Properties of Autocorrelated Processes
Economics 30331

Bill Evans
Fall 2015

Suppose we have time series data series labeled as y_t where t=1,2,3,…T (the final period). Some examples are the daily closing price of the S&500, quarterly per capita GDP, monthly unemployment, weekly movie attendance, etc.

As we will show below, many time series processes demonstrate autocorrelation, that is, there is some persistence from one period to the next in the values. For example, this month's unemployment rate is highly predictive of next periods. Likewise, the stock price for Apple today is highly predictive of tomorrow's prices.

A primary characteristic of autocorrelated processes are whether they are stationary or not. The time series is considered covariance stationary if the series has a finite second moment and exhibits three characteristics

i) $Var(y_t)$ = constant for all t
ii) $E[y_t]$ = constant for all t
iii) $Cov(y_t,y_s)=Cov(y_{t+h},y_{s+h})$ where t>s

The first characteristic says that the variance of the variable must be finite for all values of time. The second says that the mean cannot depend on time. The third characteristic says that the covariance between two points is only a function of the distance between the points (t-s) and not the point we are considering (t+h and s+h).

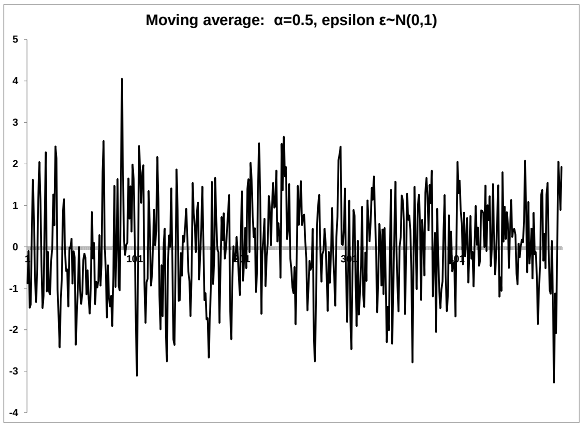
We will put aside for now the importance of establishing stationarity but in general, there are two problems with non-stationary series. First, if data is non-stationary, the underlying assumptions of our statistical tests are wrong so we cannot do things like t-tests and f-tests. Second, as we will show with some simulations, a regression of a non-stationary series on another non-stationary series tends to generate spurious correlation and high Type I error rates. We will discuss these issues in some detail in class but for now, we want to focus on establishing whether a series is stationary or not.

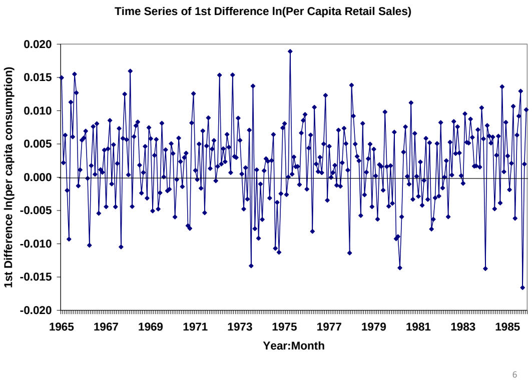
When we ask whether a series is stationary or not, we simply check the three criteria above and if one is violated, the series is non-stationary.

Example 1: A moving average representation

Suppose that $y_t = \varepsilon_t + \alpha\varepsilon_{t-1}$

Where ε_t is an independent and identically distributed error, so $E[\varepsilon_t]=0$, $Var(\varepsilon_t)=\sigma_\varepsilon^2$, $\text{cov}(\varepsilon_t,\varepsilon_{t-1})=0$ and $\alpha<1$. A graph of this process is below with some random variables selected for ε_t and assuming $\sigma_\varepsilon^2=1$. Note that this model looks a lot like the first difference in monthly real per capita retail sales!





To check stationarity, we need simply go through the conditions above. For ii) note that $E[y_t]=0$ because $E[\varepsilon_t]=0$ and $E[\varepsilon_{t-1}]=0.$ For condition i), the variance is a little harder --

$$\begin{aligned}Var(y_t) &= E[(y_t - E(y_t))^2] = E[y_t{}^2] = E[(\varepsilon_t + \alpha\varepsilon_{t-1})^2] \\ &= E[\varepsilon_t^2 + \alpha^2\varepsilon_{t-1}^2 + 2\varepsilon_t\varepsilon_{t-1}] = E[\varepsilon_t^2] + E[\alpha^2\varepsilon_{t-1}^2] + 2E[\varepsilon_t\varepsilon_{t-1}] \\ &= (1+\alpha^2)\sigma_\varepsilon^2\end{aligned}$$

