Time Series Analysis Projet

VEE Winter 2016

by ZHAO AI SUN

Introduction

The purpose of this project is to illustrate how we can analyze the time series by using Excel and SAS software.

First, I will explain how to use *Descriptive analysis* command in Excel to have a quickly look at the time series concerning Summary Statistic, Normality test and white noise tests, ACF and PACF charts. Secondly, I will look at the trend, seasonality and stationarity of time series by verifying graphics and running the different tests, such as Mann-Kendall trend test, Dickey-Fuller test, Phillips-Perron test, KPSS test. Thirdly, I will demonstrate how to do time series transformation. Then, the more more difficult part of this project is to build some ARIMA models for time series and choose a model for forcasting. Finally, I will forecast the values of time series at future times by SAS.

Contents

Part 1 A quick look at time series GDP

Opening *XLSTAT / XLSTAT-Time / descriptive analysis* command in Excel, I have a quick look at the given time series: Quarterly GDP from year 1947 to 2007.

The first table displays the summary statistics. Then the Normality test and white noise tests table is displayed. At the end, I see Descriptive analysis (GDP).

Summary Statistic

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
GDP	244	0	244	1567.966	11675.714	5344.860	2916.979

Normality test and white noise tests

Results of the analysis of the GDP series:

Statistic	DF	Value	p-value
Jarque-Bera	2	19.855	< 0.0001
Box-Pierce	6	1330.151	< 0.0001
Ljung-Box	6	1360.173	< 0.0001
McLeod-Li	6	1324.125	< 0.0001
Box-Pierce	12	2443.713	< 0.0001
Ljung-Box	12	2527.957	< 0.0001
McLeod-Li	12	2406.583	< 0.0001

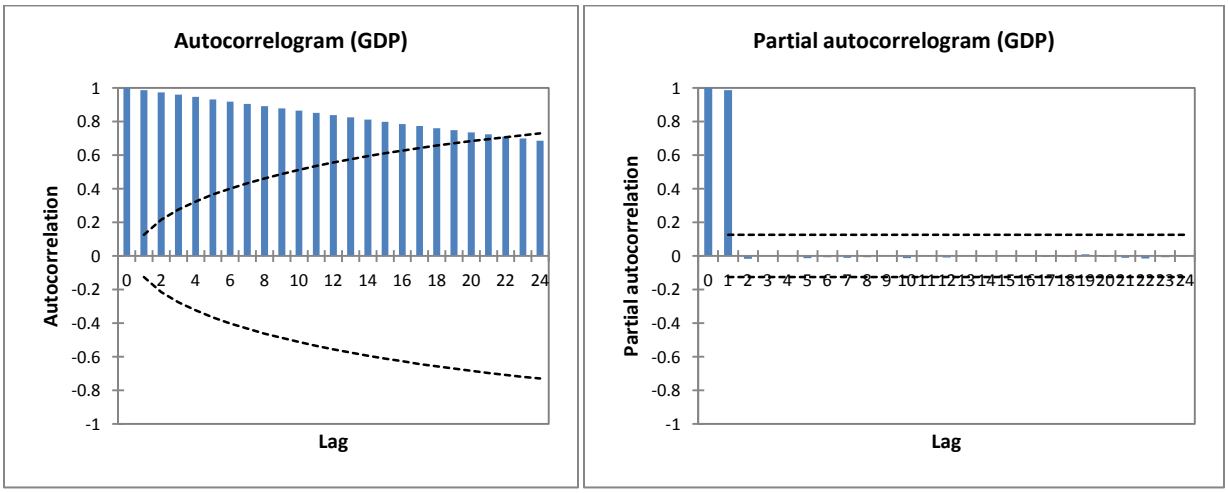
- The Jarque-Bera test is a normality test, based on the skewness and kurtosis coefficients. The higher the value of the Chi-square statistic, the more unlikely the null hypothesis that the data are normally distributed. Here the p-value, which corresponds to the probability of being wrong when rejecting the null hypothesis, is less than 0.0001. With an alpha=0.05 significance level, we should reject the null hypothesis.

- The three other three tests (Box-Pierce, Ljung-Box, McLeod-Li) are computed at different time lags. They allow to test if the data could be assumed to be a white noise or not. These tests are also based on the Chi-square distribution. They all agree that the data cannot be assumed to be generated by a white noise process. While the sorting of the data has no influence on the Jarque-Bera test, it does have an influence on the three other tests which are particularly suited for time series analysis.

Descriptive analysis (GDP):

Lag	Autocorrelation	Standard error	Lower bound (95%)	Upper bound (95%)	Partial autocorrelation	Standard error
0	1.000	0.000			1.000	0.000
1	0.987	0.064	-0.125	0.125	0.987	0.064
2	0.973	0.110	-0.215	0.215	-0.016	0.064
3	0.960	0.141	-0.276	0.276	-0.002	0.064
4	0.946	0.165	-0.324	0.324	-0.002	0.064
5	0.933	0.186	-0.365	0.365	-0.012	0.064
6	0.919	0.205	-0.401	0.401	-0.007	0.064
7	0.906	0.221	-0.433	0.433	-0.012	0.064
8	0.892	0.236	-0.462	0.462	-0.007	0.064
9	0.879	0.249	-0.488	0.488	-0.002	0.064
10	0.865	0.261	-0.512	0.512	-0.013	0.064
11	0.852	0.273	-0.535	0.535	-0.001	0.064
12	0.838	0.284	-0.556	0.556	-0.008	0.064
13	0.825	0.294	-0.575	0.575	0.000	0.064
14	0.812	0.303	-0.594	0.594	-0.004	0.064
15	0.799	0.312	-0.611	0.611	0.001	0.064
16	0.786	0.320	-0.627	0.627	-0.002	0.064
17	0.773	0.328	-0.643	0.643	-0.005	0.064
18	0.761	0.335	-0.657	0.657	-0.005	0.064
19	0.748	0.342	-0.671	0.671	0.009	0.064
20	0.736	0.349	-0.684	0.684	-0.005	0.064
21	0.724	0.355	-0.696	0.696	-0.010	0.064
22	0.711	0.361	-0.708	0.708	-0.014	0.064
23	0.699	0.367	-0.719	0.719	-0.008	0.064
24	0.686	0.372	-0.730	0.730	-0.004	0.064

