Autocorrelation

In many applications using time series data, systematic patterns in the behaviour of the residuals can lead to problems with OLS estimation

Autocorrelation

In many applications using time series data, systematic patterns in the behaviour of the residuals can lead to problems with OLS estimation

Given the model

$$Y_t = b_0 + b_1 X_t + u_t$$

Autocorrelation

In many applications using time series data, systematic patterns in the behaviour of the residuals can lead to problems with OLS estimation

Given the model

$$Y_t = b_0 + b_1X_t + u_t$$

for OLS to give unbiased and efficient (minimum variance) estimates then one of the Gauss-Markov conditions is that the value of the residual in one time period should give no indication about the value of the residual in any other period

Autocorrelation

In many applications using time series data, systematic patterns in the behaviour of the residuals can lead to problems with OLS estimation

Given the model

$$Y_t = b_0 + b_1 X_t + u_t$$

for OLS to give unbiased and efficient (minimum variance) estimates then one of the Gauss-Markov conditions is that the value of the residual in one time period should give no indication about the value of the residual in any other period

$Cov(u_t\ u_{t-j}) = 0$ for **any** (unspecified) lag length j

Autocorrelation

In many applications using time series data, systematic patterns in the behaviour of the residuals can lead to problems with OLS estimation

Given the model

$$Y_t = b_0 + b_1X_t + u_t$$

for OLS to give unbiased and efficient (minimum variance) estimates then one of the Gauss-Markov conditions is that the value of the residual in one time period should give no indication about the value of the residual in any other period

$$Cov(u_t\ u_{t-j}) = 0 \quad \text{for } \textbf{any} \text{ (unspecified) lag length } j$$

Autocorrelation exists when this assumption no longer holds, so that

Autocorrelation

In many applications using time series data, systematic patterns in the behaviour of the residuals can lead to problems with OLS estimation

Given the model

$$Y_t = b_0 + b_1X_t + u_t$$

for OLS to give unbiased and efficient (minimum variance) estimates then one of the Gauss-Markov conditions is that the value of the residual in one time period should give no indication about the value of the residual in any other period

$$Cov(u_t\ u_{t-j}) = 0 \quad \text{for } \textbf{any} \text{ (unspecified) lag length } j$$

Autocorrelation exists when this assumption no longer holds, so that

$$Cov(u_t\ u_{t-j}) \neq 0 \qquad \text{for any (unspecified) lag length } j$$

This could be caused by
inertia in omitted economic variables (multiplier working through)

Given $C = b_0 + b_1Y + e$

Given $C = b_0 + b_1Y + e$

We know $\Delta e \rightarrow \Delta C \rightarrow \Delta Y \rightarrow \Delta C$ etc

So the values of C are serially correlated
(a shock to C in period 1 leads to related shocks to C in future periods)

Given $C = b_0 + b_1Y + e$

We know $\Delta e \rightarrow \Delta C \rightarrow \Delta Y \rightarrow \Delta C$ etc

So the values of C are serially correlated
(a shock to C in period 1 leads to related shocks to C in future periods)

and if consumption is omitted from a model this must lead to serial (auto) correlation in the error term

Given $C = b_0 + b_1Y + e$

We know $\Delta e \rightarrow \Delta C \rightarrow \Delta Y \rightarrow \Delta C$ etc

So the values of C are serially correlated
(a shock to C in period 1 leads to related shocks to C in future periods)

and if consumption is omitted from a model this must lead to serial (auto) correlation in the error term

$$Y_t = \delta_0 + \delta_1P_t + \delta_2C_t + u_t$$

Given $C = b_0 + b_1Y + e$

We know $\Delta e \rightarrow \Delta C \rightarrow \Delta Y \rightarrow \Delta C$ etc

So the values of C are serially correlated
(a shock to C in period 1 leads to related shocks to C in future periods)

and if consumption is omitted from a model this must lead to serial (auto) correlation in the error term

$$Y_t = \delta_0 + \delta_1 P_t + \delta_2 C_t + u_t$$

becomes

$$Y_t = \delta_0 + \delta_1 P_t + \{\delta_2 C_t + u_t\}$$

Given $C = b_0 + b_1Y + e$

We know $\Delta e \rightarrow \Delta C \rightarrow \Delta Y \rightarrow \Delta C$ etc

So the values of C are serially correlated
(a shock to C in period 1 leads to related shocks to C in future periods)

and if consumption is omitted from a model this must lead to serial (auto) correlation in the error term

$Y_t = \delta_0 + \delta_1P_t + \delta_2C_t + u_t$

becomes

$Y_t = \delta_0 + \delta_1P_t + \{\delta_2C_t + u_t\} = Y_t = \delta_0 + \delta_1P_t + v_t$

where $v_t = \delta_2C_t + u_t$

Given $C = b_0 + b_1 Y + e$

We know $\Delta e \rightarrow \Delta C \rightarrow \Delta Y \rightarrow \Delta C$ etc

So the values of C are serially correlated
(a shock to C in period 1 leads to related shocks to C in future periods)

and if consumption is omitted from a model this must lead to serial (auto) correlation in the error term

$Y_t = \delta_0 + \delta_1 P_t + \delta_2 C_t + u_t$

becomes

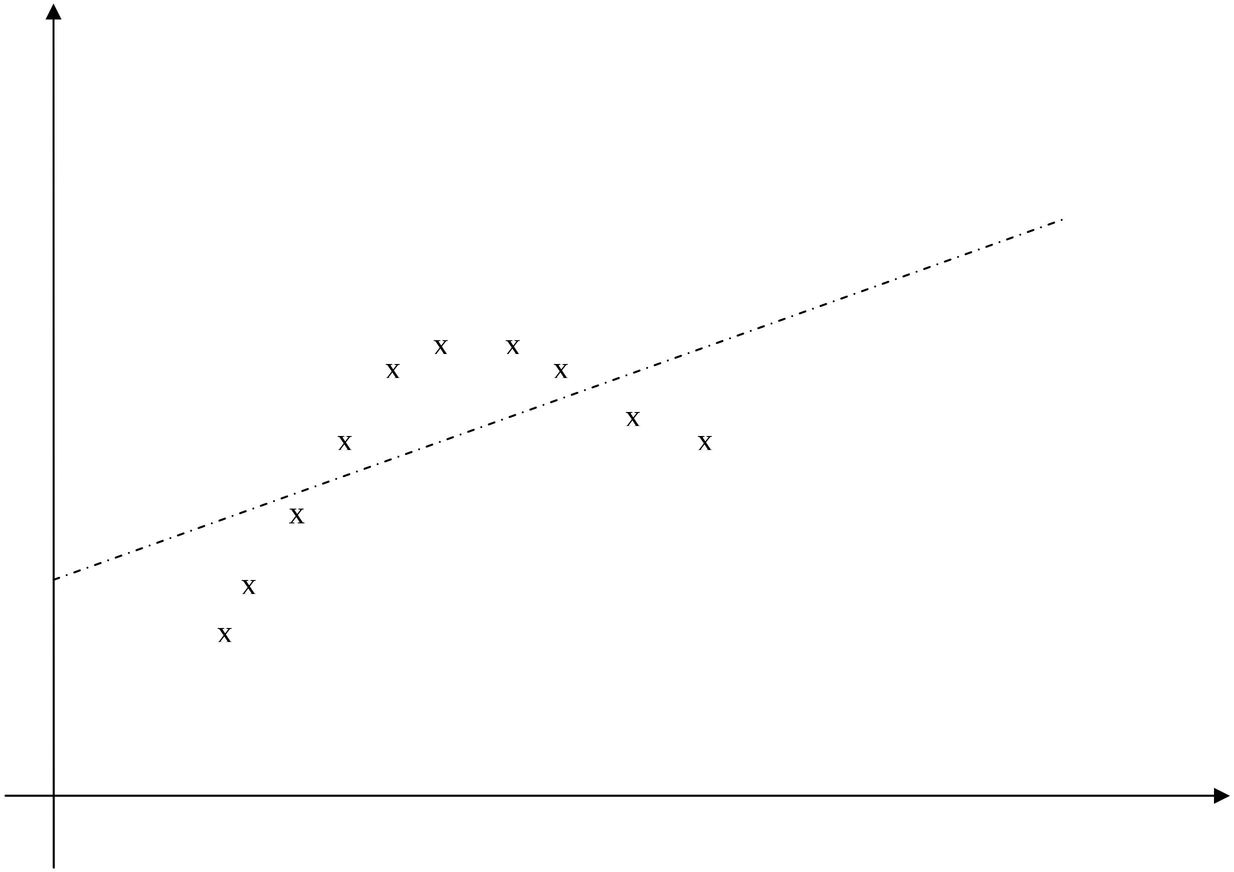
$Y_t = \delta_0 + \delta_1 P_t + \{\delta_2 C_t + u_t\} = Y_t = \delta_0 + \delta_1 P_t + v_t$

where $v_t = \delta_2 C_t + u_t$

and if $Cov(C_t, C_{t+1}) \neq 0$ then $Cov(v_t\ v_{t+1}) \neq 0$
there is autocorrelation in the error terms

