Time-Series Modelling
ECON20222 - Lecture 9

Ralf Becker and Martyn Andrews

April 2022

Aim for today

- Understand the basic features of time-series data
- Understand autocorrelation
- Understand the difference between stationary and non-stationary data and the different consequences of dealing with these
- Understand how to build dynamic models that can be used for forecasting

Purpose of time-series modelling (TS modelling)

There are typically three things econometricians want to achieve with time-series modelling

1. **Establishing causal relationships** between time-series.
   This is very difficult and in general causal relationships are more difficult to establish with TS modelling. It is not impossible but getting convincing exogenous variation is difficult.

2. Understanding the **dynamics in relationships between variables**.
   Questions like, "If the Central Bank changes the base rate, how long will it take for this to carry through to mortgage rates?" This is perfectly possible as long as we don't make strong causal statements (the CB may change base rates because mortgage rates are very low!!!)

3. **Forecasting one or several time series**.
   This is possibly the most common purpose of TS modelling. We will focus on this.

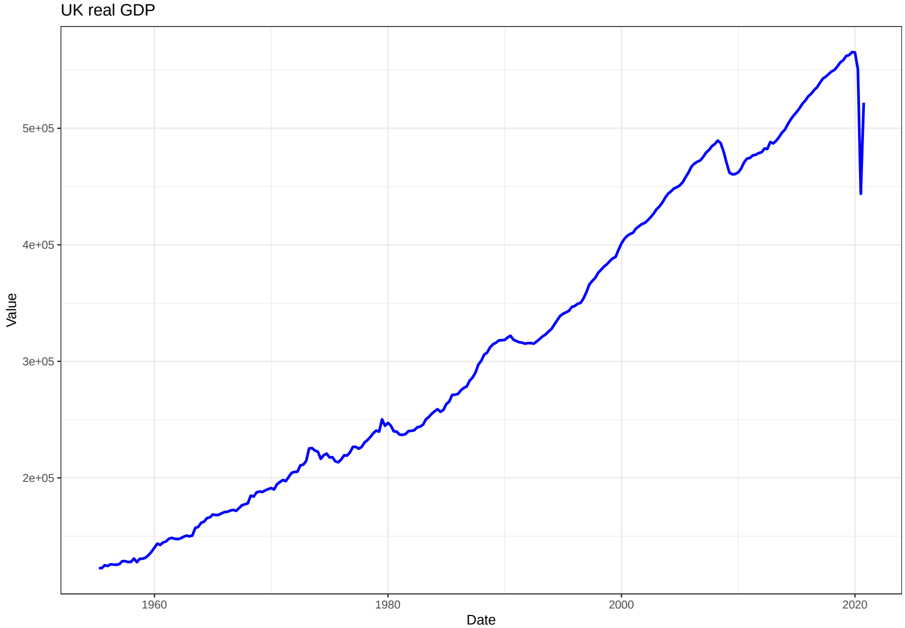
Import some data into R

```
rGDP <- pdfetch_ONS("ABMI","UKEA")
periodicity(rGDP)   # check data frequency
names(rGDP) <- "real GDP" # give a sensible name

# keep all the data including 2020-Q3
# this was the last observation available at the time this was
# remove this line if you want to use updated data
rGDP <- rGDP["/2020-9"]
```

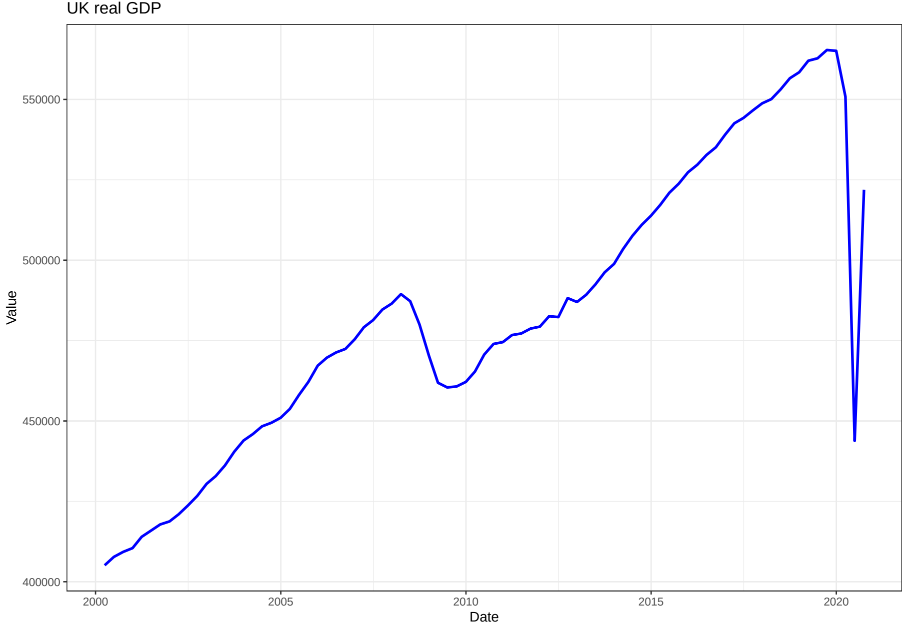
pdfetch functions allow you to directly tap a umber of large data depositories:

- Bundesbank
- Office for National Statistics (ONS)
- Eurostats
- FRED, etc

An example


UK real GDP
Value
5e+05
4e+05
3e+05
2e+05
1960
1980
2000
2020
Date

An example - focus on later periods



The autocorrelation function (ACF)

Recall the correlation coefficient between two random variables z_i and p_i for two cross-sectional variables (index i)

$$Corr(z_i, p_i) = \frac{Cov(z_i, p_i)}{sd(z_i) sd(p_i)}$$

- a measure that expresses the strength of relationship between z_i and p_i
- it takes values in the interval $[-1, 1]$

The autocorrelation function (ACF)

Consider the $y_t = rGDP_t$ time series

- we now use the subscript t
- the subscript goes from $t = 1, ..., T$ where T indicates how many observations we have
- here observations are quarterly (but other series can have other frequencies: e.g. annual, monthly, weekly, daily, hourly, etc.)
- here observations are from Q1 1955 to Q3 2019 (259 observations)

The ACF expresses how observations are correlated to observations 1, 2, 3 or k observations prior.

How can you calculate a correlation of a series with itself?

The autocorrelation function (ACF)

Let's consider the time series y_t and the series one period prior, y_{t-1}. We also call y_{t-1} a one period lag of y_t.

Observation	y_t	y_{t-1}
0	y_{1955Q1}	NA
1	y_{1955Q2}	y_{1955Q1}
2	y_{1955Q3}	y_{1955Q2}
3	y_{1955Q4}	y_{1955Q3}
$\vdots$	$\vdots$	$\vdots$
253	y_{2019Q2}	y_{2019Q1}
254	y_{2019Q3}	y_{2019Q2}

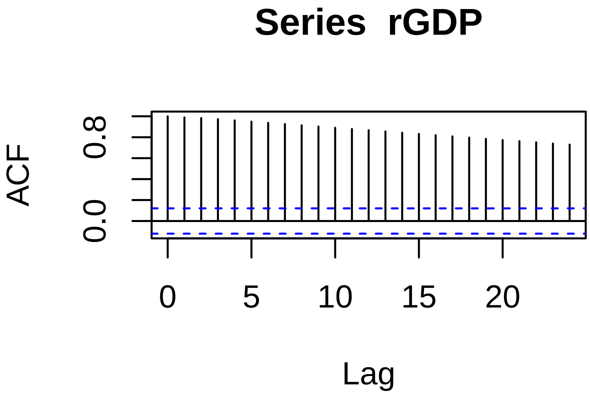
Now we have "two" series for which we can calculate a correlation coefficient. We call this the first order autocorrelation coefficient ρ_1.

An ACF is a collection of autocorrelation coefficients calculated for longer lags k, ρ_k.

The autocorrelation function (ACF)

In R this ACF is easily calculated using the `acf` function.

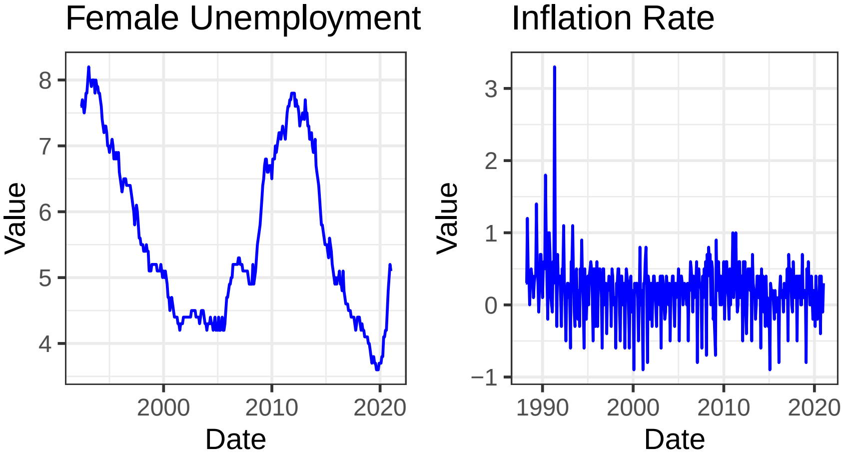
```
temp_acf <- acf(rGDP)
```



The autocorrelation function (ACF)