ACF and PACF of original variable GDP

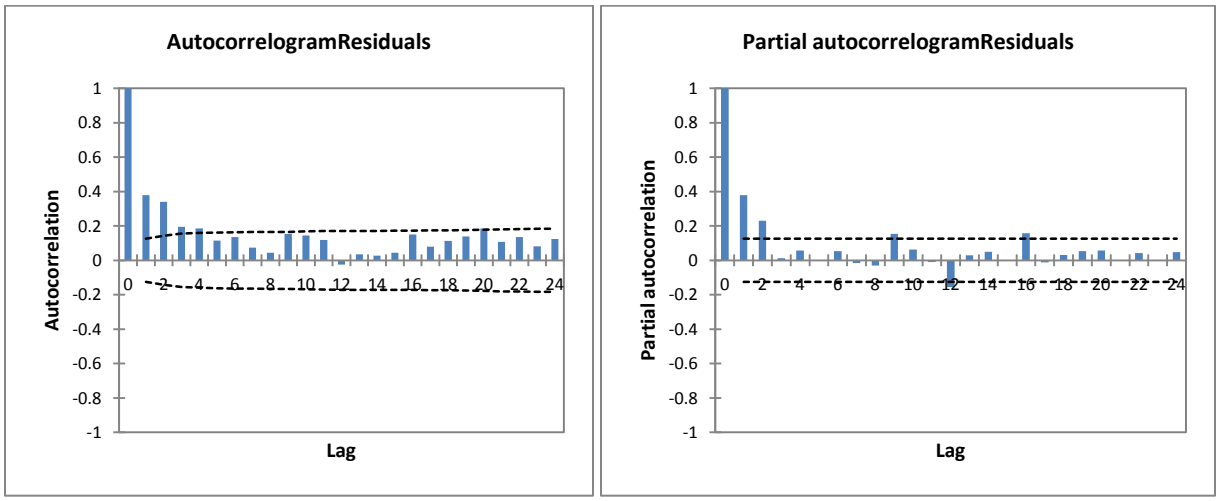


- These two bar charts display the evolution of the autocorrelation function (ACF) and of the partial autocorrelation function (PACF). The 95% confidence intervals are also displayed.
- How to identify non-stationary series by ACF?

The ACF of stationary data drops to zero relatively quickly, while the ACF of non-stationary data decreases slowly. For non-stationary data, the value of *r*1 is often large and positive. In this case, for variable GDP, ACF declines very slowly, which indicates non-stationarity.

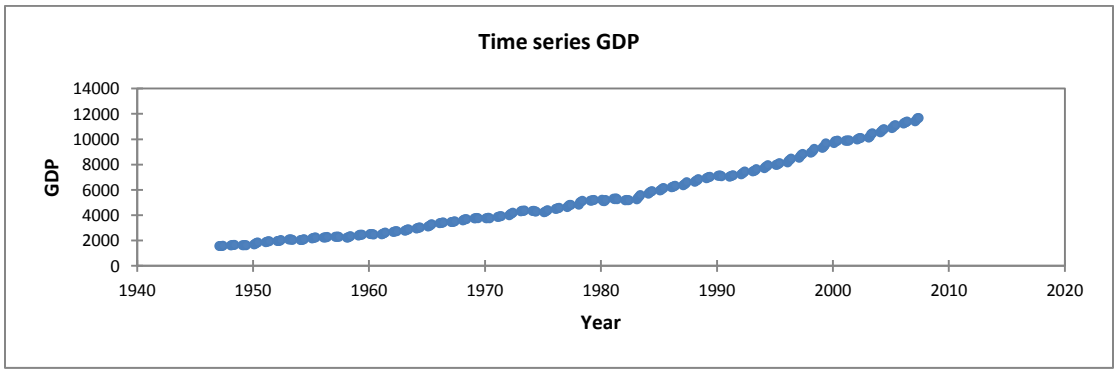
- ACF up to about 22 lags are individually statistically significant from zero, because they are all outside the 95 percent confidence bounds. PACF drops dramatically, it cuts off after 2nd lag.
- According to the patterns displaying in the chart of PACF, it looks like an AR(2) model.
- I can identify a clear lag 1 autocorrelation, but I can't identify a seasonality because it does not have seasonal behavior.

ACF and PACF of Residuals



For the residuals, I can identify a clear lag 1 autocorrelation. ACF tails off and it cuts off after first 5 lags. PACF drops dramatically and cuts off after 3rd lag.

Part 2 Trend



From graphic, we can see variable GDP is increasing with the time, so time series GDP displays a positive/upward secular trend. I run Mann-Kendall trend test in EXCEL to double check it.

Mann-Kendall trend test / Two-tailed test (GDP):

Kendall's tau	0.987
S	29274.000
Var(S)	1623942.000
p-value (Two-tailed)	< 0.0001
alpha	0.05

The exact p-value could not be computed. An approximation has been used to compute the p-value.

Test interpretation:
H0: There is no trend in the series
Ha: There is a trend in the series
As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis Ha.
The risk to reject the null hypothesis H0 while it is true is lower than 0.01%.

Part 3 Seasonality

Definition: Time series displays very regular patterns: seasonal peaks and trends, which are called seasonality.
Time series GDP doesn't show seasonal peaks and it only displays an increasing trend.
Therefore, time series GDP does not have a seasonality.