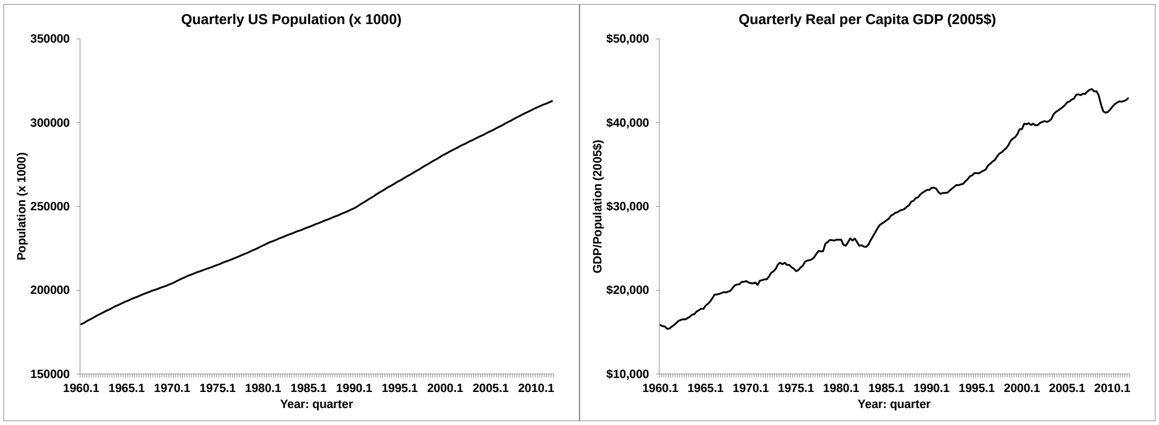
For condition iii), we note that $\mathrm{cov}(y_t, y_{t-1}) = E[(y_t - E[y_t])(y_{t-1} - E[y_{t-1}])] = E[y_t y_{t-1}]$ because $E[y_t] = E[y_{t-1}] = 0$. Note as well that $y_{t-1} = \varepsilon_{t-1} + \alpha\varepsilon_{t-2}$ so

$$\begin{aligned}\mathrm{cov}(y_t, y_{t-1}) &= E[y_t y_{t-1}] = E[(\varepsilon_t + \alpha\varepsilon_{t-1})(\varepsilon_{t-1} + \alpha\varepsilon_{t-2})] \\ &= E[\varepsilon_t\varepsilon_{t-1} + \alpha\varepsilon_{t-1}^2 + \alpha\varepsilon_t\varepsilon_{t-2} + \alpha^2\varepsilon_{t-1}\varepsilon_{t-2}] = E[\alpha\varepsilon_{t-1}^2] = \alpha\sigma_\varepsilon^2\end{aligned}$$

You can show that $\mathrm{cov}(y_t, y_{t-h}) = \alpha^h\sigma_\varepsilon^2$ for h>1 so the Cov(y_t,y_s) is only a function of x-s, not x or s.

Example 2: A linear time trend

Consider a linear model of the form $y_t = \beta_0 + t\beta_t + \varepsilon_t$ where $E[\varepsilon_t]=0$, $Var(\varepsilon_t) = \sigma_\varepsilon^2$ and $\mathrm{cov}(\varepsilon_t, \varepsilon_{t-1}) = 0$. The time series for quarterly population in the US monthly or quarterly real per capita GDP look like a linear time trend.



Note from the start that

$$E[y_t] = E[\beta_0 + t\beta_t + \varepsilon_t] = \beta_0 + t\beta_t + E[\varepsilon_t] = \beta_0 + t\beta_t$$

and hence, the expected value of y_t is a function of time so we violate condition ii) above and hence, this is a non-stationary series. This model is called "trends stationary" because the model can be made stationary by "de-trending."

Example 3: An AR(1) process.

The series for y_t is autocorrelated – which means that current values are correlated with the past. The process can be very complicated or rather simple. In this case, we will consider the simplest time of autocorrelated process – AR(1) – autocorrelation or order 1 where the variable is only correlated with a one-period lag.

