Menzie Chinn
Spring 2020

A Cookbook Approach to Time Series Analysis and Forecasting using Stata
PA819, Lecture 24

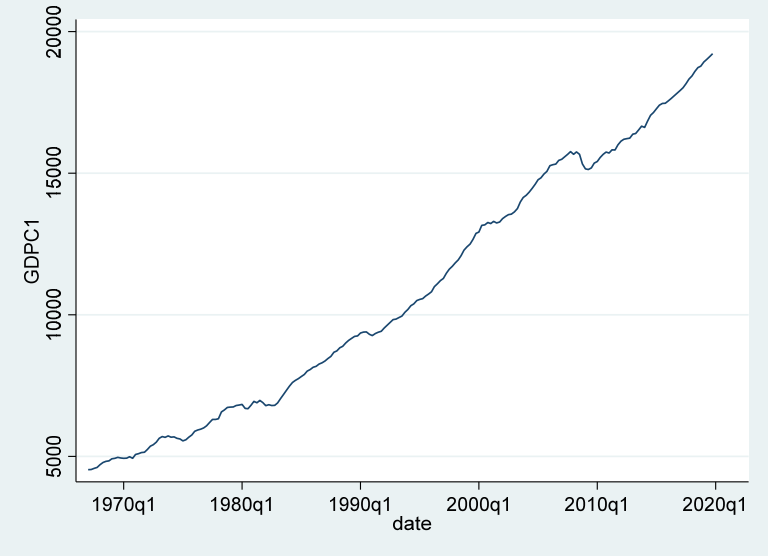
The Problem

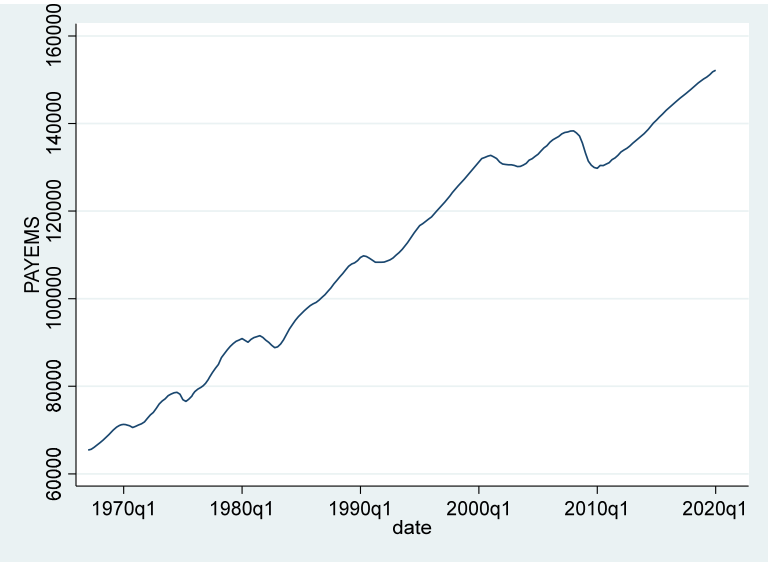
Suppose you want to analyze data for the United States over the 1967q1-2020q1 period, with two objectives: (1) find the relationship between the nonfarm payroll employment (or employment for short) and real gross domestic product (GDP), and (2) forecast GDP, using (2a) employment, and (2b) using only the past history of GDP.

Data

Nonfarm payroll employment is in 000's, seasonally adjusted (PAYEMS), sourced from Bureau of Labor Statistics. The original data is reported on a monthly basis; I've converted to quarterly basis by taking the average of the monthly data. Real, or inflation adjusted, GDP is in billions of Chained 2012$, Seasonally Adjusted at Annual Rates (GDPC1), sourced from Bureau of Economic Analysis. Both of these series can be retrieved from the St. Louis Fed data app, FRED: https://fred.stlouisfed.org/ . These two series are in a Stata data set `us_gdp_empl.dta` .

