Homework assignment #3
Time Series Data: Forecasting

MPP-C6: Statistics 2

Dr. Nicolas Koch
koch@mcc-berlin.net
http://moodle.hertie-school.org/course/view.php?id=1192

24 September 2015

Disclaimer: This document sketches some brief responses to the questions of the homework assignment. It serves to provide students with some guidance. The document may include some flaws - those should be reported to me as students go through the responses.

Project Description

You are working as consultant for the IMF Research Department. The department prepares the World Economic Outlook (WEO) that portrays the world economy in the near and medium context, with projections for key macroeconomic indicators, such as GDP and inflation, for up to four years into the future. The department would like to review their forecast methodologies. You are tasked to evaluate basic time series models, which are used as benchmark against which more complicated forecasting models are assessed. Your focus is on quarterly data (from 1947:Q1 to 2009:Q4) on two main macroeconomic series for the United States:

- GDP: Quarterly values of real GDP for the United States in billions of chained (2005) Dollars seasonally adjusted.
- $TBill$: Quarterly values of the rate on 3-month Treasury Bills.

Use the file UsMacro.dta for this exercise.

Questions

1. **Compute $Y_t = In(GDP_t)$, the logarithm of real GDP and ΔY_t the quarterly growth rate of GDP.**

Interactive Stata example

```
. gen lgdp = log(gdp)
. gen dgdp = D.lgdp
(1 missing value generated)
```

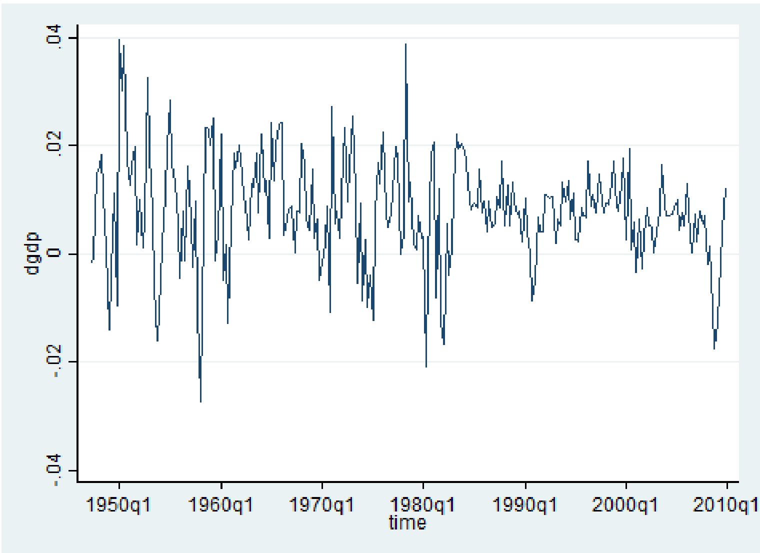
Note: Stata has several time series operators, described at *help tsvarlist*. The prefixe *S.* is particularly useful for quarterly data. For instance *S4.dgdp* will refer to quarter-over-quarter GDP growth rate, comparing the observation to that of the same quarter in the previous year.

2. **Look at the plot of the quarterly growth rate of GDP. Is this time series a good candidate for time series forecasts? Explain.**

Interactive Stata example

```
. twoway line dgdp time
```

Figure 1: Quarterly growth rate of GDP



In general, the plot is dominated by rather short-lived swings in the quarterly growth rate of GDP. However, there is some serial correlation in growth rate, indicating that the last quarter's growth rate contains information about this quarter's growth. If there is serial correlation, then you can predict the GDP growth using past quartely growth rates. In sum, the variable is a reasonable candidate for time series forecasts.

Note however: Simple AR forecast will probably perform bad in those periods that are characterized by strong surprise movements, e.g. early 2000.

Excursus Sesonality: *Many time series exhibit seasonality. A seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week). Series that do display seasonal patterns are often seasonally adjusted before they are reported for public use. Our GDP data is indeed reported in seasonally adjusted form (at annual rate).*

However, in some cases our variables of interest may only be available in their raw form. In cases where we do not have seasonally adjusted data, we may either de-seasonalize each series or we may include a set of seasonal dummies in a regression of y on x and test for joint significance.