- The GDP series is strongly upward trending
- this is comon for many macroeconomic series
- this results in an ACF which has large ρ_k for fairly large values of k ($\rho_8 = 0.908$)
- we call this a persistant series

Two further examples



These are data at a monthly frequency

The ACF


Unemployment Rate
ACF
0.0
0.4
0.8
0
10
20
Lag
Inflation
ACF
0.0
0.4
0.8
0
10
20
Lag

The ACF

- the ACF shows that the unemployment rate is also very persistant and the ACF only slowly converges to 0. $\Rightarrow$ a series can be oersitant without time trend

- The inflation rate is not persistant and the autocorrelation quickly drops towards 0.

- But there are peaks of autocorrelation at frequencies 6 and 12 indicating some seasonal variation.

The ACF tells us something about how informative today's observation is for that in 1, 2, 3 or k periods ahead.

- If the ACF decays quickly to 0 then todays info is not very valuable for forecasting long into the future
- If the ACF decays slowly then todays info is valuable for future observations

Stationary and Nonstationary Series

The ACF expresses how persistant a series is.

- A series that is extremely persistant is called a **nonstationary** series.
- A series that is not very persistant is called a **stationary** series.

Here: rGDP and unemployment rate are nonstationary. Inflation rate is stationary.

- In general series with a time-trend are nonstationary
- Some series without time trend are also nonstationary (e.g. female unemployment rate)
- **BUT** there is a huge grey area inbetween.

Formal statistical tests exist (e.g. Augmented Dickey-Fuller test) to decide (but they can be contradictory) and are not dealt with here. Here we eye-ball the series and look at how slowly the ACF converges to 0.

Transformations

An important time-series transformation we consider is that of differencing a series.

Observation	y_t	y_{t-1}	Δy_t
1	y_{1955Q2}	y_{1955Q1}	$y_{1955Q2} - y_{1955Q1} = \Delta y_{1955Q2}$
2	y_{1955Q3}	y_{1955Q2}	$y_{1955Q3} - y_{1955Q2} = \Delta y_{1955Q3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Often we are actually much more interested in the difference of a series rather than the level. GDP is a case in point, the growth rate is what we are really interested in!

The GDP growth rate can be approximated for small growth rates by (assuming that y_t is the GDP series)

$$gGDP_t = \frac{y_t - y_{t-1}}{y_{t-1}} \texttt{ or} \tag{1}$$

$$gGDP_t = ln(y_t) - ln(y_{t-1}) \tag{2}$$

ACF of differenced series

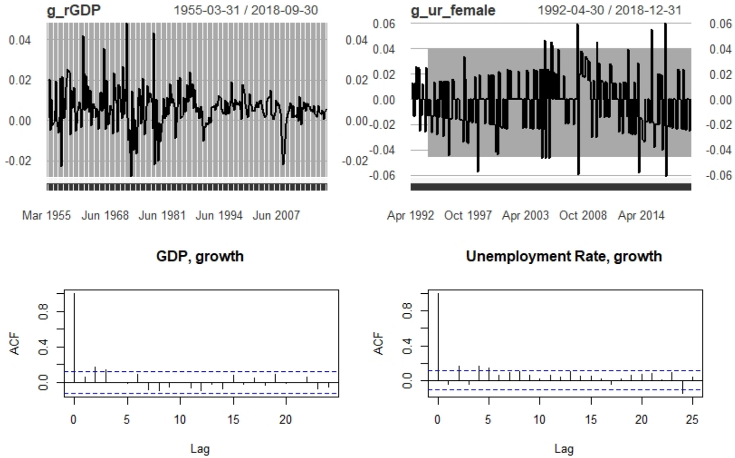


Figure 1: TS and ACF plots for growth in GDP and unemployment rate, ONS

Differencing can turn nonstationary series into a stationary series.

A simple regression

Cross-Section Data

$$y_t = \alpha + \beta x_t + u_t \tag{3}$$

$$\begin{aligned} E(u) &= 0 \qquad (4) \\ E(u|x) &= 0 \qquad (5) \end{aligned}$$

this implied that x_t was exogenous.