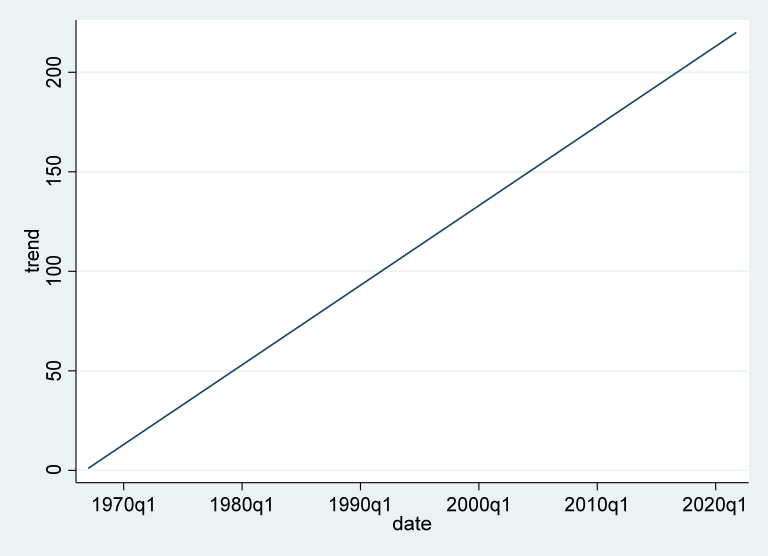
The two figures below depict the two series (use command "`tsline GDPC1`")





Some Data Preliminaries

One might wish to characterize how fast each series is growing. One way to do this for GDPC1 is to regress the logged GDPC1 (call it LGDPC1) on a time trend. Define a time trend variable called "trend" using the command "`gen trend=_n`".



```
. reg LGDPC1 trend

      Source |       SS           df       MS      Number of obs   =       212
-------------+----------------------------------   F(1, 210)       =  19431.38
       Model |  39.6689278         1  39.6689278   Prob > F        =    0.0000
    Residual |  .428712383       210  .002041488   R-squared       =    0.9893
-------------+----------------------------------   Adj R-squared   =    0.9893
       Total |  40.0976402       211   .19003621   Root MSE        =    .04518

------------------------------------------------------------------------------
      LGDPC1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       trend |   .0070683   .0000507   139.40   0.000     .0069684    .0071683
       _cons |   8.446179   .0062284  1356.08   0.000       8.4339    8.458457
------------------------------------------------------------------------------
```

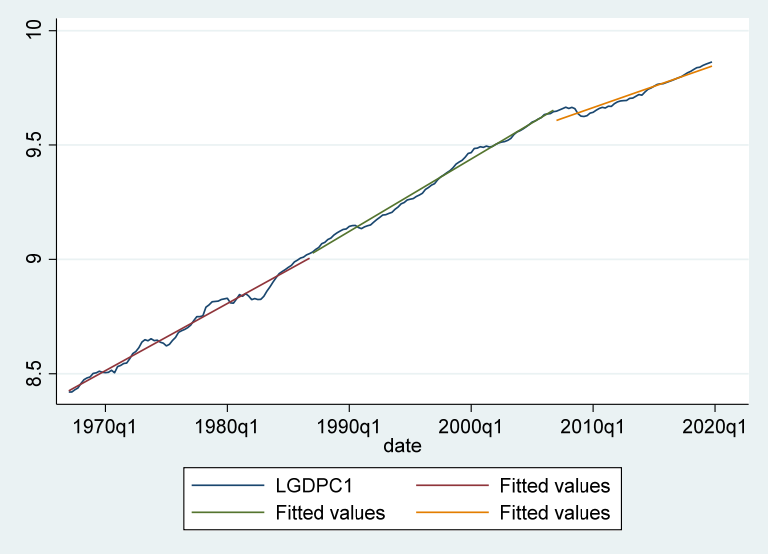
The coefficient on "trend" is the rate of change per quarter, so multiply by 4 to get 0.0283 (2.83%), which is the annual growth rate of real GDP.