How can you deseasonalize data?

a) Regress the series on seasonal dummies

b) Save the residuals

c) Add the mean of the original series

Interactive Stata example

```
// for loop to create dummies
. forvalues i=1/3 {
generate qseas‘i’ = (quarter(dofq(time)) == ‘i’)
}

. reg dgdp qseas*

Source   |     SS        df       MS              Number of obs =     251
-------------+------------------------            F(  3,    247) =    0.50
Model    |.000150157      3  .000050052           Prob > F      =  0.6831
Residual|.024760942    247  .000100247            R-squared     =  0.0060
-------------+------------------------            Adj R-squared = -0.0060
Total    |.024911099    250  .000099644           Root MSE      =  .01001

------------------------------------------------------------------------------
dgdp     | Coef.      Std. Err.       t     P>|t|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
qseas1  |.0018989    .0017911     1.06    0.290    -.0016289    .0054267
qseas2  |.0018877    .0017839     1.06    0.291    -.0016259    .0054014
qseas3  |.0012717    .0017839     0.71    0.477     -.002242    .0047853
_cons   |.0066829    .0012614     5.30    0.000     .0041984    .0091675
-------------------------------------------------------------------------------


. predict double dgdpSA, residual
(1 missing value generated)

. qui mean dgdp
. scalar mu = r(mean)

. replace dgdpSA = dgdpSA + mu
(251 real changes made, 251 to missing)
```

3. Estimate the mean of the quarterly growth rate of GDP (ΔY_t) for the sample period 2000:Q1–2009:Q4. Express the mean growth rate in percentage points at an annual rate. Repeat the calculations over the 1947:Q1–1999:Q4 sample period.

The command *tin* is a useful function here: we can specify dates using this function to restrict the estimation sample.

For period 2000:Q1–2009:Q4

Interactive Stata example

```
. mean(dgdp) if tin(2000q1,2009q4)

Mean estimation     Number of obs     =      40
-----------------------------------------------
       Mean     Std. Err.  [95% Conf. Interval]
-------------+---------------------------------
dgdp|.0041805  .0011058  .0019437    .0064173
-----------------------------------------------

* mean growth rate in percentage points at an annual rate
. gen ydgdp = 400*dgdp
(1 missing value generated)

. mean(ydgdp) if tin(2000q1,2009q4)

Mean estimation     Number of obs     =      40
-----------------------------------------------
      Mean      Std. Err.  [95% Conf. Interval]
-------------+---------------------------------
ydgdp|1.672201 .4423388  .7774865    2.566916
-----------------------------------------------
```

For period 1947:Q1–1999:Q4

Interactive Stata example

```
. mean(ydgdp) if tin(,1999q4)

Mean estimation     Number of obs     =     211
-----------------------------------------------
      Mean      Std. Err.  [95% Conf. Interval]
-------------+---------------------------------
ydgdp|3.463444 .2839517  2.903683    4.023205
-----------------------------------------------
```

4. **Estimate the first four autocorrelations of ΔY_t. Given your estimates, what kind of autoregressive model (lag order) seems most promising for forecasting? Explain.**

   The command *corrgram* tabulates and plots autocorrelations. It also provides the portmanteau Q statistics of tests for autocorrelation in the time series.

   Interactive Stata example

```
. corrgram dgdp, lags(4)

                                          -1       0       1 -1       0       1
LAG       AC       PAC      Q      Prob>Q  [Autocor.]      [Partial Autocor]
-------------------------------------------------------------------
1      0.3673    0.3676   34.277  0.0000          |--             |--
2      0.2120    0.0898   45.74   0.0000          |-              |
3     -0.0004   -0.1245   45.74   0.0000          |               |
4     -0.0883   -0.0895   47.746  0.0000          |               |
```

   Alternatively, you can calculate the autocorrelations by hand (only the first numerical column is of interest for us).

