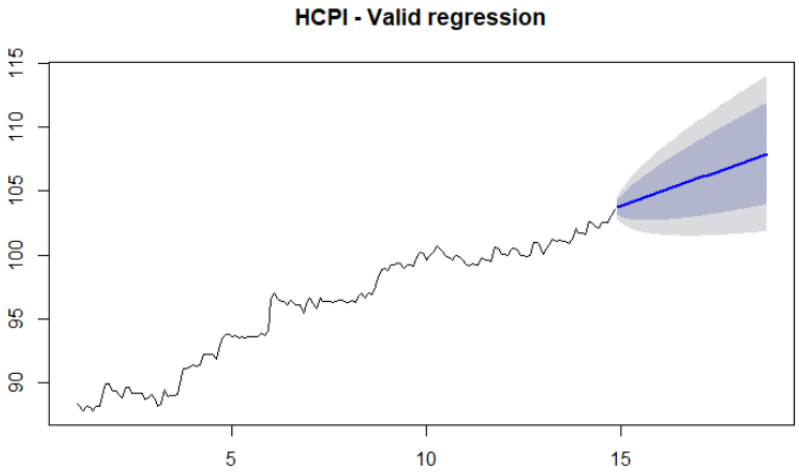
In this case 1st differencing of the timeseries are removing the autocorrelation and the heteroskedasticity. And we can therefor proceed with the valid prediction for the forecast.

```
data:  linmod1
BP = 5.2131, df = 3, p-value = 0.1568

> dwtest(linmod1)

        Durbin-Watson test

data:  linmod1
DW = 1.9713, p-value = 0.4263
alternative hypothesis: true autocorrelation is greater than 0
```



The residuals from the differentiated regression are not completely white noise (some significant spikes), but the DWtest and the BPtest both indicated that there was no autocorrelation and heteroskedasticity.

5. Valid regression prediction



6. Unit root and cointegration

As tested in assignment 1, unit roots is detected in all of the timeseries.

Unit root/persistence/non stationary series:
If a unit root exists, it exists in the irregular component in a series. When something gets shocked it will remain "not the same" for infinite periods (never gets back to the mean again) meaning that the shock ($b_i > 1$) has permanent impact on the data.

Cointegration

4 non-stationary series should not be regressed upon each other due to high correlation based on their beta (trend). BUT some series might be individually stochastically trending, but a combination of them will have a long run equilibrium. This is cointegration. Cointegration makes the "jump" between the two timeseries will never deviate to far from each other. Therefor it is easier to regress the variables upon each other and see the true relationship between the two variables.

Testing for co-integration

1. Both $Y_t \; and \; x_t$ are I(1) - *unit root/non stationary* -
   because otherwise e(0) condition won't be satisfied
2. $y_t = ßx_T + e_t$
   Hypothesis: H_0: ß = 0
   ß should not be zero = non stationary
3. $\hat{e}_t \; is \; I(0) \; - stationary$

Spurious regression

1. Both $Y_t \; and \; x_t$ are I(1) (Unit root)
2. $y_t = ßx_T + e_t$
   Hypothesis: H_0: ß = 0
3. $\hat{e}_t \; is \; I(1) \; - they \; are \; divergent$

Spurious regression is where correlation is due to coincidence and this is not intended for regression analysis.

To check for cointegration you can either use the Phillips-Ouliaris (2-step Engle-granger) or the Engle Granger.

Both testing methods have the following hypothesis:
H_0: No cointegration
H_1: Cointegration

```
        Phillips-Ouliaris Cointegration Test

data:  combinedvector
Phillips-Ouliaris demeaned = -39.589, Truncation lag parameter = 1, p-value = 0.01427

                    Augmented Dickey-Fuller Test

            data:  resid(linmod)
            Dickey-Fuller = -4.069, Lag order = 5, p-value = 0.01
            alternative hypothesis: stationary
```

Both tests show a low p-value meaning that we can reject H0 with a 95% confidence interval. This means that there is cointegration in the timeseries combined. This means that the cointegrated coefficient estimate will converge to the true value (which is nonzero) and indicate a long run equilibrium relationship. It is very likely that a change or movement in one of the variables will have something to do with each other.

Cointegration are making the 4 non-stationary series = stationary.

7. VAR setting

We do not need to differentiate due to stationarity of the series combined.

To calculate the number of lags to include, an VARselect has been conducted to determine the amount of lags to include in the VAR setting. In this case the BIC/SC parameter choose 1 as the number of lags to include.

Cholesky ordering
->Check relative exodegenity
Exogenous: Not determined by others - driven by iid shocks
Endogenous: Determined by one or more variables - as a function of some indicators

They are all endogenous variables, but the food equation should be in first ordering due to most significant lags on the other series.