One problem with this approach is that many time series are "nonstationary" – that is they do not have a true deterministic trend. This is likely true of GDPC1; I estimate the trend over three subsamples: 1967q1-1986q4, 1987q1-2006q4, 2007q1-2019q4. Notice how the trend growth (per quarter) in red is different in each subsample.

```
. reg LGDPC1 trend if tin(1967q1, 1986q4)

      Source |       SS           df       MS      Number of obs   =        80
-------------+----------------------------------   F(1, 78)        =   3991.05
       Model |  2.29798672         1  2.29798672   Prob > F        =    0.0000
    Residual |  .044911278        78  .000575786   R-squared       =    0.9808
-------------+----------------------------------   Adj R-squared   =    0.9806
       Total |    2.342898        79  .029656937   Root MSE        =      .024

------------------------------------------------------------------------------
      LGDPC1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       trend |   .0073394   .0001162    63.17   0.000     .0071082    .0075707
       _cons |   8.417857   .0054163  1554.18   0.000     8.407074    8.428639
------------------------------------------------------------------------------

. reg LGDPC1 trend if tin(1987q1, 2006q4)

      Source |       SS           df       MS      Number of obs   =        80
-------------+----------------------------------   F(1, 78)        =   8996.00
       Model |  2.66981575         1  2.66981575   Prob > F        =    0.0000
    Residual |  .023148686        78  .000296778   R-squared       =    0.9914
-------------+----------------------------------   Adj R-squared   =    0.9913
       Total |  2.69296444        79  .034088157   Root MSE        =    .01723

------------------------------------------------------------------------------
      LGDPC1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
       trend |    .007911   .0000834    94.85   0.000     .0077449     .008077
       _cons |   8.386654   .0102335   819.53   0.000     8.366281    8.407028
------------------------------------------------------------------------------

. reg LGDPC1 trend if tin(2007q1, 2019q4)

      Source |       SS           df       MS      Number of obs   =        52
-------------+----------------------------------   F(1, 50)        =    669.08
       Model |  .254170413         1  .254170413   Prob > F        =    0.0000
    Residual |  .018994094        50  .000379882   R-squared       =    0.9305
-------------+----------------------------------   Adj R-squared   =    0.9291
       Total |  .273164506        51  .005356167   Root MSE        =    .01949

------------------------------------------------------------------------------
      LGDPC1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       trend |   .0046583   .0001801    25.87   0.000     .0042966      .00502
       _cons |   8.857689   .0336954   262.88   0.000      8.79001    8.925369
------------------------------------------------------------------------------
```

As additional data is added, the estimate of the trend changes. I plot the (log) of real GDP, and the various trends over the subperiods.

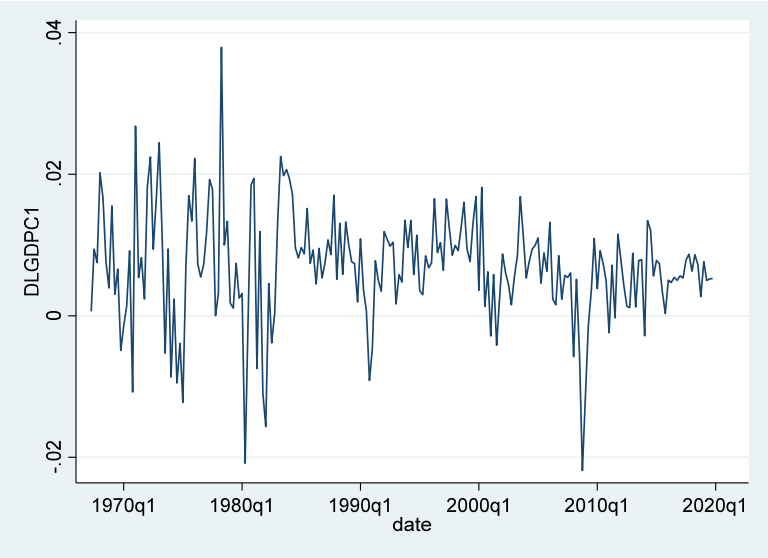


Roughly speaking, log real GDP is a random walk (with drift α), written as:

$$y_t = \alpha + y_{t-1} + \varepsilon_t$$

A formal approach to confirming this view involves using what are called unit root tests – the Augmented Dickey Fuller test is the most commonly used, and is included in Stata (command "`dfuller`"). If the series has these characteristics it is hard to include in regular regressions, particularly if the other variables in the regression do not have the same characteristics.

In order to render this variable usable ("stationary"), first differencing usually works (i.e., subtracting a lagged value). When the variable is in logs, then the first difference is the growth rate of that variable. Below I plot the first difference of log real GDP ("`gen DLGDPC1 = LGDPC1-LGDPC1[_n-1]`").



A value of "0.02" means GDP grew by 2% quarter-on-quarter; on an annualized basis, that's 8%.