Define the AR(1) process to be

(1) $$y_t = \rho y_{t-1} + \varepsilon_t$$

At the start, we have to make the assumption that $|\rho| < 1$. It will become clear later on why we have to make this assumption. We will maintain many of the original assumptions about the errors, namely that

$E[\varepsilon_t] = 0$ and $Var[\varepsilon_t] = \sigma_\varepsilon^2$ for all t and $\mathrm{cov}[\varepsilon_t, y_{t-1}] = 0$

The autocorrelated process described in (1) and the assumption that $|\rho| < 1$ means that "shocks" to the time series in one period will eventually "die out" in the series. To demonstrate this point, suppose there is some "shock" such that y_t is unusually high. How much of that will persist into the future? Note that the series for time period t+1 is defined as

(2) $$y_{t+1} = \rho y_t + \varepsilon_{t+1}$$

If we make the assumption that $E[y_t] = 0$ and $Var[y_t] = \sigma_y^2$ for all t, then

(3) $$\mathrm{cov}(y_{t+1}, y_t) = E[y_{t+1} y_t] = E[(\rho y_t + \varepsilon_{t+1}) y_t] = E[\rho y_t^2] + E[\varepsilon_{t+1} y_t]$$

Because $E[\rho y_t^2] = \rho\sigma_y^2$ and $E[\varepsilon_{t+1} y_t]$ =0, then

(4) $\operatorname{cov}(y_{t+1}, y_t) = \rho\sigma_y^2$

Note that

(5) $Corr(y_{t+1}, y_t) = \dfrac{\operatorname{cov}(y_{t+1}, y_t)}{(Var(y_{t+1})Var(y_t))^{0.5}} = \dfrac{\rho\sigma_y^2}{\sigma_y\sigma_y} = \rho$

Not consider something how long the shock persists h periods in the future. Write the definition of y_{t+h} as

(6) $y_{t+h} = \rho y_{t+h-1} + \varepsilon_{t+h}$

And note that we can write y_{t+h-1} as

(7) $y_{t+h-1} = \rho y_{t+h-2} + \varepsilon_{t+h-1}$

And substituting this into (6), we get

(8) $y_{t+h} = \rho y_{t+h-1} + \varepsilon_{t+h} = \rho[\rho y_{t+h-2} + \varepsilon_{t+h-1}] + \varepsilon_{t+h} = \rho^2 y_{t+h-2} + \rho\varepsilon_{t+h-1} + \varepsilon_{t+h}$

Doing this again, we know that

(9) $y_{t+h-2} = \rho y_{t+h-3} + \varepsilon_{t+h-2}$

So (8) can be written as

(10)

$$y_{t+h} = \rho^2 y_{t+h-2} + \rho\varepsilon_{t+h-1} + \varepsilon_{t+h} = \rho^2[\rho y_{t+h-3} + \varepsilon_{t+h-2}] + \rho\varepsilon_{t+h-1} + \varepsilon_{t+h} = \rho^3 y_{t+h-3} + \rho^2\varepsilon_{t+h-2} + \rho\varepsilon_{t+h-1} + \varepsilon_{t+h}$$

If we continue to make these substitutions, then we will eventually write (10) as

(11) $y_{t+h} = \rho^h y_t + \rho^{h-1}\varepsilon_{t+1} + \rho^{h-2}\varepsilon_{t+2} + ...\rho\varepsilon_{t+h-1} + \varepsilon_{t+h}$

Looking at the covariance between y_{t+h} and y_t

(12) $\operatorname{cov}(y_{t+h}, y_t) = E(y_{t+h}, y_t) = E[(\rho^h y_t^2 + \rho^{h-1}\varepsilon_{t+1} y_t + \rho^{h-2}\varepsilon_{t+2} y_t + ...\rho\varepsilon_{t+h-1} y_t + \varepsilon_{t+h} y_t]$

Note that $E[\rho^h y_t^2] = \rho^h\sigma_y^2$ and $E[\varepsilon_{t+m} y_t]$ =0 for all m>0, then

(13) $\text{cov}(y_{t+h}, y_t) = \rho^h \sigma_Y^2$

Note from (13) that the covariance is a function of the distance between t and t+h but not t, so criteria iii) for weak stationarity is satisfied. Recognizing the definition of correlation coefficients

(14) $Corr(y_{t+h}, y_t) = \dfrac{\text{cov}(y_{t+h}, y_t)}{(Var(y_{t+h})Var(y_t))^{0.5}} = \dfrac{\rho^h \sigma_y^2}{\sigma_y \sigma_y} = \rho^h$

Example 4: A Highly persistent series – the random walk

Consider the AR(1) process $y_t = \rho y_{t-1} + \varepsilon_t$ and relax the assumption that $|\rho| < 1$. The model then can be re-written as