Roots of the characteristic polynomial, stability is when each root is below 1. In this case we have a proper VAR specification, if it was above 1, the VAR will give results that is not equal to the dataset

In the coefficients we are interested in the dependencies between the series itself
The Health estimation results shows that the health is significant at a 5% confidence interval on its own first lags and the others are not significant.

```
VAR Estimation Results:
========================
Endogenous variables: health, com, edu, food
Deterministic variables: const
Sample size: 119
Log Likelihood: -360.484
Roots of the characteristic polynomial:
0.9971 0.9535 0.8866 0.7535
Call:
VAR(y = z, p = 1, type = "const")


Estimation results for equation health:
=======================================
health = health.l1 + com.l1 + edu.l1 + food.l1 + const

           Estimate Std. Error t value Pr(>|t|)
health.l1  0.836442   0.058777  14.231   <2e-16 ***
com.l1     0.015028   0.019913   0.755   0.4520
edu.l1     0.081284   0.041567   1.956   0.0530 .
food.l1   -0.007881   0.027741  -0.284   0.7769
const      7.593513   4.235304   1.793   0.0756 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.4409 on 114 degrees of freedom
Multiple R-Squared: 0.9886,     Adjusted R-squared: 0.9882
F-statistic:  2461 on 4 and 114 DF,  p-value: < 2.2e-16
```

```
Estimation results for equation com:
====================================
com = health.l1 + com.l1 + edu.l1 + food.l1 + const

          Estimate Std. Error t value Pr(>|t|)
health.l1 -0.03667    0.10371  -0.354    0.724
com.l1     0.94332    0.03611  26.123   <2e-16 ***
edu.l1    -0.03034    0.06620  -0.458    0.648
food.l1    0.02660    0.05087   0.523    0.602
const      0.44612    0.37628   1.186    0.238
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.008104 on 114 degrees of freedom
Multiple R-Squared: 0.9795,     Adjusted R-squared: 0.9787
F-statistic:  1359 on 4 and 114 DF,  p-value: < 2.2e-16
```

```
Estimation results for equation edu:
====================================
edu = health.l1 + com.l1 + edu.l1 + food.l1 + const

          Estimate Std. Error t value Pr(>|t|)
health.l1  0.05262    0.05898   0.892    0.374
com.l1    -0.03185    0.01998  -1.594    0.114
edu.l1     0.93775    0.04171  22.481   <2e-16 ***
food.l1    0.03377    0.02784   1.213    0.228
const      1.03324    4.25016   0.243    0.808
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.4425 on 114 degrees of freedom
Multiple R-Squared: 0.9979,     Adjusted R-squared: 0.9978
F-statistic: 1.335e+04 on 4 and 114 DF,  p-value: < 2.2e-16


Estimation results for equation food:
====================================
food = health.l1 + com.l1 + edu.l1 + food.l1 + const

          Estimate Std. Error t value Pr(>|t|)
health.l1 -0.11321    0.06389  -1.772  0.07908 .
com.l1     0.05395    0.02165   2.493  0.01412 *
edu.l1     0.12641    0.04518   2.798  0.00604 **
food.l1    0.89492    0.03016  29.677  < 2e-16 ***
const      3.83087    4.60381   0.832  0.40709
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.4793 on 114 degrees of freedom
Multiple R-Squared: 0.9924,     Adjusted R-squared: 0.9921
F-statistic:  3701 on 4 and 114 DF,  p-value: < 2.2e-16
```

```
        Portmanteau Test (asymptotic)

data:  Residuals of VAR object var
Chi-squared = 184.99, df = 144, p-value = 0.01202

> arch.test(var)

        ARCH (multivariate)

data:  Residuals of VAR object var
Chi-squared = 521.95, df = 500, p-value = 0.2404
```

The test conducted to see if autocorrelation and heteroskedasticity was present showed that both was present in the data and therefore the VAR was not valid. The next step is to transform the variables independent in an ARCH setting to get rid of autocorrelation and heteroskedasticity and then run the VAR setting again.

Transforming the VAR

We managed to get rid of both autocorrelation and heteroskedasticity at a 5% confidence interval meaning we now have a valid regression and can proceed.

```
Diagnostic Tests:
        Jarque Bera Test

data:  Residuals
X-squared = 1.9815, df = 2, p-value = 0.3713


        Box-Ljung test

data:  Squared.Residuals
X-squared = 0.078106, df = 1, p-value = 0.7799
```