Estimating the GDP-Employment Relationship

If I want to estimate a relationship between GDP and employment, instead of running a regression between LGDPC1 and LPAYEMS, I run a regression of DLGDPC1 and DLPAYEMS. (One *can* estimate a relationship between LGDPC1 and LPAYEMS, but only under very special conditions – when the two variables are both integration (requiring first differencing to render stationary) and are "cointegrated"; see Chinn *(JPAM*, 1991), https://www.ssc.wisc.edu/~mchinn/Beware_of_Econometricians.pdf for details.) Here is the regression in growth rates.

```
. reg DLGDPC1 DLPAYEMS

      Source |       SS           df       MS      Number of obs   =       211
-------------+----------------------------------   F(1, 209)       =    159.87
       Model |  .005578854         1  .005578854   Prob > F        =    0.0000
    Residual |  .007293461       209  .000034897   R-squared       =    0.4334
-------------+----------------------------------   Adj R-squared   =    0.4307
       Total |  .012872315       210  .000061297   Root MSE        =    .00591

------------------------------------------------------------------------------
     DLGDPC1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    DLPAYEMS |   1.000834   .0791559    12.64   0.000     .8447878     1.15688
       _cons |   .0028544   .0005147     5.55   0.000     .0018397    .0038691
------------------------------------------------------------------------------
```

This output means that each one percent increase in employment is associated with one percent increase in GDP. Over each year, GDP drifts higher by 0.01142 (= 0.0028544 x 4). This can be interpreted as productivity growth.

There's a possibility that the estimated errors are serially correlated, so that the calculated standard errors are wrong. A Durbin-Watson test checks for first-order serial correlation (that is, is residual this period correlated with the residual in the preceding period).

```
. estat dwatson

Durbin-Watson d-statistic(  2,   211) =  1.826087
```

A value between 1.7 and 2.3 is often take to mean serial correlation is not too substantial. However, if we wanted to account for the serial correlation in the calculation of the standard errors, we can use "Newey-West robust standard errors". To estimate OLS and get these robust standard errors, use:

```
. newey DLGDPC1 DLPAYEMS, lag(3)

Regression with Newey-West standard errors      Number of obs     =        211
maximum lag: 3                                  F(  1,       209) =     105.42
                                                Prob > F          =     0.0000

------------------------------------------------------------------------------
             |             Newey-West
     DLGDPC1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    DLPAYEMS |   1.000834    .097476    10.27   0.000     .8086719    1.192996
       _cons |   .0028544   .0006651     4.29   0.000     .0015433    .0041655
------------------------------------------------------------------------------
```

The "lag(3)" option indicates I assume the maximum autocorrelation is 3 quarters (which makes sense for quarterly data). Notice the Newey-West standard errors are typically larger than the standard.

Forecasting

We're going to examine two ways of forecasting LGDPC1. The first is using only the history of LGDPC1; the second is using a related series that extends later in time than LGDPC1 (notice the employment series goes up to 2020q1, while GDP only goes up to 2019q4.

Forecasting One Series

The first approach is to use a "time series model". Consider a series, y_t; many macroeconomic series can be modeled as:

$$\Delta y_t = \alpha + \varphi \Delta y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

