Econometrics Lecture 6

EC2METRIE

Dr. Anna Salomons

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

18 December 2016

This class

- **Time series models**
  - What are time-series data?
  - Different types of time series models
  - Some issues arising in time-series models, including **serial correlation**

- **Studenmund**
  - Section 7.3
  - Chapter 9 (Serial Correlation) excluding sections 9.3.2 and 9.3.3
  - Note: slides contain additional materal not covered in Studenmund

Different types of data

- **Cross-sectional data**: different units observed in one time period
  - Units can be individuals, firms, regions, countries, ...
  - The number of observations is equal to the number of observed units.
  - Weeks 1-5 and week 8 of this course
- **Time-series data**: one unit of observation over different time periods.
  - Time periods can be days, weeks, months, years, ..
  - The number of observations is equal to the number of time periods.
  - Weeks 6-7 of this course.

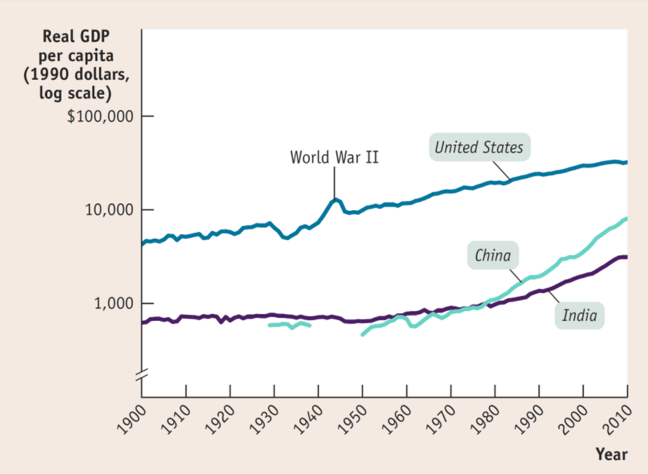
Different types of data

- **Panel data**: a combination of cross-sectional and time-series data, i.e. the same units followed over time.
  - E.g. a number of countries followed over different time periods; a number of individuals followed over different time periods.
  - The number of observations is equal to the number of time periods $\times$ the number of observed units.
  - Not part of this course!

Important examples of time series in economics

- **Long-run economic growth**
- There has been a sustained rise over time in real GDP per capita since the Industrial Revolution for advanced economies (e.g. US, The Netherlands)
- In more recent decades, some developing economies (e.g. China, India) have experienced strong long-run growth.

3 time series of long-run economic growth, 1900-2010

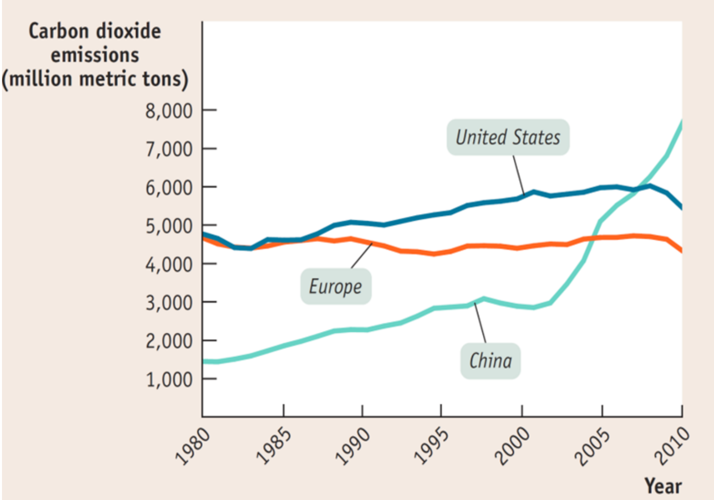


source: Krugman & Wells, Economics, 3rd Ed.

Important examples of time series in economics

- Recent economic growth has lead to **climate change**
- In particular, an unsustainable rise in carbon dioxide emissions over time
- It is unsustainable because there is no market (yet) for pollution in contrast to e.g. oil of which the price rises if oil supply gets limited.

3 time series of CO2 emissions, 1980-2010

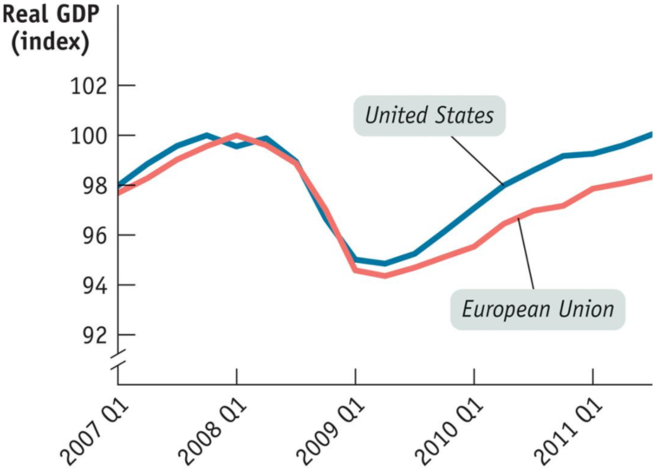


source: Krugman & Wells, Economics, 3rd Ed.

Important examples of time series in economics

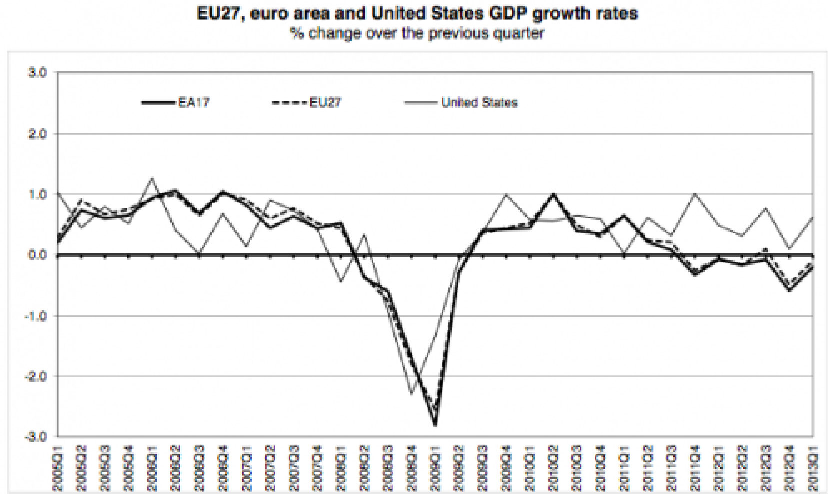
- **The Great Recession of 2007-2009**
- The largest decrease in GDP and the largest increase in unemployment since the Great Depression in the 1930s.
- Recovery is faster in the US than in Europe

2 time series of real GDP, 2007Q1-2011Q3



source: Krugman & Wells, Economics, 3rd Ed.

3 time series of GDP growth, 2005Q1-2013Q1



Cross-sectional model

An example of a **cross-sectional model**:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + .. + \beta_k X_{ki} + \varepsilon_i$$

- **Variation comes from different units of observation**, denoted by i
- i can denote individuals, firms, regions, countries, houses,...: the number of observations n is equal to the sum of i
- This is what we used so far in this course.

Cross-sectional models in economics: examples

Examples of questions that economists answer with **cross-sectional models**:

- How high is the return to one more year of education? (i=individual workers)
- Does harsher sentencing reduce crime? (i=regions)
- What is the effect of the number of bedrooms on housing prices? (i=houses)
- Are firms with more unionized workers less profitable? (i=firms)
- What is the effect of investment on economic growth? (i=countries)

Times-series model

An example of a **time series model**:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + .. + \beta_k X_{kt} + \varepsilon_t$$

- **Variation comes from different time periods**, denoted by t
- t can denote days, months, years,...
- The number of observations n is equal to the sum of t

Why are time series models different?

- Time series data have a natural **temporal ordering**, unlike cross-sectional data, in which there is no natural ordering of the observations
  - E.g. cross-sectional data of people's wages and their respective education levels: the order in which the individuals' data is entered is irrelevant.
- The ordering means we can specify **different types of models** (static or dynamic).
- This **ordering also leads to a number of issues** (we discuss some of these this week, and others in week 7)

Different types of time series models

- **Static time series model**

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t$$

- Distributed lag model

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

- Autoregressive distributed lag model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{1t} + \beta_3 X_{1t-1} + \varepsilon_t$$

Static time series model

- **Static time series model**

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t$$

- This can also be written as

$$\begin{aligned} Y_{t=1} &= \beta_0 + \beta_1 X_{1t=1} + \varepsilon_{t=1} \\ Y_{t=2} &= \beta_0 + \beta_1 X_{1t=2} + \varepsilon_{t=2} \\ & \ldots \\ Y_{t=n} &= \beta_0 + \beta_1 X_{1t=n} + \varepsilon_{t=n} \end{aligned}$$

- In a static model, **all effects are contemporaneous**: X has an impact on Y in the same time period t.