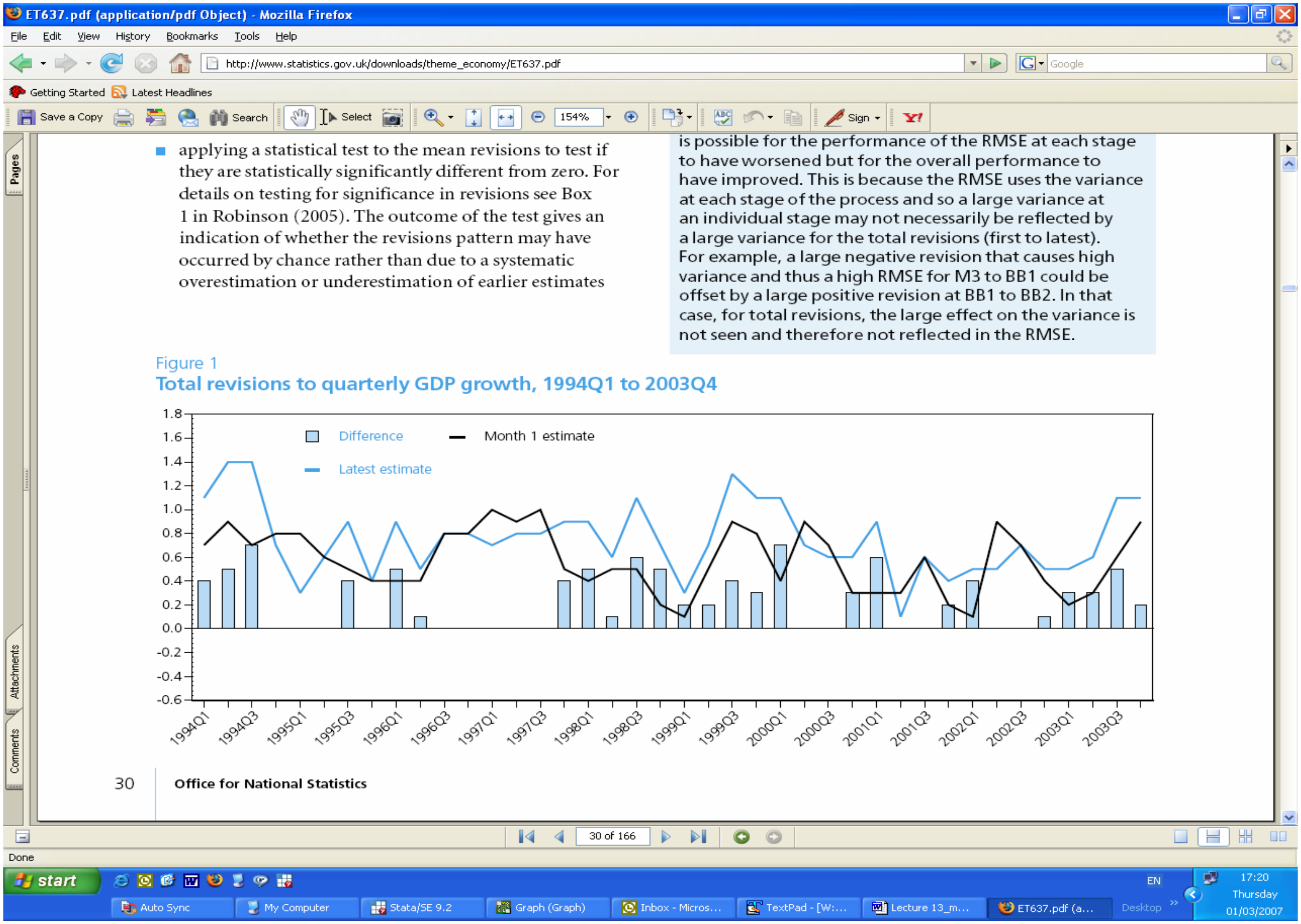
This could be caused by

- inertia in omitted economic variables (multiplier working through)
- incorrect functional form



+ve residuals followed by +ve
-ve residuals followed by –ve

This could be caused by
-inertia in omitted economic variables (multiplier working through)
-incorrect functional form
-data interpolation/revision



- applying a statistical test to the mean revisions to test if they are statistically significantly different from zero. For details on testing for significance in revisions see Box 1 in Robinson (2005). The outcome of the test gives an indication of whether the revisions pattern may have occurred by chance rather than due to a systematic overestimation or underestimation of earlier estimates

is possible for the performance of the RMSE at each stage to have worsened but for the overall performance to have improved. This is because the RMSE uses the variance at each stage of the process and so a large variance at an individual stage may not necessarily be reflected by a large variance for the total revisions (first to latest). For example, a large negative revision that causes high variance and thus a high RMSE for M3 to BB1 could be offset by a large positive revision at BB1 to BB2. In that case, for total revisions, the large effect on the variance is not seen and therefore not reflected in the RMSE.

Figure 1
Total revisions to quarterly GDP growth, 1994Q1 to 2003Q4

when series are revised they tend to be revised in blocks of time, so any residuals in a block will tend to be correlated

In order to deal with this, need to make an assumption about how to model this correlation in the residuals

Can model this process

$$u_t = \rho u_{t-1} + e_t \qquad -1 <= \rho <= 1$$

Can model this process

$$u_t = \rho u_{t-1} + e_t \qquad -1 <= \rho <= 1$$

The coefficient ρ indicates the strength of the relationship between the residuals

Can model this process

$$u_t = \rho u_{t-1} + e_t \qquad\qquad -1 <= \rho <= 1$$

The coefficient ρ indicates the strength of the relationship between the residuals

If ρ=0 then residuals are not autocorrelated

Can model this process

$$u_t = \rho u_{t-1} + e_t \qquad -1 \le \rho \le 1$$

The coefficient ρ indicates the strength of the relationship between the residuals

If ρ=0 then residuals are not autocorrelated

If ρ>0 then residuals are positively autocorrelated
(+ve residuals tend to be followed by +ve residuals and –ve by –ve)

Can model this process

$$u_t = \rho u_{t-1} + e_t \qquad -1 <= \rho <= 1$$

The coefficient ρ indicates the strength of the relationship between the residuals

If $\rho=0$ then residuals are not autocorrelated

If $\rho>0$ then residuals are positively autocorrelated
(+ve residuals tend to be followed by +ve residuals and –ve by –ve)

If $\rho<0$ then residuals are negatively autocorrelated
(+ve residuals tend to be followed by -ve residuals and –ve by +ve)

Does it matter?

Does it matter?

We know OLS estimation gives

$$\hat{\beta} = \frac{Cov(X,y)}{Var(X)} = \beta + \frac{Cov(X,u)}{Var(X)}$$

(sub in y= a+ BX+u)

Does it matter?