Part 4 Stationality

A stationary series is roughly horizontal with constant variance. It has no patterns predictable in the long-term. That means a stationary process has the property that the mean, variance and autocorrelation structure do not change over time.
Any time series without a constant mean over time is non stationary. A constant value makes a stationary time series.
Time series GDP has an increasing trend, so it's not stationary.

After opening XLSTAT in EXCEL, I select the *XLSTAT / XLSTAT-Time / Unit root and stationarity tests* command. I run the following three tests in EXCEL to check whether time series GDP is stationary or not. All the results show that it is not stationary.

1)Dickey-Fuller test (ADF(stationary) / k: 6 / GDP):

Tau (Observed value)	-0.130
Tau (Critical value)	-0.855
p-value (one-tailed)	0.991
alpha	0.05

Test interpretation:
H0: There is a unit root for the series.
Ha: There is no unit root for the series. The series is stationary.
As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.
The risk to reject the null hypothesis H0 while it is true is 99.13%.

2)Phillips-Perron test (PP(no intercept) / Lag: Short / GDP):

Tau (Observed value)	12.804
Tau (Critical value)	-1.942
p-value (one-tailed)	1.000
alpha	0.05

Test interpretation:
H0: There is a unit root for the series.
Ha: There is no unit root for the series. The series is stationary.
As the computed p-value is greater than the significance level alpha=0.05, one cannot reject the null hypothesis H0.
The risk to reject the null hypothesis H0 while it is true is 100.00%.

3)KPSS test (Level / Lag Short / GDP):

Eta (Observed value)	5.959
Eta (Critical value)	0.453
p-value (one-tailed)	< 0.0001
alpha	0.05

Test interpretation:
H0: The series is stationary.
Ha: The series is not stationary.
As the computed p-value is lower than the significance level alpha=0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis Ha.
The risk to reject the null hypothesis H0 while it is true is lower than 0.01%.

Part 5 Time Series Transformation

In order to improve the normality of the data, we want to perform two transformations:

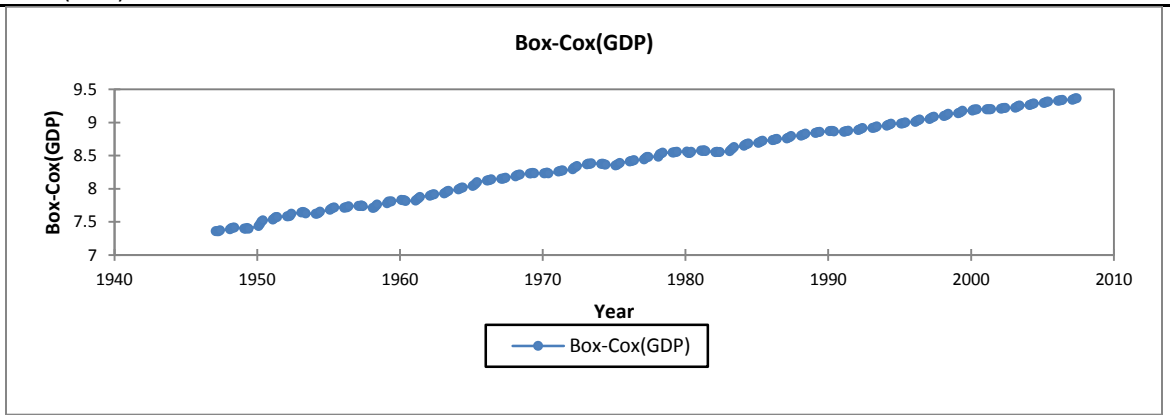
First, stabilize the increasing variability of the time series

We use the **Box-Cox transformation (log transformation)** in order to remove the increasing variability of the original data of GDP.

This can be done using the *Time series transformation* tool in Excel.

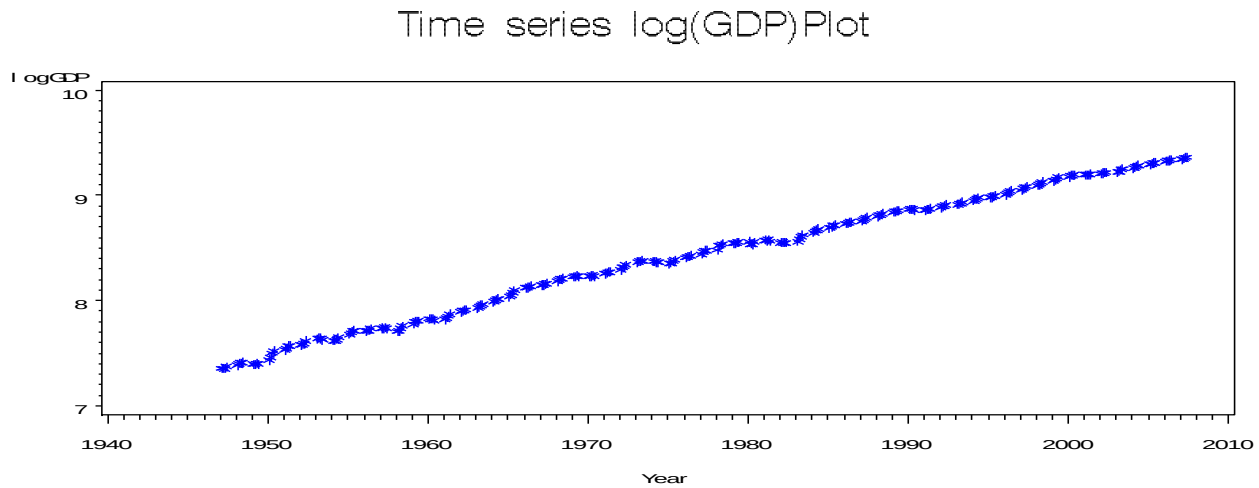
Summary statistics:

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
Box-Cox(GDP)	244	0	244	7.358	9.365	8.424	0.584



This can also be done using SAS.

```
  proc gplot data=data;
plot logGDP*Year;
symbol1 v=star c=blue;
title"Time series log(GDP)Plot";
run;
```



The transformed variable Log(GDP) displays an increasing trend, so it's not stationary.