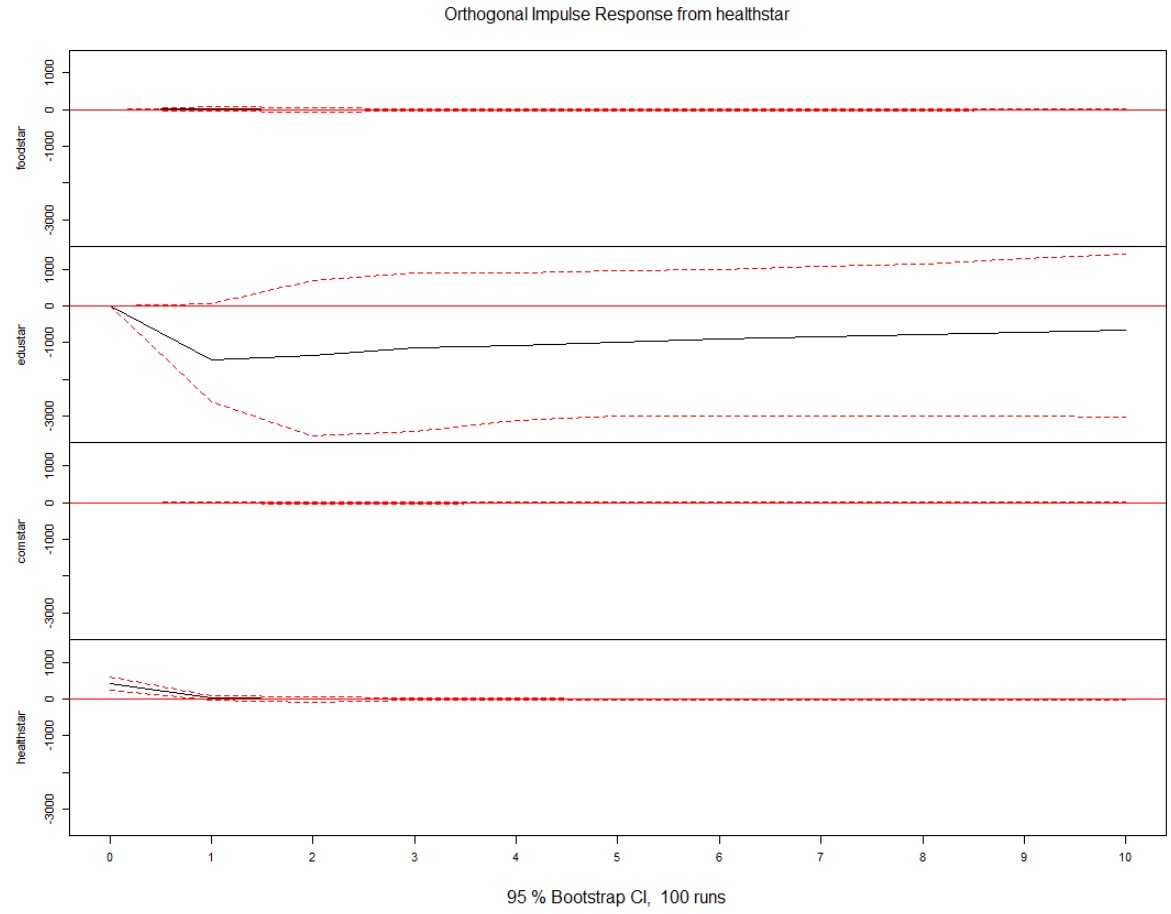
PLEASE NOTE:

One of the roots of characteristic polynomial are above 1, and therefor this is not a valid regression. However, I will make the explanations to show the understanding behind IR and FEVD.

Impulse responses

The impulse responses are plottet for each timeseries with a 95% confidence interval, showing if shocks will have impact on the variables. Significant shocks are the areas with the confident interval not containing the zero lines on the plot.