We know OLS estimation gives

$$\hat{\beta} = \frac{Cov(X, y)}{Var(X)} = \beta + \frac{Cov(X, u)}{Var(X)}$$

And autocorrelation does not affect the assumption that Cov(X,u)=0 so OLS remains unbiased in the presence of autocorrelation

But …

Can show that variance of the OLS estimate of β in presence of autocorrelation is

$$Var(\hat{\beta}^{OLS}_{autocorr}) = Var\left(\frac{Cov(X, y \quad)}{Var(X)}\right)$$

Can show that variance of the OLS estimate of β in presence of autocorrelation is

$$Var(\hat{\beta}^{OLS}_{autocorr}) = Var\left(\frac{Cov(X, y)}{Var(X)}\right)$$

$$= Var\left(\beta + \frac{Cov(X, u)}{Var(X)}\right)$$

Can show that variance of the OLS estimate of β in presence of autocorrelation is

$$Var(\hat{\beta}^{OLS}_{autocorr}) = Var\left(\frac{Cov(X, y)}{Var(X)}\right)$$

$$= Var\left(\beta + \frac{Cov(X, u)}{Var(X)}\right)$$

$$= \frac{\sigma_u^2}{N * Var(X)} + 2\frac{\sigma_u^2}{(N * Var(X))^2} \sum_{t=1}^{T-1} \sum_{j=1}^{T-t} \rho^j X_t X_{t+j}$$

(where j and t are just different time periods within the period covered by the sample t = 1, 2…T time periods)

Can show that variance of the OLS estimate of β in presence of autocorrelation is

$$Var(\hat{\beta}^{OLS}_{autocorr}) = Var\left(\frac{Cov(X, y)}{Var(X)}\right)$$

$$= Var\left(\beta + \frac{Cov(X, u)}{Var(X)}\right)$$

$$= \frac{\sigma_u^2}{N * Var(X)} + 2\frac{\sigma_u^2}{(N * Var(X))^2} \sum_{t=1}^{T-1} \sum_{j=1}^{T-t} \rho^j X_t X_{t+j}$$

(where j and t are just different time periods within the period covered by the sample t = 1, 2…T time periods)

The first term is just the equation for the OLS variance of the coefficient estimate in the absence of autocorrelation so

Can show that variance of the OLS estimate of β in presence of autocorrelation is

$$Var(\hat{\beta}^{OLS}_{autocorr}) = Var\left(\frac{Cov(X, y\ \ )}{Var(X)}\right)$$

$$= Var\left(\beta + \frac{Cov(X,u)}{Var(X)}\right)$$

$$= \frac{\sigma_u^2}{N * Var(X)} + 2\frac{\sigma_u^2}{(N * Var(X))^2}\sum_{t=1}^{T-1}\sum_{j=1}^{T-t}\rho^j X_t X_{t+j}$$

(where j and t are just different time periods within the period covered by the sample t = 1, 2…T time periods)

The first term is just the equation for the OLS variance of the coefficient estimate in the absence of autocorrelation so

$$Var\left(\hat{\beta}^{OLS}_{auto}\right) = Var\left(\hat{\beta}^{OLS}_{uncorrected}\right) + 2F(\rho)$$

and F(ρ) is the complex second term which depends on the value of ρ
Given

$$\frac{\sigma_u^2}{N * Var(X)} + 2\frac{\sigma_u^2}{(N * Var(X))^2} \sum_{t=1}^{T-1} \sum_{j=1}^{T-t} \rho^j X_t X_{t+j}$$

Given

$$\frac{\sigma_u^2}{N*Var(X)} + 2\frac{\sigma_u^2}{(N*Var(X))^2}\sum_{t=1}^{T-1}\sum_{j=1}^{T-t}\rho^j X_t X_{t+j}$$

If ρ=0 then the 2nd term is zero and back to the original formula for the variance

Given

$$\frac{\sigma_u^2}{N*Var(X)} + 2\frac{\sigma_u^2}{(N*Var(X))^2}\sum_{t=1}^{T-1}\sum_{j=1}^{T-t}\rho^j X_t X_{t+j}$$

If ρ=0 then the 2[nd] term is zero and back to the original formula for the variance

If $\rho \neq$ 0 then can see $Var\left(\overset{\wedge}{\beta}{}_{auto}^{OLS}\right) \neq Var\left(\overset{\wedge}{\beta}{}_{uncorrected}^{OLS}\right)$

Given

$$\frac{\sigma_u^2}{N*Var(X)}+2\frac{\sigma_u^2}{(N*Var(X))^2}\sum_{t=1}^{T-1}\sum_{j=1}^{T-t}\rho^j X_t X_{t+j}$$

If ρ=0 then the 2nd term is zero and back to the original formula for the variance

If $\rho \neq$ 0 then can see $Var\left(\hat{\beta}_{auto}^{OLS}\right) \neq Var\left(\hat{\beta}_{uncorrected}^{OLS}\right)$

Indeed if ρ > 0

(ie positive autocorrelation - most common form of autocorrelation)

Given

$$\frac{\sigma_u^2}{N*Var(X)} + 2\frac{\sigma_u^2}{(N*Var(X))^2}\sum_{t=1}^{T-1}\sum_{j=1}^{T-t}\rho^j X_t X_{t+j}$$

If ρ=0 then the 2nd term is zero and back to the original formula for the variance

If $\rho \neq$ 0 then can see $Var\left(\overset{\wedge}{\beta}{}_{auto}^{OLS}\right) \neq Var\left(\overset{\wedge}{\beta}{}_{uncorrected}^{OLS}\right)$

Indeed if ρ > 0
(ie positive autocorrelation - most common form of autocorrelation)

then $Var\left(\overset{\wedge}{\beta}{}_{auto}^{OLS}\right) > Var\left(\overset{\wedge}{\beta}{}_{uncorrected}^{OLS}\right)$

Given

$$\frac{\sigma_u^2}{N*Var(X)} + 2\frac{\sigma_u^2}{(N*Var(X))^2}\sum_{t=1}^{T-1}\sum_{j=1}^{T-t}\rho^j X_t X_{t+j}$$

If ρ=0 then the 2nd term is zero and back to the original formula for the variance

If $\rho \neq$ 0 then can see $Var\left(\overset{\wedge}{\beta}{}^{OLS}_{auto}\right) \neq Var\left(\overset{\wedge}{\beta}{}^{OLS}_{uncorrected}\right)$

Indeed if ρ > 0
(ie positive autocorrelation - most common form of autocorrelation)

then $Var\left(\overset{\wedge}{\beta}{}^{OLS}_{auto}\right) > Var\left(\overset{\wedge}{\beta}{}^{OLS}_{uncorrected}\right)$

so in general uncorrected OLS will ***underestimate*** the true variance in the presence of autocorrelation

so the t values on the OLS estimates will be ***larger*** than should be

so the t values on the OLS estimates will be ***larger*** than should be

so might conclude variables are statistically significant when they are not (type I error)

so the t values on the OLS estimates will be ***larger*** than should be

so might conclude variables are statistically significant when they are not (type I error)

- OLS may be unbiased but it is no longer efficient – in the sense that the variance of the estimates is wrong.

so the t values on the OLS estimates will be ***larger*** than should be

so might conclude variables are statistically significant when they are not (type I error)

- OLS may be unbiased but it is no longer efficient – in the sense that the variance of the estimates is wrong.

Lecture 17

- testing for autocorrelation
- correcting for autocorrelation
- Dynamic Models (short run long run effects)

Autocorrleation: Cov($u_t u_{t-1}$) ≠ 0
OLS may be unbiased but it is no longer efficient – in the sense that the variance of the estimates is wrong.

$$\frac{\sigma_u^2}{N*Var(X)} + 2\frac{\sigma_u^2}{(N*Var(X))^2}\sum_{t=1}^{T-1}\sum_{j=1}^{T-t}\rho^j X_t X_{t+j}$$

If $\rho \neq$ 0 then can see $Var\left(\hat{\beta}_{auto}^{OLS}\right) \neq Var\left(\hat{\beta}_{uncorrected}^{OLS}\right)$

If ρ > 0 (ie positive autocorrelation - most common form of autocorrelation) in general uncorrected OLS will ***underestimate*** the true variance in the presence of autocorrelation

so the t values on the OLS estimates will be ***larger*** than should be

so might conclude variables are statistically significant when they are not (type I error)

- **Testing for Autocorrelation**

Given time series data

$$Y_t = b_0 + b_1 X_t + u_t \qquad t = 1, 2\ldots T \quad (1)$$

the model may be subject to autocorrelation

Testing for Autocorrelation

Given time series data

$$Y_t = b_0 + b_1X_t + u_t \qquad t = 1, 2\ldots T \quad (1)$$

the model may be subject to autocorrelation

Accepted way of testing is to specify a functional form for the persistence (correlation) in the residuals over time and test to see whether this specification is seen in the data.

Suppose we assume that the residuals follow a 1st Order Autoregressive process AR(1)

$$u_t = \rho u_{t-1} + e_t \qquad -1 <= \rho <= 1 \quad (2)$$

So movement in current value of residuals is related to last period's value and a current period random component (e_t)

Suppose we assume that the residuals follow a 1st Order Autoregressive process AR(1)

$$u_t = \rho u_{t-1} + e_t \qquad -1 <= \rho <= 1 \quad (2)$$

So movement in current value of residuals is related to last period's value and a current period random component (e_t)

Good practice to test by simply plotting the residuals over time and looking at the pattern.