Time-Series Data

$$ur_t = \alpha + \beta\ rGDP_t + u_t \tag{6}$$

A simple regression

let's look at the end of the data table

```
# we multiply by 100 to express in percentage points, i.e. 0.5 is 0.5%
reg_data$d_lgdp <- 100*diff(log(reg_data$real.GDP))
reg_data$d_lur <- 100*diff(log(reg_data$ur_female.Close))
tail(reg_data,10)
```

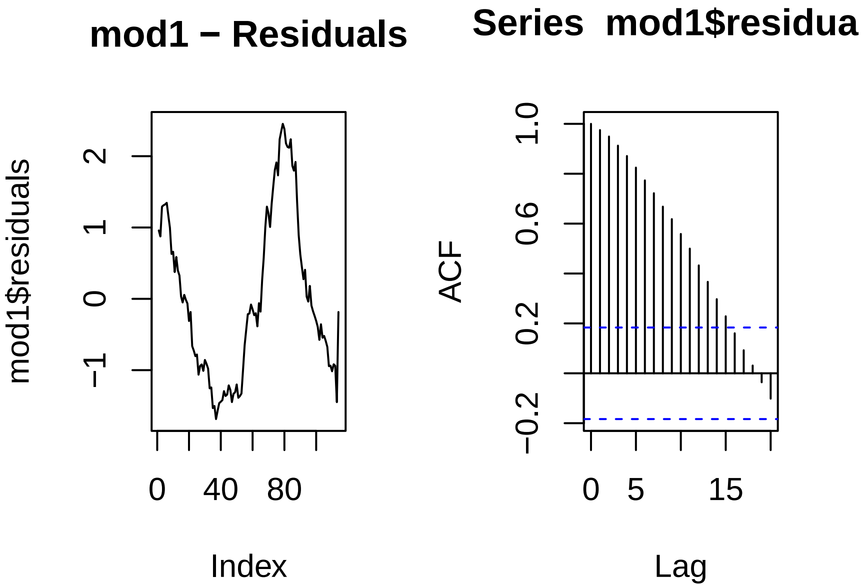
```
           real.GDP ur_female.Close       d_lgdp     d_lur
2018-09-30   556581             4.1   0.62973998 -2.409755
2018-12-31   558448             4.0   0.33487950 -2.469261
2019-03-31   562033             3.7   0.63990594 -7.796154
2019-06-30   562779             3.7   0.13264441  0.000000
2019-09-30   565362             3.6   0.45792228 -2.739897
2019-12-31   565109             3.7  -0.04476011  2.739897
2020-03-31   550835             3.8  -2.55833241  2.666825
2020-06-30   443817             4.2 -21.60229935 10.008346
2020-09-30   521910             4.8  16.20828439 13.353139
2020-12-31       NA             5.1           NA  6.062462
```

A simple regression

```
mod1 <- lm(ur_female.Close~real.GDP,data = reg_data)
stargazer_HC(mod1)
```

```
=========================================================
                          Dependent variable:
                    -------------------------------------
                            ur_female.Close
---------------------------------------------------------
real.GDP                      -0.00001***
                               (0.00000)

Constant                        9.420***
                                (0.531)

---------------------------------------------------------
Observations                      114
R2                               0.216
Adjusted R2                      0.209
Residual Std. Error         1.146 (df = 112)
F Statistic              30.825*** (df = 1; 112)
=========================================================
Note:                       *p<0.1; **p<0.05; ***p<0.01
                    Robust standard errors in parenthesis
```

A simple regression - residual autocorrelation



A simple regression - testing for residual autocorrelation

We can again apply a hypothesis test the Breusch-Godfrey test (`bgtest`). The null hypothesis is that there is no autocorrelation.

```
bgtest(mod1,order=4)
```

```
    Breusch-Godfrey test for serial correlation of order up to

data:  mod1
LM test = 108.96, df = 4, p-value < 2.2e-16
```

A simple regression - HAC standard errors

When we estimate a regression which has autocorrelated error terms we need to apply a different formula to calculate standard errors for coefficients in a regression model (autoregressive heteroscedasticity consistent - HAC).

- They are called Newey-West standard errors.
- They are implemented in `stargazer_HAC.r`.
- This will not change the coefficient estimates.
- Will change the standard errors to the coefficients and hence inference (which will be incorrect if you don't use them).
- If you have time-series data, in doubt, use Newey-West standard errors
- But crucial problems remain (see next slides)

In this example the standard errors only change marginally and hence are not shown here.