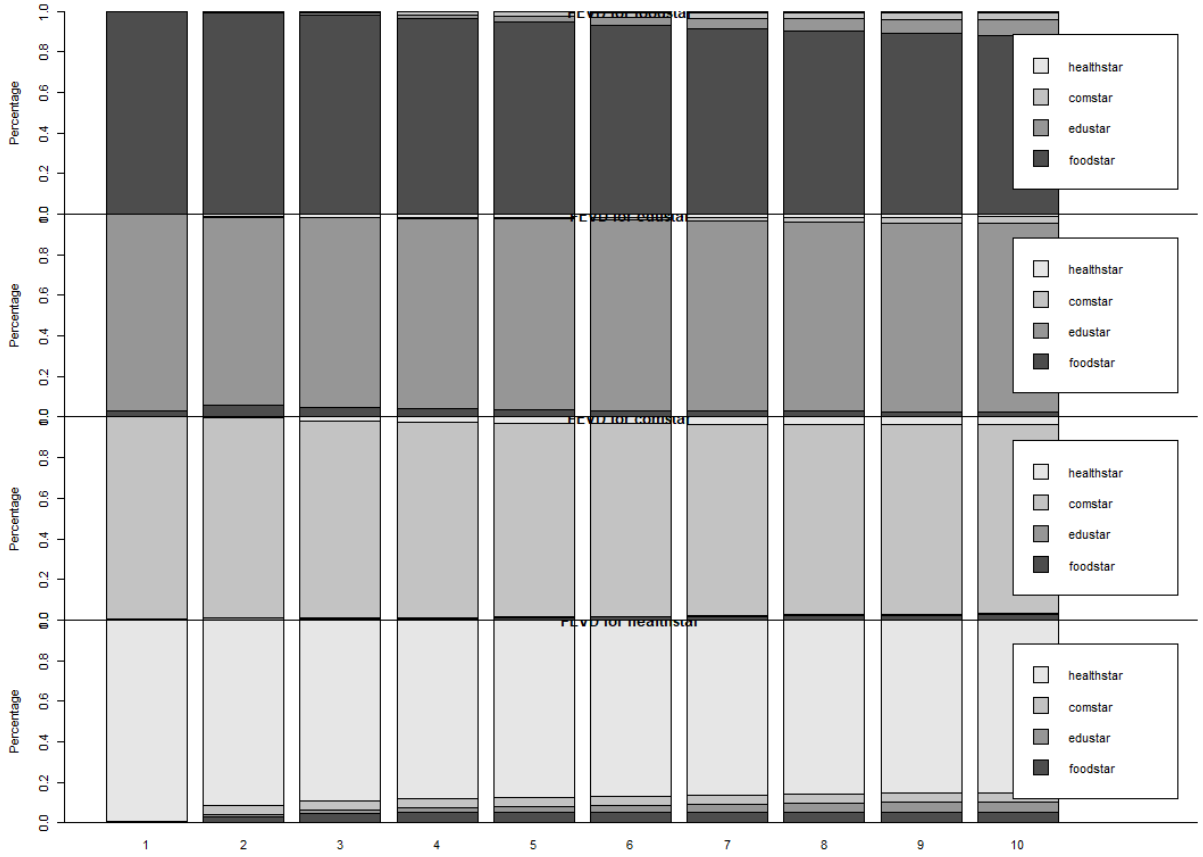
The impulse response on HCPI will have impact on the short run on itself and until the 1st lag on regulation. This will mean on short and medium run it will have an impact on HCPI itself. Shocks on HCPI would have no impact on the other variables on either short, medium or long-run.



Forecast error variance decomposition

This is the forecasted error variance for 10 periods ahead. On the FEVD plot you can tell how much of the variance that is explained by other variables.

In this case if you look at **FEVD for Health (Bottom)** you can see that until the 2nd lag, all of its variance is explained by itself. This means that if you should forecast until 2nd lag on Health you could probably use an AR structure with the Principle of parsimony in mind. If you go beyond the second lag and VAR approach is justified.



Forecast

The forecast will not be showed due to the invalid VAR setting.

8. Executive summary

Executive summary

Through the previous steps in the report, this executive summary will explain key points to asses when looking at Health Consumer Price Index (HCPI) and Education Consumer Price Index (ECPI), Food consumer Price Index (FCPI) and Communication Consumer Price Index (CCPI). All the provided recommendations are based on the findings in the conducted technical report.

The environment impacts the Industrial production

The HCPI are having a long-term trend upwards for the past 13. Seasonal patterns are not detected throughout the period, this means that the consumers are paying more for Health than they did 13 years ago. The HCPI goes very well in hand with the all other Consumer Price Indexes which could indicate growing prices in general.

Increased HCPI in the future

The HCPI are based on projection estimated to keep rising for the next 20 months, this will as stated before make the consumer pay even more than today for Health. If we are looking at all the variables combined it seems like they have a long-term relationship meaning that it is very likely that a change or movement in one of the variables will have impact on HCPI.

HCPI and the other variables

The variables are very well interconnected meaning that a 1 unit increase in the ECPI would cause a 0.61 increase in the HCPI. But a 1 unit increase in the FCPI would cause a decline in HCPI on -0.18. While a 1 unit increase in CCPI would cause a 0.19 increase in HCPI. This would indicate that a rise in the CCPI could make the HCPI reach higher index levels. This is very useful information when channeling this assessment to the ministry of finance. Because this shows the effect of regulations on certain areas would cause a cascade of reactions to the other consumer price indexes.