Second, remove the autocorrelations by differencing the time series

In order to remove the trend and the seasonal component, we use the differencing method. Differencing helps to stabilize the mean. There are two types of differencing involved:

- **Ordinary differencing**

The differenced series is the *change* between each observation in the original series: $\nabla Y_t = Y_t - Y_{t-1}$. The differenced series will have only T-1 values since it is not possible to calculate a difference ∇Y_1 for the first observation.

The first differences are the change between **one observation and the next**.

- **Seasonal differencing**

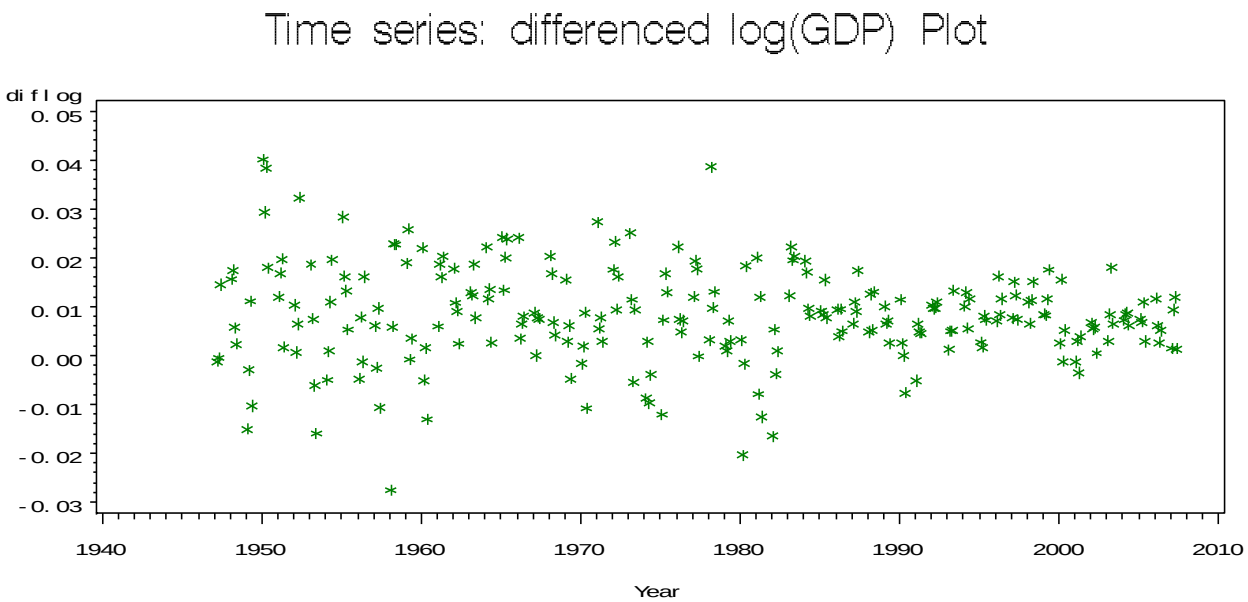
A seasonal difference is the difference between an observation and the corresponding observation from the previous year. $\nabla Y_t = Y_t - Y_{t-m}$ where m = number of seasons. For example: for monthly data m = 12; for quarterly data m = 4.

Seasonal differences are the change between **one year to the next**.

Application:

For the transformed time series GDP (log GDP), it does not have a seasonality, therefore I use ordinary differencing method. I select the Box-Cox transformed series, and then apply the differencing method.

```
* Plotting the data of differenced logGDP (diflog);
proc gplot data=data;
plot diflog*Year;
symbol1 v=star c=green;
title"Time series: differenced log(GDP) Plot";
run;
```



The differenced variable of Log(GDP) is stationary, although the variance decreases.

The resulting graphic shows that the differencing transformation effectively removed the trend.

Summary of two Variables: Log(GDP) and first difference of Log(GDP):

```
proc means data=data;
var logGDP diflog;
run;
```

```
                                The SAS System          21:14 Friday, March 26, 2016
                              The MEANS Procedure

Variable        N            Mean         Std Dev         Minimum         Maximum
logGDP        244       8.4235407       0.5841800       7.3575345       9.3652662
diflog        243       0.0082556       0.0097804      -0.0275253       0.0401981
```

Part 6 ARIMA Model and Forcasting

Autoregressive integrated moving average ARIMA (*p*,*d*,*q*) model, where

p= the number of autoregressive terms

d = the number of time series has to be differenced before it becomes stationary

q = the number of moving average terms

The parameters p and q are called the autoregressive and the moving average orders, respectively

A stationary ARMA (*p*,*q*) Model can be defined by the equition:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \text{---} + \phi_p Y_{t-p} - \theta_1 e_{t-1} - \theta_2 e_{t-2} + \text{---} \theta_q Y_{t-q} + e_t$$

where $e_1, e_2, \text{----} e_t$ are iid (independent identically distributed) and $e_t \sim N(0, \sigma_e^2)$ for t=1,2,---,n

- Autoregressive Process AR(p) satisfies the equation:

  $$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \text{---} + \phi_p Y_{t-p} + e_t$$

- Moving average MA (*q*) process satisfies the equation:

  $$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} + \text{---} \theta_q Y_{t-q}$$

6 A) Building ARIMA model

- **Idea: Box and Jenkins methodology**

Important Notes: Before we apply Box and Jenkins methodology, we have to make the time series stationary.

Most time series are nonstationary and must be transformed to a stationary series before the ARIMA modeling process can proceed. If the series has a nonstationary variance, taking the log of the series can help. We can compute the log values in a DATA step and then analyze the log values with PROC ARIMA.

In Part 5: I transform time series GDP to Log(GDP) first. When I plot Log(GDP), it shows non-stationary. Then I plot the first difference of logGDP (I named it diflog in SAS code), I don't observe any trend, which suggest the time series of diflog is stationary. Dickey-Fuller root test shows it is stationary, indeed.