Where

$$\Delta y_t \equiv y_t - y_{t-1}$$

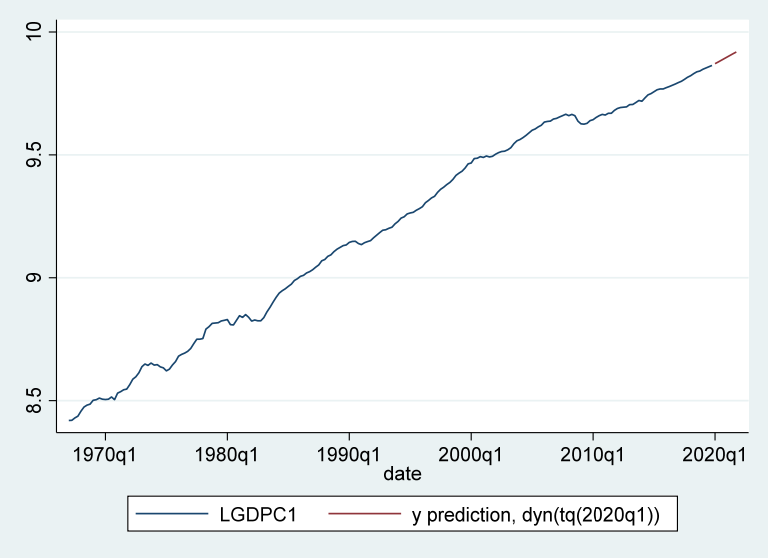
This specification is called an ARIMA(1,1,1), for a model where you have to difference once (I=1), which is related to itself lagged once (AR=1), and the error and lagged error are included (MA=1).

```
. arima LGDPC1, arima(1,1,1)

(setting optimization to BHHH)
Iteration 0:   log likelihood =  729.92198
:
Iteration 7:   log likelihood =   736.8766

ARIMA regression

Sample:  1967q2 - 2019q4                        Number of obs     =        211
                                                Wald chi2(2)      =      53.61
Log likelihood =  736.8766                      Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |                 OPG
    D.LGDPC1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
LGDPC1       |
       _cons |   .0068202   .0009087     7.51   0.000     .0050392    .0086011
-------------+----------------------------------------------------------------
ARMA         |
          ar |
         L1. |    .627217   .1450927     4.32   0.000     .3428405    .9115935
             |
          ma |
         L1. |  -.3527366   .1691097    -2.09   0.037    -.6841854   -.0212878
-------------+----------------------------------------------------------------
      /sigma |    .007361   .0002366    31.11   0.000     .0068972    .0078247
------------------------------------------------------------------------------
Note: The test of the variance against zero is one sided, and the two-sided
      confidence interval is truncated at zero.
```

One can think of forecasting the growth rate of GDP (DLGDPC1), or the (log) level of GDP (LGDPC1). In general, I'm interested in the (log) level, so in order to get that forecast, you type in right after running the ARIMA regression:

```
. predict LGDPC1_hat if tin(2020q1, 2021q4), dynamic(tq(2020q1)) y
```

This will generate a dynamic forecast starting in 2020q1 extending up to the end of the sample. A dynamic forecast is one where the forecast for 2020q2 depends on the *forecasted* value (not actual value) for DLGDPC1 in 2020q1, the forecast for 2020q3 depends on the *forecasted* value for DLGDPC1 in 2020q2, and so forth. The graph for the log GDP and dynamic forecast is shown below.

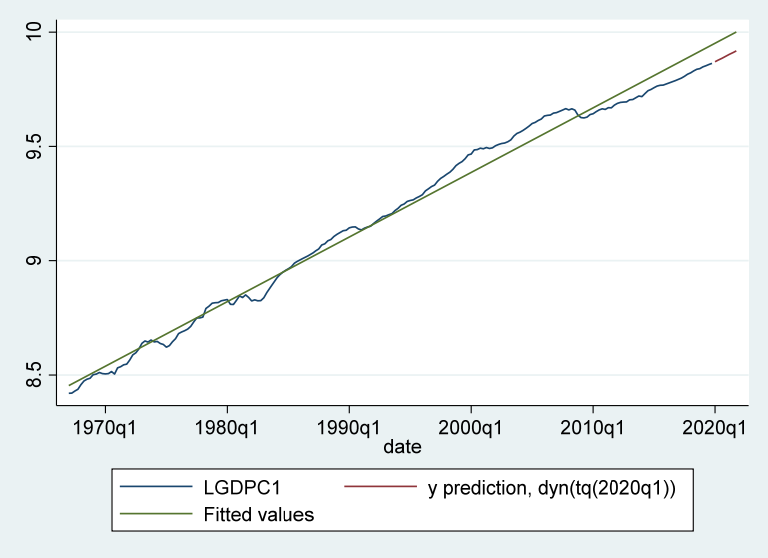


Where the blue line is the actual log level of real GDP, and the red is the dynamic forecast, using data available up to 2019Q4.