Suppose we assume that the residuals follow a 1st Order Autoregressive process AR(1)

$$u_t = \rho u_{t-1} + e_t \qquad -1 <= \rho <= 1 \quad (2)$$

So movement in current value of residuals is related to last period's value and a current period random component (e_t)

Good practice to test by simply plotting the residuals over time and looking at the pattern.

+ve AR(1) autocorrelation means +ve residual values are followed by +ve values (and –ve by –ve), $\rho > 0$

Suppose we assume that the residuals follow a 1st Order Autoregressive process AR(1)

$$u_t = \rho u_{t-1} + e_t \qquad -1 <= \rho <= 1$$

So movement in current value of residuals is related to last period's value and a current period random component (e_t)

Good practice to test by simply plotting the residuals over time and looking at the pattern.

+ve AR(1) autocorrelation means +ve residual values are followed by +ve values (and –ve by –ve), $\rho > 0$

-ve AR(1) autocorrelation means +ve residual values are followed by -ve values (and –ve by +ve), $\rho < 0$

Good practice also to use the **standardised residuals** when do this

Good practice also to use the **standardised residuals** when do this

$$u_{s\tan dardised} = \frac{\hat{u}_t}{s}$$

Good practice also to use the **standardised residuals** when do this

$$u_{s\tan dardised} = \frac{\hat{u}_t}{s}$$

ie residuals made scale invariant by dividing through by the standard error of the equation,

Good practice also to use the **standardised residuals** when do this

$$u_{s\tan dardised} = \frac{\hat{u}_t}{s}$$

ie residuals made scale invariant by dividing through by the standard error of the equation,

$$s = \sqrt{RSS / T - k}$$

Good practice also to use the **standardised residuals** when do this

$$u_{s\tan dardised} = \frac{\hat{u}_t}{s}$$

ie residuals made scale invariant by dividing through by the standard error of the equation,

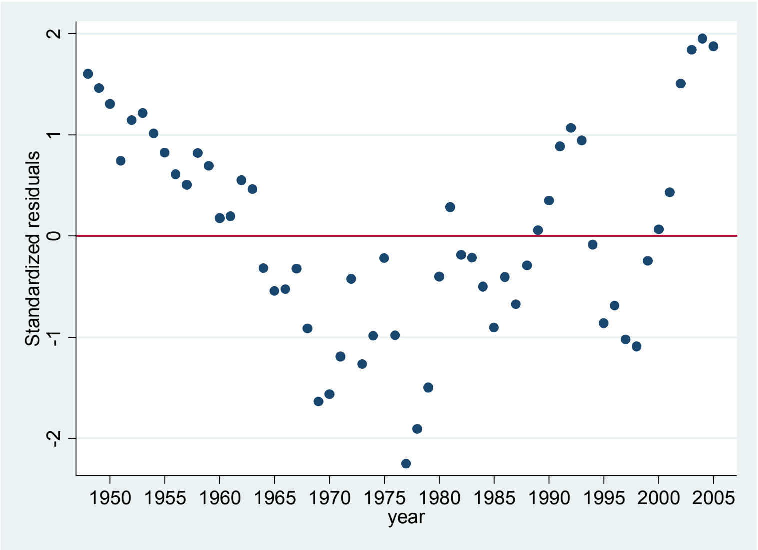
$$s = \sqrt{RSS / T - k}$$

scale invariance means can compare across different specifications or models using variables with different units of measurement or even compare residuals from models with different variables

```
. reg cons inc

      Source |       SS       df       MS              Number of obs =      58
-------------+------------------------------           F(  1,    56) =16998.53
       Model |  2.5426e+12     1  2.5426e+12           Prob > F      =  0.0000
    Residual |  8.3762e+09    56   149575459           R-squared     =  0.9967
-------------+------------------------------           Adj R-squared =  0.9967
       Total |  2.5509e+12    57  4.4753e+10           Root MSE      =   12230

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .8447215    .006479   130.38   0.000     .8317425    .8577005
       _cons |  -2376.815   4398.976    -0.54   0.591    -11189.02    6435.394
------------------------------------------------------------------------------
```

Do graphical inspection using **standardised residuals** ie divide residuals by standard error of regression, s, (given by Root MSE in Stata output above)

```
. predict stanres, rstandard        /* command to get standardised residuals */
two (scatter stanres year, yline(0) xlabel(1950(5)2006) )
```



Standardised residuals confirm general (positive) autocorrelation pattern in residuals as before. Only difference is values on y axis have changed (since are now scale invariant)

However this is useful (and advised) but is not a formal test for the problem.

However this is useful (and advised) but is not a formal test for the problem.

One common statistical for presence of AR(1) in the residuals is to compute **Durbin-Watson** statistic

$$DW = \frac{\sum_{t=2}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T}\hat{u}_t^2}$$

However this is useful (and advised) but is not a formal test for the problem.

One common statistical for presence of AR(1) in the residuals is to compute **Durbin-Watson** statistic

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2}$$

where u_t are residuals saved from OLS estimation of the model

$$Y_t = b_0 + b_1 X_t + u_t$$

However this is useful (and advised) but is not a formal test for the problem.

One common statistical for presence of AR(1) in the residuals is to compute **Durbin-Watson** statistic

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2}$$

where u_t are residuals saved from OLS estimation of the model

$$Y_t = b_0 + b_1 X_t + u_t$$

Can show that DW =2(1-$\hat{\rho}$)

However this is useful (and advised) but is not a formal test for the problem.

One common statistical for presence of AR(1) in the residuals is to compute **Durbin-Watson** statistic

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2}$$

where u_t are residuals saved from OLS estimation of the model

$$Y_t = b_0 + b_1 X_t + u_t$$

Can show that DW =2(1-$\hat{\rho}$)

where $\hat{\rho}$ is taken from an OLS regression of $\hat{u}_t = \rho \hat{u}_{t-1} + e_t$

Given DW = $2(1-\hat{\rho})$

if $\hat{\rho} = 0$ then DW = 2 and residuals are not autocorrelated

if $\hat{\rho} \rightarrow 1$ then DW $\rightarrow$ 0 and $\exists$ +ve autocorrelation

if $\hat{\rho} \rightarrow -1$ then DW $\rightarrow$ 4 and $\exists$ -ve autocorrelation

So how close to 2 does DW have to be before we can be confident of accepting null hypothesis of no autocorrelation?

Turns out that there are 2 critical values than need to compare estimated DW against: an “upper” value and a “lower” value

Turns out that there are 2 critical values than need to compare estimated DW against: an "upper" value and a "lower" value

Reject Null: there is +ve A/C.	Test inconclusive	accept Null (No Autocorrelation)		Test inconclusive	Reject Null: there is –ve A/C.
0	DW_{low}	DW_{upper}	2	$4\text{-}DW_{upper}$	$4\text{-}DW_{low}$

Turns out that there are 2 critical values than need to compare estimated DW against: an “upper” value and a “lower” value

Reject Null: there is +ve A/C.	Test inconclusive	accept Null (No Autocorrelation)		Test inconclusive	Reject Null: there is –ve A/C.
0	DW_{low}	DW_{upper}	2	$4\text{-}DW_{upper}$	$4\text{-}DW_{low}$

So

if $DW < DW_{lower}$ conclude ∃ +ve autocorrelation

Turns out that there are 2 critical values than need to compare estimated DW against: an “upper” value and a “lower” value

Reject Null: there is +ve A/C.	Test inconclusive	accept Null (No Autocorrelation)	Test inconclusive	Reject Null: there is –ve A/C.
0 DW_{low}	DW_{upper}	2	$4\text{-}DW_{upper}$	$4\text{-}DW_{low}$

So

if $DW < DW_{lower}$ conclude $\exists$ +ve autocorrelation

if $DW_{upper} < DW < 4\text{-}DW_{upper}$ residuals are not autocorrelated

Turns out that there are 2 critical values than need to compare estimated DW against: an “upper” value and a “lower” value

Reject Null: there is +ve A/C.	Test inconclusive	accept Null (No Autocorrelation)		Test inconclusive	Reject Null: there is –ve A/C.
0 DW_{low}	DW_{upper}	2		$4\text{-}DW_{upper}$	$4\text{-}DW_{low}$