➢ A quick look at the first difference of log(GDP) in EXCEL

Summary statistics:

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
dif(logGDP)	243	0	243	-0.028	0.040	0.008	0.010

Normality test and white noise tests (dif(logGDP)):

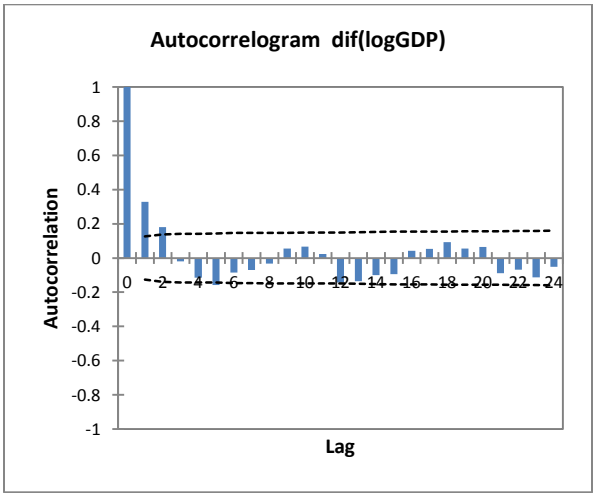
Statistic	DF	Value	p-value
Jarque-Bera	2	18.851	< 0.0001
Box-Pierce	6	45.264	< 0.0001
Ljung-Box	6	46.040	< 0.0001
McLeod-Li	6	25.707	0.000
Box-Pierce	12	53.645	< 0.0001
Ljung-Box	12	54.879	< 0.0001
McLeod-Li	12	42.394	< 0.0001

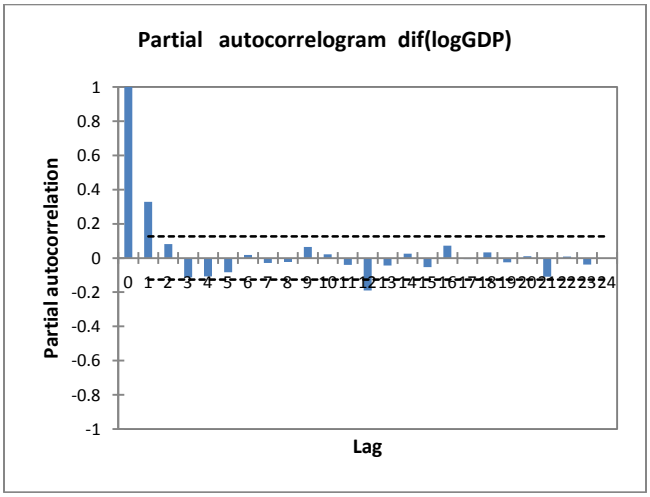
Interpretation: Time series for the first difference of log(GDP) is not white noise.

How to read time series ACF and PACF graphics ?

Model	ACF	PACF
White Noise	All zeros	All zeros
AR(p)	Exponential Decay	P significant lags before dropping to zero
MA(q)	q significant lags before dropping to zero	Exponential Decay
ARMA(p,q)	Decay after qth lag	Decay after pth lag

ACFs and PACFs for first difference of log(GDP)

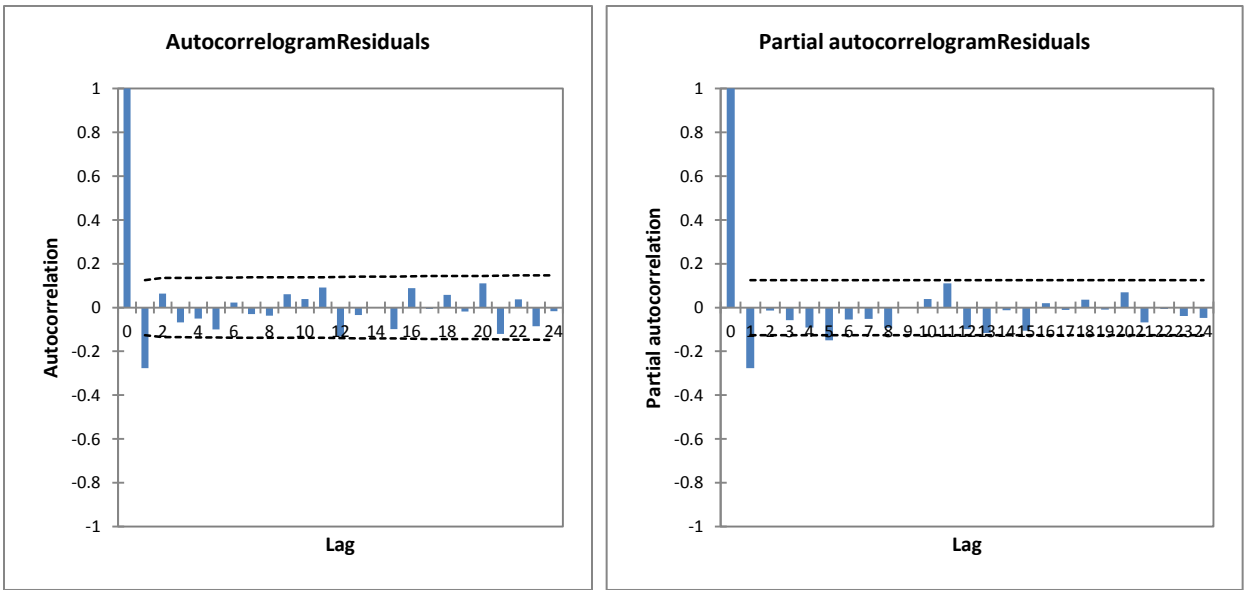




- A visual inspection of the autocorrelation function plot indicates that the *dif(logGDP)* series is stationary, since the ACF decays very fast.