CASE 2

1. Plot figures and interpretations

Facebook

Deterministic components

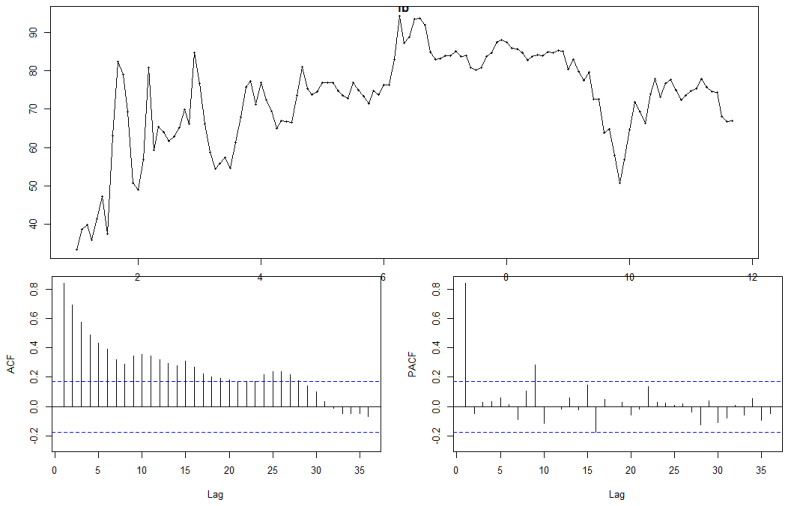
The Facebook timeseries are looking like it certainly has a trending component, and in the correlogram it seems like there is some seasonality behavior in the ACF diagram. The cyclical factor is hard to determine on this plot.

Stochastic component

The spike on the PACF diagram might indicate a unit root and some stochastic properties.

Correlogram

The spike on the PACF and the slowly decay on the ACF diagram indicates an AR(1) would be a good starting point for analysis.



The ADF test indicates a unit root meaning that an ARIMA (1,1,0) would most likely be better.

```
        Augmented Dickey-Fuller Test

data:  fb
Dickey-Fuller = -2.9613, Lag order = 5, p-value = 0.1769
alternative hypothesis: stationary
```

2. Decomposition

In a decomposition you can choose between additive or multiplicative

Additive decomposition

$$Y = T + S + C + I$$

Multiplicative decomposition

$$Y_t = T_t * S_t * C_t * I_t$$

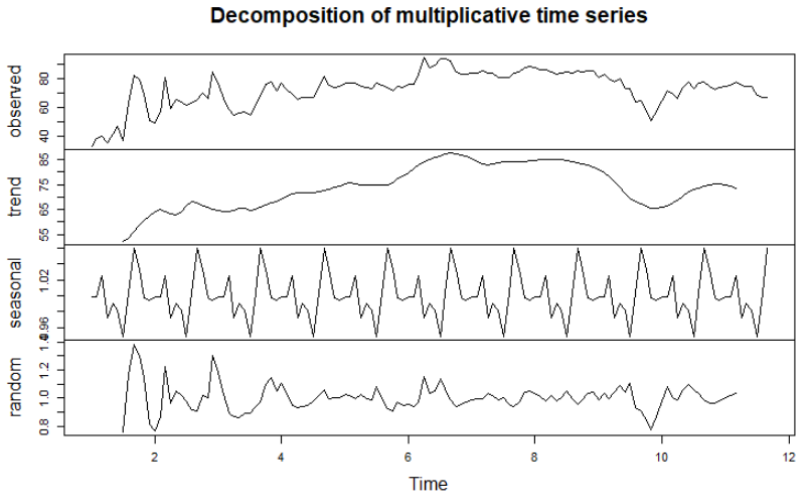
When chosing an multiplicative decomposition Y_t is not a sum, but as a product.

Observed value are the ones actually in the dataset. (like the one in assignment 1)

The trend is then isolated into one component, and in this case, it seems like there could be some trend with some clusters in the middle of the plot.

Seasonality is the 3rd component where the dataset does contain some pattern with repetitive behavior.

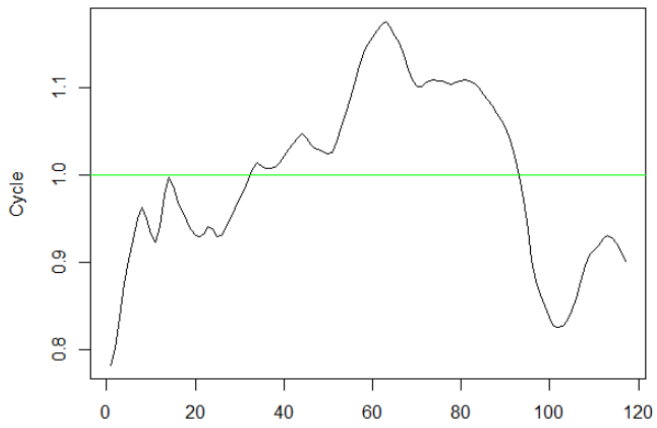
Random (stochastic) part is the 4th graph and here you have the irregular component (epsilon) where you want this component to be white noise as described earlier. On this graph it might seems like there is still some information left in the error term eg. not iid.