So

if $DW < DW_{lower}$ conclude ∃ +ve autocorrelation

if $DW_{upper} < DW < 4\text{-}DW_{upper}$ residuals are not autocorrelated

if $DW_{lower} < DW < DW_{upper}$ test is inconclusive

Turns out that there are 2 critical values than need to compare estimated DW against: an “upper” value and a “lower” value

Reject Null: there is +ve A/C.	Test inconclusive	accept Null (No Autocorrelation)		Test inconclusive	Reject Null: there is –ve A/C.
0 DW_{low}		DW_{upper}	2	$4\text{-}DW_{upper}$	$4\text{-}DW_{low}$

So

if $DW < DW_{lower}$ conclude ∃ +ve autocorrelation

if $DW_{upper} < DW < 4\text{-}DW_{upper}$ residuals are not autocorrelated

if $DW_{lower} < DW < DW_{upper}$ test is inconclusive

(and similarly for negative autocorrelation)

Turns out that there are 2 critical values than need to compare estimated DW against: an “upper” value and a “lower” value

Reject Null: there is +ve A/C.	Test inconclusive	accept Null (No Autocorrelation)	Test inconclusive	Reject Null: there is –ve A/C.
0	DW_{low}	DW_{upper} 2	$4\text{-}DW_{upper}$	$4\text{-}DW_{low}$

So

if $DW < DW_{lower}$ conclude ∃ +ve autocorrelation

if $DW_{upper} < DW < 4\text{-}DW_{upper}$ residuals are not autocorrelated

if $DW_{lower} < DW < DW_{upper}$ test is inconclusive

(and similarly for negative autocorrelation)
Note: Unfortunately the critical values vary with sample size and number of RHS variables **excluding the constant**

```
/* Example: DETECTION OF AUTOCORRELATION */

. tsset year
        time variable:  year, 1948 to 2005

. regdw cons income

      Source |       SS       df       MS              Number of obs =      58
-------------+------------------------------           F(  1,    56) =16998.53
       Model |  2.5426e+12     1  2.5426e+12           Prob > F      =  0.0000
    Residual |  8.3762e+09    56   149575459           R-squared     =  0.9967
-------------+------------------------------           Adj R-squared =  0.9967
       Total |  2.5509e+12    57  4.4753e+10           Root MSE      =   12230

------------------------------------------------------------------------------
 consumption |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .8447215    .006479   130.38   0.000     .8317425    .8577005
       _cons |  -2376.815   4398.976    -0.54   0.591    -11189.02    6435.394
------------------------------------------------------------------------------
```

Durbin-Watson Statistic = .2516256
From Tables, given T=58 and K'=1, DW_{low} = 1.55 and DW_{high} = 1.62
(k'=no. rhs variables **excluding** the constant)

So estimated value is less than DWlow. Hence reject null of no autocorrelation. Accept there exists positive 1st order autocorrelation.

Problems with Durbin-Watson

1. Existence of an inconclusive region reduces the usefulness of this test

Problems with Durbin-Watson

1 Existence of an inconclusive region reduces the usefulness of this test

2. Can show DW not valid in the presence of **lagged dependent variables or endogenous variables**

Problems with Durbin-Watson

1 Existence of an inconclusive region reduces the usefulness of this test

2 Can show DW not valid in the presence of **lagged dependent variables or endogenous variables**

$$Y_t = b_0 + \lambda Y_{t-1} + b_1 X_t + u_t \qquad \text{(A)}$$

If there are lagged dependent variables could use **Durbin's h test**

$$h = \hat{\rho}\sqrt{\frac{T}{1 - TVar(\hat{\lambda})}}$$

where T = sample size (number of time periods) and var(λ) is the estimated variance of the coefficient on the lagged dependent variable from an OLS estimation of (A)

Can show that under null hypothesis of no +ve autocorrelation

h ~ Normal(0,1)

So that Pr[-1.96<= h <=1.96] = 0.95

ie 95% chance that value of h will lie between –1.96 and + 1.96

In this case if estimated h>1.96 then can reject null of no +ve autocorrelation

But: can't compute h if

$$1 - TVar(\hat{\lambda}) < 0$$

which could happen

So need alternative measure which can always be calculated.

Breusch-Godfrey Test for AR(q)

This is in fact a general test for autocorrelation of **any** order
(ie can use it to test if residuals are correlated over more than one period)

Breusch-Godfrey Test for AR(q)

This is in fact a general test for autocorrelation of **any** order
(ie can use it to test if residuals are correlated over more than one period)

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \ldots \rho_q u_{t-q} + e_t \qquad (A)$$

Breusch-Godfrey Test for AR(q)

This is in fact a general test for autocorrelation of **any** order
(ie can use it to test if residuals are correlated over more than one period)

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \ldots \rho_q u_{t-q} + e_t \qquad \text{(A)}$$

Eg quarterly data are often correlated with values 4 periods ago (ie 1 year), monthly data with the values 12 months ago, daily data with values 7 days ago

Breusch-Godfrey Test for AR(q)

This is in fact a general test for autocorrelation of **any** order
(ie can use it to test if residuals are correlated over more than one period)

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \ldots \rho_q u_{t-q} + e_t \qquad \text{(A)}$$

Eg quarterly data are often correlated with values 4 periods ago (ie 1 year), monthly data with the values 12 months ago, daily data with values 7 days ago

So test for no autocorrelation of order q amounts to test

$$H0{:}\ \rho_1 = \rho_2 = \rho_3 = \ldots \rho_q = 0 \qquad \text{in (A) above}$$

Do this as follows:

1. Estimate original model $Y_t = b_0 + b_1 X_t + u_t$

Save residuals

1. Estimate original model $Y_t = b_0 + b_1X_t + u_t$

Save residuals

2. Regress estimated residuals on lagged values up to lag q **and** all the original RHS X variables

$$\hat{u}_t = \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + \ldots \rho_q \hat{u}_{t-1} + \gamma_1 X_t + e_t$$

1. Estimate original model $Y_t = b_0 + b_1X_t + u_t$

Save residuals

2. Regress estimated residuals on lagged values up to lag q **and** all the original RHS X variables

$$\hat{u}_t = \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + ...\rho_q \hat{u}_{t-1} + \gamma_1 X_t + e_t$$

3. Either compute the F test for the joint significance of the residuals

$$\hat{u}_{t-1} ....\hat{u}_{t-q}$$

and if $\hat{F} > F_{critical}$ **reject** null of no q order autocorrelation

1. Estimate original model $Y_t = b_0 + b_1 X_t + u_t$

Save residuals

2. Regress estimated residuals on lagged values up to lag q **and** all the original RHS X variables

$$\hat{u}_t = \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + \ldots \rho_q \hat{u}_{t-1} + \gamma_1 X_t + e_t$$

4. Either compute the F test for the joint significance of the residuals

$$\hat{u}_{t-1} \ldots \hat{u}_{t-q}$$

and if $\hat{F} > F_{critical}$ **reject** null of no q order autocorrelation
or

compute $(N-q)*R^2_{auxillary} \sim \chi^2_{(q)}$

if estimated $\chi^2 > \chi^2_{critical}$ again reject null of no q order A/c.

