Econometrics Lecture 7

EC2METRIE

Dr. Anna Salomons

Utrecht School of Economics (U.S.E.)

9 January 2017

This class

- **Time series models**
  - Dynamic models: further interpretation
  - Exogeneity
  - Spurious regression due to non-stationarity
- **Studenmund**
  - Chapter 11 (Time-Series Models) excluding section 11.3 (Granger causality)
  - Note: slides contain additional material not covered in Studenmund

Distributed lag model

In week 6, we saw a particular type of time-series model called the **distributed lag model**

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + .. + \beta_{p+1} X_{t-p} + \varepsilon_t$$

- This model allows the effect of X on Y to be spread out (i.e. distributed) over time.
- Our example was the effect of safety training on workplace accidents.

Distributed lag model: problems

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + .. + \beta_{p+1} X_{t-p} + \varepsilon_t$$

However, in practice, this model is **often problematic** because many lags of X may need to be included:

- The lagged independent variables are often strongly correlated, leading to **multicollinearity** (i.e. less significant estimates).
- The lagged independent variables take up degrees of freedom, **decreasing the precision** of our estimates (i.e. less significant estimates).

Alternative to distributed lag model: autoregressive model

- However, we can show that **instead** of including many lags of the independent variables (i.e. estimating a distributed lag model), **we can estimate a model with a lagged dependent variable** (i.e. an autoregressive model)

- This is because we can **rewrite the autoregressive (AR) model as a particular distributed lag (DL) model**

Rewriting the AR model as a DL model

Consider the simplest AR model:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \varepsilon_t$$

Let's rename the coefficient on the lagged variable λ, following Studenmund's notation:

$$Y_t = \beta_0 + \beta_1 X_t + \lambda Y_{t-1} + \varepsilon_t \tag{1}$$

Then write the equation for $t - 1$

$$Y_{t-1} = \beta_0 + \beta_1 X_{t-1} + \lambda Y_{t-2} + \varepsilon_{t-1} \tag{2}$$

Rewriting the AR model as a DL model

Substitute equation (2) into equation (1) to get:

$$Y_t = \beta_0 + \beta_1 X_t + \lambda \left(\beta_0 + \beta_1 X_{t-1} + \lambda Y_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t$$

$$Y_t = \beta_0 + \lambda\beta_0 + \beta_1 X_t + \lambda\beta_1 X_{t-1} + \lambda^2 Y_{t-2} + \varepsilon_t + \lambda\varepsilon_{t-1}$$

Writing the new intercept as β_0^* and the new error term as ε_t^*:

$$Y_t = \beta_0^* + \beta_0 X_t + \lambda\beta_0 X_{t-1} + \lambda^2 Y_{t-2} + \varepsilon_t^*$$

Rewriting the AR model as a DL model

$$Y_t = \beta_0^* + \beta_0 X_t + \lambda \beta_0 X_{t-1} + \lambda^2 Y_{t-2} + \varepsilon_t^*$$

We can repeat this process, using the expression for Y_{t-2} :

$$\begin{aligned} Y_{t-2} &= \beta_0 + \beta_1 X_{t-2} + \lambda Y_{t-3} + \varepsilon_{t-2} \\ Y_t &= \left\{ \begin{array}{c} \beta_0^* + \beta_1 X_t + \lambda \beta_1 X_{t-1} \\ +\lambda^2 \left(\beta_0 + \beta_1 X_{t-2} + \lambda Y_{t-3} + \varepsilon_{t-2} \right) + \varepsilon_t^* \end{array} \right\} \\ Y_t &= \widetilde{\beta}_0 + \beta_1 X_t + \lambda \beta_1 X_{t-1} + \lambda^2 \beta_1 X_{t-2} + \lambda^3 Y_{t-3} + \widetilde{\varepsilon}_t \end{aligned}$$

With the new intercept $\widetilde{\beta}_0$ and the new error term $\widetilde{\varepsilon}_t$.

(We can of course keep repeating this process, using the expression for Y_{t-3}, then Y_{t-4}, etcetera.)

Rewriting the AR model as a DL model

DL model:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + .. + \beta_{p+1} X_{t-p} + \varepsilon_t$$

Rewritten AR model:

$$Y_t = \widetilde{\beta}_0 + \beta_1 X_t + \lambda\beta_1 X_{t-1} + \lambda^2 \beta_1 X_{t-2} + \lambda^3 Y_{t-3} + \widetilde{\varepsilon}_t$$

If $0 < \lambda < 1$:

- The effect of X on Y is smaller for larger lags: $\beta_1 > \lambda\beta_1 > \lambda^2\beta_1$
- The speed of decline of the effect of X on Y depends on the estimated λ: the closer λ is to 1, the longer the effect of X lasts. If $\lambda = 0$, X only has a contemporaneous effect on Y (i.e. we are back in a static time series model).

Comparing the AR and DL models

DL model:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + .. + \beta_{p+1} X_{t-p} + \varepsilon_t$$

AR model:

$$Y_t = \beta_0 + \beta_1 X_t + \lambda Y_{t-1} + \varepsilon_t$$

Which do we **prefer** estimating (when we expect the effect of X on Y to diminish over time[1])? Usually **the AR model**, for 3 reasons:

- The DL model has stronger multicollinearity;
- The DL model requires the estimation of more parameters and also has a lower number of observations (due to the lags);
- The DL model typically has more serial correlation.

[1] Note that if we do not expect the effect to dimish monotonically over time, the AR model is not a substitute for the DL model.

Example: comparing the AR and DL models

Let's go back to our political economy example from week 6: testing **whether the economy matters for politicians' approval rates**.

We use time series data from the UK, 1979-1996 (the Thatcher-Major era).

Example: comparing the AR and DL models

```
variable name   storage  display    value
                type     format     label      variable label

date            float    %tm                   Date, monthly from July 1979 until December 1996
econft          float    %9.0g                 Evaluation of the economy
pmsat           float    %9.0g                 Satisfaction with prime minister

Sorted by:  econft

. tsset date
        time variable:  date, 1979m7 to 1996m12
                delta:  1 month
```

Example: comparing the AR and DL models

We will estimate and compare the following models:

- **Distributed lag model**:

$$\begin{aligned} pmsat_t &= \beta_0 + \beta_1 econ_t + \beta_2 econ_{t-1} + \beta_3 econ_{t-2} \\ &\quad + \beta_4 econ_{t-3} + \beta_5 econ_{t-4} + \varepsilon_t \end{aligned}$$

- **Autoregressive model**:

$$pmsat_t = \beta_0 + \beta_1 econ_t + \beta_2 pmsat_{t-1} + \varepsilon_t$$

Estimate of the DL model

```
.
. reg pmsat econft l.econft l2.econft  l3.econft  l4.econft date

      Source |       SS       df       MS              Number of obs =     206
-------------+------------------------------           F(  6,   199) =   23.42
       Model |  9046.35426     6  1507.72571           Prob > F      =  0.0000
    Residual |  12812.9403   199  64.3866345           R-squared     =  0.4138
-------------+------------------------------           Adj R-squared =  0.3962
       Total |  21859.2945   205  106.630705           Root MSE      =  8.0241

------------------------------------------------------------------------------
       pmsat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      econft |
         --. |   .1178607   .0551748     2.14   0.034     .0090583    .2266631
         L1. |   .0553193    .063971     0.86   0.388    -.0708286    .1814673
         L2. |   .1557988   .0639171     2.44   0.016      .029757    .2818405
         L3. |   .0364383   .0640291     0.57   0.570    -.0898243    .1627008
         L4. |   .0285457   .0550675     0.52   0.605    -.0800449    .1371364
             |
        date |  -.0845654     .00949    -8.91   0.000    -.1032794   -.0658515
       _cons |   69.08139   3.366403    20.52   0.000     62.44299    75.71979
------------------------------------------------------------------------------

. predict uhat_dl, resid
(4 missing values generated)
```

Estimate of the DL model

We see **many insignificant estimates** in the DL model: this is because the lags are strongly correlated with each other.

```
. corr econft l.econft l2.econft l3.econft l4.econft
(obs=206)
```

	econft	L. econft	L2. econft	L3. econft	L4. econft
econft					
--.	1.0000				
L1.	0.7170	1.0000			
L2.	0.5639	0.7181	1.0000		
L3.	0.4662	0.5654	0.7183	1.0000	
L4.	0.4807	0.4680	0.5669	0.7196	1.0000