- PACF drops dramatically and it cuts off. Notice that (a) the correlation at lag 1 is significant and positive, and (b) the PACF shows a sharper "cut-off" than the ACF. In particular, the PACF has only two significant spikes, while the ACF has three. Thus, the differenced series displays an AR(2) signature. If we therefore set the order of the AR term to 2--i.e., it fits an ARIMA (2,1,0) model.

ACFs and PACFs of Residuals



Both the ACF and the PACF shows sharp "cutoff". They both have two significant spikes.

6 B) ARIMA Modeling in SAS

The analysis performed by **PROC ARIMA** is divided into different stages, corresponding to the stages described by Box and Jenkins.

- Step 1 Identification.
  We use the **IDENTIFY statement** to specify the response series and identify candidate ARIMA models for it. The goal is to choose tentative p,d,q.
  The analysis of the IDENTIFY statement output usually suggests one or more ARIMA models that could be fit. Options enable you to test for stationarity and tentative ARMA order identification.

- Step 2 Estimation and diagnostic checking
  - We use the **ESTIMATE statement** to estimate the parameters of AR and MA terms included in the model. Goodness-of-fit statistics aid in comparing this model to others.

- Diagnostic checking is to see whether the chosen model fits the data reasonably. Are estimated residuals white noise? The outpour of graphical analysis also reveals the inadequacy of the ARIMA model. If the diagnostic tests indicate problems with the model, we try another model and then repeat the estimation and diagnostic checking stage.

➢ Step 3 Forecasting

We use the **FORECAST statement** to forecast future values of the time series and to generate confidence intervals for these forecasts from the ARIMA model produced by the preceding ESTIMATE statement.

One of the primary objectives of building a model for a time series is to be able to forecast the values for that series at future times. We can only predict by what happened in the past.
In this case, the forecast statement (forecast lead=12;) is quarterly forecasting for the first difference of Log(GDP) in the following 3 years.

6 C) Interpretation of SAS output:

```
proc arima data=data ;
identify var=diflog;
estimate p=1 ;
forecast lead=12;
run;
```

```
                          The SAS System        16:50 Saturday, March 27, 2016

 1                                The ARIMA Procedure


                        Name of Variable = diflog

                    Mean of Working Series    0.008256
                    Standard Deviation         0.00976
                    Number of Observations         243


                             Autocorrelations

  Lag    Covariance    Correlation    -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      Std Error

    0    0.00009526        1.00000    |                    |********************|            0
    1    0.00003127        0.32826    |                  .  |*******             |      0.064150
    2    0.00001724        0.18092    |                  .  |****                |      0.070726
    3    -1.7742E-6        -.01862    |                  .  |  .                 |      0.072605
    4    -0.0000110        -.11523    |                 .**|  .                 |      0.072625
    5    -0.0000151        -.15809    |                 ***|  .                 |      0.073373
    6    -8.0626E-6        -.08464    |                 .**|  .                 |      0.074762
    7    -6.6226E-6        -.06952    |                 . *|  .                 |      0.075155
    8    -3.0247E-6        -.03175    |                 . *|  .                 |      0.075419
    9    5.26559E-6        0.05527    |                  .  |* .                |      0.075474
```

10	6.3328E-6	0.06648	. |* .	0.075641
11	2.22373E-6	0.02334	. | .	0.075881
12	-0.0000137	-.14363	***| .	0.075910
13	-0.0000129	-.13566	***| .	0.077021
14	-9.5468E-6	-.10022	.**| .	0.077998
15	-9.0187E-6	-.09467	.**| .	0.078526
16	4.07836E-6	0.04281	. |* .	0.078994
17	5.03942E-6	0.05290	. |* .	0.079090
18	8.79402E-6	0.09231	. |**.	0.079235
19	5.34892E-6	0.05615	. |* .	0.079677
20	6.18545E-6	0.06493	. |* .	0.079839
21	-8.3818E-6	-.08799	.**| .	0.080056
22	-6.442E-6	-.06762	. *| .	0.080453
23	-0.0000108	-.11349	.**| .	0.080687
24	-4.9738E-6	-.05221	. *| .	0.081341

"." marks two standard errors

Interpretation: **ACF**

The autocorrelations decrease rapidly, indicating that the change in *logGDP* is a stationary time series.

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.22081	****| .
2	-0.12607	***| .
3	0.01940	. | .
4	0.06967	. |* .
5	0.06611	. |* .
6	-0.05412	. *| .
7	0.03392	. |* .

The ARIMA Procedure

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
8	0.08740	. |**.
9	0.01036	. | .
10	-0.05592	. *| .
11	-0.06434	. *| .
12	0.13876	. |***
13	0.03247	. |* .
14	-0.03046	. *| .
15	0.06602	. |* .
16	-0.04229	. *| .
17	0.02523	. |* .
18	-0.01907	. | .
19	0.01478	. | .
20	-0.05429	. *| .
21	0.09157	. |**.
22	-0.02560	. *| .
23	0.03081	. |* .
24	0.00296	. | .

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.32826	|. |******* |
2	0.08200	|. |**. |
3	-0.11291	|. **| . |
4	-0.10672	|. **| . |
5	-0.08305	|. **| . |
6	0.01862	|. | . |
7	-0.02896	|. * | . |
8	-0.02244	|. | . |
9	0.06406	|. |* . |
10	0.02236	|. | . |
11	-0.03957	|. * | . |
12	-0.19048	|**** | . |
13	-0.04393	|. * | . |
14	0.02610	|. |* . |
15	-0.05400	|. * | . |
16	0.07162	|. |* . |
17	-0.00656	|. | . |
18	0.03294	|. |* . |
19	-0.02516	|. * | . |
20	0.01067	|. | . |
21	-0.10955	|. **| . |
22	0.00770	|. | . |
23	-0.03846	|. * | . |
24	-0.00351	|. | . |

Interpretation: **PACF**

The PACF "cuts off" at lag *k*--then this suggests that we should try fitting an autoregressive model of order *k.* In this case, it shows that the process is order 1 autoregressive.