Example of a static model: It's the economy, stupid!

How to test whether it's the economy, stupid!

- **US presidential elections**: **how important is the state of the economy** for the proportion of the popular vote obtained by the incumbent party[1]?
- We expect that **economic conditions in the election year have an impact on people's votes**: if there is higher economic growth, more people think the incumbent party is doing a good job and will support their candidate.

$$\text{vote}_t = \beta_0 + \beta_1 \text{growth}_t + \varepsilon_t$$

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0$$

[1] The incumbent party is the party (Republican or Democratic) that won the previous election.

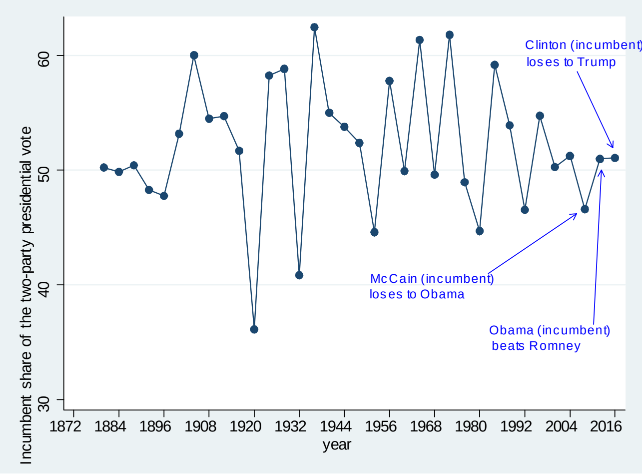
Summary of the dataset

```
                  storage   display    value
variable name       type    format     label      variable label
-----------------------------------------------------------------------------------------------
vote              float     %9.0g                 Incumbent share of the two-party presidential
                                                    vote
growth            float     %9.0g                 growth rate GDP in first three quarters of the
                                                    election year

. sum year vote growth

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        year |         35        1948     40.9878       1880       2016
        vote |         35    52.03743    5.881357     36.119     62.458
      growth |         35    .6971714    5.300731    -14.499     11.765

. tsset year
        time variable:  year, 1880 to 2016, but with gaps
                delta:  1 unit
```

Telling Stata we have time series data

- Note: when using a time-series dataset, you have to **tell Stata** that this is the case.
- You do this by using the **tsset command** (tsset stands for "time series set"), **followed by the variable which uniquely identifies time series observations** in your data.
- In this case, that variable is *year*: hence *tsset year*
  - Stata tells us *year* goes from 1880 to 2016, but with gaps. These gaps of course occur because elections are only held every 4 years.
  - That is, the variable *year* changes like this: 1880, 1884, 1888, ..., 2012, 2016.

Variation in the dependent variable



Estimates of the model

```
. 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
------------------------------------------------------------------------------

. predict yhat
```

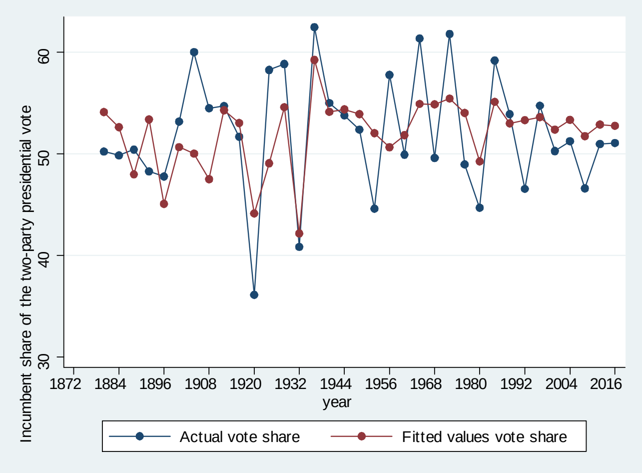
Static model: interpretation

The estimated equation is (standard errors in parentheses):

$$\text{vote}_t = \underset{(0.83)}{51.6} + \underset{(0.16)}{0.65}\,\text{growth}_t + e_t$$

- We find a **significant effect of growth$_t$ on vote$_t$**, hence we can interpret the estimated parameter.
- The **interpretation** is: when growth in the election year was 1 percentage point higher, the incumbent party obtained 0.65 percentage points more of the popular vote in that year's election.
- So the economy does have an effect on US presidential election outcomes.

Fitted values against actual values



Elections & the economy: further research

- David Autor, David Dorn, Gordon Hanson & Kaveh Majlesi have examined the effect of import competition from China on the 2016 U.S. presidential election.
- They find that Clinton would have won the election if the China trade shock would have been half the size, showing another way in which the economy matters for elections.
- For those interested in this work, see: http://economics.mit.edu/files/12418

Different types of time series models

- Static time series model: no lagged variables (of either X or Y) on the right-hand side of the equation.

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t$$

- **Distributed lag model**

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

- Autoregressive distributed lag model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{1t} + \beta_3 X_{1t-1} + \varepsilon_t$$

Different types of time series models

- **Distributed lag model**

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{1t-1} + \ldots + \beta_{p+1} X_{1t-p} + \varepsilon_t$$

- Also known as a **finite distributed lag model** if not all possible lags of the independent variable are included, e.g.

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

- In this model, **independent variables have an impact that is spread out over more than one time period**.
  - Let's look at an example in which this is a better representation of reality than a static model.

Finite distributed lag model: example

- **Workplace accidents** are common, and **cost firms a lot of money**.
- For this reason, many **firms provide safety training** for workers.
- Since safety training also costs money, firms want to know **the monetary benefit of an hour of such training**: by how much does 1 hr of training reduce the amount of money lost to workplace accidents?
- To do so, we start by estimating the following **static model** (i.e. not yet a finite distributed lag model):

$$loss_t = \beta_0 + \beta_1 safe_t + \varepsilon_t$$

Summary statistics

```
                  storage   display      value
variable name      type     format       label        variable label
--------------------------------------------------------------------------------------
time               float    %9.0g                     month
losses             double   %10.0g                    Losses from accidents in pounds per month
safety             double   %10.0g                    Nr of hours of safety training per month

. sum time loss safety

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        time |        60        30.5    17.46425          1         60
      losses |        60    74234.54    10178.21   45432.68   94022.72
      safety |        60    13.70862    15.53743          0    59.0358

. tsset time
        time variable:  time, 1 to 60
                delta:  1 unit
```

Estimates of the static model

```
. reg loss safety

      Source |       SS           df       MS      Number of obs   =        60
-------------+----------------------------------   F(1, 58)        =      0.10
       Model |  10166274.7         1  10166274.7   Prob > F        =    0.7570
    Residual |  6.1020e+09        58   105206747   R-squared       =    0.0017
-------------+----------------------------------   Adj R-squared   =   -0.0155
       Total |  6.1122e+09        59   103595892   Root MSE        =     10257

------------------------------------------------------------------------------
      losses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      safety |  -26.71627   85.94421    -0.31   0.757    -198.7523    145.3198
       _cons |   74600.78   1772.441    42.09   0.000     71052.85     78148.7
------------------------------------------------------------------------------
```

Improving on the static model

- **In the static model, we find no effect** of safety training in month t on losses due to accidents in month t.
- But maybe this is not a correctly specified model: it is likely that training in month t still has an effect in later months.
- To account for this, we should estimate a **finite distributed lag model.** Can include a different number of lags:

$$
\begin{aligned}
loss_t &= \beta_0 + \beta_1 safe_t + \beta_2 safe_{t-1} + \varepsilon_t \\
loss_t &= \beta_0 + \beta_1 safe_t + \beta_2 safe_{t-1} + \beta_3 safe_{t-2} + \varepsilon_t \\
loss_t &= \beta_0 + \beta_1 safe_t + \beta_2 safe_{t-1} + \beta_3 safe_{t-2} + \beta_4 safe_{t-3} + \varepsilon_t
\end{aligned}
$$