Test for autocorrelation in the DL model

```
. reg uhat_dl l.uhat_dl econft l.econft l2.econft  l3.econft  l4.econft date

      Source |       SS           df       MS      Number of obs   =       205
-------------+----------------------------------   F(  7,    197)  =    132.84
       Model |  10550.5442         7  1507.2206   Prob > F        =    0.0000
    Residual |  2235.24384       197  11.3464154   R-squared       =    0.8252
-------------+----------------------------------   Adj R-squared   =    0.8190
       Total |  12785.7881       204  62.6754316   Root MSE        =    3.3684

------------------------------------------------------------------------------
     uhat_dl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     uhat_dl |
         L1. |    .908084   .0297806    30.49   0.000     .8493544    .9668137
             |
      econft |
         --. |   .0076261   .0232012     0.33   0.743    -.0381285    .0533807
         L1. |  -.0040111   .0268765    -0.15   0.882    -.0570137    .0489916
         L2. |  -.0010019   .0268329    -0.04   0.970    -.0539185    .0519146
         L3. |   -.000725    .026879    -0.03   0.979    -.0537324    .0522825
         L4. |    .001059    .023153     0.05   0.964    -.0446005    .0467185
             |
        date |   .0001288   .0040074     0.03   0.974    -.0077741    .0080316
       _cons |  -.0613557   1.420912    -0.04   0.966    -2.863506    2.740794
------------------------------------------------------------------------------
```

Estimate of the AR model

```
. reg pmsat econft l.pmsat date

      Source |       SS       df       MS              Number of obs =     209
-------------+------------------------------           F(  3,   205) =  503.61
       Model |  19449.1556     3  6483.05186           Prob > F      =  0.0000
    Residual |  2638.99998   205  12.8731706           R-squared     =  0.8805
-------------+------------------------------           Adj R-squared =  0.8788
       Total |  22088.1556   208  106.193056           Root MSE      =  3.5879

------------------------------------------------------------------------------
       pmsat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      econft |   .0899079   .0177428     5.07   0.000     .0549262    .1248896
             |
       pmsat |
         L1. |   .8642697   .0284818    30.34   0.000     .8081149    .9204245
             |
        date |  -.0134321   .0047323    -2.84   0.005    -.0227622    -.004102
       _cons |   10.33292    2.37503     4.35   0.000     5.650307    15.01554
------------------------------------------------------------------------------

. predict uhat_ar, resid
(1 missing value generated)
```

Test for autocorrelation in the AR model

```
. reg uhat_ar l.uhat_ar econft l.pmsat date

      Source |       SS       df       MS              Number of obs =     208
-------------+------------------------------           F(  4,   203) =    0.01
       Model |  .414404567     4  .103601142           Prob > F      =  0.9999
    Residual |  2616.14155   203  12.8873968           R-squared     =  0.0002
-------------+------------------------------           Adj R-squared = -0.0195
       Total |  2616.55595   207  12.6403669           Root MSE      =  3.5899

------------------------------------------------------------------------------
     uhat_ar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     uhat_ar |
         L1. |  -.0004413    .077622    -0.01   0.995    -.1534901    .1526074
             |
      econft |   .0013244   .0182027     0.07   0.942    -.0345662    .0372151
             |
       pmsat |
         L1. |   .0002581   .0315456     0.01   0.993     -.061941    .0624571
             |
        date |   .0006389   .0048952     0.13   0.896    -.0090131    .0102909
       _cons |  -.2365836   2.554928    -0.09   0.926    -5.274183    4.801016
------------------------------------------------------------------------------
```

Comparison of the DL and AR models

- **Insignificant estimates in the DL model**, no insignificant estimates in the AR model (also note: 3 more observations in the AR model than in the DL model because 1 observation is lost for each lag).
- **Autocorrelation in the DL model**, no autocorrelation in the AR model.
- Hence, we **prefer the AR model**: estimates of this model indicate that:
  - the short-run effect of the economy on approval of the prime minister (cet. par. on a timetrend) is 0.09;
  - the long run effect (cet. par. on a timetrend) is $\frac{0.09}{1-0.86} = 0.64$

Comparing the AR and DL models: sidenote

- Although we usually prefer the AR model over the DR model, this is only appropriate if we expect the effect of X on Y to decline smoothly over time: the DL model is more flexible.

- For example, a DL model with 3 lags allows the contemporanous effect and thrice-lagged effect (i.e. effects of X_t and X_{t-3}) to be smallest, and the effects of the middle two lags (i.e. effects of X_{t-1} and X_{t-2}) to be largest: one example where we found such a non-monotonic effect was the effect of safety training on accidents (see week 6).

This class

Time series models

- Dynamic models: further interpretation
- **Exogeneity - note: not covered in Studenmund**
- Spurious regression due to non-stationarity

Exogeneity

- **Cross-sectional model**:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Recall assumption 3: $Corr(\varepsilon_i, X_i) = 0$, also known as **exogeneity** (since it implies X is exogenous).

- **Time-series model**:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

We now have two versions of the exogeneity assumption: **strict exogeneity** and **weak exogeneity**.

Exogeneity in time-series models

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

- **Strict exogeneity**: the error term ε at time t is uncorrelated with the explanatory variable(s) in *all* time periods.
- **Weak exogeneity**: the error term ε at time t is uncorrelated with the explanatory variable(s) in *the same* time period t.
  - Also called **contemporaneous exogeneity**
- Strict exogeneity is a much stronger assumption than weak exogeneity.

Strict exogeneity

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

- Error term ε in period t is uncorrelated with the explanatory variable(s) **in the past**, i.e.:
  - ε_t uncorrelated to X_{t-1}
  - ε_t uncorrelated to X_{t-2}
  - .. etc.
- Error term ε in period t is uncorrelated with the explanatory variable(s) **in the present**, i.e.:
  - ε_t uncorrelated to X_t
- Error term ε in period t is uncorrelated with the explanatory variable(s) **in the future**, i.e.:
  - ε_t uncorrelated to X_{t+1}
  - ε_t uncorrelated to X_{t+2}
  - .. etc.

Weak exogeneity (or contemporaneous exogeneity)

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

- Error term ε in period t is uncorrelated with the explanatory variable(s) **in the present**, i.e.:
    - ε_t uncorrelated to X_t
- Note: the error term ε in period t is allowed to be correlated with the explanatory variable(s) in the past or future.
- For **consistent estimates of parameters β with OLS** we only need contemporaneous exogeneity.

A note on consistency and unbiasedness

- In this course, we **do not distinguish between consistency and unbiasnedness**: hence, you may say that OLS estimates are unbiased under contemporaneous exogeneity.
- Strictly speaking, consistency means the estimates converge in probability to the true population parameter with increasing sample sizes. This means the sampling distribution of the estimates becomes more and more concentrated on the true population parameter as the sample size increases. As such, a consistent estimator can be though of as asymptotically unbiased.
- On the exam, you do not need to make this distinction.

This class

Time series models

- Dynamic models: further interpretation
- Exogeneity – note: not covered in Studenmund
- **Spurious regression due to non-stationarity**

Spurious regression

- In time series models, we can have a **problem** known as **spurious regression**
- Spurious regression: a **strong statistical relationship between two or more variables that is not driven by an underlying causal relationship**
- Spurious regression essentially means we get "fake results".

Spurious regression

Spurious regression can result from 2 main causes:

- Trending variables (covered in week 6)
- Non-stationary variables (other than trending variables)

Spurious regression due to trending variables

- Economic **time series often have a trend** (i.e. they increase or decrease steadily over time)– e.g. GDP, prices, employment, ...
- Just because **2 series are trending together**, we **can't assume that the relation is causal.**
- We can easily fix this problem by **controlling for the trend.**

Spurious regression due to nonstationary variables

- The **variables of a time-series regression equation need to be stationary** to obtain consistent parameter estimates of β.
- When our variables are **non-stationary**, the regression results are **spurious** (with one important exception, which we will also discuss).
- We will now define non-stationarity, diagnose it, and provide a solution that avoids spurious regression.

Stationarity: general definition

A time-series Y_t is **strictly stationary if its statistical properties are unaffected by a change of time**.

In other words, Y_t is strictly stationary if the distribution of $Y_1, Y_2, .., Y_n$ is the same as the distribution of the variable shifted by some time lag k, $Y_{1+k}, Y_{2+k}, .., Y_{n+k}$; the distribution of the variable does not depend on time t.