Note that the PACF plot has a significant spike only at lag 1, meaning that all the higher-order autocorrelations are effectively explained by the lag-1 autocorrelation.

The ARIMA Procedure

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	46.04	6	**<.0001**	0.328	0.181	-0.019	-0.115	-0.158	-0.085
12	54.88	12	**<.0001**	-0.070	-0.032	0.055	0.066	0.023	-0.144
18	68.07	18	**<.0001**	-0.136	-0.100	-0.095	0.043	0.053	0.092
24	77.56	24	**<.0001**	0.056	0.065	-0.088	-0.068	-0.113	-0.052

Interpretation: **White Noise Test**

This is an approximate statistical test of the hypothesis that none of the autocorrelations of the series up to a given lag are significantly different from 0. If this is true for all lags, then there is no information in the series to model, and no ARIMA model is needed for the series.

While testing the null hypothesis that the time series observations are uncorrelated (white noise), probability values are all highly significant; therefore the null hypothesis has been rejected. It can be concluded that the given time series is not white noise.

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0082136	0.0008827	9.31	<.0001	0
AR1,1	0.32896	0.06090	5.40	<.0001	1

Interpretation: **Parameter Estimates**

The table of parameter estimates lists the parameters in the model; for each parameter, the table shows the estimated value and the standard error and *t* value for the estimate. The table also indicates the lag at which the parameter appears in the model.

The fitted model is $Y_t = \mu + \Phi Y_{t-1} = 0.0082136 + 0.32896 Y_{t-1}$

The *t* values provide significance tests for the parameter estimates and indicate whether some terms in the model might be unnecessary.

In this case, the *t* value for MU is 9.31, for the autoregressive parameter is 5.40, both of these terms are highly significant.

Constant Estimate	0.005512
Variance Estimate	0.000086
Std Error Estimate	0.009256
AIC	**-1584.07**
SBC	**-1577.09**
Number of Residuals	243

* AIC and SBC do not include log determinant.

Interpretation: **Goodness-of-Fit Statistics**

The 'Constant Estimate' is a function of the mean term MU and the autoregressive parameters. The 'Variance Estimate' is the variance of the residual series, which estimates the innovation variance. The 'Std Error Estimate' is the square root of the 'Variance Estimate'.

When we are comparing candidate models, **smaller** AIC and BIC indicate the **better** fitting model.

Correlations of Parameter Estimates

Parameter	MU	AR1,1
MU	1.000	-0.007
AR1,1	-0.007	1.000

Interpretation: **Correlations of the Estimates**

This table can help to assess the extent to which collinearity might have influenced the results. If two parameter estimates are very highly correlated, we might consider dropping one of them from the model. It is not the case for AR(1) model.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	9.10	5	0.1052	-0.027	0.110	-0.048	-0.077	-0.123	-0.021
12	17.10	11	0.1048	-0.043	-0.035	0.057	0.053	0.058	-0.137
18	24.26	17	0.1124	-0.078	-0.040	-0.097	0.069	0.016	0.073
24	32.49	23	0.0904	0.012	0.093	-0.108	-0.010	-0.097	-0.030
30	43.61	29	**0.0400**	0.040	-0.021	0.055	0.029	0.071	-0.170
36	47.07	35	0.0836	-0.028	-0.074	0.012	0.061	0.027	0.038
42	60.52	41	**0.0252**	0.006	-0.073	-0.060	-0.021	-0.094	0.165
48	66.96	47	**0.0294**	-0.011	0.029	0.101	-0.049	0.088	-0.010

Interpretation: **Check for White Noise of Residuals**

The χ^2 test statistics for the residuals series indicate whether the residuals are uncorrelated (white noise) or contain additional information that might be used by a more complex model. The null hypothesis is "The residuals of the fitted model are uncorrelated (white noise)." If it's been accepted (probability values not significant), it can be concluded that the fitted time series model is a good fit.

In this case, some of the test statistics show that reject the no-autocorrelation hypothesis at a level of significance ((p= 0.04;p=0.0252;p=0.0294 for the last 6 lags),most of statistics show that accept the null hypothesis because the probability values not significant.

This means that the residuals are not really white noise, and so the AR(1) model is not a fully adequate model for this series.

6 D) Comparing the SAS output and choose ARIMA models

When d=1, we apply differenced variable of GDP in the following models and use ARMA(p,q) to write SAS code.

- ARIMA(1,1,1) and ARIMA(1,1,2) don't work well in SAS. We get *'Warning message: The estimation algorithm did not converge'* in *ARIMA Estimation Optimization Summary*.
- ARIMA(2,1,1) doesn't work well in SAS, either. There are inverse autocorrelations calculation results. We get all the results as we expected, but we still get *'Warning message: Estimates may not have converged'* in *ARIMA Estimation Optimization Summary*.

Comparing the SAS OUTPUT:

- Possibly choose the model with the fewest parameters.
- Examine standard errors of forecast values. Pick the model with the generally lowest standard errors for predictions of the future.