Estimates of the finite distributed lag model (1 lag)

```
. reg loss safety l.safety

      Source |       SS           df       MS      Number of obs   =        59
-------------+----------------------------------   F(  2,     56)  =     13.65
       Model |  2.0030e+09         2  1.0015e+09   Prob > F        =    0.0000
    Residual |  4.1077e+09        56  73352417.7   R-squared       =    0.3278
-------------+----------------------------------   Adj R-squared   =    0.3038
       Total |  6.1108e+09        58   105358133   Root MSE        =    8564.6

------------------------------------------------------------------------------
      losses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      safety |
         --. |  -73.53727   72.51141    -1.01   0.315    -218.7951    71.72054
         L1. |  -377.7844   72.47037    -5.21   0.000      -522.96   -232.6088
             |
       _cons |   80365.57   1862.497    43.15   0.000     76634.53     84096.6
------------------------------------------------------------------------------
```

Estimates of the finite distributed lag model (2 and 3 lags)

```
. reg loss safety l.safety l2.safety

      Source |       SS       df       MS              Number of obs =      58
-------------+------------------------------           F(  3,    54) =   39.78
       Model |  4.1990e+09     3  1.3997e+09           Prob > F      =  0.0000
    Residual |  1.9000e+09    54  35185962.8           R-squared     =  0.6885
-------------+------------------------------           Adj R-squared =  0.6712
       Total |  6.0990e+09    57   107000844           Root MSE      =  5931.8

------------------------------------------------------------------------------
      losses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      safety |
         --. |  -91.08063   50.64648    -1.80   0.078    -192.6207    10.45946
         L1. |  -431.4759   50.75049    -8.50   0.000    -533.2245   -329.7273
         L2. |  -394.7269   50.50434    -7.82   0.000     -495.982   -293.4718
             |
       _cons |   86933.93   1537.919    56.53   0.000     83850.59    90017.28
------------------------------------------------------------------------------

. reg loss safety l.safety l2.safety l3.safety

      Source |       SS       df       MS              Number of obs =      57
-------------+------------------------------           F(  4,    52) =   40.60
       Model |  4.5839e+09     4  1.1460e+09           Prob > F      =  0.0000
    Residual |  1.4679e+09    52  28228501.5           R-squared     =  0.7574
-------------+------------------------------           Adj R-squared =  0.7388
       Total |  6.0518e+09    56   108067954           Root MSE      =  5313.1

------------------------------------------------------------------------------
      losses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      safety |
         --. |     -125.9   46.24049    -2.72   0.009    -218.6884   -33.11169
         L1. |  -443.4918   45.88164    -9.67   0.000    -535.5601   -351.4236
         L2. |  -417.6089   45.73324    -9.13   0.000    -509.3794   -325.8384
         L3. |  -179.9043   46.25205    -3.89   0.000    -272.7158   -87.09274
             |
       _cons |   90402.22   1643.183    55.02   0.000     87104.93    93699.51
------------------------------------------------------------------------------
```

Estimates of the finite distributed lag model (4 lags)

```
. reg loss safety l.safety l2.safety l3.safety l4.safety

      Source |       SS       df       MS              Number of obs =      56
-------------+------------------------------           F(  5,    50) =   31.64
       Model |  4.5802e+09     5   916048673           Prob > F      =  0.0000
    Residual |  1.4475e+09    50  28950924.3           R-squared     =  0.7599
-------------+------------------------------           Adj R-squared =  0.7358
       Total |  6.0278e+09    55   109596174           Root MSE      =  5380.6

------------------------------------------------------------------------------
      losses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      safety |
         --. |  -131.9943   47.43609    -2.78   0.008    -227.2725   -36.71608
         L1. |  -449.8597   47.55659    -9.46   0.000    -545.3799   -354.3395
         L2. |  -422.5183   46.77785    -9.03   0.000    -516.4744   -328.5623
         L3. |  -187.1041   47.64089    -3.93   0.000    -282.7936   -91.41453
         L4. |  -27.77104    47.6619    -0.58   0.563    -123.5028    67.96071
             |
       _cons |   91173.32    1949.85    46.76   0.000     87256.93     95089.7
------------------------------------------------------------------------------
```

- The 4th lag has an insignificant effect- we prefer the model with 3 lags.
- (More official selection criteria for the number of lags exist, but these are not part of this course.)

Including lagged variables in Stata

- We can include lagged variables in Stata by using the time-series operator L.
- To be able to do this, we need to have first told Stata our dataset is a time-series dataset (i.e. we must have used the *tsset* command)
- Stata also has time-series operators for leads (F.) and differences (D.) - and all the different operators can be combined.

Time-series operators in Stata

Operator	**Meaning**
L.	1-period lag ($L.X_t = X_{t-1}$)
L1.	1-period lag ($L1.X_t = X_{t-1}$)
L2.	2-period lag ($L2.X_t = X_{t-2}$)
...	
F.	1-period lead ($F.X_t = X_{t+1}$)
F1.	1-period lead ($F1.X_t = X_{t+1}$)
F2.	2-period lead ($F2.X_t = X_{t+2}$)
...	
D.	1-period difference ($D.X_t = X_t - X_{t-1}$)
D1.	1-period difference ($D.X_t = X_t - X_{t-1}$)
L.D.	1-period lagged difference ($L.D.X_t = X_{t-1} - X_{t-2}$)

Finite distributed lag model: interpretation

$$loss_t = \widehat{\beta}_0 + \widehat{\beta}_1 safe_t + \widehat{\beta}_2 safe_{t-1} + \widehat{\beta}_3 safe_{t-2} + \widehat{\beta}_4 safe_{t-3} + e_t$$

$$loss_t = \left\{ \begin{array}{l} \underset{(18365)}{80365} - \underset{(46.2)}{125}\, safe_t - \underset{(45.8)}{443}\, safe_{t-1} \\ - \underset{(45.7)}{417}\, safe_{t-2} - \underset{(46.3)}{180}\, safe_{t-3} + e_t \end{array} \right\}$$

- The **short-run effect** is given by $\widehat{\beta}_1$: the short-run effect of safety training on losses due to accidents is -125, i.e. one more hour of safety training reduces accident losses by £125 in the short run.

Finite distributed lag model: interpretation

$$
\begin{aligned}
loss_t &= \widehat{\beta}_0 + \widehat{\beta}_1 safe_t + \widehat{\beta}_2 safe_{t-1} + \widehat{\beta}_3 safe_{t-2} + \widehat{\beta}_4 safe_{t-3} + e_t \\
loss_t &= \left\{ \begin{array}{l} \underset{(18365)}{80365} - \underset{(46.2)}{125}\, safe_t - \underset{(45.8)}{443}\, safe_{t-1} \\ - \underset{(45.7)}{417}\, safe_{t-2} - \underset{(46.3)}{180}\, safe_{t-3} + e_t \end{array} \right\}
\end{aligned}
$$

- The **long-run effect** is found by letting **all time-varying variables achieve their equilibrium values**, i.e. they no longer vary over time.

Finite distributed lag model: finding the long-run effect

- If all time-varying variables achieve their equilibrium values, they no longer vary over time- we drop the time subscript and introduce the * superscript to denote equilibrium:

$$\begin{aligned}
\widehat{loss}_t &= \widehat{\beta}_0 + \widehat{\beta}_1 safe_t + \widehat{\beta}_2 safe_{t-1} + \widehat{\beta}_3 safe_{t-2} + \widehat{\beta}_4 safe_{t-3} \\
\widehat{loss}^* &= \widehat{\beta}_0 + \widehat{\beta}_1 safe^* + \widehat{\beta}_2 safe^* + \widehat{\beta}_3 safe^* + \widehat{\beta}_4 safe^* \\
\widehat{loss}^* &= \widehat{\beta}_0 + \left(\widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3 + \widehat{\beta}_4\right) safe^*
\end{aligned}$$

- Hence the **long-run effect** is $\left(\widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_3 + \widehat{\beta}_4\right) = -125 - 443 - 417 - 180 = -1165$
- Interpretation: each hour of safety training decreases losses from workplace accidents by £1165 in the long run.