Detrending using a trend dummy variable

You can also measure the cyclical behavior by making a Cycle factor = ratio of CMA to CMAT where CF>1 the depersonalized value is above the long-term trend of the data CF<1 the depersonalized value is below the long-term trend of the data.

But you have to remove the trend and seasonality and this is to remove short term fluctuations (seasonal patterns and irregular variations) to focus on longer-term trend and cycles.



Long run trend

You can see on the plot that in the beginning the monthly user percentage data on facebook are below the long run trend, while in the middle it is above. And at the end it is again below the long run trend. This backs up my intuition about the clustering behavior in the middle of the plot.

3. VAR setting

To treat the variables in a VAR setting they are both combined into a vector with a log functions (removing extreme values/outliers). But before that you will need to balance the timeseries and make sure they all are stationary. Otherwise you will run into a suppression from the unit root and dominate the stationary series.

LinkedIn are the only variable not containing a unit root, and therefor the 2 variables are differentiated to balancing the VAR. When I made the VAR the ROCP where all below one, but in the VAR there was detected heteroskedasticity and it must therefor be transformed in a GARCH setting.

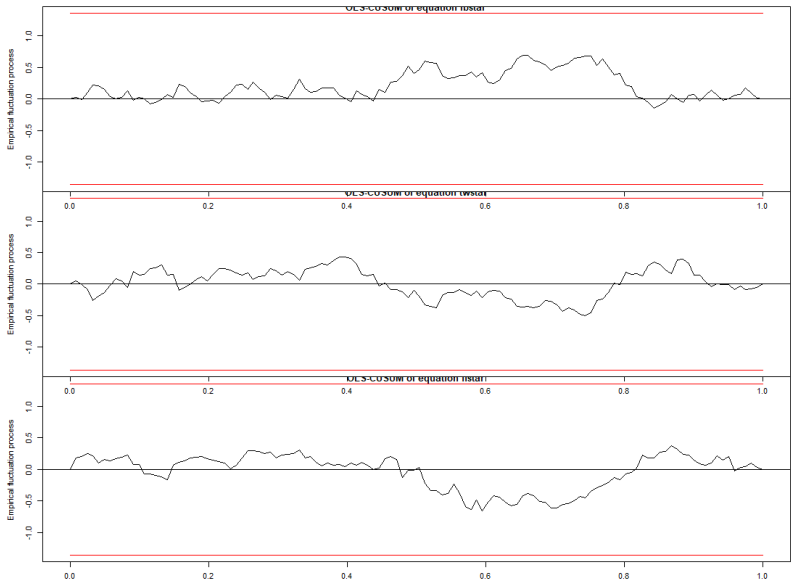
After the GARCH transformation of each variable it has removed the heteroskedasticity and it is now a valid regression. The lag chosen is 3. BIC stated 1 lag but it did not remove the autocorrelation entirely, however AIC stated 3 and it removed the autocorrelation.

```
> serial.test(varcase2, lags.pt=10, type="PT.asymptotic")

        Portmanteau Test (asymptotic)

data:  Residuals of VAR object varcase2
Chi-squared = 76.52, df = 63, p-value = 0.1178

> arch.test(varcase2)

        ARCH (multivariate)

data:  Residuals of VAR object varcase2
Chi-squared = 165.55, df = 180, p-value = 0.7726
```

OLS cumsum

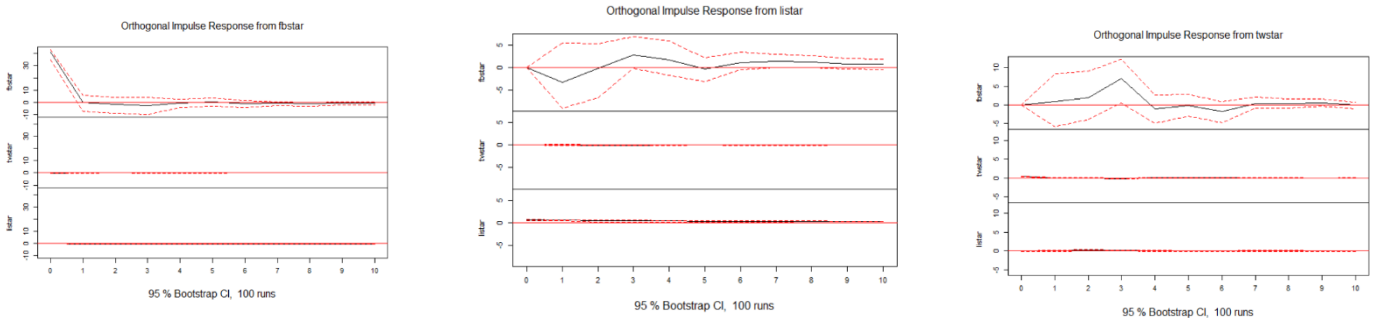
OLS-CUSUM test

Looking at the OLS-CUSUM test is testing for stationarity and OLS assumptions. The red lines are the boundaries and if the lines are within the red lines it means that the VAR is stable is it is in this case. If it was exceeding the red lines – then it's probably due to a non-stationary variable. This VAR is stable.