(15) $y_t = y_{t-1} + \varepsilon_t$

This series is called a random walk and the series has a number of important properties. Stock prices are thought to follow a random walk. Given the highly persistent nature of the series, it is easy to demonstrate the history of y as a function of the errors ε_t. Note that the observation for period t-1 can be written as

(16) $y_{t-1} = y_{t-2} + \varepsilon_{t-1}$

And substituting this into equation (15), we can re-write the equation as

(17) $y_t = y_{t-2} + \varepsilon_{t-1} + \varepsilon_t$

Noting that $y_{t-2} = y_{t-3} + \varepsilon_{t-2}$ we can re-write the equation again as

(16) $y_t = y_{t-3} + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t$

Doing this for all n observations in the series, we can write the series for y_t as

(17) $y_t = y_0 + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \varepsilon_{t-2} + ...\varepsilon_2 + \varepsilon_1$

Note that

(18) $E[y_t] = E[y_0] + E[\varepsilon_t] + E[\varepsilon_{t-1}] + E[\varepsilon_{t-2}] + E[\varepsilon_{t-2}] + ...E[\varepsilon_2] + E[\varepsilon_1]$

and because $E[\varepsilon_t] = 0$ for all t, it must be the case that

(19) $E[y_t] = E[y_0]$

For t≥1. This simple result has power implications about the future predictions of y. Suppose we want data on y_t and we want the prediction of the variable sometime in the future. In one period, we can write

(20) $y_{t+1} = y_t + \varepsilon_{t+1}$

And note that because $E[\varepsilon_{t+1}] = 0$

(21) $E[y_{t+1} \mid y_t] = y_t + E[\varepsilon_{t+1}] = y_t$

Two periods in the future, we know that

(22) $y_{t+2} = y_{t+1} + \varepsilon_{t+2} = y_t + \varepsilon_{t+1} + \varepsilon_{t+2}$

And because $E[\varepsilon_{t+1}] = E[\varepsilon_{t+2}] = 0$

(23) $E[y_{t+2} \mid y_t] = y_t + E[\varepsilon_{t+1}] + E[\varepsilon_{t+2}] = y_t$

Extending this to h periods in the future, it is easy to show that

(24) $y_{t+y} = y_t + \varepsilon_{t+h} + \varepsilon_{t+h-1} + \varepsilon_{t+h-2} + ...\varepsilon_{t+2} + \varepsilon_{t+1}$

And because $E[\varepsilon_{t+h}] = E[\varepsilon_{t+h-1}] = E[\varepsilon_{t+h-2}] = E[\varepsilon_{t+2}] = E[\varepsilon_{t+1}] = 0$ then $E[y_{t+h} \mid y_t] = y_t$

The best prediction for y any time in the future is today's y.

Another important property of the series is the variance. Note that $Var[\varepsilon_t] = \sigma_\varepsilon^2$ and we will assume that the errors are not autocorrelated where $\text{cov}[\varepsilon_t, \varepsilon_{t-h}] = 0$ for all h>0. Because y_t is a linear combination of independent random variables and y_o is a fixed parameter, $Var[y_o] = 0$. Therefore, given (24)
(25) $y_t = y_0 + \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \varepsilon_{t-2} + ...\varepsilon_2 + \varepsilon_1$, it is the case that

(26) $Var(y_t) = Var(y_0) + Var(\varepsilon_t) + Var(\varepsilon_{t-1}) + Var(\varepsilon_{t-2}) + Var(\varepsilon_{t-2}) + ...Var(\varepsilon_2) + Var(\varepsilon_1)$

Because the series is t periods long (1,2,3…t) and $Var[\varepsilon_t] = \sigma_\varepsilon^2$ for all t and $Var[y_o] = 0$

(27) $Var(y_t) = t\sigma_\varepsilon^2$

Note then that $\lim_{t\to\infty} Var(y_t) = t\sigma_\varepsilon^2 = \infty$. So this is a non-stationary series.

Testing for unit roots

Testing for a random walk is a little difficult. It is tempting to simply run a regression of y on its lag, which is a pretty good approximation. However, most of the statistics associated with OLS models assume the models are stationary and therefore, as we approach a non-stationary model, the typical standard tests we would calculate are now no longer valid. Therefore, a whole set of other statistical models have been produced to test whether a model is a random walk or not.