Different types of time series models

- Static time series model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \varepsilon_t$$

- Distributed lag model

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

- **Autoregressive distributed lag model**

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{1t} + \beta_3 X_{1t-1} + \varepsilon_t$$

Example of an autoregressive distributed lag model

- Back to political economy: rather than assessing whether the economy matters for elections, let's see if it matters for **prime minister approval ratings**.
- This time, we use UK data from the **Thatcher-Major era** (1979-1996), a time when the conservative party was in power in Britain.
- We estimate

$$pmsat_t = \beta_0 + \beta_1 pmsat_{t-1} + \beta_2 econft_t + \beta_3 econft_{t-1} + \varepsilon_t$$

Autoregressive distributed lag model: summary statistics

```
              storage  display     value
variable name   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 of an autoregressive distributed lag model

$$pmsat_t = \beta_0 + \beta_1 pmsat_{t-1} + \beta_2 econft_t + \beta_3 econft_{t-1} + \varepsilon_t$$

This is an **autoregressive distributed lag model**:

- **Distributed lag** since it includes a **lag of the independent variable**;
- **Autoregressive** since it includes a **lag of the dependent variable** (the dependent variable is regressed on its own lag, i.e. autoregressive).

Why estimate an autoregressive distributed lag model?

$$pmsat_t = \beta_0 + \beta_1 pmsat_{t-1} + \beta_2 econft_t + \beta_3 econft_{t-1} + \varepsilon_t$$

We want to estimate an autoregressive distributed lag model because:

- **Distributed lag:** we expect that **effects are not all contemporaneous,** i.e. past economic conditions can impact current approval rates;
- **Autoregressive:** we expect that **effects are not instantaneous**, i.e. the adjustment of approval rates to economic conditions does not occur within one time period.

Autoregressive distributed lag model

$$pmsat_t = \beta_0 + \beta_1 pmsat_{t-1} + \beta_2 econft_t + \beta_3 econft_{t-1} + \varepsilon_t$$

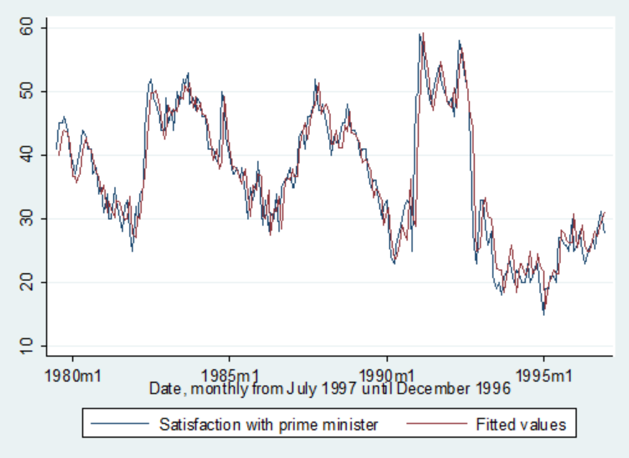
Meaning of the size of the autoregressive coefficient β_1:

- β_1 large (close to 1): slow adjustment of the approval rate
- β_1 small (close to 0): rapid adjustment of the approval rate
- $\beta_1 = 0$: instantaneous adjustment of the approval rate (i.e. no autoregressive component needs to be included)
- $\beta_1 = 1$: OLS yields biased estimates of β_1,i.e. $E(\widehat{\beta}_1) \neq 1$. We come back to this case in week 7!

Estimates of an autoregressive distributed lag model

```
. reg pmsat l.pmsat econft l.econft

      Source |       SS       df       MS              Number of obs =     209
-------------+------------------------------           F(  3,   205) =  495.92
       Model |  19413.186     3  6471.06202           Prob > F      =  0.0000
    Residual | 2674.96952   205  13.0486318           R-squared     =  0.8789
-------------+------------------------------           Adj R-squared =  0.8771
       Total | 22088.1556   208  106.193056           Root MSE      =  3.6123

------------------------------------------------------------------------------
       pmsat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       pmsat |
         L1. |   .9166358   .0258064    35.52   0.000     .8657557    .9675158
             |
      econft |
         --. |   .1148156   .0240748     4.77   0.000     .0673497    .1622815
         L1. |   -.05602     .024586    -2.28   0.024    -.1044939   -.0075461
             |
       _cons |   3.551733   1.050092     3.38   0.001     1.481369    5.622097
------------------------------------------------------------------------------
```

Fitted versus actual values



Autoregressive distributed lag model: interpretation

$$pmsat_t = \widehat{\beta}_0 + \widehat{\beta}_1 pmsat_{t-1} + \widehat{\beta}_2 econft_t + \widehat{\beta}_3 econft_{t-1} + e_t$$

$$pmsat_t = \left\{ \begin{array}{c} \underset{(1.050)}{3.55} + \underset{(0.026)}{0.92}\, pmsat_{t-1} + \underset{(0.024)}{0.11}\, econft_t \\ - \underset{(0.025)}{0.06}\, econft_{t-1} + e_t \end{array} \right\}$$

- **Short-run effect** is given by the estimated β_2 : when the economy is evaluated 1 point higher, satisfaction with the prime minister increases by 0.11 point.

Autoregressive distributed lag model: finding the long-run effect

$$pmsat_t = \widehat{\beta}_0 + \widehat{\beta}_1 pmsat_{t-1} + \widehat{\beta}_2 econft_t + \widehat{\beta}_3 econft_{t-1} + e_t$$

Long-run effect is found by letting all time-varying variables achieve their equilibrium values, i.e. they no longer vary over time:

$$pmsat^* = \widehat{\beta}_0 + \widehat{\beta}_1 pmsat^* + \widehat{\beta}_2 econft^* + \widehat{\beta}_3 econft^* + e^*$$

$$\begin{aligned}
pmsat^* - \widehat{\beta}_1 pmsat^* &= \widehat{\beta}_0 + \widehat{\beta}_2 econft^* + \widehat{\beta}_3 econft^* + e^* \\
\left(1 - \widehat{\beta}_1\right) pmsat^* &= \widehat{\beta}_0 + \left(\widehat{\beta}_2 + \widehat{\beta}_3\right) econft^* + e^* \\
pmsat^* &= \frac{\widehat{\beta}_0}{1 - \widehat{\beta}_1} + \frac{\widehat{\beta}_2 + \widehat{\beta}_3}{1 - \widehat{\beta}_1} econft^* + e^*
\end{aligned}$$

Autoregressive distributed lag model: interpretation

- **Long-run effect** is:

$$\widehat{pmsat^*} = \frac{\widehat{\beta}_0}{\left(1-\widehat{\beta}_1\right)} + \frac{\left(\widehat{\beta}_2+\widehat{\beta}_3\right)}{\left(1-\widehat{\beta}_1\right)} econft^*$$

- Hence the long-run effect is given by $\frac{\left(\widehat{\beta}_2+\widehat{\beta}_3\right)}{\left(1-\widehat{\beta}_1\right)} = \frac{0.1148-0.0560}{1-0.9166} \approx 0.71$
- In the long run, when voters the economy is evaluated 1 point higher, satisfaction with the prime minister increases by 0.71 point.

Some issues specific to time series models

- Here we discuss a few **issues that arise when we estimate time series models**
- These are all specific to time series datasets, i.e. we did not yet encounter them in cross-sectional analysis.
- The remaining issues are discussed in week 7

Some issues specific to time series models

1. An introduction to **spurious regression** (*more in week 7!*)
2. Seasonality
3. Serial correlation

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 due to trending variables

- Economic **time series often have a trend** (i.e. increase or decrease steadily over time)
- Just because **2 series are trending together**, we **can't assume that the relation is causal**
- Often, both will be trending because of other unobserved factors
- Even if those factors are unobserved, we can control for them by directly **controlling for the trend**

Spurious regression: an example

```
. descr marriages beer

              storage  display     value
variable name   type   format      label      variable label
---------------------------------------------------------------------------
marriages       double %10.0g                 UK marriages, in thousands
beer            long   %10.0g                 UK home consumption of beer, in hectoli

. sum marriages beer

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   marriages |        14    327.5429    31.70272      286.1        392
        beer |        14    60483.14    2479.284      57007      65303

. tsset year
        time variable:  year, 1989 to 2002
                delta:  1 unit
```

Spurious regression: an example

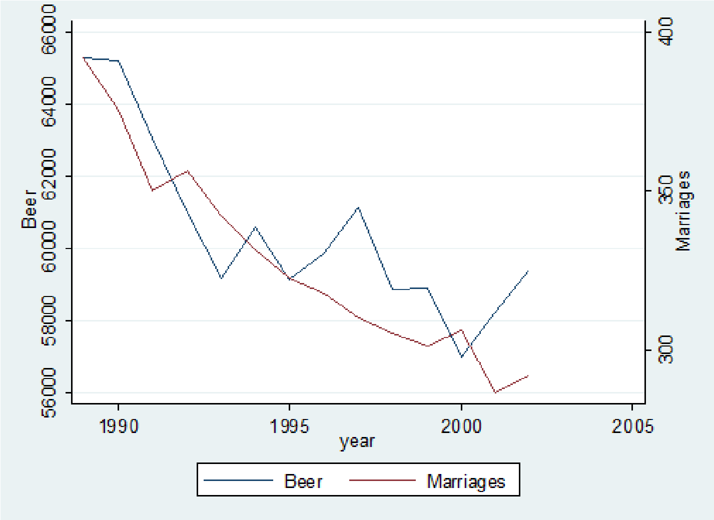
```
. reg marriages beer