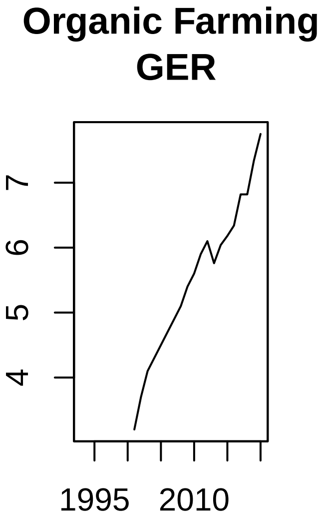
Spurious Regression

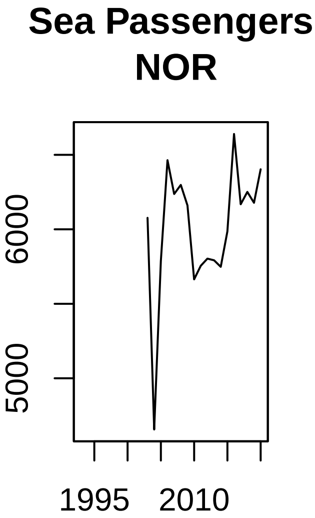
We will explore what can happen if we run a regression involving nonstationary variables.

Let's get some datasets fron EUROSTAT (using `pdfetch_EUROSTAT`).

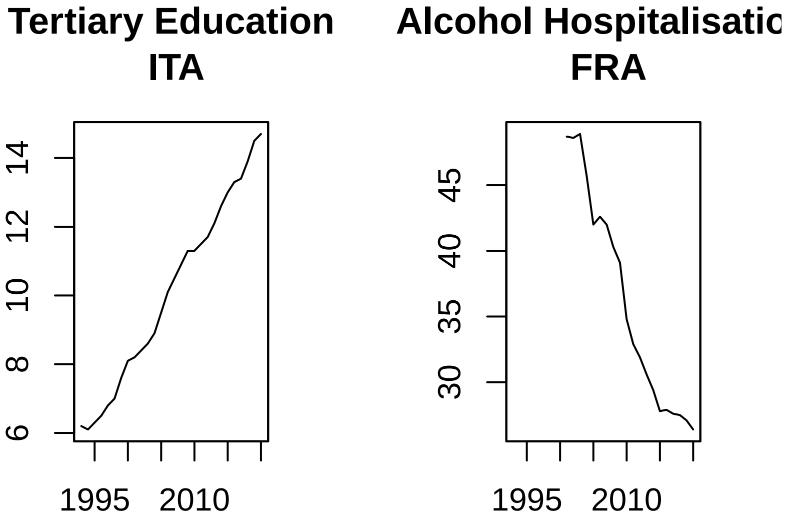
- % of agricultural area Total fully converted and under conversion to organic farming in Germany [`Org`]
- Thousands of passengers travelling to and from Norway by boat [`Pass`]
- % of population with tertiary education in Italy [`Tert`]
- Hospital Discharges, Alcoholic liver disease (in Thousands) in France [`Alc`]

A set of time-series





A set of time-series



Spurious Regression - An example

$$Alc_t = \alpha + \beta\ Org_t + u_t \tag{7}$$

```
mod_sr <- lm(Alc.Disease.FR~Organic.Farming.GE, data = data_sr)
stargazer_HAC(mod_sr)
```

```
=========================================================
                             Dependent variable:
                    -------------------------------------
                               Alc.Disease.FR
---------------------------------------------------------
Organic.Farming.GE                -6.411***
                                   (0.482)

Constant                          71.529***
                                   (2.727)

---------------------------------------------------------
Observations                          20
R2                                  0.908
Adjusted R2                         0.902
Residual Std. Error            2.556 (df = 18)
F Statistic                 176.665*** (df = 1; 18)
=========================================================
Note:                         *p<0.1; **p<0.05; ***p<0.01
                    Robust standard errors in parenthesis
```

Spurious Regression

All possible combinations of simple regressions between the four variables.