1. Estimate original model $Y_t = b_0 + b_1X_t + u_t$

Save residuals

2. Regress estimated residuals on lagged values up to lag q **and** all the original RHS X variables

$$\hat{u}_t = \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + ... \rho_q \hat{u}_{t-1} + \gamma_1 X_t + e_t$$

5. Either compute the F test for the joint significance of the residuals

$$\hat{u}_{t-1} .... \hat{u}_{t-q}$$

and if $\hat{F} > F_{critical}$ **reject** null of no q order autocorrelation

or

compute $(N-q)*R^2_{auxillary} \sim \chi^2_{(q)}$

if estimated $\chi^2 > \chi^2_{critical}$ again reject null of no q order A/c.
(intuitively if lagged residuals are significant this gives a high R^2)

Useful test since

a)generalises to any order autocorrelation wish to test

Useful test since

a) generalises to any order autocorrelation wish to test

b)is robust to inclusion of lagged dependent variables

Useful test since

a) generalises to any order autocorrelation wish to test

b) is robust to inclusion of lagged dependent variables

But

1. Since this is a test of joint significance may not be able to distinguish **which** lagged residual is important

Useful test since

a) generalises to any order autocorrelation wish to test

b) is robust to inclusion of lagged dependent variables

But
1. Since this is a test of joint significance may not be able to distinguish **which** lagged residual is important

2. Test is only valid asymptotically (ie in large samples)

Example: Breusch-Godfrey Test For Autocorrelation

```
. reg cons income

      Source |       SS       df       MS              Number of obs =      61
-------------+------------------------------           F(  1,    59) =24523.30
       Model |  3.8562e+12     1  3.8562e+12           Prob > F      =  0.0000
    Residual |  9.2775e+09    59   157245264           R-squared     =  0.9976
-------------+------------------------------           Adj R-squared =  0.9976
       Total |  3.8655e+12    60  6.4424e+10           Root MSE      =   12540
------------------------------------------------------------------------------
        cons |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .8474867   .0054118   156.60   0.000     .8366576    .8583157
       _cons |  -1849.363   4109.675    -0.45   0.654     -10072.8    6374.078
------------------------------------------------------------------------------
```

Now save residuals and lag them

```
. predict res, resid

. sort year
g res1=l.res
(1 missing value generated)
. g res2=l2.res
(2 missing values generated)
. g res3=l3.res
(3 missing values generated)
. g res4=l4.res
(4 missing values generated)
```

/* note number of missing values, since u_{t-1} does not exist for 1st observation in data set (no information before 1955)... u_{t-4} does not exist for first 4 observations in data set */

/* Now do Breusch-Godfrey test for residuals of AR(1) manually */

```
. reg reshat res1

      Source |       SS       df       MS              Number of obs =      57
-------------+------------------------------           F(  1,    55) =  167.21
       Model |  6.0235e+09     1  6.0235e+09           Prob > F      =  0.0000
    Residual |  1.9813e+09    55  36023373.4           R-squared     =  0.7525
-------------+------------------------------           Adj R-squared =  0.7480
```

```
       Total |  8.0048e+09    56   142943278           Root MSE      =  6001.9
------------------------------------------------------------------------------
      reshat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        res1 |   .8735296   .0675529    12.93   0.000     .7381505    1.008909
       _cons |  -1.550637   795.3959    -0.00   0.998     -1595.56    1592.458
------------------------------------------------------------------------------
```

Regress residuals from original regression on residual lagged one period and original rhs variables (doesn't matter if include constant or not – results the same asymptotically)

Given the output from this auxiliary regression

compute $(N-q)*R^2_{aux} = (58-1)*.7525 = 42.9$

(Note N-q = number of observations in auxiliary regression)

This statistic has a chi-squared distribution with 1 degree of freedom (equal to the number of lags tested)

From tables $\chi^2_{critical}$ at 5% level = 3.84

So estimated $\chi^2 > \chi^2_{critical}$, so **reject** null that residuals are **uncorrelated** over one year to the next.

```
For test of AR(4)

. reg reshat res1 res2 res3 res4

      Source |       SS       df       MS              Number of obs =      54
-------------+------------------------------           F(  4,    49) =   40.44
       Model |  5.6161e+09     4  1.4040e+09           Prob > F      =  0.0000
    Residual |  1.7013e+09    49  34720229.7           R-squared     =  0.7675
-------------+------------------------------           Adj R-squared =  0.7485
       Total |  7.3174e+09    53   138064871           Root MSE      =  5892.4
------------------------------------------------------------------------------
      reshat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        res1 |   1.201768   .1408626     8.53   0.000     .9186938    1.484842
        res2 |  -.5499785   .2159075    -2.55   0.014    -.9838609   -.1160961
```

```
        res3 |   .3375441   .2161457     1.56   0.125    -.0968169    .7719051
        res4 |  -.1504679   .1424541    -1.06   0.296    -.4367402    .1358044
       _cons |   -11.1964   811.6147    -0.01   0.989    -1642.197    1619.804
------------------------------------------------------------------------------
```

and $(N-q)*R^2_{aux} = 54*.7675 = 41.4$

This statistic has a chi-squared distribution with 4 degree of freedom (equal to the number of lags tested)

From tables $\chi^2_{critical\ (4)}$ at 5% level = 9.5

So estimated $\chi^2 > \chi^2_{critical}$, so again **reject** null that residuals are **not correlated** over 4 year periods

```
/* Stata will do this automatically */

. bgtest                /* default is test for AR(1)   */
Breusch-Godfrey LM statistic:  43.49101  Chi-sq( 1)  P-value =  4.3e-11

. bgtest, lags(4)
Breusch-Godfrey LM statistic:  41.74953  Chi-sq( 4)  P-value =  1.9e-08
```

Remember the test is a test of the **joint** significance of the residuals, so it may not be able to pinpoint the exact lag structure. In the example above, inspection of the t statistics on each lagged residual in the auxiliary regression suggests that the 1^{st} and 2^{nd} order residuals are doing all the work, so that a test for AR(2) would seem more appropriate.

With monthly data may wish to test up to AR(12) – observations one year ago may influence current value, particularly if there is a seasonal pattern to the data, (try using the data set bop.dta to test this)

Similarly with quarterly data may wish to test for AR(4)

Autocorrelation

exists when this assumption no longer holds, so that

$Cov(u_t\, u_{t-j}) \neq 0$ for any (unspecified) lag length j

and in general uncorrected OLS will ***underestimate*** the true variance in the presence of autocorrelation so t values will tend to be **larger** than they should be

Test for presence using **Breusch-Godfrey** test

$$\hat{u}_t = \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + \ldots \rho_q \hat{u}_{t-1} + \gamma_1 X_t + e_t$$

Either compute the F test for the joint significance of the residuals

$$\hat{u}_{t-1} \ldots \hat{u}_{t-q}$$

and if $\hat{F} > F_{critical}$ **reject** null of no q order autocorrelation

or

compute $(N-q)*R^2_{auxillary} \sim \chi^2_{(q)}$

if estimated $\chi^2 > \chi^2_{critical}$ again reject null of no q order A/c.

What to do about autocorrelation?

If tests show it exists then

What to do about autocorrelation?

If tests show it exists then

- try a different functional form,

What to do about autocorrelation?

If tests show it exists then

- try a different functional form,

- add/delete variables

What to do about autocorrelation?

If tests show it exists then

- try a different functional form,

- add/delete variables

- increase the interval between observations (eg monthly to quarterly)

What to do about autocorrelation?

If tests show it exists then

- try a different functional form,

- add/delete variables

- increase the interval between observations (eg monthly to quarterly)

- fix up the standard errors

Example: Model Misspecification

Sometimes autocorrelation in residuals can be caused by incorrect functional form in your model or (effectively the same thing) the omission of relevant variables

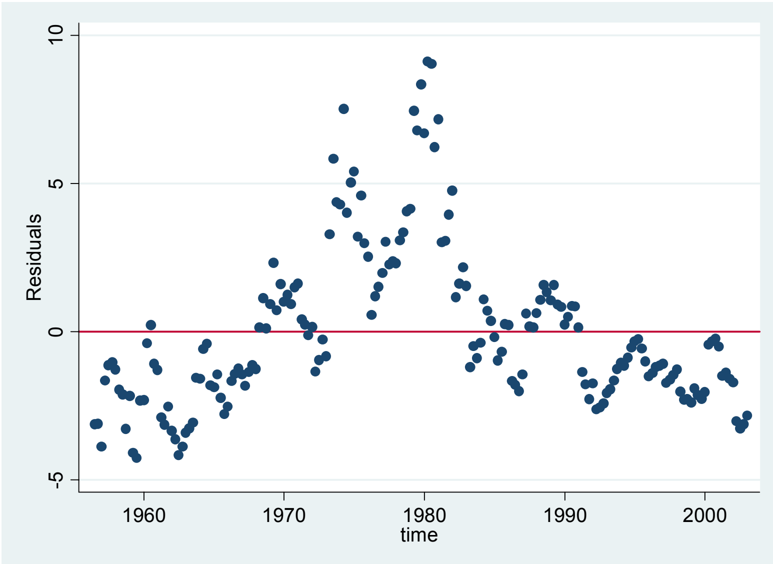
The data set *gdpus.dta* contains quarterly time series data on US GDP growth and inflation over the period 1956:q1 to 2002:q4