      Source |       SS       df       MS              Number of obs =      14
-------------+------------------------------           F(  1,    12) =   33.53
       Model |  9622.21377     1  9622.21377           Prob > F      =  0.0001
    Residual |  3443.60051    12  286.966709           R-squared     =  0.7364
-------------+------------------------------           Adj R-squared =  0.7145
       Total |  13065.8143    13  1005.06264           Root MSE      =   16.94

------------------------------------------------------------------------------
   marriages |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        beer |   .0109734    .001895     5.79   0.000     .0068444    .0151023
       _cons |  -336.1602   114.7072    -2.93   0.013    -586.0857   -86.23473
------------------------------------------------------------------------------
```

Spurious regression: an example

$$marriages_t = \beta_0 + \beta_1 beer_t + \varepsilon_t$$

- From the estimated model, it looks like an increase in beer consumption significantly increases the number of marriages in the UK.
- But this makes no sense– this is an example of **spurious regression** as we'll demonstrate.
- In particular, a spurious relationship appears because both the number of marriages and beer consumption **follow a timetrend**– that is, they change steadily over time.

Both variables follow a timetrend



Both variables follow a timetrend

```
. reg marriages year

      Source |       SS       df       MS              Number of obs =      14
-------------+------------------------------           F(  1,    12) =  146.65
       Model |  12077.5286     1  12077.5286           Prob > F      =  0.0000
    Residual |  988.28567     12  82.3571392           R-squared     =  0.9244
-------------+------------------------------           Adj R-squared =  0.9181
       Total |  13065.8143    13  1005.06264           Root MSE      =  9.0751

------------------------------------------------------------------------------
   marriages |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -7.286154   .6016722   -12.11   0.000    -8.597085   -5.975223
       _cons |   14867.06   1200.639    12.38   0.000     12251.09    17483.03
------------------------------------------------------------------------------

. reg beer year

      Source |       SS       df       MS              Number of obs =      14
-------------+------------------------------           F(  1,    12) =   24.82
       Model |  53863943.8     1  53863943.8           Prob > F      =  0.0003
    Residual |    26045118    12   2170426.5           R-squared     =  0.6741
-------------+------------------------------           Adj R-squared =  0.6469
       Total |  79909061.7    13   6146850.9           Root MSE      =  1473.2

------------------------------------------------------------------------------
        beer |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -486.5846   97.67464    -4.98   0.000    -699.3994   -273.7698
       _cons |    1031463   194910.2     5.29   0.000       606790     1456135
------------------------------------------------------------------------------
```

Spurious regression

- The previous slide shows that both the number of marriages and beer consumption **follow a trend**.
  - The number of marriages decreases by 7286 per year.
  - Beer consumption decreases by 486 hectolitres per year.
- As such, a regression of the number of marriages on beer consumption produces **spurious results if we do not control for a timetrend**.
  - Essentially, the timetrend is an omitted variable since time has a significant effect on the dependent variable (marriages) AND is correlated with the independent variable (beer consumption).
- **Solution: include a timetrend in the model**.

Fixing spurious regression by including the timetrend

```
. reg marriages beer year

      Source |       SS       df       MS              Number of obs =      14
-------------+------------------------------           F(  2,    11) =   84.50
       Model |  12267.3227     2  6133.66134           Prob > F      =  0.0000
    Residual |  798.491611    11  72.5901465           R-squared     =  0.9389
-------------+------------------------------           Adj R-squared =  0.9278
       Total |  13065.8143    13  1005.06264           Root MSE      =    8.52

------------------------------------------------------------------------------
   marriages |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        beer |   .0026995   .0016695     1.62   0.134     -.000975    .0063739
        year |  -5.972634   .9894252    -6.04   0.000    -8.150345   -3.794924
       _cons |   12082.66   2058.108     5.87   0.000     7552.797    16612.53
------------------------------------------------------------------------------
```

Spurious regression

- **Once we control for a timetrend**, there is **no significant relationship** between beer consumption and the number of marriages (as we would expect).
- Thus: when we estimate time-series regressions, we **should always check whether a time-trend belongs in the equation**- we should then include it to avoid spurious regression.
- In week 7, we will see that spurious regression is more generally caused by **nonstationarity** in timeseries variables- a timetrend is just one example of this.
    - In week 7, we will also see how to test for nonstationarity; and correct for it.

Spurious regression - another example

- The **South-African AIDS epidemic** has been pinpointed as having an important effect on child mortality there.
- We can investigate this in the following time-series model:

$$cmortality_t = \beta_0 + \beta_1 aids_t + \varepsilon_t$$

Spurious regression - another example

variable name	variable label
cmortality	Probability that a child dies before the age of 5, rate per 100 live births.
aids	Number of people living with HIV per 100 population of age group 15-49.

```
. sum cmortality aids
```

Variable	Obs	Mean	Std. Dev.	Min	Max
cmortality	22	6.710909	.890382	4.67	7.93
aids	22	12.46818	6.609185	.7	18.1

```
. tsset year
        time variable:  year, 1990 to 2011
                delta:  1 unit
```

Spurious regression - another example

```
. reg cmortality aids

      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  1,    20) =    5.46
       Model |  3.56898549     1  3.56898549           Prob > F      =  0.0300
    Residual |  13.0793976    20  .653969881           R-squared     =  0.2144
-------------+------------------------------           Adj R-squared =  0.1751
       Total |  16.6483831    21  .792780149           Root MSE      =  .80868

------------------------------------------------------------------------------
  cmortality |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aids |   .0623756   .0267006     2.34   0.030     .0066791    .1180721
       _cons |   5.933199    .374905    15.83   0.000      5.15116    6.715237
------------------------------------------------------------------------------
```

Spurious regression - another example

- From the regression results we see that when the fraction of 15-49 year olds with AIDS increases by 1 percentage point, the probability that a child dies before the age of 5 increases by 0.06 percentage points.

- However, it is **possible that this result is spurious** because we have not yet checked what happens when we include a timetrend in the equation.

- To do so, estimate the following model:

$$cmortality_t = \beta_0 + \beta_1 aids_t + \beta_2 t + \varepsilon_t$$

Spurious regression - another example

```
. reg cmortality aids year

      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  2,    19) =   35.70
       Model |  13.1489127     2  6.57445634           Prob > F      =  0.0000
    Residual |  3.49947045    19  .184182655           R-squared     =  0.7898
-------------+------------------------------           Adj R-squared =  0.7677
       Total |  16.6483831    21  .792780149           Root MSE      =  .42917

------------------------------------------------------------------------------
  cmortality |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        aids |   .2688616   .0319454     8.42   0.000      .201999    .3357241
        year |  -.2344923   .0325141    -7.21   0.000    -.3025451   -.1664395
       _cons |   472.4605    64.6878     7.30   0.000     337.0674    607.8537
------------------------------------------------------------------------------
```

Spurious regression - another example

- From the regression results we see that the **timetrend is significant**, indicating a decrease in the probability that a child dies before the age of 5 of 0.23 percentage points per year, controlling for AIDS.
- More importantly, we **now get a much stronger effect of AIDS on child mortality**: when the fraction of 15-49 year olds with AIDS increases by 1 percentage point, the probability that a child dies before the age of 5 increases by 0.27 percentage points.
- This again illustrates the importance of **checking whether a timetrend should be included** in the model, **to avoid spurious regression** results.

Some issues specific to time series models

1. An introduction to spurious regression (*more in week 7!*)
2. **Seasonality**
3. Serial correlation

Seasonality

- **Seasonality** or a **seasonal effect** is a systematic and calendar related effect in a time series.
- Example: the sharp increase in most retail time series around December during the Christmas period
- Sometimes we are interested in the seasonal effects themselves, and sometimes we want to say something about the non-seasonal characteristics of the time-series, meaning we want to control for seasonality.
- Both these aims can be achieved by **including dummies for the relevant time period(s)**.

An example of seasonality: icecream consumption

We have a dataset of icecream consumption for a number of years, and we observe the consumption per capita in each month for these years.