Table 1: Regression statistics

Dep.Var		Exp. Var Org	Pass	Tert	Alc
Org	$\hat{\gamma}_1$		0.001	0.587***	-0.134***
	$se_{\hat{\gamma}_1}$		(0.001)	(0.026)	(0.009)
	R^2		0.143	0.954	0.923
Pass	$\hat{\gamma}_1$	159.463		106.734*	-15.061
	$se_{\hat{\gamma}_1}$	(97.459)		(53.471)	(15.086)
	R^2	0.143		0.199	0.062
Tert	$\hat{\gamma}_1$	1.644***	0.002*		-0.232***
	$se_{\hat{\gamma}_1}$	(0.074)	(0.001)		(0.014)
	R^2	0.965	0.199		0.938
Alc	$\hat{\gamma}_1$	-6.878***	-0.004	-4.043***	
	$se_{\hat{\gamma}_1}$	(0.483)	(0.004)	(0.252)	
	R^2	0.923	0.062	0.938	

Note: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$,
Newey-West standard errors in parenthesis

(Unmasking of a) Spurious Regression

One way how you can unmask the spuriousness, if both series are trending is to include a time trend

```
mod_sr2 <- lm(Alc.Disease.FR~Organic.Farming.GE+index(data_sr),
              data = data_sr)
```

or estimate a model in the differences of variables rather than the levels

```
mod_sr3 <- lm(diff(Alc.Disease.FR)~diff(Organic.Farming.GE),
              data = data_sr)
```

(Unmasking of a) Spurious Regression

```
stargazer_HAC(mod_sr,mod_sr2,mod_sr3,type_out = "text", omit.stat = "f")
```

```
=============================================================================
                                     Dependent variable:
                         ----------------------------------------------------
                               Alc.Disease.FR          diff(Alc.Disease.FR)
                             (1)             (2)                (3)
-----------------------------------------------------------------------------
Organic.Farming.GE        -6.411***         1.784
                           (0.475)         (1.333)

index(data_sr)                             -0.005***
                                           (0.0005)

diff(Organic.Farming.GE)                                       1.393
                                                              (0.948)

Constant                  71.529***       95.476***          -1.507***
                           (2.738)         (1.684)            (0.461)

-----------------------------------------------------------------------------
Observations                 20              20                 19
R2                          0.908           0.954              0.041
Adjusted R2                 0.902           0.948              -0.015
Residual Std. Error      2.556 (df = 18) 1.863 (df = 17)  1.353 (df = 17)
=============================================================================
Note:                                            *p<0.1; **p<0.05; ***p<0.01
                                   Newey-West standard errors in parenthesis
```

Spurious Regression - A summary

If you estimate a regression between two nonstationary series:

- Do not mistake very significant coefficients or large R^2 values for evidence of a substantial link between two series
- when regressing non-stationary series (in particular but not only when the series have a time-trend) will deliver a spurious correlation
- When series have a time-trend you can include a time trend variable to test whether there is correlation beyond the common time trend
- When series are nonstationary but don't have a time-trend then you should consider estimating a regression in differences (recall differencing can turn nonstationary series into stationary ones)

For all these reasons, when considering time-series data, we need to **use stationary data. If not estimated coefficients will be, in general, neither unbiased nor consistent**.

A simple regression - but better

From the discussion on spurious regressions we have learned that estimating a model in differences can protect you against spurious regressions.

$$\Delta ur_t = \alpha + \beta\ \Delta rGDP_t + u_t$$

```
# Data (real.GDP, ur_female.Close) are in reg_data
# we multiply by 100 to express in percentage points,
# i.e. 0.5 is 0.5% or 0.005
reg_data$d_lgdp <- 100*diff(log(reg_data$real.GDP))
reg_data$d_lur <- 100*diff(log(reg_data$ur_female.Close))
mod4 <- lm(d_lur~d_lgdp,data = reg_data)
```

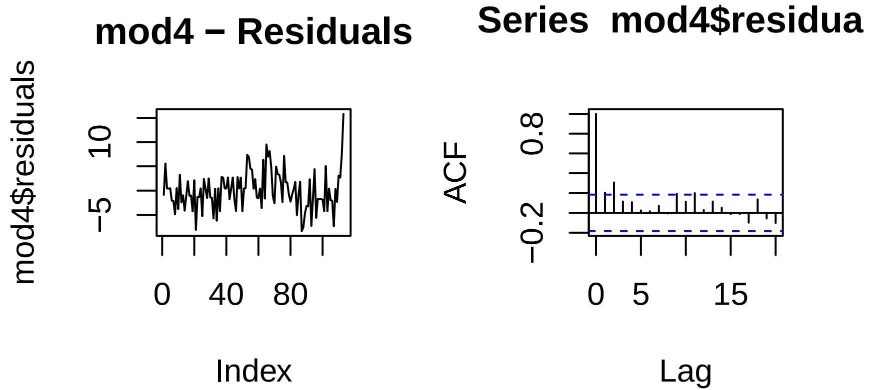
A simple regression - but better

```
===========================================================
                            Dependent variable:
                    ---------------------------------------
                                  d_lur
-----------------------------------------------------------
d_lgdp                           -0.130
                                 (0.144)

Constant                         -0.360
                                 (0.382)

-----------------------------------------------------------
Observations                      113
R2                               0.007
Adjusted R2                      -0.002
Residual Std. Error          4.008 (df = 111)
F Statistic                0.818 (df = 1; 111)
===========================================================
Note:                         *p<0.1; **p<0.05; ***p<0.01
                    Robust standard errors in parenthesis
```