- Compare models with regard to statistics such as the MSE (the estimate of the variance of the w_t), AIC and SBC. Lower values of AIC and SBC of these statistics are desirable.
- Compare the parameter estimates: The *t* values provide significance tests
- Check for White Noise of Residuals

6D-1) Comparing SAS OUTPUT: parameter estimates

	ARIMA(1,1,0)	ARIMA(0,1,1)	ARIMA(2,1,0)	ARIMA(0,1,2)	ARIMA(2,1,1)	ARIMA(2,1,2)	ARIMA(3,1,2)	ARIMA(3,1,0)	ARIMA(0,1,3)
constant	0.0082136	0.0082377	0.0081962	0.0082057	0.0083308	0.0082111	0.0082715	0.0082062	0.0081891
L1,AR	0.32896		0.302		1.31176	1.32688	0.99895	0.31126	
L2,AR			0.08173		-0.31764	-0.72867	0.08560	0.11614	
L3,AR							-0.21391	-0.11355	
L1,MA		0.25853		0.29090	1	1.05445	0.71232		0.30695
L2,MA				0.19980		-0.55545	0.18731		0.23689
L3,MA									0.07418
AIC	-1584.07	-1577.78	-1583.69	-1585.43	-1581.67	-1587.36	-1584.61	-1584083	-1584.6
SBC	-1577.09	-1570.8	-1573.21	-1574.95	-1567.69	-1569.89	-1563.65	-1570.86	-1570.62

Notes for the above table: The *t* values provide significance tests for the parameter estimates and indicate whether some terms in the model might be unnecessary. All the parameters colored in yellow show these terms are not significant.

By Comparing SAS OUTPUT of parameter estimates, I won't choose the following models: ARIMA(2,1,0), ARIMA(3,1,2), ARIMA(3,1,0), ARIMA(0,1,3), because those terms corresponding to the parameters colored in yellow are not necessary.

6D-2) SAS output of *Autocorrelation Check of Residuals*

- ARIMA(1,1,0) model:

The output shows that the residuals are not really white noise, so this model is not a fully adequate model for this series.

- ARIMA(0,1,1) model:

All of the test statistics show that reject the no-autocorrelation hypothesis at a level of significance. The residuals are not white noise, so this model is not a good fit.

Conclusion: These two models are not our choice.

6D-3) Both ARIMA(3,1,1) and ARIMA(1,1,3) models:

- SAS OUTPUT of *Autocorrelation Check of Residuals*

  All of the test statistics show that reject the no-autocorrelation hypothesis at a level of significance. The residuals are not white noise, so these models are not good fits.

- SAS OUTPUT of parameter estimates: some terms corresponding to the parameters are not significant, which means they are not necessary.

Conclusion: These two models are not our choice.

6D-4) ARIMA(0,1,2) and ARIMA(2,1,2)

At the end, we have these two options left; I check every output from SAS:

- *Parameter estimates* show they are significant.
- *Autocorrelation Check of Residuals* show the residuals are white noise, so these models are good fits.
- *Autocorrelation Check for White Noise* show the given time series are not white noise.
- *Goodness-of-Fit Statistics* show very similar results for AIC and SBC values.
- *Correlations of Parameter Estimates*

Correlations of Parameter Estimates

Parameter	MU	MA1,1	MA1,2	AR1,1	AR1,2
MU	1.000	0.010	-0.001	0.009	-0.007
MA1,1	0.010	1.000	-0.446	0.929	-0.782
MA1,2	-0.001	-0.446	1.000	-0.230	0.763
AR1,1	0.009	**0.929**	-0.230	1.000	-0.730
AR1,2	-0.007	-0.782	0.763	-0.730	1.000

Interpretation: Correlations of the Parameter Estimates for ARIMA(2,1,2) Model

This table can help to assess the extent to which collinearity might have influenced the results. If two parameter estimates are very highly correlated, we might consider dropping one of them from the model.

A correlation of 0.8 or 0.9 is regarded as a high correlation, i.e., there is a very close relationship between the two variables.

AR1,1 and MA1,1 are highly correlated, we might consider dropping one of them from the model. They have positive correlation. AR1,2 and MA1,2 have relatively strong correlation, too.

AR1,2 and MA1,1 also have pretty strong correlation, we might consider dropping one of them from the model. They have negative correlation. AR1,1 and AR1,2 have very similar situation.

Conclusion: We will not consider ARIMA(2,1,2) Model.

Correlations of Parameter Estimates

Parameter	MU	MA1,1	MA1,2
MU	1.000	0.006	0.005
MA1,1	0.006	1.000	0.240

MA1,2	0.005	0.240	1.000

Interpretation: Correlations of the Estimates for ARIMA(0,1,2) Model

The result looks fine.

6D-5) Final choice for forecasting: ARIMA(0,1,2) Model

```
proc arima data=data ;
identify var=diflog;
estimate q=2 ;
forecast lead=12;
run;
```

Forecasts for variable diflog

Obs	Forecast	Std Error	95% Confidence Limits	
245	0.0065	0.0092	-0.0115	0.0246
246	0.0065	0.0096	-0.0123	0.0253
247	0.0082	0.0098	-0.0109	0.0274
248	0.0082	0.0098	-0.0109	0.0274
249	0.0082	0.0098	-0.0109	0.0274
250	0.0082	0.0098	-0.0109	0.0274
251	0.0082	0.0098	-0.0109	0.0274
252	0.0082	0.0098	-0.0109	0.0274
253	0.0082	0.0098	-0.0109	0.0274
254	0.0082	0.0098	-0.0109	0.0274
255	0.0082	0.0098	-0.0109	0.0274
256	0.0082	0.0098	-0.0109	0.0274

Interpretation: **The forecast table**

The forecast table shows for each forecast period the observation number, forecast value, standard error estimate for the forecast value, and lower and upper limits for a 95% confidence interval for the forecast.

Forecasts for ARIMA(0,1,2) Model

In this case, the forecast statement (forecast lead=12;) is for quarterly forecasting in the following 3 years, concerning the time series Log(GDP).

Conclusion

When building every part of this project, I applied the knowledge we learned from the course *Time Series Analysis.* I checked the trend, seasonality and stationarity of the time series. I apply time series transformation. I build some ARIMA models, compare every model and make a choice. Finally I use SAS to forecast the chosen time series for the following three years. I have illustrated the time series using a number of realistic results from Excel and SAS output.

Reference:

Textbook: Time Series Analysis With Applications in R *by Jonathan D. Cryer • Kung-Sik Chan*

Identifying the numbers of AR or MA terms in an ARIMA model
http://people.duke.edu/~rnau/411arim3.htm