```
. descr consumption

              storage  display     value
variable name   type   format      label      variable label
-----------------------------------------------------------------------------------------
consumption     float  %9.0g                  monthly consumption of ice cream per capita (in pints)

. tsset time
        time variable:  time, 1 to 30
                delta:  1 unit
```

To **test for seasonality in icecream consumption**, we estimate a model with dummies for all but one calendar month (to avoid perfect collinearity):

$$consumption_t = \beta_0 + \beta_1 feb_t + \beta_2 mar_t + \beta_3 apr_t + \ldots + \beta_{12} dec_t + \varepsilon_t$$

An example of seasonality: icecream consumption

```
. reg consumption feb mar apr may jun jul aug sept oct nov dec

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F( 11,    18) =    4.21
       Model |  .090411025    11  .008219184           Prob > F      =  0.0035
    Residual |  .035112333    18  .001950685           R-squared     =  0.7203
-------------+------------------------------           Adj R-squared =  0.5493
       Total |  .125523358    29  .004328392           Root MSE      =  .04417

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         feb |       .027   .0441666     0.61   0.549    -.0657905    .1197905
         mar |   .0563333   .0403184     1.40   0.179    -.0283724    .1410391
         apr |   .0653333   .0403184     1.62   0.123    -.0193724    .1500391
         may |   .0983333   .0403184     2.44   0.025     .0136276    .1830391
         jun |   .1223333   .0403184     3.03   0.007     .0376276    .2070391
         jul |   .1526667   .0403184     3.79   0.001     .0679609    .2373724
         aug |        .16   .0403184     3.97   0.001     .0752942    .2447057
        sept |      .0715   .0441666     1.62   0.123    -.0212905    .1642905
         oct |        .03   .0441666     0.68   0.506    -.0627905    .1227905
         nov |       .009   .0441666     0.20   0.841    -.0837905    .1017905
         dec |     -.0035   .0441666    -0.08   0.938    -.0962905    .0892905
       _cons |       .285   .0312305     9.13   0.000     .2193872    .3506128
------------------------------------------------------------------------------
```

An example of seasonality: icecream consumption

To **test for seasonality** in icecream consumption, we should perform an **F-test** for joint significance of the monthly dummies

```
. test feb mar apr may jun jul aug sept oct nov dec

 ( 1)  feb = 0
 ( 2)  mar = 0
 ( 3)  apr = 0
 ( 4)  may = 0
 ( 5)  jun = 0
 ( 6)  jul = 0
 ( 7)  aug = 0
 ( 8)  sept = 0
 ( 9)  oct = 0
 (10)  nov = 0
 (11)  dec = 0

       F( 11,    18) =    4.21
            Prob > F =    0.0035
```

We reject the null hypothesis of joint insignificance, which means we find **evidence of seasonality** in icecream consumption.

Some issues specific to time series models

1. An introduction to spurious regression (m*ore in week 7!*)
2. Seasonality
3. **Serial correlation**

Assumptions 1-4

OLS is unbiased estimator of parameters β if assumptions 1-4 hold:

1. **Population model is linear in parameters** (and the error term is additive).
2. **Error term has a zero population mean**: $E(\varepsilon_i) = 0$.
3. **All independent variables are uncorrelated with the error term**: $Corr(\varepsilon_i, X_i) = 0$.
4. **No perfect (multi)collinearity** between independent variables.

Assumptions 5-6

OLS is unbiased estimator of $Var(\widehat{\beta})$ if **assumptions 1-4 hold, as well as 5-6**:

5. **No serial correlation**: $Corr(\varepsilon_i, \varepsilon_j) = 0$.
6. **No heteroskedasticity**: error term has constant variance, $Var(\varepsilon_i) = \sigma^2$ (where σ^2 is a constant).

Questions about serial correlation

1. **What is serial correlation?**
2. What are the consequences of serial correlation?
3. How to test for serial correlation?
4. How to correct for serial correlation?

Serial correlation

- This week, we discuss assumption 5: no serial correlation.
- **Serial correlation** is also called **autocorrelation**, and means the error term is correlated over subsequent time periods.
- **Positive serial correlation**: $Corr(\varepsilon_t, \varepsilon_{t-1}) > 0$, the error in period t is likely to be positive if the the error in period $t-1$ is positive
  Negative serial correlation: $Corr(\varepsilon_t, \varepsilon_{t-1}) < 0$, the error in period t is likely to be negative if the error in period $t-1$ is positive
  - In economic time series, we are more likely to have positive than negative serial correlation.

What is serial correlation?

- Consider the model:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$$

- If the **error term of this model is serially correlated**, we can write:

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

  ε_t depends on ε_{t-1}; u_t is a serially uncorrelated error term with a mean of zero and a constant variance.

- This is **first order serial correlation** because the error in t depends on the error in $t-1$.

What is serial correlation?

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t \\ \varepsilon_t &= \rho\varepsilon_{t-1} + u_t \end{aligned}$$

- ρ **is the correlation between** ε_t **and** ε_{t-1} (see next 2 slides for proof)
- Thus $|\rho| \leq 1$ (since this is the range any correlation coefficient can have).
- The closer $|\rho|$ is to 1, the stronger the serial correlation.
  - If $|\rho| = 1$, OLS yields biased estimates of β_1,i.e. $E(\widehat{\beta}_1) \neq \beta_1$. We come back to this case in week 7!
- If $\rho = 0$, there is no serial correlation.

Proof of interpretation of rho (I)

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

To prove that ρ is the correlation between ε_t and ε_{t-1}, we first calculate the autocovariance, which is the covariance between ε_t and ε_{t-1} :

$$\begin{aligned}
Cov(\varepsilon_t, \varepsilon_{t-1}) &= Cov(\rho\varepsilon_{t-1} + u_t, \varepsilon_{t-1}) \\
&= Cov(\rho\varepsilon_{t-1}, \varepsilon_{t-1}) + Cov(u_t, \varepsilon_{t-1}) \\
&= \rho Cov(\varepsilon_{t-1}, \varepsilon_{t-1}) = \rho Var(\varepsilon_{t-1})
\end{aligned}$$

Proof of interpretation of rho (II)

We assume the variance of ε is the same in all time periods (assumption 6- homoskedasticity), such that $Var(\varepsilon_{t-1}) = Var(\varepsilon_t)$

$$
\begin{aligned}
Corr(\varepsilon_t, \varepsilon_{t-1}) &= \frac{Cov(\varepsilon_t, \varepsilon_{t-1})}{\sqrt{Var(\varepsilon_t)}\sqrt{Var(\varepsilon_{t-1})}} \\
&= \frac{\rho Var(\varepsilon_t)}{\sqrt{Var(\varepsilon_t)}\sqrt{Var(\varepsilon_{t-1})}} \\
&= \frac{\rho Var(\varepsilon_t)}{\sqrt{Var(\varepsilon_t)}\sqrt{Var(\varepsilon_t)}} = \rho
\end{aligned}
$$

Illustration: The demand for ice cream

- We estimate a model that relates icecream consumption to the price of icecream, income, and the monthly temperature:

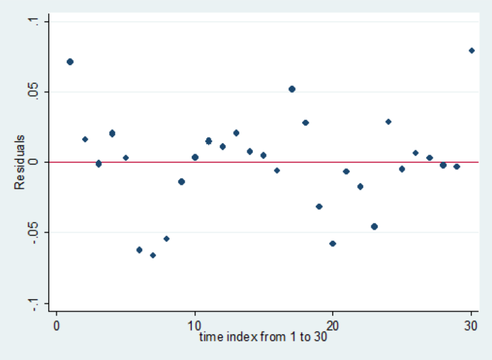
$$consumption_t = \beta_0 + \beta_1 price_t + \beta_2 income_t + \beta_3 temp_t + \varepsilon_t$$

- We predict residuals of this model, and plot them against the time variable.

Estimated icecream demand equation

```
. reg consumption price income temp

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  3,    26) =   22.17
       Model |  .090250521     3  .030083507           Prob > F      =  0.0000
    Residual |  .035272836    26  .001356648           R-squared     =  0.7190
-------------+------------------------------           Adj R-squared =  0.6866
       Total |  .125523358    29  .004328392           Root MSE      =  .03683