   Interactive Stata example

```
. corr dgdp l.dgdp l2.dgdp l3.dgdp l4.dgdp
(obs=247)

                     L.       L2.       L3.      L4.
            dgdp     dgdp     dgdp      dgdp     dgdp
-------------+------------------------------
dgdp|
--.  |     1.0000
L1.  |     0.3681   1.0000
L2.  |     0.2196   0.3704   1.0000
L3.  |     0.0026   0.2184   0.3662   1.0000
L4.  |    -0.0898   0.0031   0.2136   0.3641 1.0000
```

   The estimated autocorrelation coefficients suggest that only the growth rate in the last and second last quarter provide information about this quarter's growth (the autocorrelation for lags 3 and 4 is actually close to zero). The positive autocorrelation of the growth rate means that, on average, an increase in GDP in one quarter is associated with an increase in the next. In sum, an AR(1) or AR(2) seems most promising.

5. **Estimate an AR(1) model for ΔY_t. What is the estimated AR(1) coefficient? Is the coefficient statistically significantly different from zero?**

Interactive Stata example

```
. reg dgdp L.dgdp if tin(1955q1,2007q4),r

Linear regression                  Number of obs =     212
F(  1,    210) =    15.52
Prob > F       =   0.0001
R-squared      =   0.0882
Root MSE       =   .00864

---------------------------------------------------------
               Robust
dgdp | Coef.   Std. Err. t     P>|t|  [95% Conf. Interval]
-------------+-------------------------------------------
dgdp |
L1.   |.2957912 .075085   3.94  0.000  .1477743    .4438081
_cons|.0057121 .0009158  6.24  0.000  .0039068    .0075174
---------------------------------------------------------
```

The coefficient is ≈ 0.3 and positive, so an increase in the growth rate in the past quarter is associated with an increase in the growth rate in the current quarter.

The coefficient is also statistically highly significant (at the 1% level). Thus, the lagged growth rate is a useful predictor.

6. **Look again at the plot of the quarterly growth rate of GDP [see Figure 1 above]. You can observe that fluctuations in the US GDP growth rate declined significantly in the mid-1980s. Given this observation, you are worried that there might be a break in the AR(1) model. [Hint: Look into section 14.7 "Nonstationarity II: Breaks" of the Stock/Watson book]**

   a) **Suppose that you suspected that the intercept changed in 1984:1. How would you modify the AR(1) model to incorporate this change? How would you test for a change in the intercept? [Conceptual question]**

      You would add a binary variable, say D_t, that equals 0 for dates prior to 1984:1 and equals 1 for dates 1984:1 and beyond. If the coefficient on D_t is significantly different from zero in the regression (as judged by its t-statistic), then this would be evidence of an intercept break in 1984:1.

   b) **Suppose that you suspect that the AR coefficient changed in 1984:1. Modify your regression in Stata to incorporate this change. Is there evidence for a break?**
      The hypothesis of a break in the coefficient can be tested using a binary variable interaction regression.

Interactive Stata example

```
. gen d84 = tin(1984q1,)
. gen gdp84 = d84 * L.dgdp
(2 missing values generated)
. reg dgdp L.dgdp gdp84 if tin(1955q1,2007q4),r

Linear regression                       Number of obs =     212
F(  2,    209) =    7.69
Prob > F       =  0.0006
R-squared      =  0.0889
Root MSE       =  .00866

-------------------------------------------------------------
                 Robust
dgdp   |  Coef.    Std. Err.  t     P>|t|  [95% Conf. Interval]
-------------+-----------------------------------------------
dgdp   |
L1.    | .3024799 .0780113   3.88  0.000   .1486899  .4562698
gdp84  |-.0475781 .095905   -0.50  0.620  -.2366432  .1414871
_cons  | .0058281 .0010091   5.78  0.000   .0038389  .0078174
-------------------------------------------------------------
```

As a robustness check, we can also include the intercept break in the model.