A simple regression - but better

Let's have a look at the residuals.

```
par(mfrow=c(1,2))
plot(mod4$residuals, type = "l", main = "mod4 - Residuals")
acf(mod4$residuals)
```



- We can see that, at lag 2, there is a small amount of autocorrelation in the residuals.

Adding dynamic effects

In the model we estimated:

$$\Delta ur_t = \alpha + \beta \; \Delta rGDP_t + u_t$$

all the action happened in one time period (t).

We need to consider that effects in the economy may take some time. This fact will motivate two major generalisations.

- Include lags of the explanatory variable
- Include lags of the dependent variable

$$\begin{aligned} \Delta ur_t \;\; &= \;\; \alpha_0 + \alpha_1 \Delta ur_{t-1} + \alpha_2 \Delta ur_{t-2} + ... + \alpha_p \Delta ur_{t-p} + \\ & \quad\;\; \beta_0 \; \Delta rGDP_t + \beta_1 \; \Delta rGDP_{t-1} + ... + \beta_k \; \Delta rGDP_{t-k} + u_t \end{aligned}$$

Autoregressive Distributed Lag (ADL) models

```
mod5 <- lm(d_lur~lag(d_lur,1)+lag(d_lur,2)+
             d_lgdp+lag(d_lgdp,1)+lag(d_lgdp,2),
              data = reg_data)
```

Here we use the `lag(series, k)` function which calculates the `k` period lag of `series`.

```r
stargazer_HAC(mod4,mod5, omit.stat = "f")
```

```
================================================================
                              Dependent variable:
                    --------------------------------------------
                                       d_lur
                             (1)                    (2)
----------------------------------------------------------------
lag(d_lur, 1)                                      0.048
                                                  (0.106)

lag(d_lur, 2)                                     0.265***
                                                  (0.082)

d_lgdp                      -0.130               -0.406***
                            (0.215)               (0.092)

lag(d_lgdp, 1)                                   -0.877***
                                                  (0.146)

lag(d_lgdp, 2)                                    -0.019
                                                  (0.813)

Constant                    -0.360                 0.189
                            (0.466)               (0.476)

----------------------------------------------------------------
Observations                  113                   111
R2                           0.007                 0.326
Adjusted R2                 -0.002                 0.294
Residual Std. Error    4.008 (df = 111)      3.364 (df = 105)
================================================================
Note:                                 *p<0.1; **p<0.05; ***p<0.01
                     Newey-West standard errors in parenthesis
```

Forecasting models

Are the above models useful forecasting models?

$$\begin{aligned}\Delta ur_t &= \alpha_0 + \alpha_1 \Delta ur_{t-1} + \alpha_2 \Delta ur_{t-2} + ... + \alpha_p \Delta ur_{t-p} + \\ &\quad \beta_0\ \Delta rGDP_t + \beta_1\ \Delta rGDP_{t-1} + ... + \beta_k\ \Delta rGDP_{t-k} + u_t\end{aligned}$$

- Say you estimated the above models using data up to Q3 2018.
- Hence you have estimated coefficients
- Could you use it to forecast the value in Q4 2018?

No, if we want to forecast Δur_{2018Q4} we would need $\Delta rGDP_{2018Q4}$! But we don't have that. We would first need a forecast for $\Delta rGDP_{2018Q4}$ to then forecast Δur_{2018Q4}. We call these conditional forecasts.

Forecasting models

To build a useful forecasting model we remove all contemporaneous terms from the right hand side, such that we can produce forecasts for period t only having information at time (say) $t-1$.