A simple regression of inflation rate on 1-period lagged growth rate of GDP gives (using lagged growth to try and deal with possible endogeneity concerns)

```
. tsset TIME
        time variable:  TIME, 1 to 188

. reg  usinf dgdp1

      Source |       SS       df       MS              Number of obs =     187
-------------+------------------------------           F(  1,   185) =   27.19
       Model |  200.809075     1  200.809075           Prob > F      =  0.0000
    Residual |  1366.54753   185  7.38674343           R-squared     =  0.1281
-------------+------------------------------           Adj R-squared =  0.1234
       Total |  1567.35661   186  8.42664844           Root MSE      =  2.7179
------------------------------------------------------------------------------
       usinf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dgdp1 |  -.4748904   .0910811    -5.21   0.000    -.6545815   -.2951992
       _cons |   5.357876   .2999389    17.86   0.000     4.766135    5.949616
------------------------------------------------------------------------------

. bgtest
Breusch-Godfrey LM statistic:   161.743  Chi-sq( 1)  P-value =  4.7e-37
```

and a graph of the residuals also shows them to be (positively) autocorrelated

```
predict res, resid
two (scatter res time, yline(0) )
```



Now suppose decide to include the 1 period lag of inflation rate on the right hand side

```
. g usinf1=usinf[_n-1]
(1 missing value generated)

. reg  usinf dgdp1 usinf1

      Source |       SS       df       MS              Number of obs =     187
-------------+------------------------------           F(  2,   184) = 1030.67
       Model |  1438.91574     2  719.457868           Prob > F      =  0.0000
    Residual |  128.440873   184  .698048222           R-squared     =  0.9181
-------------+------------------------------           Adj R-squared =  0.9172
       Total |  1567.35661   186  8.42664844           Root MSE      =  .83549

------------------------------------------------------------------------------
       usinf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       dgdp1 |   .0832063   .0309768     2.69   0.008      .022091    .1443216
      usinf1 |   .9831361   .0233441    42.11   0.000     .9370796    1.029193
       _cons |  -.1340872   .1597084    -0.84   0.402    -.4491824    .1810081
------------------------------------------------------------------------------
```

```
. bgtest
Breusch-Godfrey LM statistic:  .2558797  Chi-sq( 1)  P-value =   .613
. predict resa, resid
```

Now the pattern of autocorrelation seems to have become much less noticeable in the new specification compared to the original, (though there still may be endogeneity bias in the OLS estimates of the coefficients)

Can, in principle, manipulate the data to remove autocorrelation from the residuals.

Can, in principle, manipulate the data to remove autocorrelation from the residuals.

Suppose you had

$$Y_t = b_0 + b_1 X_t + u_t \qquad (1)$$

and **assumed** AR(1) behaviour in the residuals

$$u_t = \rho u_{t-1} + e_t \qquad (2)$$

Can, in principle, manipulate the data to remove autocorrelation from the residuals.

Suppose you had

$$Y_t = b_0 + b_1X_t + u_t \qquad (1)$$

and **assumed** AR(1) behaviour in the residuals

$$u_t = \rho u_{t-1} + e_t \qquad (2)$$

(1) $\Rightarrow$ $$Y_{t-1} = b_0 + b_1X_{t-1} + u_{t-1} \qquad (3)$$

(ie relationship holds in any time period)

Can, in principle, manipulate the data to remove autocorrelation from the residuals.

Suppose you had

$$Y_t = b_0 + b_1X_t + u_t \qquad (1)$$

and **assumed** AR(1) behaviour in the residuals

$$u_t = \rho u_{t-1} + e_t \qquad (2)$$

(1) $\Rightarrow$ $$Y_{t-1} = b_0 + b_1X_{t-1} + u_{t-1} \qquad (3)$$

(ie relationship holds in any time period)

Multiplying (3) by ρ

$$\rho Y_{t-1} = \rho b_0 + \rho b_1X_{t-1} + \rho u_{t-1} \qquad (4)$$

(1) – (4)

On left hand side: $Y_t - \rho Y_{t-1}$

On left hand side: $Y_t - \rho Y_{t-1}$

On right hand side: $b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$

On left hand side: $Y_t - \rho Y_{t-1}$

On right hand side: $b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$

so

$$Y_t - \rho Y_{t-1} = b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$$

On left hand side: $Y_t - \rho Y_{t-1}$

On right hand side: $b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$

so

$$Y_t - \rho Y_{t-1} = b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$$

Or taking ρY_{t-1} to the other side

$$Y_t = b_0 - \rho b_0 + \rho Y_{t-1} + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$$

On left hand side: $Y_t - \rho Y_{t-1}$

On right hand side: $b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$

so

$$Y_t - \rho Y_{t-1} = b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$$

Or taking ρY_{t-1} to the other side

$$Y_t = b_0 - \rho b_0 + \rho Y_{t-1} + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$$

or

$$Y_t = (b_0 - \rho b_0) + \rho Y_{t-1} + b_1 X_t - \rho b_1 X_{t-1} + e_t \qquad (5)$$

On left hand side: $Y_t - \rho Y_{t-1}$

On right hand side: $b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$

so

$$Y_t - \rho Y_{t-1} = b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$$

Or taking ρY_{t-1} to the other side

$$Y_t = b_0 - \rho b_0 + \rho Y_{t-1} + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$$

or

$$Y_t = (b_0 - \rho b_0) + \rho Y_{t-1} + b_1 X_t - \rho b_1 X_{t-1} + e_t \qquad (5)$$

Since $e_t = u_t - \rho u_{t-1}$ from (2)

then if estimate (5) by OLS there should be no autocorrelation.

On left hand side: $Y_t - \rho Y_{t-1}$

On right hand side: $b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$

so

$$Y_t - \rho Y_{t-1} = b_0 - \rho b_0 + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$$

Or taking ρY_{t-1} to the other side

$$Y_t = b_0 - \rho b_0 + \rho Y_{t-1} + b_1 X_t - \rho b_1 X_{t-1} + u_t - \rho u_{t-1}$$

or

$$Y_t = (b_0 - \rho b_0) + \rho Y_{t-1} + b_1 X_t - \rho b_1 X_{t-1} + e_t \qquad (5)$$

Since $e_t = u_t - \rho u_{t-1}$ from (2)

then if estimate (5) by OLS there should be no autocorrelation.

This is called Feasible Generalised Least Squares (FGLS)

So why not do this?

So why not do this?

Ok **if** assumption (2) about the type of autocorrelation is correct. If not may make things worse.

So why not do this?

Ok **if** assumption (2) about the type of autocorrelation is correct. If not may make things worse.

Also if the X variables are endogenous then this technique is not valid

So why not do this?

Ok **if** assumption (2) about the type of autocorrelation is correct. If not may make things worse.

Also if the X variables are endogenous then this technique is not valid

So If don't know exact form of autocorrelation (highly likely) it may be preferable to fix up the OLS standard errors so they are no longer biased but remain inefficient - in the sense that if you knew the precise form of autocorrelation you could write down the exact formula for the standard errors, but you don't so do the best you can instead.

So why not do this?

Ok **if** assumption (2) about the type of autocorrelation is correct. If not may make things worse.

Also if the X variables are endogenous then this technique is not valid

So If don't know exact form of autocorrelation (highly likely) it may be preferable to fix up the OLS standard errors so they are no longer biased but remain inefficient - in the sense that if you knew the precise form of autocorrelation you could write down the exact formula for the standard errors, but you don't so do the best you can instead.

Newey-West standard errors do this and are valid in presence of lagged dependent variables and endogenous X variables **if** have large sample (ie fix-up is only valid asymptotically though it has been used on sample sizes of around 50).