Interactive Stata example

```
. reg dgdp L.dgdp d84 gdp84 if tin(1955q1,2007q4),r

Linear regression                       Number of obs =     212
F(  3,    208) =    6.23
Prob > F       =  0.0005
R-squared      =  0.0890
Root MSE       =  .00868

--------------------------------------------------------------
                  Robust
dgdp   |  Coef.     Std. Err.  t     P>|t|  [95% Conf. Interval]
-------------+------------------------------------------------
dgdp   |
L1.    | .2992613  .0870672   3.44  0.001   .1276139  .4709087
d84    |-.0002678  .0017076  -0.16  0.876  -.0036342  .0030987
gdp84  |-.0274952  .1427788  -0.19  0.847  -.3089743  .2539839
_cons  | .0059036  .0013452   4.39  0.000   .0032516  .0085557
--------------------------------------------------------------
```

There is no evidence for a break in the AR(1) coefficient. The interaction variable remains insignificant and also the AR(1) estimate is robust.

7. Estimate an AR(2) model for ΔY_t. Is the AR(2) coefficient statistically significantly different from zero? Is this model preferred to the AR(1) model?

Interactive Stata example

```
. reg dgdp L(1/2).dgdp if tin(1955q1,2007q4),r

Linear regression                          Number of obs =     212
F(  2,    209) =     8.31
Prob > F       =   0.0003
R-squared      =   0.0969
Root MSE       =   .00862

--------------------------------------------------------------
                  Robust
dgdp   | Coef.     Std. Err.  t     P>|t|  [95% Conf. Interval]
-------------+------------------------------------------------
dgdp   |
L1.    | .2668701  .0798902   3.34  0.001   .1093763  .4243639
L2.    | .0971629  .0860787   1.13  0.260  -.0725309  .2668567
_cons  | .0051502  .0011136   4.63  0.000    .002955  .0073454
--------------------------------------------------------------
```

The AR(2) coefficient is not statistically significantly different from zero, indicating that the second lag is not a useful predictor.

However, the coefficients on the two lags are jointly significantly different from zero at the 1% significance level.

Interactive Stata example

```
. test L.dgdp L2.dgdp

( 1)  L.dgdp = 0
( 2)  L2.dgdp = 0

F(  2,    209) =     8.31
Prob > F =     0.0003
```

The AR(2) also leads to a marginal improvement in the R^2.

Interactive Stata example

```
. dis "Adjusted Rsquared = " _result(8)
Adjusted Rsquared = .08820865

. qui reg dgdp L.dgdp if tin(1955q1,2007q4),r
. dis "Adjusted Rsquared = " _result(8)
Adjusted Rsquared = .0838363
```

In sum, the AR(2) could be preferred to an AR(1) but the decision, at this point, is not clear cut.

8. Estimate AR(3) and AR(4) models.

Interactive Stata example

```
. reg dgdp L(1/3).dgdp if tin(1955q1,2007q4),r

Linear regression                    Number of obs =     212
F(  3,   208) =    5.64
Prob > F      =  0.0010
R-squared     =  0.1003
Root MSE      =  .00863

-------------------------------------------------------------
                 Robust
dgdp   | Coef.    Std. Err. t     P>|t|    [95% Conf. Interval]
-------------+-----------------------------------------------
dgdp   |
L1.    |.2724849 .0803879   3.39  0.001    .1140054   .4309645
L2.    |.1136115 .0915593   1.24  0.216   -.0668916   .2941147
L3.    |-.061137 .0787244  -0.78  0.438   -.2163370   .0940630
_cons  |.0054697 .0011733   4.66  0.000    .0031565   .0077828
-------------------------------------------------------------

. dis "Adjusted Rsquared = " _result(8)
Adjusted Rsquared = .08729245
```