It's useful to compare this forecast against what you get if you assumed log real GDP could be modeled as a simple function of a time trend. I repeat the regression reported on page 3.

```
. reg LGDPC1 trend

      Source |       SS           df       MS      Number of obs   =       212
-------------+----------------------------------   F(1, 210)       =  19431.38
       Model |  39.6689278         1  39.6689278   Prob > F        =    0.0000
    Residual |  .428712383       210  .002041488   R-squared       =    0.9893
-------------+----------------------------------   Adj R-squared   =    0.9893
       Total |  40.0976402       211   .19003621   Root MSE        =    .04518

------------------------------------------------------------------------------
      LGDPC1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       trend|   .0070683   .0000507   139.40   0.000     .0069684    .0071683
       _cons |   8.446179   .0062284  1356.08   0.000       8.4339    8.458457
------------------------------------------------------------------------------
```

The prediction using the ARIMA model (red) compared to the function of time trend (green) is shown in the figure below:



The bottom line is that you have to be very careful how you treat time series data.

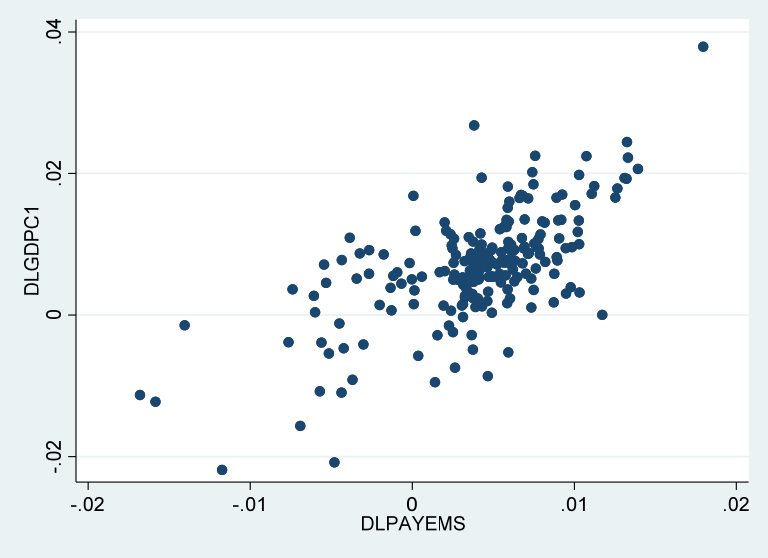
By the way, oftentimes simpler models will work, like growth rate this period depends on growth rate last quarter, i.e., ARIMA(1,1,0).

$$\Delta y_t = \alpha + \varphi \Delta y_{t-1} + \varepsilon_t$$

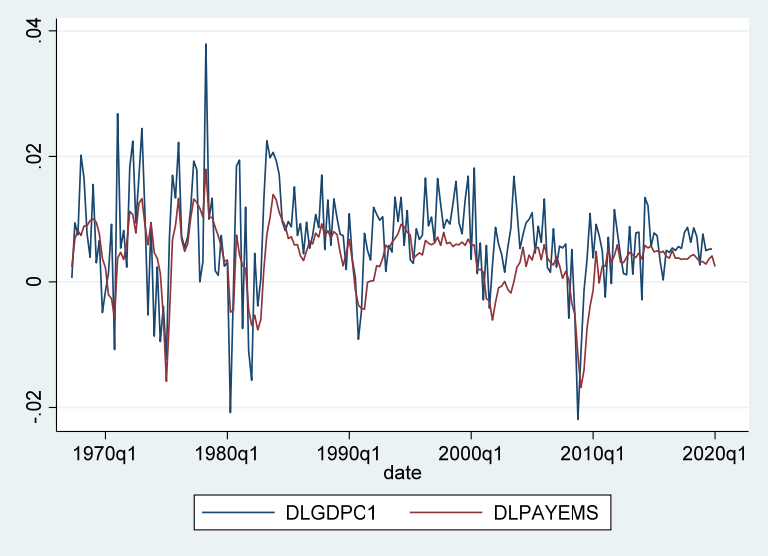
Forecasting with Multiple Series

This caveat about how you treat trends applies when you have multiple series. Suppose you want to forecast a variable that lags in reporting. For instance, in this data set, we have observation for 2020q1 for employment (PAYEMS), but the latest observation for real GDP (GDPC1) in 2019q4. One could use the relationship between LGDPC1 and LPAYEMS1, but given the uncertainty whether the two are linked in log levels, one might be better of using the relationship between growth rates (DLGDPC1 and DLPAYEMS). This relationship is shown below, using the command:

```
. graph twoway scatter DLGDPC1 DLPAYEMS
```



It's always a good idea to plot the data too:



Let's repeat the regression on page 5:

```
. reg DLGDPC1 DLPAYEMS

      Source |       SS           df       MS      Number of obs   =       211
-------------+----------------------------------   F(1, 209)       =    159.87
       Model |  .005578854         1  .005578854   Prob > F        =    0.0000
    Residual |  .007293461       209  .000034897   R-squared       =    0.4334
-------------+----------------------------------   Adj R-squared   =    0.4307
       Total |  .012872315       210  .000061297   Root MSE        =    .00591

------------------------------------------------------------------------------
     DLGDPC1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    DLPAYEMS |   1.000834   .0791559    12.64   0.000     .8447878     1.15688
```

```
       _cons |    .0028544   .0005147     5.55   0.000      .0018397    .0038691
------------------------------------------------------------------------------
```

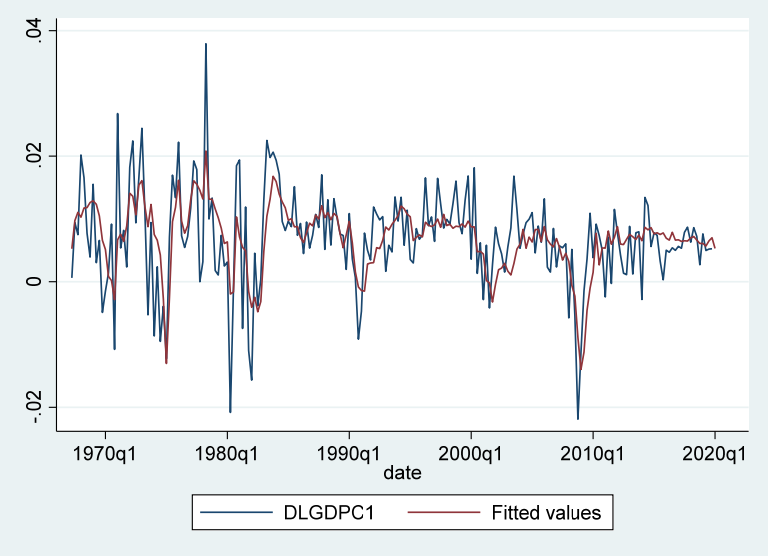
We can use this regression to fit the 2020q1 DLGDPC1, based on 2020q1 DLPAYEMS:

```
. predict DLGDPC1_static
```

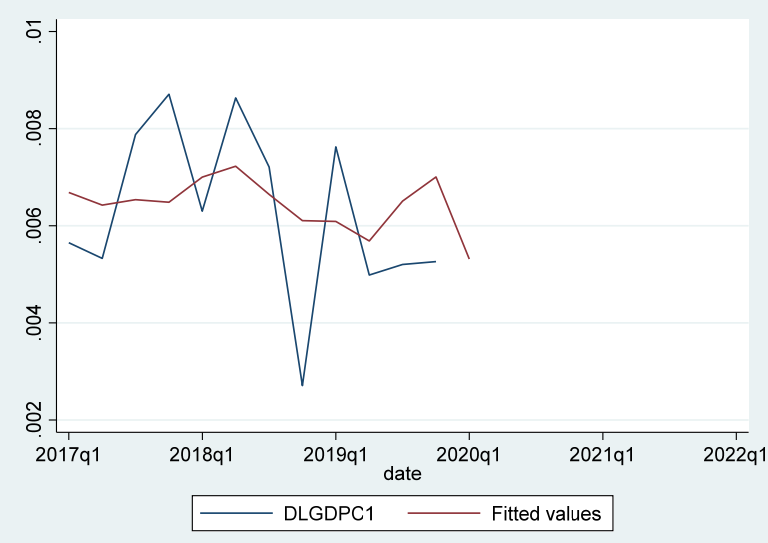
You can plot the actual (blue) and predicted (red)

series:

```
. tsline  DLGDPC1 DLGDPC1_static
```



It's hard to see how good of a fit we obtained, but here's a detail:



The fit is not particularly good, and one might want to use more variables on the right hand side of the regression, if all one wants to do is predict GDP growth.

To see what the regression line looks like with the scatterplot overlaid, you first have to estimate the standard error of prediction (which is based on the last regression estimated), then second give the plot command:

```
. predict staticse1, stdp

. twoway lfitci DLGDPC1 DLPAYEMS || scatter DLGDPC1 DLPAYEMS
```