However, to avoid spurious regression, we only need a weaker version – covariance stationarity.

Stationarity: weaker definition

A time-series Y_t is **covariance stationary if the following 3 statistical properties are unaffected by a change of time:**

- the mean: i.e. $E(Y_t)$ is constant over time
- the variance: i.e. $Var(Y_t)$ is constant over time
- the covariance: i.e. $Cov(Y_t, Y_{t+k})$ does not depend on time (it only depends on the lag length k)

When one or more of these conditions is violated, a time-series is non-stationary.

Examples of stationarity

- White noise time series
- First-order stable autoregressive time series

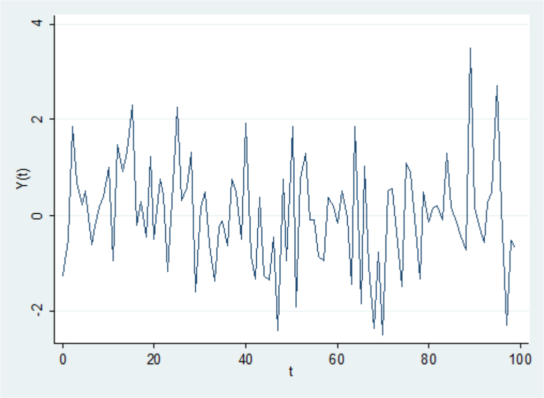
Example of stationarity: white noise

- **White noise** time series is the simplest example of stationary process:

$$Y_t = \varepsilon_t$$

- ε_t is i.i.d. (independently and identically distributed, with a mean of 0 and a variance of σ^2): this means that, no matter what the date is, the distribution of ε_t remains the same.
- Why is this **stationary**?
  - $E(Y_t) = 0$
  - $Var(Y_t) = \sigma^2$
  - $Cov(Y_t, Y_{t+k}) = 0$
- White noise is uninteresting in itself, but it illustrates the concept of stationarity nicely, and is the building block of other models (as we'll see).

Example of stationarity: white noise



$$Y_t = \varepsilon_t \qquad \varepsilon_t \backsim N(0,1)$$

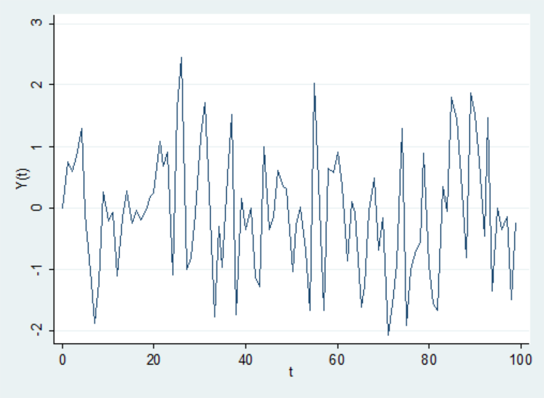
Example of stationarity: stable AR(1) time-series

- Y_t is a stationary variable if it follows a **first-order autoregressive process**, AR(1), as long as $|\rho| < 1$:

$$\begin{aligned} Y_t &= \rho Y_{t-1} + \varepsilon_t \qquad (\varepsilon_t \text{ is i.i.d}) \\ \text{where } |\rho| &< 1 \end{aligned}$$

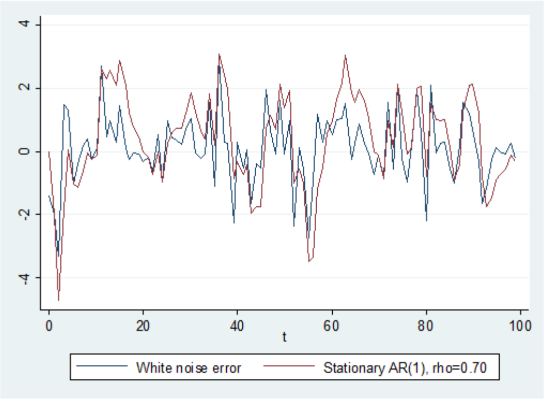
- This is called a **stable AR(1) process:** it means Y_t is modeled as a weighted average of past observations plus a white noise error.
- Stationary processes are also said to be **I(0)**, which stands for integrated of order 0.

Example of stationarity: stable AR(1) time-series



$$Y_t = 0.25Y_{t-1} + \varepsilon_t \qquad \varepsilon_t \backsim N(0,1)$$

Example of stationarity: stable AR(1) time-series



$$Y_t = 0.70 Y_{t-1} + \varepsilon_t \qquad \varepsilon_t \backsim N(0,1)$$

Example of stationarity: stable AR(1) time-series

$$Y_t = \rho Y_{t-1} + \varepsilon_t \qquad \text{where } |\rho| < 1$$

- The model is as follows for the different periods:
  - $Y_t = \rho Y_{t-1} + \varepsilon_t$
  - $Y_{t-1} = \rho Y_{t-2} + \varepsilon_{t-1}$
  - ...
  - $Y_1 = \rho Y_0 + \varepsilon_1$
- We can prove that this is a **stationary time-series** by showing that the mean and variance are constant over time, and that the covariance does not depend on time.

Stable AR(1) time-series: mean

Using repeated substitution, we rewrite Y_t in terms of Y_0 and the error terms

$$\begin{aligned}
Y_t &= \rho Y_{t-1} + \varepsilon_t \\
\text{using that } Y_{t-1} &= \rho Y_{t-2} + \varepsilon_{t-1} \\
Y_t &= \rho \left(\rho Y_{t-2} + \varepsilon_{t-1} \right) + \varepsilon_t \\
&= \rho^2 Y_{t-2} + \rho \varepsilon_{t-1} + \varepsilon_t \\
\\
\text{using that } Y_{t-2} &= \rho Y_{t-3} + \varepsilon_{t-2} \\
Y_t &= \rho^2 \left(\rho Y_{t-3} + \varepsilon_{t-2} \right) + \rho \varepsilon_{t-1} + \varepsilon_t \\
&= \rho^3 Y_{t-3} + \rho^2 \varepsilon_{t-2} + \rho \varepsilon_{t-1} + \varepsilon_t \\
&\dots \\
Y_t &= \rho^t Y_0 + \rho^{t-1} \varepsilon_1 + \rho^{t-2} \varepsilon_2 + .. + \rho \varepsilon_{t-1} + \varepsilon_t
\end{aligned}$$

Stable AR(1) time-series: mean

$$\begin{aligned} Y_t &= \rho^t Y_0 + \rho^{t-1}\varepsilon_1 + \rho^{t-2}\varepsilon_2 + .. + \rho\varepsilon_{t-1} + \varepsilon_t \\ &= \rho^t Y_0 + \sum_{i=0}^{t-1} \rho^i \varepsilon_{t-i} \end{aligned}$$

We now take the expected value to find the mean, $E(Y_t)$

$$\begin{aligned} E(Y_t) &= E(\rho^t Y_0 + \rho^{t-1}\varepsilon_1 + \rho^{t-2}\varepsilon_2 + .. + \rho\varepsilon_{t-1} + \varepsilon_t) \\ &= E(\rho^t Y_0) + E(\rho^{t-1}\varepsilon_1 + \rho^{t-2}\varepsilon_2 + .. + \rho\varepsilon_{t-1} + \varepsilon_t) \\ &= \rho^t E\left(Y_0\right) + \sum_{i=0}^{t-1} \rho^i E\left(\varepsilon_{t-i}\right) \\ &= \rho^t E(Y_0) + 0 \text{ (since the mean of } \varepsilon \text{ is zero for all } t) \\ &= \rho^t E(Y_0) \end{aligned}$$

- If $E(Y_0) = 0$, $E(Y_t) = 0$, hence the mean is constant.

Stable AR(1) time-series: variance

We saw that the mean of a stable AR(1) time-series is constant over time- let's turn to the variance, $Var(Y_t)$

$$
\begin{aligned}
Var(Y_t) &= Var(\rho Y_{t-1} + \varepsilon_t) \qquad \text{where } |\rho| < 1 \\
& \quad \text{using that } Y_{t-1} \text{ and } \varepsilon_t \text{ are uncorrelated} \\
&= Var(\rho Y_{t-1}) + Var(\varepsilon_t) \\
&= \rho^2 Var(Y_{t-1}) + \sigma^2
\end{aligned}
$$

Stable AR(1) time-series: variance

$$\begin{aligned}
Var(Y_t) &= \rho^2 Var(Y_{t-1}) + \sigma^2 \\
\text{Assuming that } Var(Y_t) &= Var(Y_{t-1}) \text{ (i.e. assuming cov. stationarity)} \\
Var(Y_t) &= \rho^2 Var(Y_t) + \sigma^2 \\
(1-\rho^2)\, Var(Y_t) &= \sigma^2 \\
Var(Y_t) &= \frac{\sigma^2}{1-\rho^2}
\end{aligned}$$

- If $|\rho| < 1$: $Var(Y_t)$ is constant.
- If $|\rho| \geq 1$, the derived formula does not apply and the variance is not constant.