In absence of autocorrelation we know OLS estimate of variance on any coefficient is

$$Var(\hat{\beta}_{ols}) = \frac{s^2_u}{N * Var(X)}$$

In presence of autocorrelation, can show the Newey-West standard errors (unbiased but inefficient) are

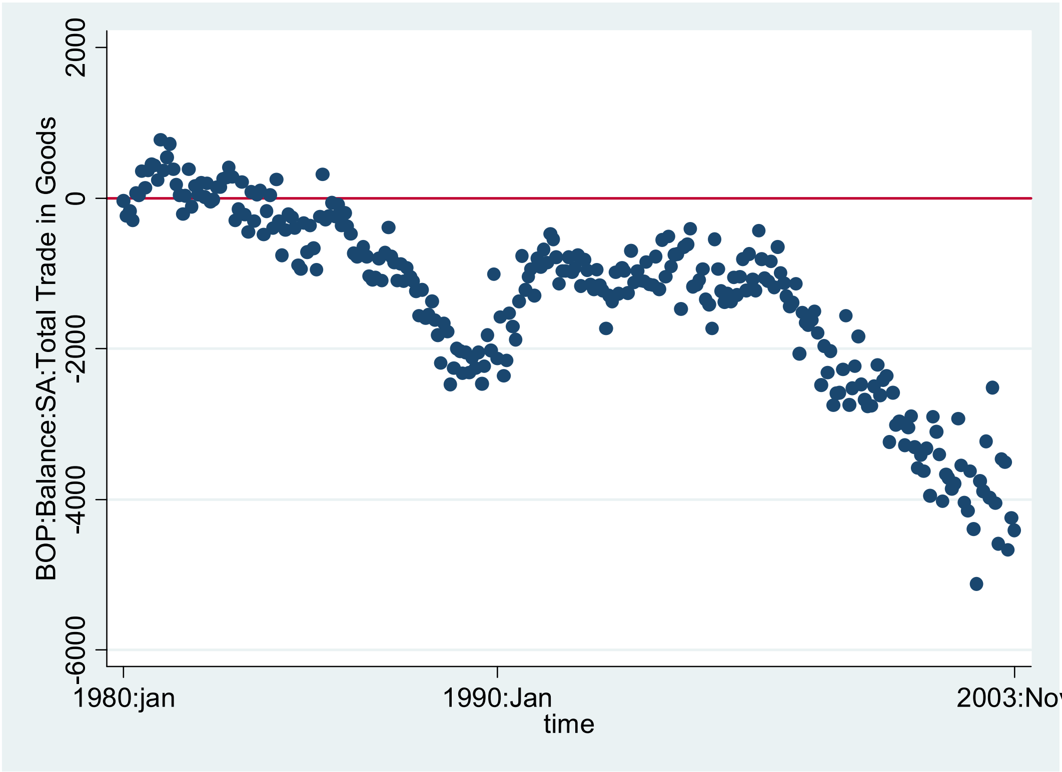
$$Var(\hat{\beta}_{ols}) = \frac{Var(\hat{\beta}_{ols})}{s_u} * v$$

where v is a (complicated) function of the maximum number of lags you believe could be correlated with current residuals (in annual data 1 or 2 lags should be enough)

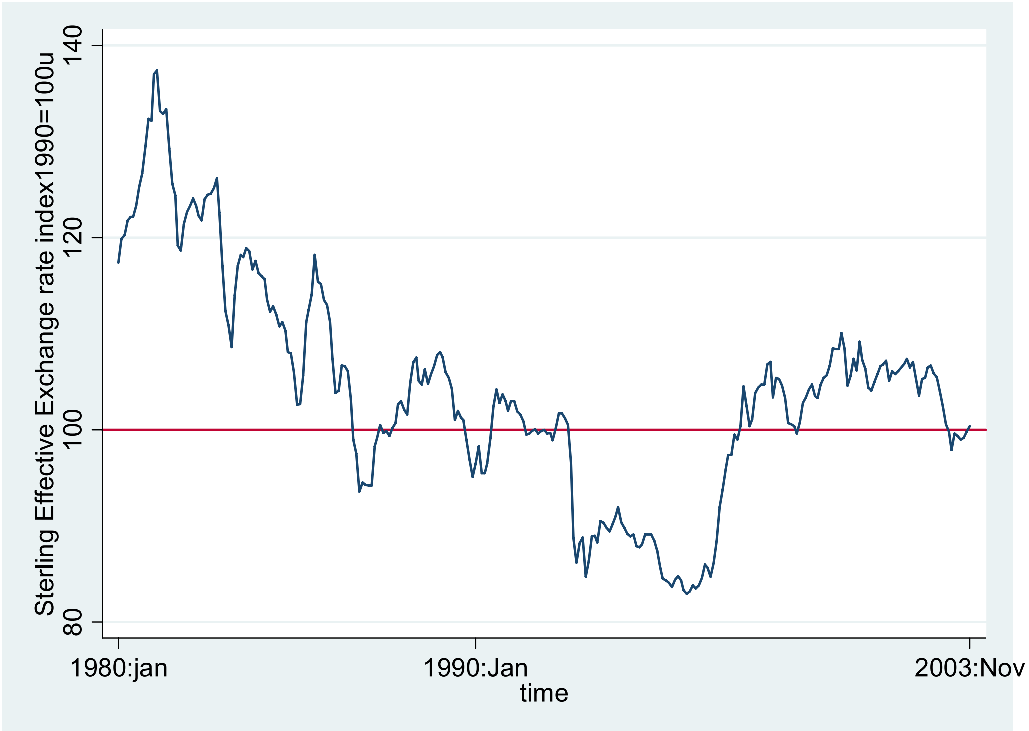
Example

The data set *bop.dta* contains monthly time series data on the UK Balance of Trade in goods (measured in £billion) and the effective exchange rate – a weighted average of sterling's value against a basket of foreign currencies which is centred on the value 100 in January 1990. A value above zero indicates the Balance of Trade is in surplus, a value below zero indicates a deficit. A value > 100 indicates sterling has appreciated, a value<100 indicates that sterling has depreciated

```
two (scatter bop time, yline(0) xlabel(1 121 287, value) )
```



```
two (scatter  xchanger time, yline(100) xlabel(1 121 287, value) )
```



Can see that over time the balance of trade has deteriorated, whilst the value of sterling has generally been high over the same period

To see if the 2 events are related, run a simple regression of the trade balance on sterling's value (lagged by one month to reduce endogeneity concerns)

```
. u bop                              /* read data in */

. tsset year                         /* declare data is time series */
        time variable:  year, 55 to 99

. sort time
. g xchange1=xchange[_n-1]           /* set up 1 period lag of exchange rate */
```

```
. reg bop xchange1
      Source |       SS       df       MS              Number of obs =     286
-------------+------------------------------           F(  1,   284) =   28.08
       Model |  37414954.7     1  37414954.7           Prob > F      =  0.0000
    Residual |   378396350   284  1332381.51           R-squared     =  0.0900
-------------+------------------------------           Adj R-squared =  0.0868
       Total |   415811305   285  1458987.03           Root MSE      =  1154.3
------------------------------------------------------------------------------
         bop |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    xchange1 |   31.56233   5.956083     5.30   0.000     19.83866      43.286
       _cons |  -4646.285   623.5585    -7.45   0.000    -5873.667   -3418.902
```

regressions suggests a high exchange rate is positively correlated with the trade balance
Check for 1st and 12th order autocorrelation (this is monthly data so residuals could be related to last month's value and/or previous 12 month's value)

```
. bgtest,lags(1)
Breusch-Godfrey LM statistic:  249.0974  Chi-sq( 1)  P-value =  4.1e-56

. bgtest,lags(12)
Breusch-Godfrey LM statistic:  250.9979  Chi-sq(12)  P-value =  8.5e-47
```

Results suggest presence of both types of autocorrelation

To fix up standard errors using newey west procedure

```
. newey bop xchange1, lag(1)
Regression with Newey-West standard errors          Number of obs  =       286
maximum lag : 1                                     F(  1,   284)  =     23.81
                                                    Prob > F       =    0.0000
------------------------------------------------------------------------------
             |             Newey-West
         bop |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    xchange1 |   31.56233   6.468508     4.88   0.000     18.83003    44.29463
       _cons |  -4646.285   674.1623    -6.89   0.000    -5973.274   -3319.296
```

and to allow for an AR(12) process

```
. newey bop xchange1, lag(12)
Regression with Newey-West standard errors          Number of obs  =       286
maximum lag : 12                                    F(  1,   284)  =      4.23
                                                    Prob > F       =    0.0406
------------------------------------------------------------------------------
             |             Newey-West
         bop |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    xchange1 |   31.56233   15.34144     2.06   0.041     1.364981    61.75968
       _cons |  -4646.285   1593.395    -2.92   0.004    -7782.646   -1509.923
```

Note the coefficients are unchanged but the standard errors are different, (uncorrected OLS t statistics much larger as expected, particularly when compared to the model that allows for autocorrelation of up to 12 lags)

How can fix up by specifying number of lags, if test is supposed to account for unknown form of autocorrelation? - can be shown that Newey-West test works for unknown forms of autocorrelation as long as number of lags is allowed to rise with number of observations/type of data (monthly, daily etc)