Start with the basic AR model

(28) $y_t = \alpha + \rho y_{t-1} + \varepsilon_t$

Which is stationary so long as $|\rho| < 1$. We are interested in testing the null $H_0 : \rho = 1$ against the alternative $H_a : \rho < 1$. To test this, subtract y_{t-1} from both sides

(29) $y_t - y_{t-1} = \Delta y_t = \alpha + \rho y_{t-1} - y_{t-1} + \varepsilon_t = \alpha + (\rho - 1) y_{t-1} + \varepsilon_t = \alpha + \theta y_{t-1} + \varepsilon_t$

Note that if $\rho = 1,$ then the coefficient on θ will equal 0. The transformation of the model from (28) to (29) allows us to proceed with a hull $H_0 : \theta = 0$ against the null $H_0 : \theta < 0$. The null is now that the model is non-stationary and if we cannot reject the null, we cannot evidence that the model is a stationary process.

We can construct the standard t-tests on $\hat{\theta}$ but this is no longer normally distributed in large samples. Dickey and Fuller (1979) have demonstrated what the distribution of this test statistic looks like under the null $H_0 : \rho = 1$ or $H_0 : \theta = 0$.

Using the DJIA data from class, we first regress ln closing prices on a lag. Note that the coefficient on the lag is $\hat{\rho} =$ 0.9999874. When we transform the model, we get a coefficient on $\hat{\theta} =$ -0.0000126. The estimate t-statistic on this is -0.19. To get the critical values for the null, we ask for them by typing

```
dfuller ln_close
```

which produces the same results we have just constructed and the critical values for the one-tailed test $H_0 : \theta = 0$. The 5% critical value is -2.86 so we cannot reject the null the data is a non-stationary.

```
. * test for random walk
. * run a regression of change ln(closing price)
. * on one period lag
. reg ln_close ln_close_1

      Source |       SS       df       MS              Number of obs =   14361
-------------+------------------------------           F(  1, 14359) =       .
       Model |  18001.7298     1  18001.7298           Prob > F      =  0.0000
    Residual |  1.18771395 14359  .000082716           R-squared     =  0.9999
-------------+------------------------------           Adj R-squared =  0.9999
       Total |  18002.9175 14360  1.25368507           Root MSE      =  .00909

------------------------------------------------------------------------------
    ln_close |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  ln_close_1 |   .9999874   .0000678        .   0.000     .9998545     1.00012
       _cons |   .0003808   .0004995     0.76   0.446    -.0005984      .00136
------------------------------------------------------------------------------

. test ln_close_1==1

 ( 1)  ln_close_1 = 1

       F(  1, 14359) =    0.03
            Prob > F =    0.8520

.
.
. * now run model where null is transformed into 0
. reg d_ln_close ln_close_1

      Source |       SS       df       MS              Number of obs =   14361
-------------+------------------------------           F(  1, 14359) =    0.03
       Model |  2.8777e-06     1  2.8777e-06           Prob > F      =  0.8520
    Residual |  1.18771395 14359  .000082716           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0001
       Total |  1.18771682 14360   .00008271           Root MSE      =  .00909

------------------------------------------------------------------------------
  d_ln_close |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  ln_close_1 |  -.0000126   .0000678    -0.19   0.852    -.0001455    .0001202
       _cons |   .0003808   .0004995     0.76   0.446    -.0005984      .00136
------------------------------------------------------------------------------
```

```
.
. * now get dickey fuller test
. dfuller ln_close

Dickey-Fuller test for unit root                   Number of obs   =     14361

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -0.187            -3.430            -2.860            -2.570
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.9401

.
.
. * get the lag of the 1st difference
. gen d_ln_close_1=d_ln_close[_n-1]
(2 missing values generated)

.
. * run a regression of the 1st difference on its lag
. reg d_ln_close d_ln_close_1

      Source |       SS       df       MS              Number of obs =   14360
-------------+------------------------------           F(  1, 14358) =   69.43
       Model |  .005716048     1  .005716048           Prob > F      =  0.0000
    Residual |  1.18199835 14358  .000082323           R-squared     =  0.0048
-------------+------------------------------           Adj R-squared =  0.0047
       Total |   1.1877144 14359  .000082716           Root MSE      =  .00907

------------------------------------------------------------------------------
  d_ln_close |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
d_ln_close_1 |   .0693733   .0083254     8.33   0.000     .0530544    .0856921
       _cons |   .0002686   .0000758     3.55   0.000     .0001201    .0004171
------------------------------------------------------------------------------
```