Stable AR(1) time-series: covariance

We saw that the mean and the variance of a stable AR(1) time-series are constant over time- lastly, let's turn to its covariance, $Cov(Y_t, Y_{t+k})$. As we showed before we can write Y_t as:

$$Y_t = \rho Y_{t-1} + \varepsilon_t = \rho^t Y_0 + \sum_{i=0}^{t-1} \rho^i \varepsilon_{t-i}$$

Then we can write Y_{t+k} as:

$$\begin{aligned}
Y_{t+k} &= \rho Y_{t+k-1} + \varepsilon_{t+k} \\
&= \rho \left(\rho Y_{t+k-2} + \varepsilon_{t+k-1} \right) + \varepsilon_{t+k} \\
&= \rho^2 Y_{t+k-2} + \rho \varepsilon_{t+k-1} + \varepsilon_{t+k} \\
& \quad \ldots \\
&= \rho^k Y_t + \sum_{i=0}^{k-1} \rho^i \varepsilon_{t+k-i}
\end{aligned}$$

Stable AR(1) time-series: covariance

Now let's consider the covariance between Y_t and Y_{t-k}:

$$Cov\left(Y_t, Y_{t-k}\right) = Cov\left(Y_t, \rho^k Y_t + \sum_{i=0}^{k-1} \rho^i \varepsilon_{t+k-i}\right)$$

Note that $Cov\left(Y_t, \sum_{i=0}^{k-1} \rho^i \varepsilon_{t+k-i}\right) = 0$ since the error term from periods other than t does not covariance with Y from period t :

$$\begin{aligned} Cov\left(Y_t, Y_{t-k}\right) &= Cov\left(Y_t, \rho^k Y_t\right) \\ &= \rho^k Cov\left(Y_t, Y_t\right) \\ &= \rho^k Var\left(Y_t\right) \end{aligned}$$

Stable AR(1) time-series: covariance

We now have that

$$Cov\left(Y_t, Y_{t-k}\right) = \rho^k Var\left(Y_t\right)$$

Using our previously derived expression, $Var\left(Y_t\right) = \frac{\sigma^2}{1-\rho^2}$:

$$Cov\left(Y_t, Y_{t-k}\right) = \frac{\rho^k \sigma^2}{1-\rho^2}$$

From this is can be seen that we need $|\rho| < 1$. And then we do indeed have that $Cov\left(Y_t, Y_{t-k}\right)$ does not depend on time but only on the lag length k.

Examples of non-stationarity

- Trending time series
- Random walk time series

Example of non-stationarity: trending variable

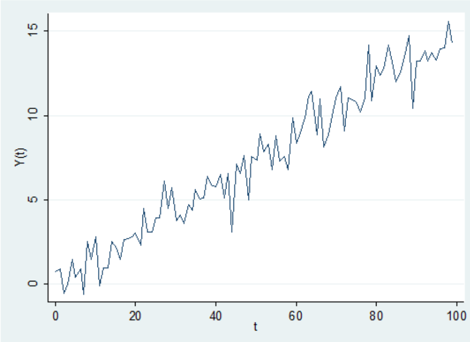
- Y_t is a **trending variable** if:

$$Y_t = \alpha t + \varepsilon_t \qquad (\varepsilon_t \text{ is i.i.d})$$

- When a variable follows a time-trend, it is **non-stationary** because its mean increases ($\alpha > 0$) or decreases ($\alpha < 0$) steadily over time

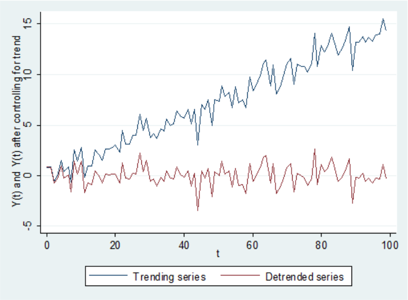
Example of non-stationarity: trending variable

In this picture, the mean increases steadily over time:



Example of non-stationarity: trending variable

Thankfully, we can easily **fix this type of non-stationarity by controlling for a timetrend**:



After controlling for the trend ("detrending"), Y_t has become stationary.

Trending variable: mean

- To see that a trending variable has a mean that is not constant over time, consider the following trending time series variable, Y_t

$$Y_t = \alpha t + \varepsilon_t$$

  Write its mean:

$$\begin{aligned} E(Y_t) &= E(\alpha t + \varepsilon_t) \\ &= E(\alpha t) + E(\varepsilon_t) \\ &= \alpha t + 0 = \alpha t \end{aligned}$$

- Clearly, the mean of this series is dependent on t, i.e. it is non-stationary.

Trending variable: variance

- We can also easily find the variance of a trending time series variable, Y_t

$$\begin{aligned} Var(Y_t) &= E\left[Y_t - E(Y_t)\right]^2 \\ &= E\left[\alpha t + \varepsilon_t - \alpha t\right]^2 \\ &= E\left[\varepsilon_t\right]^2 = 0 \end{aligned}$$

- Hence, the variance of this series is constant over time.
- However, since the mean is not constant over time (see previous slide), Y_t is still a non-stationary time series!

Example of non-stationarity: random walk

- Y_t is a non-stationary variable if it follows a **random walk**

$$\begin{aligned} Y_t &= \rho Y_{t-1} + \varepsilon_t \qquad (\varepsilon_t \text{ is i.i.d}) \\ \text{where } \rho &= 1, \text{ i.e.:} \\ Y_t &= Y_{t-1} + \varepsilon_t \qquad (\varepsilon_t \text{ is i.i.d}) \end{aligned}$$

- If Y_t follows a random walk, then the value of Y tomorrow is the value of Y today, plus an unpredictable (i.i.d) disturbance ε_t.

Example of non-stationarity: random walk

- Y_t is a non-stationary variable if it follows a **random walk**

$$Y_t = Y_{t-1} + \varepsilon_t$$

- A variable that follows a random walk is also said to have a **unit root,** or be **I(1)**, which stands for integrated of order 1.
- This is very important for economic applications since **many macro-economic time series are random walks**!

Example of non-stationarity: random walk

This figure shows 2 random walks with $Y_0 = X_0 = 0$.

Example of non-stationarity: random walk

- The previous figure showed random walks of the following form:

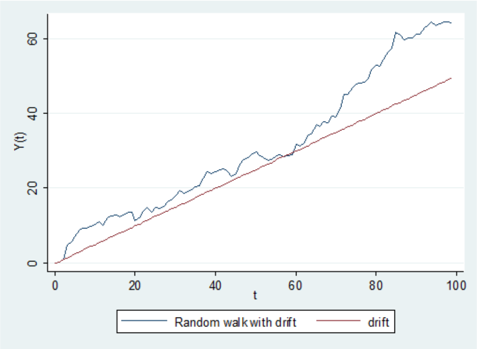
$$Y_t = Y_{t-1} + \varepsilon_t \qquad (\varepsilon_t \text{ is i.i.d})$$

- More generally, we can write:

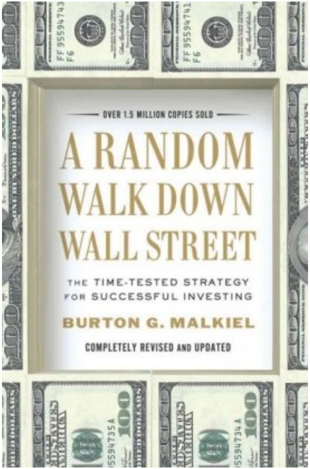
$$Y_t = \alpha + Y_{t-1} + \varepsilon_t \qquad (\varepsilon_t \text{ is i.i.d})$$

- The variable Y_t follows a random walk,
  - Without drift if $\alpha = 0$
  - With drift if $\alpha \neq 0$

Random walk with drift



Economic example of random walk: stock market prices



Random walk: mean

To see that a random walk is non-stationary, we should look at its mean, variance and covariance over time.