We remove the $\beta_0 \; \Delta rGDP_t$ term from our model:

```
mod6 <- lm(d_lur~lag(d_lur,1)+lag(d_lur,2)+
             lag(d_lgdp,1)+lag(d_lgdp,2),data = reg_data)
```

Forecasting models

	Dependent variable:		
	d_lur		
	(1)	(2)	(3)
lag(d_lur, 1)		0.048	0.072
		(0.106)	(0.101)
lag(d_lur, 2)		0.265***	0.282***
		(0.082)	(0.083)
d_lgdp	-0.130	-0.406***	
	(0.215)	(0.092)	
lag(d_lgdp, 1)		-0.877***	-0.567***
		(0.146)	(0.088)
lag(d_lgdp, 2)		-0.019	-0.512***
		(0.813)	(0.128)
Constant	-0.360	0.189	0.188
	(0.466)	(0.476)	(0.335)
Observations	113	111	112
R2	0.007	0.326	0.285
Adjusted R2	-0.002	0.294	0.259
Residual Std. Error	4.008 (df = 111)	3.364 (df = 105)	3.473 (df = 107)
Note:			*p<0.1; **p<0.05; ***p<0.01
			Newey-West standard errors in parenthesis

Autoregressive models

- We are interested in forecasting the unemployment rate changes, Δur_t, but we still need observations for $rGDP_{t-1}$ and further lags.
- Realistically there may be other series we may want to consider: e.g. interest rate, inflation, wages, etc.
- Can we forecast Δur_t with nothing else but the history of Δur_t?

Yes, we call these **autoregressive (AR) models**

$$\Delta ur_t = \alpha_0 + \alpha_1 \Delta ur_{t-1} + \alpha_2 \Delta ur_{t-2} + ... + \alpha_p \Delta ur_{t-p} + u_t$$

We "merely" need to chose the lag length p!

Autoregressive models

For starters we use $p = 2$ as above.

```
mod7 <- lm(d_lur~lag(d_lur,1)+lag(d_lur,2),data = reg_data)
```

Autoregressive models

	Dependent variable:			
	d_lur			
	(1)	(2)	(3)	(4)
lag(d_lur, 1)		0.048	0.072	0.207**
		(0.106)	(0.101)	(0.104)
lag(d_lur, 2)		0.265***	0.282***	0.343***
		(0.082)	(0.083)	(0.079)
d_lgdp	-0.130	-0.406***		
	(0.215)	(0.092)		
lag(d_lgdp, 1)		-0.877***	-0.567***	
		(0.146)	(0.088)	
lag(d_lgdp, 2)		-0.019	-0.512***	
		(0.813)	(0.128)	
Constant	-0.360	0.189	0.188	-0.132
	(0.466)	(0.476)	(0.335)	(0.376)
Observations	113	111	112	112
R2	0.007	0.326	0.285	0.180
Adjusted R2	-0.002	0.294	0.259	0.165
Residual Std. Error	4.008 (df = 111)	3.364 (df = 105)	3.473 (df = 107)	3.686 (df = 109)

Note: *p<0.1; **p<0.05; ***p<0.01

Newey-West standard errors in parenthesis

Lag Length - Information Criterion

The last piece in the puzzle is how we determine the best lag length

- Every additional lag will improve the in-sample fit of your model
  $\Rightarrow$ Should we include as many lags as possible?
- Will "over-fit" the model (to features in the data which are not generic but specific to the in-sample period)
  $\Rightarrow$ Tends to make the forecasts more erratic/volatile.

This trade-off needs to be optimised. $\Rightarrow$ **Information Criteria**

Lag Length - Information Criterion

Trade-off between too many variables in model (bad) v in-sample fit (good). Information criteria quantify this trade-off.

One example **Akaike Information Criterion (AIC)**

In R:

Where

- `mod6` and `mod6_4` are the ADL models with 2 and 4 lags respectively
- `mod7` and `mod7_4` are the AR models with 2 and 4 lags respectively

Lag Length - Information Criterion

Table 2: AIC for ADL and AR with 2 and 4 lags

Model	N. of para	AIC
ADL (2 lags)	6	559.1177
AR (2 lags)	4	558.2149
ADL (4 lags)	10	557.2884
AR (4 lags)	6	552.9934

Note: Number of parameters includes linear parameters and residual variance

The optimal model (as per the trade off in the AIC criterion) is the model with the **lowest AIC**.

Here AR(4).

Summary

We learned that

- the ACF encapsulates how persistant a time-series is
- Time-series which are very persistant are called nonstationary
- Using nonstationary series in simple regressions can lead to very misleading (spurious) results
- Estimating models with either a time-trend or in differences can protect you against misleading results
- To build forecasting models we need to ensure that we use explanatory variables which are available at the time of forecasting
- AR models can be a convenient tool for forecasting
- Information criteria can help us to select the right model for forecasting