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -1.044413    .834357    -1.25   0.222    -2.759458    .6706323
      income |   .0033078   .0011714     2.82   0.009     .0008999    .0057156
        temp |   .0062252    .000802     7.76   0.000     .0045767    .0078737
       _cons |   .3079846   .2657777     1.16   0.257    -.2383292    .8542985
------------------------------------------------------------------------------

. predict uhat, resid
```

Residuals for ice cream demand equation



This shows that positive and negative residuals group together over time: this is what **positive serial correlation** looks like.

Questions about serial correlation

1. What is serial correlation?
2. **What are the consequences of serial correlation?**
3. How to test for serial correlation?
4. How to correct for serial correlation?

Consequences of serial correlation

- **Assumption 5 is not satisfied** and therefore OLS formulas for $\widehat{\sigma}^2$ and therefore $se(\widehat{\beta}_k)$ are incorrect.
- Hence, t-test and F-tests are invalid: **cannot perform hypothesis tests**.
- Typically, we too often conclude that the estimated parameters are significant.
- However, **serial correlation does not cause bias in the estimated coefficients- with one important exception** (see next slide)!

Consequences of serial correlation in a model with a lagged dependent variable

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{1t} + \varepsilon_t \\
\varepsilon_t &= \rho \varepsilon_{t-1} + u_t
\end{aligned}
$$

When there is **serial correlation** (i.e. if $\rho \neq 0$) **in a time-series model containing a lagged dependent variable**, not only are the $se(\widehat{\beta}_k)$ incorrect, the $\widehat{\beta}_k$ **are also biased**!

Consequences of serial correlation in a model with a lagged dependent variable

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{1t} + \varepsilon_t \\
\varepsilon_t &= \rho \varepsilon_{t-1} + u_t
\end{aligned}
$$

- Recall that OLS assumption 3 states that the **correlation between the error term ε_t and all independent variables is zero**.
- But it turns out Y_{t-1} (one of the independent variables) will always be correlated with ε_t if there is serial correlation. (See proof on next slide)
- **A violation of assumption 3** means we get **biased estimates of the coefficients**!

Consequences of serial correlation in a model with a lagged dependent variable

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{1t} + \varepsilon_t \qquad \varepsilon_t = \rho\varepsilon_{t-1} + u_t$$

To see why $\rho \neq 0$ causes $Corr(\varepsilon_t, Y_{t-1}) \neq 0$:

- We can write the correlation between ε_t and Y_{t-1} as:

$$\begin{aligned} Corr(\varepsilon_t, Y_{t-1}) &= Corr(\rho\varepsilon_{t-1} + u_t, Y_{t-1}) \\ &= Corr(\rho\varepsilon_{t-1}, Y_{t-1}) \end{aligned}$$

- But Y_{t-1} and ε_{t-1} are correlated by definition, as seen by rewriting the model for $t-1$

$$Y_{t-1} = \beta_0 + \beta_1 Y_{t-2} + \beta_2 X_{1t-1} + \varepsilon_{t-1}$$

This proves that ε_t is correlated with Y_{t-1} if $\rho \neq 0$ (i.e. if there is serial correlation), which means we obtain **biased coefficient estimates**.

Consequences of serial correlation: summary

- Serial correlation in static or finite distributed lag models:
  - Biased standard errors
- Serial correlation in autoregressive models- i.e. models with a lagged dependent variable[2]:
  - Biased standard errors;
  - Biased coefficient estimates.

[2] Side note- they may or may not also contain a distributed lag: as soon as a lagged dependent variable is included, these consequences hold.

Questions about serial correlation

1. What is serial correlation?
2. What are the consequences of serial correlation?
3. **How to test for serial correlation?**
4. How to correct for serial correlation?

How to test for serial correlation?

There are 2 commonly used tests for serial correlation:

- **Durbin-Watson test**
  - This test is will not be on the exam- you may skip that part of Studenmund.
  - Here, a few slides about it are included because statistical software often reports this test (and you may see it in research papers).
- **Breusch-Godfrey test** (resembles Breusch-Pagan test for heteroskedasticity)
  - This is the test we will use in this course.

The Durbin-Watson test (DW-test)

- The test statistic is

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=2}^{T} e_t^2}$$

- This has the property:

$$d \approx 2(1-\rho)$$

  - If $\rho = 0$ then $d = 2$: no serial correlation
  - If $\rho \to 1$ then $d \to 0$: strong positive serial correlation

The Durbin-Watson test (DW-test)

- The hypotheses are

$$
\begin{aligned}
H_0 &: \rho = 0 \qquad \text{no serial correlation} \\
H_A &: \rho \neq 0 \qquad \text{serial correlation}
\end{aligned}
$$

- Drawback of DW test is that it **cannot be for models with a lagged dependent variable** because it requires the **assumption that all regressors are strictly exogenous** (we discuss what this means in week 7).
- The advantage of the **Breusch-Godfrey test** is that it **can be used for all types of time-series models**: the assumption required for this test is that **all regressors are contemporaneously exogenous** (which we discuss in week 7 as well).

The Durbin-Watson test (DW-test): demand for icecream example

```
. reg consumption price income temp

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  3,    26) =   22.17
       Model |  .090250521     3  .030083507           Prob > F      =  0.0000
    Residual |  .035272836    26  .001356648           R-squared     =  0.7190
-------------+------------------------------           Adj R-squared =  0.6866
       Total |  .125523358    29  .004328392           Root MSE      =  .03683

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -1.044413    .834357    -1.25   0.222    -2.759458    .6706323
      income |   .0033078   .0011714     2.82   0.009     .0008999    .0057156
        temp |   .0062252    .000802     7.76   0.000     .0045767    .0078737
       _cons |   .3079846   .2657777     1.16   0.257    -.2383292    .8542985
------------------------------------------------------------------------------

. dwstat

Durbin-Watson d-statistic(  4,    30) =  1.021169
```

The Breusch-Godfrey test (BG-test)

We want to test for first-order serial correlation in the following time series model:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + .. + \beta_k X_{kt} + \varepsilon_t \qquad (1) \\ \varepsilon_t &= \rho \varepsilon_{t-1} + u_t \qquad \text{with } |\rho| < 1 \end{aligned}$$

1. Estimate equation (1)
2. Compute the residuals e_t from the estimated equation
3. Regress e_t on the 1-period lagged residual e_{t-1} and all independent variables from equation 1 $X_{1t}, X_{2t}..X_{kt}$
4. Use the t-value on e_{t-1} to test

$$\begin{aligned} H_0 &: \rho = 0 \qquad \text{no 1st-order serial correlation} \\ H_A &: \rho \neq 0 \qquad \text{1st-order serial correlation} \end{aligned}$$

(Note that the structure of the BG test for serial correlation resembles that of the BP test for heteroskedasticity)

Example: The demand for ice cream

$$consumption_t = \beta_0 + \beta_1 price_t + \beta_2 income_t + \beta_3 temp_t + \varepsilon_t$$

- We estimated this model for the demand for icecream, and want to know **whether the error term in this equation is serially correlated**.
- When inspecting the residuals of these estimated model, we saw a pattern indicative of positive serial correlation- however, such an analysis is never conclusive.
- To determine whether there is serial correlation, we should use the **Breusch-Godfrey test for serial correlation**.

BG test for 1st order serial correlation: steps 1 and 2

```
. reg consumption price income temp

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(3, 26)        =     22.17
       Model |  .090250521         3  .030083507   Prob > F        =    0.0000
    Residual |  .035272836        26  .001356648   R-squared       =    0.7190
-------------+----------------------------------   Adj R-squared   =    0.6866
       Total |  .125523358        29  .004328392   Root MSE        =    .03683

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -1.044413    .834357    -1.25   0.222    -2.759458    .6706323
      income |   .0033078   .0011714     2.82   0.009     .0008999    .0057156
        temp |   .0062252    .000802     7.76   0.000     .0045767    .0078737
       _cons |   .3079846   .2657777     1.16   0.257    -.2383292    .8542985
------------------------------------------------------------------------------

. predict uhat, resid
```

BG test for 1st order serial correlation: step 3

```
. reg uhat l.uhat price income temp

      Source |       SS           df       MS      Number of obs   =        29
-------------+----------------------------------   F(  4,      24) =      1.16
       Model |  .004878776         4  .001219694   Prob > F        =    0.3523
    Residual |  .025197463        24  .001049894   R-squared       =    0.1622
-------------+----------------------------------   Adj R-squared   =    0.0226
       Total |  .030076239        28  .001074151   Root MSE        =     .0324