- Let's start with the mean:

$$
\begin{aligned}
Y_t &= Y_{t-1} + \varepsilon_t \\
E(Y_t) &= E(Y_{t-1} + \varepsilon_t) \\
\text{using that } Y_{t-1} &= Y_{t-2} + \varepsilon_{t-1} \\
E(Y_t) &= E(Y_{t-2} + \varepsilon_{t-1} + \varepsilon_t) \\
&= E(Y_{t-3} + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t) \\
&\ldots \\
&= E(Y_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_{t-1} + \varepsilon_t) \\
&= E(Y_0) = Y_0
\end{aligned}
$$

- Thus the mean of a random walk is constant over time.

Random walk: variance

The mean of a random walk is constant- however, we will show that this is not true for the variance:

$$\begin{aligned}
Var(Y_t) &= Var(Y_0 + \varepsilon_1 + \varepsilon_2 ... + \varepsilon_{t-1} + \varepsilon_t) \\
&\quad \text{since the } \varepsilon \text{ are independent from each other} \\
&\quad \text{and from } Y_0 \text{ we can write this as} \\
&= Var(Y_0) + Var(\varepsilon_1) + Var(\varepsilon_2) + ... + Var(\varepsilon_{t-1}) + Var(\varepsilon_t) \\
&= Var(\varepsilon_1) + Var(\varepsilon_2) + ... + Var(\varepsilon_{t-1}) + Var(\varepsilon_t)
\end{aligned}$$

$$\begin{aligned}
\text{we know that } Var(\varepsilon_1) &= Var(\varepsilon_2) = .. = Var(\varepsilon_t) = \sigma^2 \\
&\quad \text{hence} \\
Var(Y_t) &= t\sigma^2
\end{aligned}$$

Random walk: variance

$$Var(Y_t) = t\sigma^2$$

- This shows that the variance of Y_t becomes larger and larger over time.
  - Period 1: σ^2; period 2: $2\sigma^2$; period 3: $3\sigma^2$.. etc
- We conclude that the variance of a random walk is not constant over time: a **random walk is non-stationary**.

Consequences of non-stationarity

- **Non-stationarity leads to spurious regression:** we cannot interpret such a regression in a meaningful way.
- We can easily see this from a simple simulation exercise:
  - Independently generate two random walks, $X_t = X_{t-1} + u_t$ and $Y_t = Y_{t-1} + v_t$, and estimate the model

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

  - Since X_t and Y_t are independently generated, there should be no statistically significant relationship between them...

Random walks Y and X

A spurious regression

```
. reg y x

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   42.07
       Model |  193.144207     1  193.144207           Prob > F      =  0.0000
    Residual |  449.967349    98  4.59150356           R-squared     =  0.3003
-------------+------------------------------           Adj R-squared =  0.2932
       Total |  643.111556    99  6.49607632           Root MSE      =  2.1428

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.3743418   .0577172    -6.49   0.000    -.4888797   -.259804
       _cons |    2.38343   .3337904     7.14   0.000     1.721034    3.045826
------------------------------------------------------------------------------
```

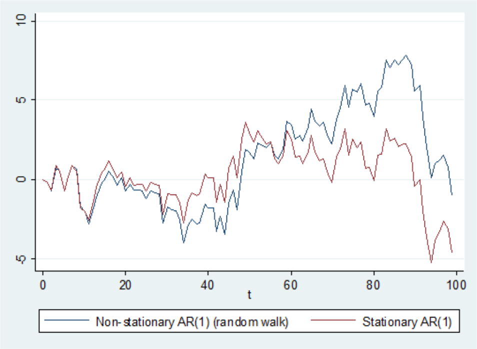
But it looks like there is a significant relationship between X and Y– however, this is **spurious since X and Y are non-stationary variables**!

Non-stationarity

- We have **defined non-stationarity** and seen examples of stationary and non-stationary time series.
- We have seen how **non-stationarity leads to spurious regression.**
- This means we need a way to **diagnose** whether our variables are non-stationary, and if we find that they are, we need a **solution**.

Stationary or non-stationary?

This figure shows that it's hard to tell the difference between a random walk and a stable AR(1) process with rho close to 1 just by looking at the data (especially when there are not many time periods).. we need a formal test!



Diagnosis of non-stationarity

- **Informal diagnosis** (not conclusive!):
  - Breusch-Godfrey test for autocorrelation: likely to have non-stationarity when the estimated ρ is close to 1.
  - Prais-Winsten estimation: likely to have non-stationarity when the estimated ρ is close to 1.
  - (Durbin-Watson statistic for a static model: likely to have non-stationarity when the statistic is close to 0).
- **Formal diagnosis**:
  - **Dickey-Fuller test for unit root** performed separately **for all variables** of the regression equation

Dickey-Fuller test

- We start with an AR(1) model:

$$Y_t = \alpha + \rho Y_{t-1} + \varepsilon_t$$

- The variable Y_t follows a random walk if $\rho = 1$
  - Without drift if $\alpha = 0$, with drift if $\alpha \neq 0$
- To determine whether Y_t is I(1) (i.e. has a unit root), we need a way of **testing whether $\rho = 1$**

Dickey-Fuller test

We rewrite the AR(1) model as follows:

$$
\begin{aligned}
Y_t &= \alpha + \rho Y_{t-1} + \varepsilon_t \\
& \quad \text{subtracting } Y_{t-1} \text{ from both sides} \\
Y_t - Y_{t-1} &= \alpha + \rho Y_{t-1} - Y_{t-1} + \varepsilon_t \\
Y_t - Y_{t-1} &= \alpha + (\rho - 1) Y_{t-1} + \varepsilon_t \\
\Delta Y_t &= \alpha + (\rho - 1) Y_{t-1} + \varepsilon_t \\
\\
\text{denote } \theta &= (\rho - 1) \\
\Delta Y_t &= \alpha + \theta Y_{t-1} + \varepsilon_t
\end{aligned}
$$

Dickey-Fuller test

Our rewritten AR(1) model:

$$\Delta Y_t = \alpha + \theta Y_{t-1} + \varepsilon_t \qquad \text{where } \theta = \rho - 1$$

- ΔY_t is the first difference (which is I(0) even if Y_t is I(1), as we will see later)
- We will test:

$$\begin{aligned} H_0 &: \theta = 0 && \Leftrightarrow \rho = 1 \text{ (unit root)} \\ H_A &: \theta < 0 && \Leftrightarrow \rho < 1 \text{ (no unit root)} \end{aligned}$$

- Problem: t-statistic for $\widehat{\theta}$ does not follow a t-distribution under H_0 since Y_{t-1} is I(1).
- Instead, we use Dickey-Fuller critical values

Dickey-Fuller critical values

Critical values of DF test without timetrend

Signif. level	1%	5%	10%
Critical value	-3.43	-2.89	-2.57

Critical values of DF test with timetrend

Signif. level	1%	5%	10%
Critical value	-3.96	-3.41	-3.12

These will be provided at the exam.

Dickey-Fuller test: outline

1. Determine whether Y_t follows a timetrend by estimating[2]:

$$Y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

2. Estimate

$$\begin{aligned} \Delta Y_t &= \alpha + \theta Y_{t-1} + \varepsilon_t \text{ if no timetrend found} \\ \Delta Y_t &= \alpha + \theta Y_{t-1} + \delta t + \varepsilon_t \text{ if a timetrend found} \end{aligned}$$

$$H_0 : \theta = 0 \quad H_A : \theta < 0$$

3. Compare the t-statistic on $\widehat{\theta}$ to the appropriate critical value from the DF table, DF_c.
4. Reject H_0 if $t < DF_c$, in which case Y_t does not have a unit root.

Repeat this for the independent variable(s) in the regression equation!

[2] Typically, we also (additionally) use visual inspection of the time series to look for evidence of a timetrend.

Testing for non-stationarity: an example

- Let's look at an example of a Dickey-Fuller test (we will use a significance level of 5% throughout).
- Going back to our first political economy example from last week: US election outcomes and the state of the economy.
- We want to estimate the following time-series model:

$$vote_t = \beta_0 + \beta_1 growth_t + \varepsilon_t$$

where the share of the vote captured by the incumbent party is related to economic growth in the election year.