Interactive Stata example

```
. reg dgdp L(1/4).dgdp if tin(1955q1,2007q4),r

Linear regression                      Number of obs =     212
F(  4,   207) =    4.30
Prob > F      =  0.0023
R-squared     =  0.1021
Root MSE      =  .00864

--------------------------------------------------------------
                 Robust
dgdp   | Coef.    Std. Err.  t      P>|t|   [95% Conf. Interval]
-------------+------------------------------------------------
dgdp   |
L1.    | .2693964  .0795022   3.39  0.001    .1126586  .4261342
L2.    | .1183090  .0935451   1.26  0.207   -.0661143  .3027322
L3.    |-.0486959  .0788254  -0.62  0.537   -.2040995  .1067077
L4.    |-.0442365  .0834590  -0.53  0.597   -.2087751   .120302
_cons  | .0057154  .0012539   4.56  0.000    .0032433  .0081874
--------------------------------------------------------------


. dis "Adjusted Rsquared = " _result(8)
Adjusted Rsquared = .08470964
```

9. **Read section 14.5 "Lag Length Selection Using Information Criteria" of the Stock and Watson book. Using the estimated AR(1)-AR(4) models, use the BIC to choose the number of lags in the AR model. How many lags does the AIC choose?**

   The postestimation command *estat ic* delivers AIC and BIC values.

   A convenient way to compare different models in terms of their information criteria is to use the *eststo* command. It simplifies making regression tables from stored estimates. You need to install the command first, e.g. via findit eststo.

   Interactive Stata example

```
. eststo clear
. eststo ar1: qui reg dgdp L(1/1).dgdp if tin(1955q1,2007q4),r
. eststo ar2: qui reg dgdp L(1/2).dgdp if tin(1955q1,2007q4),r
. eststo ar3: qui reg dgdp L(1/3).dgdp if tin(1955q1,2007q4),r
. eststo ar4: qui reg dgdp L(1/4).dgdp if tin(1955q1,2007q4),r

. est stat ar1 ar2 ar3 ar4

-------------------------------------------------------
Model | Obs  ll(null)  ll(model) df  AIC        BIC
-------------+-----------------------------------------
one   | 212  697.5823  707.3673  2  -1410.735  -1404.021
two   | 212  697.5823  708.3804  3  -1410.761  -1400.691
three | 212  697.5823  708.7823  4  -1409.565  -1396.138
four  | 212  697.5823  708.9936  5  -1407.987  -1391.204
-------------------------------------------------------
Note:  N=Obs used in calculating BIC; see [R] BIC note
```

   Important: model with the lowest IC is preferred.

   BIC: AR(1) and AIC: AR(2)

10. **Let $R_t = ln(TBill_t)$ denote the logarithm of the interest rate for 3-month treasury bills and ΔR_t the quarterly growth rate of treasury bills. Estimate an ADL(1,4) model for ΔY_t using lags of ΔR_t as additional predictors. Comparing the ADL(1,4) model to the AR(1) model, by how much has the adjusted R^2 changed?**

   Interactive Stata example

```

. gen ltbill = log(tbill)
. gen dtbill = D.ltbill
(1 missing value generated)

. reg dgdp L.dgdp L(1/4).dtbill if tin(1955q1,2007q4),r


```

```
Linear regression                             Number of obs =      212
F(  5,   206) =    7.21
Prob > F      =  0.0000
R-squared     =  0.1663
Root MSE      =  .00835

--------------------------------------------------------------
                  Robust
dgdp     | Coef.    Std. Err. t       P>|t|  [95% Conf. Interval]
-------------+------------------------------------------------
dgdp     |
L1.      | .2329648 .0783186   2.97  0.003    .078556  .3873737

dtbill |
L1.      | .0079146 .0046685   1.70  0.092  -.0012895  .0171187
L2.      |-.0123303 .0057237  -2.15  0.032  -.0236147 -.0010458
L3.      | .0042966 .0046757   0.92  0.359  -.0049218   .013515
L4.      |-.0174902 .0050049  -3.49  0.001  -.0273576 -.0076227
_cons    | .0063399 .0009458   6.70  0.000   .0044753  .0082045
-------------------------------------------------------------

. dis "Adjusted Rsquared = " _result(8)
Adjusted Rsquared = .14608674

. qui reg dgdp L.dgdp if tin(1955q1,2007q4),r
. dis "Adjusted Rsquared = " _result(8)
Adjusted Rsquared = .0838363
```

The R^2 of the ADL (1,4) regression is 0.15, a solid improvement over 0.084 for the AR(1).