------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        uhat |
         L1. |   .3998496   .1970353     2.03   0.054    -.0068114    .8065105
             |
       price |   .1513481   .7495092     0.20   0.842    -1.395563    1.698259
      income |   .0005283   .0010725     0.49   0.627    -.0016853    .0027418
        temp |  -.0000289   .0007412    -0.04   0.969    -.0015587    .0015008
       _cons |  -.0875828   .2439211    -0.36   0.723    -.5910112    .4158457
------------------------------------------------------------------------------
```

BG test for 1st order serial correlation: step 4

- The **p-value for the lagged residual is statistically significant** at the 10% level.
- Hence we reject the null hypothesis of no serial correlation and **conclude there is first-order serial correlation** in the demand for icecream model. (Note that we have chosen $\alpha = 0.10$ since the dataset is quite small.)

Breusch-Godfrey test for second-order serial correlation

The BG test can also be used to test for second-order (or in general, higher-order) serial correlation. Second order serial correlation implies:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + u_t$$

The test procedure is the same, except:

- In **step 3** we regress ε_t on the independent variables from the original model, ε_{t-1} (one-period lagged residual) as well as ε_{t-2} (two-period lagged residual).
- In **step 4** we must use a **joint significance test** (i.e. F-test) for ε_{t-1} and ε_{t-2} to test:

$$\begin{aligned} H_0 &: \rho_1 = \rho_2 = 0 && \text{no 2nd order serial correlation} \\ H_A &: H_0 \text{ not true} && \text{2nd order serial correlation} \end{aligned}$$

BG test for second-order serial correlation: step 3

```
. reg uhat l.uhat l2.uhat price income temp

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  5,    22) =    0.98
       Model |  .005423863     5  .001084773           Prob > F      =  0.4505
    Residual |  .024291415    22  .001104155           R-squared     =  0.1825
-------------+------------------------------           Adj R-squared = -0.0033
       Total |  .029715278    27  .001100566           Root MSE      =  .03323

------------------------------------------------------------------------------
        uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        uhat |
         L1. |   .5085876   .2435505     2.09   0.049     .0034947     1.01368
         L2. |  -.2158795   .2525786    -0.85   0.402    -.7396954    .3079364
             |
       price |   .1706399   .7723084     0.22   0.827     -1.43103     1.77231
      income |   .0007143   .0011562     0.62   0.543    -.0016835    .0031122
        temp |   .0003236   .0008768     0.37   0.716    -.0014947    .0021419
       _cons |  -.1117705   .2517041    -0.44   0.661    -.6337729    .4102319
------------------------------------------------------------------------------

. test l.uhat l2.uhat

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

       F(  2,    22) =    2.18
            Prob > F =    0.1365
```

BG test for second-order serial correlation: step 4

- The **p-value for the F-test is not statistically significant** at the 10% level.
- Hence we **do not reject the null hypothesis of no second-order serial correlation**.
- To sum up, we do find evidence of first-order serial correlation (see previous BG test) but not of second-order serial correlation in the demand for icecream model.

Questions about serial correlation

1. What is serial correlation?
2. What are the consequences of serial correlation?
3. How to test for serial correlation?
4. **How to correct for serial correlation?**

How to correct for serial correlation

Different solutions are possible:

- Use **a different estimator**: Cochrane-Orcutt/**Prais-Winsten Generalized Least Squares** (GLS)
- Use **Newey-West standard errors**
- Alternatively: include more lags of the variables in your model - we will see an example of this in week 7

GLS: Cochrane-Orcutt/Prais-Winsten

- Stata command *prais* instead of *reg*
- **Adjusts for first-order serial correlation**.
- **Transforms the dependent and independent variables** to eliminate first-order serial correlation from the equation (see Appendix slides).
- Due to this transformation, **both the estimated parameters and their standard errors may change**.

Example of Prais-Winsten: demand for icecream

```
. prais consumption price income temp

Iteration 0:   rho = 0.0000
Iteration 1:   rho = 0.4006
..etc..
Iteration 41:   rho = 0.8002

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

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  3,    26) =   14.35
       Model |  .044945961     3  .014981987           Prob > F      =  0.0000
    Residual |  .027154356    26  .001044398           R-squared     =  0.6234
-------------+------------------------------           Adj R-squared =  0.5799
       Total |  .072100317    29  .002486218           Root MSE      =  .03232

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -1.048854    .759751    -1.38   0.179    -2.610545    .5128362
      income |  -.0008022   .0020458    -0.39   0.698    -.0050074    .0034029
        temp |   .0053173   .0012795     4.16   0.000     .0026872    .0079474
       _cons |   .6815345   .2903562     2.35   0.027     .0846987     1.27837
-------------+----------------------------------------------------------------
         rho |   .8002264
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.021169
Durbin-Watson statistic (transformed) 1.846795
```

Newey-West standard errors

- Stata command *newey* instead of *reg*, and the number of autocorrelated lags also have to be specified, *e.g. newey y x1 x2, lags(1)* for first-order autocorrelation
- **Adjusts for serial correlation of any specified order**.
- **Only the standard errors change**, the estimated parameters remain the same.
- Side note: Newey-West standard errors **correct for both serial correlation and heteroskedasticity**, and are for that reason also known as HAC (Heteroskedasticity and Autocorrelation Consistent) standard errors.

Example of Newey-West standard errors: demand for icecream

```
. newey consumption price income temp, lag(1)

Regression with Newey-West standard errors          Number of obs  =        30
maximum lag: 1                                       F(  3,     26)  =     18.76
                                                     Prob > F        =    0.0000
```

consumption	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
price	-1.044413	.8874728	-1.18	0.250	-2.868639	.7798133
income	.0033078	.0012065	2.74	0.011	.0008278	.0057877
temp	.0062252	.0008364	7.44	0.000	.0045059	.0079444
_cons	.3079846	.2955309	1.04	0.307	-.2994879	.9154572

Example of Newey-West standard errors: demand for icecream

- The **Newey-West standard errors are higher than the OLS standard errors**.
- Since we found first-order autocorrelation in the original model, we specified , *lag(1)*
  - Had we found second-order autocorrelation in the original model, we should have specified , *lag(2)*

Serial correlation: summary

- **Serial correlation**: violates OLS assumption 5
- **Consequences**: biased standard errors, and additionally, biased coefficient estimates in autoregressive models.
- **Diagnosis**: Breusch-Godfrey test.
- **Solution**: GLS estimation or Newey-West standard errors.

Things to do for your project paper this week

- Download a time series dataset from the course website (whichever one you want, doesn't have to be related to the cross-sectional one)– choose a dependent variable and 1 independent variable.
- Determine whether your dependent and independent variables follow a timetrend (to avoid spurious regression)
- Estimate a model with lagged dependent and independent variables (and a timetrend, if you find this to be necessary) and interpret the estimated parameters (short-run vs long-run effect).
- Test for serial correlation.
- If necessary, re-estimate the model, adjusting for serial correlation.

GLS: Cochrane-Orcutt/Prais-Winsten

Take the following model:

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

Let's say the errors of this model are first-order serially correlated:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

where u_t is an error term that meets all OLS assumptions.

GLS: Cochrane-Orcutt/Prais-Winsten

$$
\begin{aligned}
Y_t &= \beta_0 + \beta_1 X_t + \varepsilon_t \\
\varepsilon_t &= \rho \varepsilon_{t-1} + u_t
\end{aligned}
$$

First, write the model for the period $t - 1$:

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

Next, subtract ρY_{t-1} from both sides of the original model:

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (X_t - \rho X_{t-1}) + \varepsilon_t - \rho \varepsilon_{t-1}$$

GLS: Cochrane-Orcutt/Prais-Winsten

We now have

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (X_t - \rho X_{t-1}) + \varepsilon_t - \rho \varepsilon_{t-1}$$

Note that

$$\begin{aligned} \varepsilon_t &= \rho \varepsilon_{t-1} + u_t \\ &\Leftrightarrow \varepsilon_t - \rho \varepsilon_{t-1} = u_t \end{aligned}$$

Hence

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (X_t - \rho X_{t-1}) + u_t$$

GLS: Cochrane-Orcutt/Prais-Winsten

This gives the following transformed equation:

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (X_t - \rho X_{t-1}) + u_t$$

Note that this transformation has removed serial correlation from the model as u_t is a serially uncorrelated error term.