Testing for non-stationarity: an example

Here are the estimates from the model (we saw them in week 6 already):

```
. reg vote growth

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(1, 33)      =   17.22
       Model |  403.196156     1  403.196156           Prob > F      =  0.0002
    Residual |  772.876026    33  23.4204856           R-squared     =  0.3428
-------------+------------------------------           Adj R-squared =  0.3229
       Total |  1176.07218    34  34.5903583           Root MSE      =  4.8395

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      growth |   .6496552   .1565751     4.15   0.000     .3311008    .9682097
       _cons |   51.58451   .8252712    62.51   0.000     49.90548    53.26353
------------------------------------------------------------------------------
```

Testing for non-stationarity: an example

- We estimated this model:

$$vote_t = \beta_0 + \beta_1 growth_t + \varepsilon_t$$

- But, we now know that if $vote_t$ and $growth_t$ have a unit root (i.e. are non-stationary), the estimation results are spurious.

- Hence, we need to apply the Dickey-Fuller test on both $vote_t$ and $growth_t$

Testing for non-stationarity in vote: step 1

Does the variable $vote_t$ follow a time trend? These estimation results suggest not. This means we will apply the DF test without a time-trend.

```
. reg vote time

      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  1,    32) =    0.00
       Model |  .135941673     1  .135941673           Prob > F      =  0.9519
    Residual |  1174.95278    32  36.7172743           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0311
       Total |  1175.08872    33  35.6087491           Root MSE      =  6.0595

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     time_id |   .0064452   .1059242     0.06   0.952    -.2093154    .2222058
       _cons |   51.95339   2.125095    24.45   0.000     47.62471    56.28206
------------------------------------------------------------------------------
```

Testing for non-stationarity in vote: step 2

We now estimate $\Delta vote_t = \alpha + \theta vote_{t-1} + \varepsilon_t$:

```
. reg d.vote l.vote

      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  1,    31) =   56.04
       Model |  1943.36749     1  1943.36749           Prob > F      =  0.0000
    Residual |  1075.10269    31   34.680732           R-squared     =  0.6438
-------------+------------------------------           Adj R-squared =  0.6323
       Total |  3018.47018    32  94.3271932           Root MSE      =   5.889

------------------------------------------------------------------------------
      D.vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        vote |
         L1. |  -1.286681   .1718849    -7.49   0.000    -1.637243   -.9361199
             |
       _cons |   67.05805   9.013585     7.44   0.000     48.67472    85.44138
------------------------------------------------------------------------------
```

Testing for non-stationarity in vote: steps 3 & 4

Critical values of DF test (no tt)

Signif. level	1%	5%	10%
Critical value	-3.43	-2.89	-2.57

- We find a t-stat of -7.49, which we should compare with a critical value from the DF table (without timetrend).
- Since $-7.49 < -2.89$, we reject H_0.
- Conclusion: the variable *vote* does not contain a unit root, it is **stationary.**
- However, we still need to **perform the same test for the independent variable**, *growth*, before we can conclude that our estimated model is not spurious!

Testing for non-stationarity in growth: step 1

Does the variable $growth_t$ follow a time trend? These estimation results suggest not (using $\alpha = 0.05$). This means we will apply the DF test without a time-trend.

```
. reg growth time

      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  1,     32) =    3.31
       Model |  89.398334     1   89.398334           Prob > F      =  0.0783
    Residual |  864.673245    32  27.0210389           R-squared     =  0.0937
-------------+------------------------------           Adj R-squared =  0.0654
       Total |  954.071579    33    28.91126           Root MSE      =  5.1982

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     time_id |   .1652817    .090868     1.82   0.078    -.0198104    .3503739
       _cons |  -2.227695   1.823031    -1.22   0.231    -5.941089    1.485698
------------------------------------------------------------------------------
```

Testing for non-stationarity in growth: step 2

We now estimate $\Delta growth_t = \alpha + \theta growth_{t-1} + \varepsilon_t$:

```
. reg d.growth l.growth

      Source |       SS       df       MS              Number of obs =      33
-------------+------------------------------           F(  1,    31) =   40.19
       Model |  1203.85833     1  1203.85833           Prob > F      =  0.0000
    Residual |  928.693982    31  29.9578704           R-squared     =  0.5645
-------------+------------------------------           Adj R-squared =  0.5505
       Total |  2132.55231    32  66.6422597           Root MSE      =  5.4734

------------------------------------------------------------------------------
    D.growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      growth |
         L1. |  -1.124387   .1773714    -6.34   0.000    -1.486138   -.7626355
             |
       _cons |   .6449846   .9592054     0.67   0.506    -1.311328    2.601297
------------------------------------------------------------------------------
```

Testing for non-stationarity in growth: steps 3 & 4

Critical values of DF test (no tt)

Signif. level	1%	5%	10%
Critical value	-3.43	-2.89	-2.57

- We find a t-stat of -6.34, which we should compare with a critical value from the DF table (without timetrend).
- Since $-6.34 < -2.89$, we reject H_0.
- Conclusion: the variable *growth* does not contain a unit root, it is **stationary.**

Testing for non-stationarity: an example

- We established that both $vote_t$ and $growth_t$ are stationary variables.
- This means that estimates of the following model are **not spurious** (i.e. we can interpret them):

$$vote_t = \beta_0 + \beta_1 growth_t + \varepsilon_t$$

Testing for non-stationarity: another example

- In last week's tutorial, we related **log military employment** to **log real defense spending** in the following model:

$$lmilemp_t = \beta_0 + \beta_1 lrdefs_t + \beta_2 t + \varepsilon_t$$

- Let's **test whether these results are spurious** by checking whether $lmilemp_t$ and $lrdefs_t$ have unit roots or not.

- We will first consider the informal evidence, and then turn to the formal DF test.

Non-stationarity: informal evidence from BG autocorrelation test

```
. reg lmilemp lrdefs t

      Source |       SS       df       MS              Number of obs =     306
-------------+------------------------------           F(  2,   303) = 2293.87
       Model |  9.31914466     2  4.65957233           Prob > F      =  0.0000
    Residual |  .615487553   303  .002031312           R-squared     =  0.9380
-------------+------------------------------           Adj R-squared =  0.9376
       Total |  9.93463222   305  .032572565           Root MSE      =  .04507

------------------------------------------------------------------------------
     lmilemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lrdefs |   5.629571   .1329232    42.35   0.000     5.368002    5.891141
           t |  -.0017988   .0000298   -60.38   0.000    -.0018575   -.0017402
       _cons |   5.178614   .0671496    77.12   0.000     5.046475    5.310752
------------------------------------------------------------------------------

. predict uhat, resid
(6 missing values generated)

. reg uhat l.uhat lrdefs t

      Source |       SS       df       MS              Number of obs =     305
-------------+------------------------------           F(  3,   301) = 3509.68
       Model |  .57921549     3   .19307183           Prob > F      =  0.0000
    Residual |  .01655836   301  .000055011           R-squared     =  0.9722
-------------+------------------------------           Adj R-squared =  0.9719
       Total |  .59577385   304  .001959782           Root MSE      =  .00742

------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        uhat |
         L1. |   .9699198    .009454   102.59   0.000     .9513154    .9885241
             |
      lrdefs |   .0057768    .021877     0.26   0.792    -.0372744    .0488281
           t |  -9.35e-06   4.93e-06    -1.90   0.059     -.000019    3.49e-07
       _cons |  -.0010425   .0110507    -0.09   0.925    -.0227889    .0207039
------------------------------------------------------------------------------
```

Non-stationarity: informal evidence from Prais-Winsten estimation

```
Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =     306
-------------+------------------------------           F(  2,   303) = 2585.69
       Model |  .18337247     2  .091686235            Prob > F      =  0.0000
    Residual |  .01074411   303  .000035459            R-squared     =  0.9447
-------------+------------------------------           Adj R-squared =  0.9443
       Total |  .19411658   305  .000636448            Root MSE      =  .00595

------------------------------------------------------------------------------
     lmilemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lrdefs |   1.685521   .2902187     5.81   0.000     1.114422    2.256621
           t |  -.0009674    .000306    -3.16   0.002    -.0015696   -.0003653
       _cons |   7.048933     .18197    38.74   0.000     6.690848    7.407018
-------------+----------------------------------------------------------------
         rho |   .9984566
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    0.028235
Durbin-Watson statistic (transformed) 0.659513
```

Non-stationarity: informal evidence

- We see that the **autocorrelation coefficient appears to be 1**- this indicates non-stationarity.
  - Intuitively: if $\varepsilon_t = \varepsilon_{t-1} + u_t$, shocks from the past never die out, which means the series does not obtain an equilibrium - it is a random walk.
- However, we could just have a stationary AR(1) process with very high autocorrelation coefficients (i.e. close to 1, but not 1).
- Therefore, we need to **use the formal DF test**.