NOTE: only the FB will be examined due to time constraints.

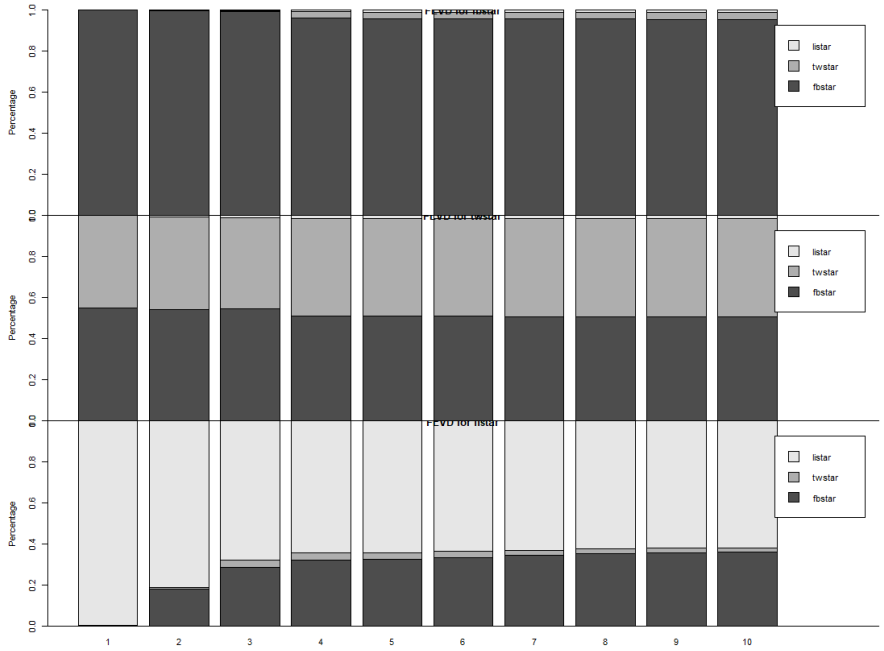
You can see that shocks on FB will have an impact on itself until the first lag (short term), the shock on FB will not impact Twitter and LinkedIn



FEVD

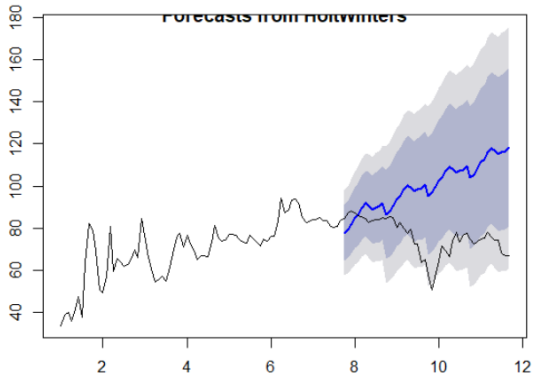
This is the forecasted error variance for 10 periods ahead. On the FEVD plot you can tell how much of the variance that is explained by other variables.

In this case if you look at **FEVD for FB (top)** you can see that until the 3rd lag, all of its variance is mainly explained by itself. Forecasting until 3rd lag on FB an AR structure is desirable. If you go beyond the third lag the VAR approach is justified.



4. Smoothing method forecast

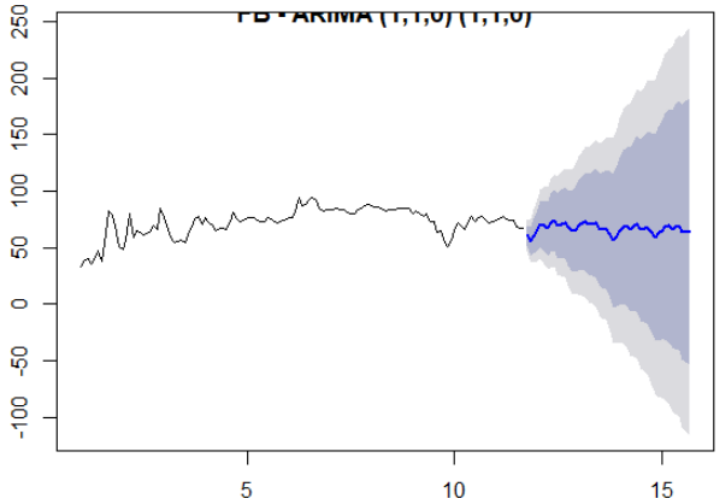
Using a Holt-Winthers smoothing method to forecast would be appropriate, which adjust for seasonal and trending behavior because it was detected in assignment 1 & 2. The projection is below and the RMSE is 3.55.




```
> accuracy(fcast, outsampl)
                    ME      RMSE       MAE        MPE      MAPE      MASE      ACF1 Theil's U
Training set -1.761612 10.319641 7.457651 -2.873590 10.314576 1.5522626 0.4801818        NA
Test set      2.687231  3.550087 3.184992  3.226167  3.819093 0.6629358 0.1979827  4.887312
> |
```