Let's also formally test whether the lagged growth rates of treasury bills (jointly) help to predict GDP growth. Here we use the command *testparm* that allows us to use the lag operator (results are identical to *test* command).

Interactive Stata example

```
. testparm L(1/4).dtbill

( 1)  L.dtbill = 0
( 2)  L2.dtbill = 0
( 3)  L3.dtbill = 0
( 4)  L4.dtbill = 0

F(  4,   206) =    4.78
Prob > F =    0.0010
```

11. **Construct out-of-sample forecasts using the AR(1) model for the two-year holdout period (i.e. 2008:1 to 2009:4). Produce also out-of-sample forecasts using the ADL(1,4) model.**

First, in-sample estimation of the the AR(1):

Interactive Stata example

```
. reg dgdp L.dgdp if tin(1955q1,2007q4),r

Linear regression                      Number of obs =     212
F(  1,    210) =    15.52
Prob > F       =   0.0001
R-squared      =   0.0882
Root MSE       =   .00864

------------------------------------------------------------
dgdp | Coef.     Std. Err.  t     P>|t|  [95% Conf. Interval]
-------------+----------------------------------------------
dgdp |
L1.   |.2957912  .075085    3.94  0.000  .1477743  .4438081
_cons|.0057121  .0009158   6.24  0.000  .0039068  .0075174
```

Then, out-of-sample forecast. Note that the command predict can be used to make in-sample or out-of-sample predictions.

Interactive Stata example

```
. predict far1 if tin(2008q1,)
(option xb assumed; fitted values)
(244 missing values generated)
```

Now the ADL(1,4) forecast:

Interactive Stata example

```
. reg dgdp L.dgdp L(1/4).dtbill if tin(1955q1,2007q4),r

Linear regression                         Number of obs =     212
F(  5,    206) =     7.21
Prob > F       =   0.0000
R-squared      =   0.1663
Root MSE       =   .00835
----------------------------------------------------------------
dgdp   | Coef.     Std. Err.  t      P>|t|  [95% Conf. Interval]
-------------+--------------------------------------------------
dgdp   |
L1.    | .2329648  .0783186   2.97   0.003   .078556    .3873737
dtbill|
L1.    | .0079146  .0046685   1.70   0.092  -.0012895   .0171187
L2.    |-.0123303  .0057237  -2.15   0.032  -.0236147  -.0010458
L3.    | .0042966  .0046757   0.92   0.359  -.0049218    .013515
L4.    |-.0174902  .0050049  -3.49   0.001  -.0273576  -.0076227
_cons  | .0063399  .0009458   6.70   0.000   .0044753   .0082045
```

```
Interactive Stata example
. predict fadl14 if tin(2008q1,)
(option xb assumed; fitted values)
(244 missing values generated)
```

12. What is the distinction between a forecast and a predicted value? [Conceptual question]

Key: The forecast is not an OLS predicted value!

A predicted value refers to the value of Y predicted (using a regression) for an observation in the sample used to estimate the regression. In contrast, the forecast is made for some date beyond the data set used to estimate the regression, so the data on the actual value of the forecasted dependent variable are not in the sample used to estimate the regression.

Predicted values are "in sample". Forecasts are "out-of sample".

13. Look at a plot of your two forecasts from (11.) and the true value of the quarterly growth rate of GDP. Are your forecasts biased?

```
Interactive Stata example
. preserve
. keep if tin(2008q1,)
. twoway line dgdp far1 fadl14 time legend(label(1 "Actual growth rate")
label(2 "AR forecast") label(3 "ADL forecast"))
. restore
```

The forecast are generally biased upwards. This is particularly true for the ADL. Only for 2009:3 the forecast errors of the models are very low.

Figure 2: Forecasts vs. Actual growth rate