Testing for non-stationarity in lmilemp: step 1

The variable $lmilemp_t$ follows a time trend- hence we will apply the DF test with a time-trend.

```
. reg lmilemp t

      Source |       SS       df       MS              Number of obs =     307
-------------+------------------------------           F(  1,   305) =  405.25
       Model |  5.67667381     1  5.67667381           Prob > F      =  0.0000
    Residual |  4.27239674   305  .014007858           R-squared     =  0.5706
-------------+------------------------------           Adj R-squared =  0.5692
       Total |  9.94907054   306  .032513302           Root MSE      =  .11835

------------------------------------------------------------------------------
     lmilemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.0015344   .0000762   -20.13   0.000    -.0016844   -.0013844
       _cons |   8.012511   .0135428   591.64   0.000     7.985862     8.03916
------------------------------------------------------------------------------
```

Testing for non-stationarity in lmilemp: step 2

We now estimate $\Delta lmilemp_t = \alpha + \theta lmilemp_{t-1} + \delta t + \varepsilon_t$:

```
. reg d.lmilemp l.lmilemp t

      Source |       SS           df       MS      Number of obs   =       306
-------------+----------------------------------   F(2, 303)       =      0.10
       Model |  8.0761e-06         2  4.0381e-06   Prob > F        =    0.9027
    Residual |   .0119543        303  .000039453   R-squared       =    0.0007
-------------+----------------------------------   Adj R-squared   =   -0.0059
       Total |  .011962376       305  .000039221   Root MSE        =    .00628

------------------------------------------------------------------------------
   D.lmilemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lmilemp |
         L1. |  -.0013762   .0030437    -0.45   0.651    -.0073656    .0046132
             |
           t |  -2.06e-06   6.21e-06    -0.33   0.741    -.0000143    .0000102
       _cons |   .0102389   .0244051     0.42   0.675    -.0377861    .0582639
------------------------------------------------------------------------------
```

Testing for non-stationarity in lmilemp: step 3 & 4

Critical values of DF test (with tt)

Signif. level	1%	5%	10%
Critical value	-3.96	-3.41	-3.12

- We find a t-stat of -0.45, which we should compare with a critical value from the DF table (with timetrend).
- Since $-0.45 > -3.41$, we do not reject H_0.
- Conclusion: the variable $lmilemp_t$ contains a unit root, it is **non-stationary.**

Testing for non-stationarity in lrdefs: step 1

The variable $lrdefs_t$ follows a time trend- hence we will apply the DF test with a time-trend.

```
. reg lrdefs t

      Source |       SS       df       MS              Number of obs =     307
-------------+------------------------------           F(  1,   305) =   14.09
       Model |  .005331851     1  .005331851           Prob > F      =  0.0002
    Residual |  .115419171   305  .000378424           R-squared     =  0.0442
-------------+------------------------------           Adj R-squared =  0.0410
       Total |  .120751022   306  .000394611           Root MSE      =  .01945

------------------------------------------------------------------------------
      lrdefs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |    .000047   .0000125     3.75   0.000     .0000224    .0000717
       _cons |   .5035234   .0022368   225.11   0.000     .4991219    .5079249
------------------------------------------------------------------------------
```

Testing for non-stationarity in lrdefs: step 2

We now estimate $\Delta lrdefs_t = \alpha + \theta lrdefs_{t-1} + \delta t + \varepsilon_t$:

```
. reg d.lrdefs l.lrdefs t

      Source |       SS       df       MS              Number of obs =     306
-------------+------------------------------           F(  2,   303) =    2.50
       Model |  6.8575e-06     2  3.4288e-06           Prob > F      =  0.0834
    Residual |  .000414763   303  1.3689e-06           R-squared     =  0.0163
-------------+------------------------------           Adj R-squared =  0.0098
       Total |  .000421621   305  1.3824e-06           Root MSE      =  .00117

------------------------------------------------------------------------------
    D.lrdefs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lrdefs |
         L1. |   .0000393   .0034506     0.01   0.991    -.0067509    .0068294
             |
           t |   1.69e-06   7.73e-07     2.19   0.029     1.71e-07    3.21e-06
       _cons |  -.0001844    .001743    -0.11   0.916    -.0036144    .0032456
------------------------------------------------------------------------------
```

Testing for non-stationarity in lrdefs: step 3 & 4

Critical values of DF test (with tt)

Signif. level	1%	5%	10%
Critical value	-3.96	-3.41	-3.12

- We find a t-stat of 0.01, which we should compare with a critical value from the DF table (with timetrend).
- Since $0.01 > -3.41$, we do not reject H_0.
- Conclusion: the variable $lrdefs_t$ contains a unit root, it is **non-stationary.**

Testing for non-stationarity: another example

- In last week's tutorial, we related log military employment to log real defense spending in the following model:

$$lmilemp_t = \beta_0 + \beta_1 lrdefs_t + \beta_2 t + \varepsilon_t$$

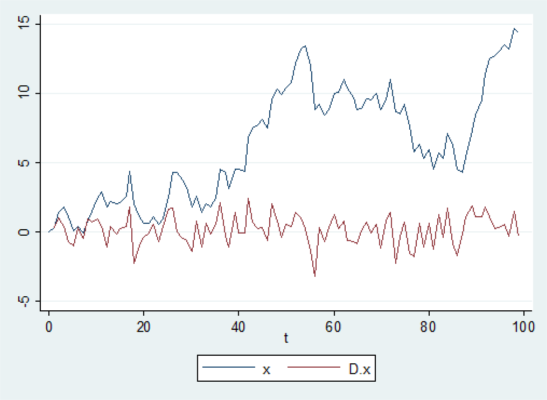
- However, estimation results are **spurious** since **lmilemp$_t$ and lrdefs$_t$ have unit roots**.
- (Note that results would also be spurious if only one of the two variables had a unit root)
- We now turn to **solving the problem of non-stationarity**.

Solution to non-stationarity

- When non-stationarity is caused by a trend, we can just control for the trend (see slides from last week).
- However, when the variable contains a unit root, we need a different solution: **first-differencing**.
- First-differencing a variable that contains a unit root often **makes it stationary**.

First-differencing illustrated

Consider how first-differencing a random walk makes it stationary:



First-differencing: our example

- We found that **both the log of military employment and the log of real defense spending are non-stationary**.
- This means we should **first-difference them both**, and estimate the following model (note that the timetrend in the original model becomes a constant- see last week's tutorial):

$$\Delta lmilemp_t = \beta_0 + \beta_1 \Delta lrdefs_t + \varepsilon_t$$

- We can now again estimate β_1, the elasticity of military employment to real defense spending, and this estimate is **not spurious**.

First-differenced estimates

```
. reg d.lmilemp d.lrdefs

      Source |       SS           df       MS      Number of obs   =       305
-------------+----------------------------------   F(  1,    303)  =     33.62
       Model |  .001192997         1  .001192997   Prob > F        =    0.0000
    Residual |  .010752309       303  .000035486   R-squared       =    0.0999
-------------+----------------------------------   Adj R-squared   =    0.0969
       Total |  .011945306       304  .000039294   Root MSE        =    .00596

------------------------------------------------------------------------------
   D.lmilemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lrdefs |
         D1. |    1.68285   .2902387     5.80   0.000     1.111711    2.253989
             |
       _cons |  -.0009573   .0003423    -2.80   0.005    -.0016308   -.0002838
------------------------------------------------------------------------------
```

An overview of non-stationarity – so far..

- Whenever we estimate time-series regressions, we should check whether our variables are non-stationary. The consequence of non-stationarity is **spurious regression.**

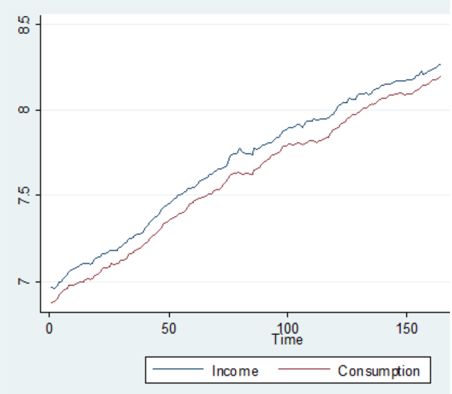
- **Diagnosis** of non-stationarity can be done with a **Dickey-Fuller test**.