5. Competing forecast

Choosing an ARIMA (1,1,0) (1,1,0) due to the seasonal behavior in the decomposition. This has an RMSE on 24.42.



```
> accuracy(fcast10, outsampl)
                    ME      RMSE       MAE         MPE      MAPE     MASE         ACF1 Theil's U
Training set -0.2356718  6.473941  4.548621 -0.5882369  6.319654 1.137777  0.004880454        NA
Test set     24.1824094 24.419627 24.182409 28.4703191 28.470319 6.048910 -0.063698631  61.83707
```

6. Forecast combination

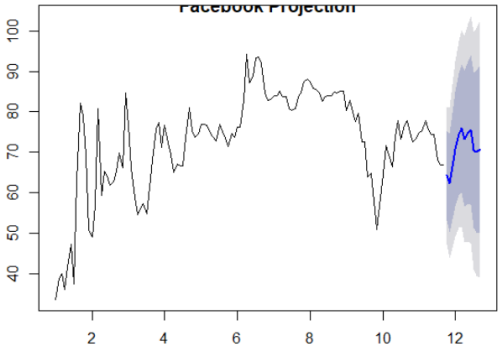
The forecast combination of these two methods shows a higher RMSE than in the moving average method meaning that we have not lowered the RMSE by combining the methods.

```
> accuracy(combfcast, outsampl)
                ME     RMSE      MAE      MPE     MAPE       ACF1 Theil's U
Test set 7.832386 8.490685 7.832386 8.970971 8.970971 -0.3075201   5.63347
```

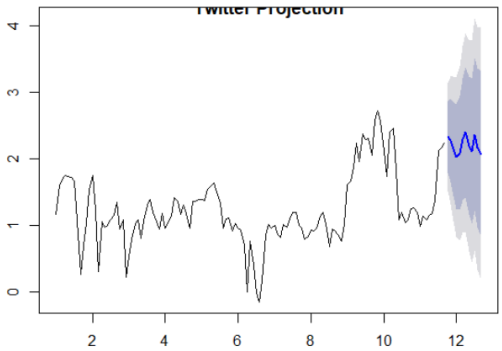
7. 12 month projection for all variables

A Holt-Winthers approach to all variables has been chosen to make a forecast based on the earlier assessment and the lowest RMSE and therefor it is justified.

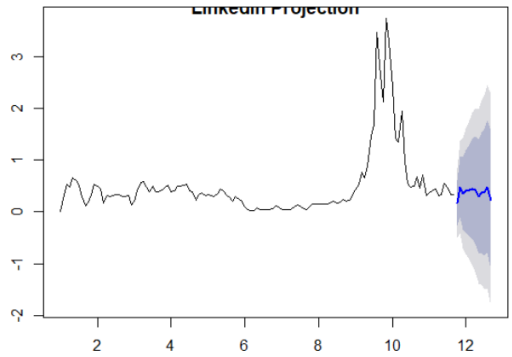
FB projection



Twitter projection



LinkedIn projection



8. Executive Summary

Through the previous assignments report, this executive summary will explain which channels to asses by the Government Bodies to help determine whether regulation is needed on Social Media or not. All the provided recommendations are based on the findings in the conducted technical report.

Facebook

Facebook seems to have some seasonal patterns, showing that a repetitive behavior reoccurs. This could be due to lower usage in the summer than during winter. The Facebook user percentage are currently below the long run trend, indicating that the user numbers are not vastly increasing. This could be due to some of the leakage of personal information Facebook has been accused for.

Facebook and Twitter are projected to increase in monthly user percentage while LinkedIn will decrease

The percentage of monthly users is predicted to increase in the next 12 months for both Facebook and Twitter. Facebook are having a steeper increase in the user percentage which indicates that regulations on Facebook are targeting an increased amount of the consumers in the future.

Forecasting method for projection

After trying different approaches, it is concluded that to achieve the most precise outcome of the forecast, an exponential smoothing approach consisting of a Holt-Winthers analysis produce the best result. To improve the model specifications, it is believed that testing other forecast methods potentially could higher the prediction outcome even further.