- The **solution is first-differencing** the non-stationarity variable(s).
    - BUT: we are missing **one important exception** here, **when we do not need to first-difference.**

The exception: cointegration

- If both the dependent and the independent variable are non-stationary, the **error term may still be stationary**.
- If this is the case, the **dependent and independent variable** are said to be **cointegrated**.
- This is good news, since when X and Y are cointegrated, the original regression results are **not spurious** and we **do not need to first-difference**.

An economic example of cointegration: income and consumption



An economic example of cointegration: income and consumption

- Log consumption and log income are **each non-stationary**: these time-series are "wandering".
- However, **they are wandering together**– the difference between these two series does not get arbitrarily large (in contrast to our 2 independently generated random walks we saw before!).
- This is because income and consumption have an equilibrium relationship– they cannot drift too far apart from because economic forces will act to restore the equilibrium relationship.
- This means the association between them is **not spurious** even though both variables are non-stationary!.

Cointegration: diagnosis

- Visual inspection is not a good way to determine whether time series are cointegrated– we need to use the **Dickey-Fuller cointegration test**.

- This test examines **whether the error term is stationary**– this should be the case if the two series are cointegrated since it indicates the series do no wander apart.

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_t \ (+\beta_2 t) \ + \varepsilon_t \\ \varepsilon_t &= Y_t - \beta_0 - \beta_1 X_t \ (-\beta_2 t) \end{aligned}$$

- The test uses the **residuals** e_t (as errors are unobserved).

The Dickey-Fuller cointegration test

This test should only be **performed on 2 variables that have been found to be non-stationary**!

1. Estimate the relationship between Y_t and X_t (include a timetrend if $\widehat{\beta}_2$ is significant):

$$Y_t = \beta_0 + \beta_1 X_t \ (+\beta_2 t) \ + \varepsilon_t$$

2. Generate the residual, e_t
3. Regress the differenced residual onto the lagged residual:

$$\Delta e_t = \gamma_0 + \gamma_1 e_{t-1} + u_t$$

The Dickey-Fuller cointegration test

4. Compare the t-statistic to the critical value from the DF cointegration (DFC) table (with or without a timetrend, depending on the model from step 1), to test:

$$
\begin{aligned}
H_0 &: \gamma_1 = 0 \quad \text{no cointegration} \\
H_A &: \gamma_1 < 0 \quad \text{cointegration}
\end{aligned}
$$

5. Reject H_0 if $t < DFC_c$: if H_0 is rejected, we conclude Y_t and X_t are cointegrated.

Critical values of DF test of cointegration

Critical values of DF cointegration test
(residual from model without timetrend)

Signif. level	1%	5%	10%
Critical value	-3.90	-3.34	-3.04

Critical values of DF cointegration test
(residual from model with timetrend)

Signif. level	1%	5%	10%
Critical value	-4.33	-3.78	-3.50

These will be provided at the exam.

Cointegration and model specification

- If we **reject** H_0 (i.e. we find Y_t and X_t to be cointegrated), we can estimate the model **in levels**:

$$Y_t = \beta_0 + \beta_1 X_t \ (+\beta_2 t) \ + \varepsilon_t$$

- If however we **cannot reject** H_0, we have to **first-difference** the non-stationary variables Y_t and X_t, i.e. use ΔY_t and ΔX_t.

Testing for cointegration: our example

- Remember we found that both the log of military employment and the log of real defense spending are non-stationary.
- This means we should take first differences of these two variables– unless they are cointegrated!
- So let's **test for cointegration** to determine whether the equation should be estimated in levels (i.e. in original units) or in first differences.

Testing for cointegration: step 1 & 2

```
. reg lmilemp lrdefs t

      Source |       SS       df       MS              Number of obs =     306
-------------+------------------------------           F(  2,   303) = 2293.87
       Model |  9.31914466     2  4.65957233           Prob > F      =  0.0000
    Residual |  .615487553   303  .002031312           R-squared     =  0.9380
-------------+------------------------------           Adj R-squared =  0.9376
       Total |  9.93463222   305  .032572565           Root MSE      =  .04507

------------------------------------------------------------------------------
     lmilemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lrdefs |   5.629571   .1329232    42.35   0.000     5.368002    5.891141
           t |  -.0017988   .0000298   -60.38   0.000    -.0018575   -.0017402
       _cons |   5.178614   .0671496    77.12   0.000     5.046475    5.310752
------------------------------------------------------------------------------

. predict uhat, resid
(6 missing values generated)
```

Testing for cointegration: step 3

```
. reg d.uhat l.uhat

      Source |       SS       df       MS              Number of obs =     305
-------------+------------------------------           F(  1,   303) =   10.07
       Model |  .000556898     1  .000556898           Prob > F      =  0.0017
    Residual |  .016757188   303  .000055304           R-squared     =  0.0322
-------------+------------------------------           Adj R-squared =  0.0290
       Total |  .017314086   304  .000056954           Root MSE      =  .00744

------------------------------------------------------------------------------
      D.uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        uhat |
         L1. |   -.03008   .0094792    -3.17   0.002    -.0487333   -.0114267
             |
       _cons |   .0004591   .0004258     1.08   0.282    -.0003789    .001297
------------------------------------------------------------------------------
```

Testing for cointegration: step 4 & 5

Critical values of DF cointegration test

(residual from model with timetrend)

Signif. level	1%	5%	10%
Critical value	-4.33	-3.78	-3.50

- The found t-stat is -3.17: since -3.17>-3.78, we do not reject H_0.
- We conclude that $lmilemp_t$ and $lrdefs_t$ are **not cointegrated**.
- This means we should use first differences, $\Delta lmilemp_t$ and $\Delta lrdefs_t$.

Estimation with first differences

```
. reg d.lmilemp d.lrdefs

      Source |       SS       df       MS              Number of obs =     305
-------------+------------------------------           F(  1,   303) =   33.62
       Model |  .001192997     1  .001192997           Prob > F      =  0.0000
    Residual |  .010752309   303  .000035486           R-squared     =  0.0999
-------------+------------------------------           Adj R-squared =  0.0969
       Total |  .011945306   304  .000039294           Root MSE      =  .00596

------------------------------------------------------------------------------
   D.lmilemp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lrdefs |
         D1. |    1.68285   .2902387     5.80   0.000     1.111711    2.253989
             |
       _cons |  -.0009573   .0003423    -2.80   0.005    -.0016308   -.0002838
------------------------------------------------------------------------------
```

An overview of non-stationarity

A standard sequence of steps for avoiding spurious regression:

1. Test all variables for unit roots (i.e. non-stationarity) using the appropriate version of the Dickey–Fuller test.
2. If the variables don't have unit roots, estimate the equation in its original units (Y and X).
3. If the variables both have unit roots, test the residuals of the equation for cointegration using the Dickey–Fuller test.
4. If the variables both have unit roots but are not cointegrated, then estimate the equation in first differences (ΔX and ΔY).
5. If the variables both have unit roots and also are cointegrated, then estimate the equation in its original units (Y and X).

Note: if only Y or only X is non-stationary, there is no need to test for cointegration, and you should first-difference only the non-stationary variable.

Model specification

- After having checked for stationarity, you know whether the equation should be in levels (Y_t) or first-differences (ΔY_t).
- You can then add lagged (dependent) variables (starting with a broad model), which in the case of an equation in first-differences means lagged differenced (ΔY_{t-1}) variables.
- Test for autocorrelation using the Breusch-Godfrey test: if you find any, include more lags.

Things to do for your project paper this week

- Using Dickey-Fuller tests, determine whether the variables in your equation are stationary or non-stationary – if both variables are non-stationary, also test for cointegration using a Dickey-Fuller test on the residual.
- Based on results from the Dickey-Fuller tests, determine whether you need to estimate the model as a level equation or in first differences.
- To help you with this, the document "spurious.pdf" has been uploaded to Blackboard.

Project paper – some announcements

- This is the **last week of material for your project paper**– next week's topic is not part of the paper.
- That means **this week** is also the **last meeting with your project advisor**– make good use of it!
- The **deadline** for submission of the project paper is **Friday January 20th** (midnight). Two submissions are required:
  - By email to your project advisor;
  - Online to the Ephorus website to detect plagiarism.
  - See course manual for exact procedure and requirements.
- Grades for the project paper are made known at the same time as the exam grades– your project advisor cannot give you results before that time.